Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: A systematic review and suggestions for improvement

Peter C. Austin, PhD

Supplemental material is available online.

Objective: I conducted a systematic review of the use of propensity score matching in the cardiovascular surgery literature. I examined the adequacy of reporting and whether appropriate statistical methods were used.

Methods: I examined 60 articles published in the *Annals of Thoracic Surgery*, *European Journal of Cardio-thoracic Surgery*, *Journal of Cardiovascular Surgery*, and the *Journal of Thoracic and Cardiovascular Surgery* between January 1, 2004, and December 31, 2006.

Results: Thirty-one of the 60 studies did not provide adequate information on how the propensity score-matched pairs were formed. Eleven (18%) of studies did not report on whether matching on the propensity score balanced baseline characteristics between treated and untreated subjects in the matched sample. No studies used appropriate methods to compare baseline characteristics between treated and untreated subjects in the propensity score-matched sample. Eight (13%) of the 60 studies explicitly used statistical methods appropriate for the analysis of matched data when estimating the effect of treatment on the outcomes. Two studies used appropriate methods for some outcomes, but not for all outcomes. Thirty-nine (65%) studies explicitly used statistical methods that were inappropriate for matched-pairs data when estimating the effect of treatment on outcomes. Eleven studies did not report the statistical tests that were used to assess the statistical significance of the treatment effect.

Conclusions: Analysis of propensity score-matched samples tended to be poor in the cardiovascular surgery literature. Most statistical analyses ignored the matched nature of the sample. I provide suggestions for improving the reporting and analysis of studies that use propensity score matching.

From the Institute for Clinical Evaluative Sciences, the Department of Public Health Sciences, University of Toronto, and the Department of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada.

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 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 in part by a New Investigator award from the Canadian Institutes of Health Research (CIHR).

Received for publication April 16, 2007; accepted for publication July 31, 2007.

Address for reprints: Peter C. Austin, PhD, Institute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Ave, Toronto, Ontario M4N 3M5, Canada (E-mail: peter.austin@ices.on.ca).

J Thorac Cardiovasc Surg 2007;134:1128-35
0022-5223/\$32.00

Copyright © 2007 by The American Association for Thoracic Surgery

doi:10.1016/j.jtcvs.2007.07.021

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 score is a subject's probability of receiving a specific treatment conditional on the observed covariates.¹⁻³ Matching on the propensity score allows one to balance measured variables between treated and untreated subjects.^{1,2,4} However, matching on the propensity score can still result in unmeasured variables being imbalanced between treated and untreated subjects.^{5,6}

There are three commonly used propensity score methods: covariate adjustment using the propensity score, stratification on the propensity score, and propensity score matching.⁷ Earlier studies have shown that propensity score matching results in the comparison of treated and untreated subjects who are more similar than does stratification on the propensity score.^{6,7} 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.⁸ A systematic review of

articles that employed propensity-score matching and were published in the literature between 1996 and 2003 found that a high proportion of articles contained errors in the application of propensity-score matching.⁹ Propensity score matching is frequently used in the cardiovascular surgery literature. The objective of the current study was twofold: first, to systematically examine the use of propensity score matching in the cardiovascular surgery literature; second, to provide recommendations to cardiovascular surgery researchers on the implementation of propensity score matching.

Statistical Methods for Propensity Score–Matched Samples

There are 4 different steps in a propensity score–matched analysis. First, one must specify the propensity score model. Second, one must construct the propensity score–matched sample. Third, one must assess the degree to which matching on the propensity score has resulted in a matched sample in which the distribution of measured baseline variables is similar between treated and untreated subjects. Fourth, one must estimate the effect of the treatment or exposure on the outcomes under consideration. Each of these steps is described in the subsequent subsections.

