Primer on Statistical Interpretation or Methods

Report Card on Propensity-Score Matching in the Cardiology Literature From 2004 to 2006
A Systematic Review

Peter C. Austin, PhD

Background—Propensity-score matching is frequently used in the cardiology literature. Recent systematic reviews have found that this method is, in general, poorly implemented in the medical literature. The study objective was to examine the quality of the implementation of propensity-score matching in the general cardiology literature.

Methods and Results—a total of 44 articles published in the American Heart Journal, the American Journal of Cardiology, Circulation, the European Heart Journal, Heart, the International Journal of Cardiology, and the Journal of the American College of Cardiology between January 1, 2004, and December 31, 2006, were examined. Twenty of the 44 studies did not provide adequate information on how the propensity-score–matched pairs were formed. Fourteen studies did not report whether matching on the propensity score balanced baseline characteristics between treated and untreated subjects in the matched sample. Only 4 studies explicitly used statistical methods appropriate for matched studies to compare baseline characteristics between treated and untreated subjects. Only 11 (25%) of the 44 studies explicitly used statistical methods appropriate for the analysis of matched data when estimating the effect of treatment on the outcomes. Only 2 studies described the matching method used, assessed balance in baseline covariates by appropriate methods, and used appropriate statistical methods to estimate the treatment effect and its significance.

Conclusions—Application of propensity-score matching was poor in the cardiology literature. Suggestions for improving the reporting and analysis of studies that use propensity-score matching are provided. (Circ Cardiovasc Qual Outcomes. 2008;1:62-67.)

Key Words: statistics ■ matching ■ observational study ■ propensity score ■ propensity-score matching

There is a growing interest in using observational studies to examine the effect of exposures or treatments on outcomes; however, nonrandomized studies can be subject to confounding when subjects who receive the treatment differ systematically from untreated subjects. Confounding occurs when factors or characteristics that influence treatment selection are also associated with the outcome of interest. Propensity-score methods increasingly are being used to reduce the impact of confounding in the estimation of causal treatment effects with nonrandomized data.

The propensity score is defined as a subject’s probability of receiving a specific treatment conditional on the observed covariates.1,2 Subjects with the same propensity score have the same distribution of measured covariates, independent of treatment status; however, treated and untreated subjects with the same propensity score can have different distributions of unmeasured variables.4,5 Propensity-score matching is a commonly used propensity-score method; however, its use requires appropriate statistical methods.

A recent review published in the statistical literature reported that propensity-score matching was frequently performed poorly in articles published in medical journals between 1996 and 2003.7 Of the 47 studies examined, only 2 fully described the methods used and applied appropriate statistical methods for the analysis of the propensity-score–matched sample. A similar review of the use of propensity-score matching in the cardiovascular surgery literature (Annals of Thoracic Surgery, European Journal of Cardiothoracic Surgery, Journal of Cardiovascular Surgery, and Journal of Thoracic and Cardiovascular Surgery) between 2004 and 2006 found that of 60 articles examined, none completely and accurately described the statistical methods employed and used appropriate statistical methods for matched data.8

The objective of the present study was to systematically examine the use of propensity-score matching in the general cardiology literature. Suggestions are provided for the conducting and reporting of studies that use propensity-score matching.
Propensity-Score Matching

The propensity score is frequently estimated by use of a logistic regression model in which treatment status is regressed on baseline characteristics. The estimated propensity score is the model-based predicted probability of exposure to the treatment. Once the propensity score has been estimated for each subject, treated and untreated subjects are matched on the propensity score. Frequently, nearest-neighbor matching within a specified caliper width is used. With this method, the first randomly selected treated subject is matched to the untreated subject with the closest propensity score within a specified range (the caliper width). If multiple untreated subjects are equally close to this treated subject, then 1 of these untreated subjects is randomly selected for matching to this treated subject. This process is repeated until all possible matches have been formed. If for a given treated subject, no available untreated subject lies within the specified caliper width, then that treated subject is excluded from the matched sample. Similarly, unmatched untreated subjects are excluded from the matched sample. Another common approach is to attempt to match treated and untreated subjects on the first 5 digits of the propensity score. If this cannot be done for a specific treated subject, then matching is attempted on the first 4 digits. If no suitable untreated subject exists, then matches are attempted on the first 3, first 2, and finally, the first digit of the propensity score. This method is referred to as 5→1 digit matching. The above strategies all use a “greedy” matching algorithm, regardless of the caliper width that is used. The term “greedy” implies that at a given step of the matching process, the nearest untreated subject is selected for matching to a given treated subject, even if that untreated subject would have served better as a match for a subsequent treated subject. An alternative to the use of greedy matching is to use “optimal” matching, in which matched pairs are formed to minimize the sum of the pairwise differences in the propensity score.