Specifying the Propensity Score Model

I briefly provide some guidance on specifying the propensity score model. Rosenbaum and Rubin¹ demonstrated that the propensity score is a balancing score: conditioning on the propensity score results in treated and untreated subjects having similar distributions of baseline variables. Ho and associates¹⁰ refer to the *propensity score tautology* as the fact that one has correctly specified the propensity score model when matching on the propensity score results in a matched sample in which treated and untreated subjects have similar distributions of measured baseline variables. This reflects Rosenbaum and Rubin's² use of an iterative process for specifying the propensity score. Recently, along with Grootendorst and Anderson, I⁶ published an article on variable selection for propensity score models. I demonstrated that including variables that are related to exposure but that are independent of the outcome in the propensity score model can result in the formation of fewer matched pairs. This can result in less precise estimates of the treatment effect. Including only the confounders of the treatment–outcome relationship, or all the variables associated with the outcome, resulted in the formation of a greater number of matched pairs and more precise estimates of treatment effect. Furthermore, it was shown that the receiver operating characteristic curve area (c-statistic) of the propensity score model did not provide any information about whether important confounders had been omitted. I propose the following steps in specifying the propensity score model: First, derive a list of measured baseline variables that are likely

related to exposure and/or the outcome. The variables in this list can be selected from reviews of the literature, from prior studies, and from expert opinion. Importantly, the list should only include variables measured at baseline, before exposure. The list should not include the outcome, nor should it include variables in the causal pathway. Second, derive an initial propensity score model by including all variables in the list as main effects. Third, assess whether matching on the propensity score results in a matched sample in which measured baseline variables are balanced between treated and untreated subjects. Fourth, in the event of imbalance, modify the propensity score, possibly by using methods described in more detail by Rosenbaum and Rubin.² The third and fourth steps can be repeated iteratively until the baseline variables are balanced between treated and untreated subjects. If the final propensity score model contains variables that are associated with treatment but that are independent of the outcome, then one can examine whether, by dropping these variables, one can form a larger number of propensity score–matched pairs without increasing systematic differences in prognostic variables between treated and untreated subjects.

Propensity Score Matching

The propensity score is usually estimated by a logistic regression model in which treatment (yes/no; 1/0) is regressed on baseline characteristics. The estimated propensity score is then the predicted probability of exposure to the treatment from the logistic regression model. Once the propensity score has been estimated for each subject, treated and untreated subjects are matched on the propensity score. Typically, nearest neighbor matching within a specified caliper width is used. By this method, treated subjects are randomly sorted. Then, the first treated subject is matched to the untreated subject with the closest propensity score within a specified range (the caliper width). If no untreated subject has a propensity score that lies within a specified caliper width of the treated subject's propensity score, then that treated subject is left unmatched and is not used in subsequent analyses. Matching without replacement is usually employed: once an untreated subject has been matched to a given treated subject, this untreated subject is no longer considered as a possible match for subsequent treated subjects. This process is then repeated until all possible matches have been formed. Because the propensity score is a probability, it takes on values that lie between 0 and 1. If two subjects have propensity scores of 0.12345678 and 0.12345123, then these subjects have propensity scores that match on the first five digits (0.12345). Another common approach is to attempt to match treated and untreated subjects on the first five digits of the propensity score. If no match is found for a specific treated subject, then matching is attempted on the first four digits. If no suitable untreated

subject exists, then matches are attempted on the first three, first two, and finally, the first digit of the propensity score.¹¹ I refer to this method as 5→1 digit matching. A recent study found that matching on the logit of propensity score, using calipers of width 0.2 of the standard deviation of the logit of the propensity score, tended to have superior performance compared with other competing methods that are used in the medical literature. [Austin PC. The performance of different propensity-score matching methods used in the medical literature. Under review.]