Statistical Methods for Assessing Balance

The performance of propensity-score matching is assessed by the degree to which matching on the estimated propensity score results in a matched sample in which the distribution of measured baseline characteristic is similar between treated and untreated subjects. In prior systematic reviews, greater details were provided on appropriate methods for comparing balance in matched samples. Both Imai et al and Ho et al suggest that statistical hypothesis testing is not appropriate. There are 2 reasons for this. First, balance is a property of a sample and not of a hypothetical superpopulation. Second, significance testing and probability values are influenced by sample size. Matching results in a diminished sample size. Thus, when statistical hypothesis testing is used to assess balance in measured variables between treated and untreated subjects in the matched sample, apparent improvements in balance may occur solely because of a reduction in sample size.

A popular method for assessing balance in measured baseline covariates that is appropriate for use in matched samples is the standardized difference, which is the difference in means in units of the pooled standard deviation (SD).

Unlike probability values, the standardized difference is not confounded with sample size, and thus, balance in the initial sample can be compared with that in the matched sample. The standardized difference can also be used to compare the relative balance of variables measured in different units.

Estimating the Treatment Effect

The final analytic step is to estimate the effect of treatment on the outcome. Propensity-score matching entails matching treated and untreated subjects with similar propensity scores. Rosenbaum and Rubin demonstrated that conditional on the propensity score, measured baseline covariates are independent of treatment assignment. Therefore, on average, treated and untreated subjects with the same propensity score do not differ systematically from one another. In observational studies, confounding results in treated subjects differing systematically from untreated subjects. Therefore, matched treated and untreated subjects will, on average, be more similar than randomly selected treated and untreated subjects. Thus, by design, the propensity-score–matched sample does not consist of independent observations but of pairs of subjects who share a correlation in their baseline covariates and thus likely have correlated outcomes. Appropriate statistical methods for estimating the magnitude and statistical significance of the effect of treatment on outcomes are described in the prior reviews. The performance of a selection of statistical methods for use of the propensity score to estimate the effect of treatment on binary, time-to-event, and count outcomes is described elsewhere.14–16

Survey of Propensity-Score Matching in the Cardiology Literature

Identification of Published Articles That Used Propensity-Score Matching

A search strategy similar to that of recently published systematic reviews of propensity-score methods was used. Both PubMed and the Science Citation Index were searched to identify studies that used propensity-score matching. PubMed was used to identify articles with the keyword “propensity.” The search was restricted to articles published between January 1, 2004, and December 31, 2006, in the following general cardiology journals: American Heart Journal, American Journal of Cardiology, Circulation, European Heart Journal, Heart, International Journal of Cardiology, and Journal of the American College of Cardiology. In addition, the Science Citation Index was used to search for articles that cited at least 1 of several important papers on
propensity-score methods (see a prior review for greater details on the search strategy). The combined search identified 140 articles. A total of 140 articles were examined, and only those that used propensity-score-matching were selected. One study that used propensity-score matching only to generate a sample for subsequent simulation analyses and 2 articles that reported using matching but that actually used stratification on the propensity score were excluded. The search resulted in the identification of 44 studies that used propensity-score matching in the following journals: American Heart Journal (6 articles), American Journal of Cardiology (16 articles), Circulation (12 articles), European Heart Journal (4 articles), International Journal of Cardiology (2 articles), and Journal of the American College of Cardiology (4 articles).

Abstraction of Analytic Methods in Propensity-Score-Matched Samples

The following information was abstracted from each of the published articles: (1) the method by which propensity-score-matched pairs were formed; (2) whether the authors assessed the balance in measured baseline characteristics between treated and untreated subjects in the matched sample, and when the authors did so, whether appropriate statistical methods were used; and (3) the statistical methods used to assess the effect of treatment on the outcome and whether the method was appropriate for matched data. The same criteria for evaluating articles were used as in the prior reviews of propensity-score matching.

Results of Systematic Review

Propensity-Score Matching

Twenty-four studies fully described how matches were formed. Among these studies, 11 used 5–1 digit matching, 1 study matched on the logit of the propensity score using calipers of width 0.2 SDs of the logit of the propensity score, and other studies matched on the propensity score using the following calipers: 0.02 (3 studies) and 0.01 (8 studies). One study required exact matching on the propensity score. Twenty (45%) of the studies did not report the manner by which propensity-score-matched pairs were formed.

Assessing Balance Between Treated and Untreated Subjects

Fourteen studies (32%) did not report whether matching on the propensity score resulted in a matched sample in which the distribution of baseline variables was similar between treated and untreated subjects. Seven additional studies reported that balance was achieved but did not report a table that allowed the reader to compare the distribution of baseline characteristics between treated and untreated subjects in the matched sample. The remaining 23 studies (52%) reported a table in which the distribution of baseline characteristics was compared between treated and untreated subjects.

Of the 30 studies (7 that did not report a table and 23 that did) that reported comparing the distribution of baseline characteristics between treated and untreated subjects, 19 used statistical significance testing, 4 used both standardized differences and significance testing, 3 relied on visual comparison, and 4 did not report what was done. Thus, no studies used only methods appropriate for matched data. Each of the studies that used standardized differences supplemented this method with the use of statistical significance testing. We acknowledge, as a limitation, that in the instance when significance testing that accounted for the matched nature of sample was done in addition to reporting of standardized differences, this insertion may have been either due to convention or at the journal’s request.

Estimating the Effect of Treatment or Exposure on the Outcome

Eleven (25%) of the 44 articles explicitly stated that methods appropriate for the analysis of matched data were used in estimating the treatment effect and its statistical significance. These studies used McNemar’s test, regression models estimated with generalized estimating equation methods to account for the matched-pairs nature of the sample, Cox proportional hazards regression stratified on the matched pairs, the signed rank test, and methods that used the bootstrap. Three additional studies used methods appropriate for correlated data for some outcomes but not for others.

Twenty-one studies (48%) explicitly used inappropriate statistical methods to assess the statistical significance of the effect of treatment on the outcomes. Common errors included the use of the log-rank test to compare Kaplan-Meier survival curves in the matched sample, use of Cox proportional hazards models in the matched sample, use of logistic regression in the matched sample, use of $\chi^2$ tests to compare proportions in the matched sample, and use of Wilcoxon rank sum tests or standard $t$ tests to compare continuous variables in the matched sample. Nine studies did not adequately describe the statistical methods that were used to compare the outcome between treated and untreated subjects. Frequently, these studies were comparing proportions or Kaplan-Meier survival curves between treated and untreated subjects.

Discussion

The objective of the present study was to critically examine the use of propensity-score matching in the general cardiology literature. Only 4 of the 44 studies compared the distribution of baseline variables between treated and untreated subjects in the matched sample and explicitly documented the use of appropriate statistical methods to assess whether measured characteristics were balanced between treated and untreated subjects in the matched sample. These four studies used standardized differences, which are appropriate in this context. Eleven (25%) of the 44 studies explicitly used appropriate statistical methods for all analyses that examined the impact of treatment on outcomes. Overall, only 2 studies described the matching method that was used and explicitly used appropriate statistical methods both for assessing balance in the matched sample and for assessing the statistical significance of the treatment effect; however,
both of these studies, in addition to using standardized differences to assess baseline balance, also used statistical hypothesis testing that did not account for the matched nature of the sample. The remaining 42 articles were subject to either incomplete reporting of methods or inappropriate statistical methods. Below are recommendations for the design, analysis, and reporting of studies that use propensity-score matching. These recommendations are summarized in the Table.

### Describe the Matching Method
The method by which the propensity-score–matched pairs were formed should be described explicitly. This allows other researchers to replicate the study methods. It is insufficient to state only that greedy matching was done or that nearest-neighbor matching was used. A recent study found that matching on the logit of the propensity score with a caliper width of 0.2 of the SD of the logit of the propensity score tended to have superior performance to other methods that were identified as being used in the medical literature.65 Furthermore, methods based on matching on the logit of the propensity score with calipers that are a proportion of the SD have stronger theoretical justification.66

### Report the Balance of Baseline Variables Between Treated and Untreated Subjects in the Matched Sample
The propensity score is a balancing score: Matching treated and untreated subjects on the propensity score allows one to mimic some of the characteristics of a randomized, controlled trial, researchers are encouraged to report risk differences or relative risks rather than odds ratios.