The term “greedy matching” refers to any matching algorithm in which, at a specific step in the matching process, the nearest untreated subject is matched to the treated subject in question, even if that untreated subject would have been a better match for a subsequent treated subject.¹² The term “greedy” does not provide any information about the calipers that were used in the matching process. The alternative to using a “greedy” approach is to use an “optimal” matching strategy that makes matches so as to minimize a weighted average of the within-pair distance over all possible matches.¹² In the above, I have assumed matching without replacement. One can also use matching with replacement, in which an untreated subject can serve as a match for more than one treated subject. Thus, it is possible to have multiple matched pairs, each consisting of a different treated subject, but each consisting of the same untreated subject. When matching with replacement is used, statistical methods for estimating the treatment effect must take into account the lack of independence in outcomes for the same untreated subject that is contained in multiple matched sets. Throughout the remainder of the manuscript, I will assume that matching without replacement is being used.

Statistical Methods for Assessing Balance

The propensity score is a balancing score: matching on the true propensity score results in a matched sample in which the distribution of each baseline characteristic is similar between treated and untreated subjects. In reality, except in controlled experiments, one does not know the true propensity score, and it must be estimated from the data. Ho and colleagues¹⁰ refer to the *propensity score tautology* as the fact that one knows that one has properly specified the propensity score model when matching on the estimated propensity score balances baseline variables between treated and untreated subjects. Thus, an important component of any propensity score–matched analysis is comparing the balance in baseline variables between treated and untreated subjects.

There is clear consensus among statisticians as to the inappropriateness of using significance test to compare the distribution of baseline covariates between different arms of a randomized controlled trial.¹³⁻¹⁹ Imai, King, and Stuart²⁰

have proposed two criteria for appropriate methods for comparing baseline variables between treated and untreated subjects in observational studies. First, because balance is a property of a sample, and not of a hypothetical super-population, the measure for assessing balance must be a property of the sample. Second, the method for assessing balance must not be influenced by the size of the sample. Both of these criteria rule out the use of significance testing for assessing balance in baseline variables between treated and untreated subjects. The second criterion is important, for if the method to assess balance were influenced by sample size, then the matched sample may appear to have better balance solely because of the reduction in sample size that results from matching. Several authors have proposed the use of standardized differences for assessing balance in observational studies.^{5,6,21,22} The standardized difference is defined as follows:

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

where $\bar{x}_{treatment}$ and $\bar{x}_{control}$ are the mean of the variable among the treated and untreated subjects, respectively, while $s_{treatment}^2$ and $s_{control}^2$ are the sample standard deviation of covariate in the treated and untreated subjects, respectively. Ho and associates¹⁰ describe other possible measures to assess balance, including quantile–quantile plots.

Estimating the Treatment Effect

The need to account for the matched nature of the sample. The propensity score–matched sample was created by matching pairs of subjects with a similar propensity score. Therefore, treated and untreated subjects within the same matched pair have a similar propensity score. Thus, these treated and untreated subjects within the same matched pair have baseline variables that come from the same multivariate distribution.¹ Randomly chosen treated and untreated subjects are likely to differ systematically in their baseline variables. Hence, treated and untreated subjects within the same matched pair are, on average, more similar than two randomly selected treated and untreated subjects. Because outcomes are influenced by baseline characteristics (otherwise there would be no confounding), then outcomes are more similar within the matched pair than between randomly selected treated and untreated subjects. This within-pair homogeneity means that subjects within the same matched pair are not independent. Therefore, by construction, the propensity score–matched sample does not consist of independent observations. The need to account for matching in the statistical analyses is well described in the epidemiology literature.^{23,24}

Statistical methods for estimating the treatment effect.

The final analytic step is to estimate the effect of treatment on the outcomes. This must be done in a manner that accounts for the matched nature of the propensity score–matched sample. The statistical significance of the effect of exposure on continuous outcomes can be assessed by a paired *t* test²⁵ or the Wilcoxon signed rank test.²⁶ Proportions can be compared by the McNemar test for correlated binary proportions, or extensions thereof for categorical variables with more than two levels.²⁷ Agresti and Min²⁸ describe methods for estimating relative risks and odds ratios in matched samples and for constructing appropriate confidence intervals. Kaplan–Meier survival curves can be compared by a test described by Klein and Moeschberger.²⁹ As a brief description of this test, let D_1 denote the number of matched pairs in which the treated subject experiences the event first, while D_2 denotes the number of matched pairs in which the untreated subject experiences the event first. The test statistic is as follows:

$$\frac{D_1 - D_2}{\sqrt{D_1 + D_2}}$$

which has a standard normal distribution under the null hypothesis and when the number of matched pairs is large (the matched pairs in which the smaller of the two times is a censored observation makes no contribution to the test statistic). The log–rank test, which assumes independent strata, is not appropriate for comparing survival curves in matched samples.^{29,30} The standard Cox regression model is not appropriate for matched-pairs data, as it assumes independent samples.³¹ However, for analyzing survival data, using a Cox proportional hazards models stratified on the matched pairs would be appropriate.³² Another approach would be to use a Cox proportional hazards model with robust standard errors that account for the clustering in matched pairs.³³ Similarly, conventional logistic regression would not be appropriate.³⁴ However, for analyzing binary outcomes, conditional logistic regression or logistic regression models estimated using generalized estimating equation methods take into account the matched nature of the data can be used.³⁵

Survey of Propensity Score Matching in the Cardiovascular Surgery Literature

Identification of Published Articles Using Propensity Score Matching

I used a search strategy similar to that of a recently published systematic review of propensity score methods in the medical literature.³⁶ I used both PubMed and the Science Citation Index to identify studies that used propensity score matching. I identified studies that included the keyword *propensity*, using a keyword search in PubMed. I restricted my search to articles published between January 1, 2004,

and December 31, 2006, in the following cardiovascular surgery journals: *Annals of Thoracic Surgery*, *European Journal of Cardio-thoracic Surgery*, *Journal of Cardiovascular Surgery*, *Journal of Heart and Lung Transplant*, and the *Journal of Thoracic and Cardiovascular Surgery*. Using the Science Citation Index, I also searched for articles that cited one of the important papers on propensity score methods.^{1,2,37–40} The combined search identified 115 articles. I then examined these 115 articles and selected only those that used propensity score matching. The combined search strategy resulted in the identification of 60 studies^{E1–E60} that used propensity score matching in the following journals: *Annals of Thoracic Surgery* (31 articles), *European Journal of Cardio-thoracic Surgery* (9 articles), *Journal of Cardiovascular Surgery* (1 article), and the *Journal of Thoracic and Cardiovascular Surgery* (19 articles).

Abstraction of Analytic Methods in Propensity Score–Matched Samples

I abstracted the following information 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 baseline characteristics between treated and untreated subjects in the matched sample. When the authors compared balance of baseline variables between treated and untreated subjects, I examined the methods that the authors used.
3. The statistical methods used to assess the effect of treatment on the outcome, and whether this method was appropriate for matched-pairs data.

Results of Systematic Review

I critically examined 60 articles published in the cardiovascular surgery literature between 2004 and 2006 that used propensity score matching. I report our results separately for each of the three items that were abstracted.

Propensity Score Matching

Seventeen (28%) of the studies did not report the manner by which propensity score–matched pairs were formed. An additional 4 reported that greedy matching was used, and 10 stated that nearest neighbor matching was used. As noted earlier, the use of the term “greedy matching” does not provide any details about the required similarity of the propensity score between matched treated and untreated subjects. Similarly, the use of the term “nearest neighbor matching” is also uninformative, because it does not provide any details on the caliper widths that were used in the matching process. Taken to the extreme, if a caliper width was not used, then a matched untreated subject would be able to be found for each treated subject, because there is no specification that their propensity scores are required to be

similar. In total, 31 (52%) of the studies did not provide adequate information about how the matching was done. This has important consequences in that it does not permit other researchers to reproduce the studies' methods. Among studies that fully describe how matches were formed, 20 used 5→1 digit matching, 1 study matched on the logit of the propensity score using calipers of width 0.2 standard deviations of the logit of the propensity score, and other studies matched on the propensity score using the following calipers: 0.1 (1 study), 0.05 (2 studies), 0.02 (1 study), 0.015 (1 study), 0.01 (2 studies), and 0.001 (1 study).