This constitutes the third systematic review of propensity-score matching that I have conducted. The first review was of articles that were published in an earlier era (1996–2003).7 In that earlier review, 4% of articles used appropriate methods to compare baseline characteristics between treated and untreated subjects. Furthermore, 28% of studies used appropriate statistical methods to estimate the significance of the effect of treatment on outcomes.7 In the present review, there was no meaningful improvement in statistical practice in a more recent era. The second review was of articles published in the same era as in the present review (2004–2006) but was restricted to articles in the cardiovascular surgical literature.8 In the cardiovascular surgery literature, no studies used appropriate methods to compare baseline characteristics between treated and untreated subjects, whereas 9% of articles in the cardiology literature used appropriate methods. In the cardiovascular surgery literature, 13% of studies explicitly used appropriate statistical methods to estimate the effect of treatment on outcomes, whereas 25% of studies in the cardiology literature used appropriate methods. Therefore, the implementation of propensity-score matching appears to be of somewhat higher quality in the cardiology literature than in the cardiovascular surgery literature; however, there still remains substantial room for improvements in both literatures. One of the primary objectives of the present systematic review is educational: to highlight and describe good statistical practice in the context of propensity-score matching.

In the present review, explicit criteria were used to assess the statistical methods used in studies that employ propensity-score matching. In particular, on the basis of arguments by Imai et al11 and Ho et al12 that significance testing is inappropriate for assessing balance in baseline covariates in matched samples, it was required that methods be used that were not based on statistical hypothesis testing. In addition, it was required that tests of the effect of treatment on outcomes should account for the matched-pairs nature of the sample. Although strong arguments can be made for both of these differences, rather than hypothesis testing, for assessing balance in baseline covariates between treated and untreated subjects. The receiver operating characteristic curve area (c-statistic) of the propensity-score model is not a measure of the goodness-of-fit of the propensity-score model. Indeed, recent research has shown that the receiver operating character-

<table>
<thead>
<tr>
<th>Table. Key Components for Studies That Use Propensity-Score Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Explicitly describe the matching method used.</td>
</tr>
<tr>
<td>2. Explicitly compare and report the balance in baseline characteristics between treated and untreated subjects. Do not use statistical tests of hypothesis.</td>
</tr>
<tr>
<td>3. Use statistical methods that account for the lack of independence induced by matching on the propensity score when estimating the statistical significance of the effect of treatment on outcomes.</td>
</tr>
</tbody>
</table>
criteria, disagreement persists within the statistical community about acceptable practices for studies that use propensity-score matching. For example, in response to an earlier review, Hansen argued that the use of balance tests that incorporate hypothesis testing is acceptable. In a similar commentary, Hill acknowledged that matching on the propensity score induces dependence within the matched sample and that a matched-pairs analysis is an appropriate way to account for this lack of independence; however, she expressed disagreement with the requirement that a matched-pairs analysis is the only way to account for this lack of independence. The reader is referred elsewhere for the response to these commentaries.

Summary
In conclusion, propensity-score matching tended to be poorly implemented in the cardiology literature. The majority of studies ignored the matched nature of the propensity-score–matched sample in the subsequent analyses. Suggestions for improving the analysis of propensity-score–matched samples and for improving the reporting of these analyses have been provided.

Disclosures
The Institute for Clinical Evaluative Sciences 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 author, 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. Dr Austin is supported by a Career Investigator Award from the Heart and Stroke Foundation of Ontario.

References
31. Kandzari DE, Turett RL, Edar JP, Jollis JG. Comparison of long-term (seven year) outcomes among patients undergoing percutaneous coronary
42. Bhatheja R; Francis GS; Pothier CE; Lauer MS. Heart rate response during dipyridamole stress as a predictor of mortality in patients with normal myocardial perfusion and normal electrocardiograms. Am J Cardiol. 2005;95:1159–1164.
50. Ting HH, Raveendran G, Lenno RJ; Long KH; Singh M; Wood DL; Gersh BJ; Rihal CS; Holmes DR. A total of 1,007 percutaneous coronary interventions without onsite cardiac surgery. J Am Coll Cardiol. 2006;47:1713–1721.
Primer on Statistical Interpretation or Methods Report Card on Propensity-Score Matching in the Cardiology Literature From 2004 to 2006: A Systematic Review

Peter C. Austin

doi: 10.1161/CIRCOUTCOMES.108.790634

Circulation: Cardiovascular Quality and Outcomes is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2008 American Heart Association, Inc. All rights reserved.
Print ISSN: 1941-7705. Online ISSN: 1941-7713

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circoutcomes.ahajournals.org/content/1/1/62

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation: Cardiovascular Quality and Outcomes can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at: http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation: Cardiovascular Quality and Outcomes is online at: http://circoutcomes.ahajournals.org//subscriptions/