Assessing Balance Between Treated and Untreated Subjects

Eleven (18%) studies did not report whether matching on the propensity score resulted in a matched sample in which the distribution of baseline characteristics was similar between treated and untreated subjects. One additional study reported that balance was achieved, but did not report a table comparing the distribution of baseline characteristics between treated and untreated subjects in the matched sample. The remaining 48 studies (80%) reported a table in which the distribution of baseline characteristics was compared between treated and untreated subjects.

Of the 49 studies that reported comparing the distribution of baseline characteristics between treated and untreated subjects, 47 studies used statistical significance testing, 1 study relied on visual comparison, and 1 study did not report the methods that were used. Of the 47 studies that used statistical significance testing, 1 study explicitly stated that correct statistical methods were used for matched-pairs data,^{E28} 35 (58%) studies *explicitly* used statistical hypothesis testing methods that did not incorporate the matched nature of the sample, and 1 study used appropriate statistical hypothesis tests for some variables but inappropriate statistical hypothesis tests for other variables. Ten studies did not report what statistical tests were used to compare the distribution of baseline characteristics between treated and untreated subjects—only significance levels were reported, but not the statistical tests used to obtain these significance levels. Importantly, none of the studies reported using appropriate methods for assessing balance in baseline variables between treated and untreated subjects. As discussed in the “Statistical Methods for Assessing Balance” section, statistical hypothesis testing (regardless of whether it accounts for the matched nature of the sample) is not appropriate for comparing baseline balance of measured covariates. No studies reported using standardized differences or other comparable methods.

Estimating the Effect of Treatment or Exposure on the Outcome

Eight (13%) of the 60 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,^{E8} regression models estimated generalized estimating equation methods to account for the matched-pairs nature of the sample,^{E16,E36,E39,E46} Cox proportional hazards regression stratified on the matched pairs,^{E42} and conditional logistic regression.^{E43,E54} Two additional studies used methods appropriate for correlated data for some outcomes, but not for other outcomes.^{E33,E49}

Thirty-nine (65%) studies explicitly used inappropriate statistical methods for assessing the statistical significance of the effect of treatment on the outcomes. 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 χ^2 tests to compare proportions in the matched sample, and using Wilcoxon rank sum tests or standard t tests to compare continuous variables in the matched sample. Eleven studies did not describe the statistical methods that were used to compare the outcome between treated and untreated subjects.^{E5,E12,E14,E15,E19,E20,E22,E24,E28,E44,E45} In general, these studies were comparing proportions, means, medians, or Kaplan–Meier survival curves between treated and untreated subjects.

Discussion

The objective of the current study was to critically examine the use of propensity score matching in the cardiovascular surgery literature. None of the 60 studies compared the distribution of baseline variables between treated and untreated subjects in the matched sample and explicitly documented using appropriate statistical methods to assess whether measured characteristics were balanced between treated and untreated subjects in the matched sample. Eight (13%) of the 60 studies explicitly used appropriate statistical methods for all analyses examining the impact of treatment on outcomes. I make the following recommendations for the design, analysis, and reporting of studies that use propensity score matching. I summarize my recommendations for the implementation of propensity score matching in Table E1.

Describing the Matching Method

The method by which the propensity score–matched pairs were formed should be explicitly described. This allows other researchers to replicate the study methods. It is insufficient to state that either “greedy” matching or “nearest neighbor” matching was used. If calipers of a fixed width were used, then this should be explicitly described. If 5→1 digit matching was used, then this should be stated explicitly. A recent study found that matching on the logit of the propensity score using calipers of width 0.2 of the standard deviation of the logit of the propensity score tended

to have superior performance compared with other competing methods. [Austin PC. The performance of different propensity-score matching methods used in the medical literature. Under review.] Furthermore, this method also has stronger theoretical justification.⁴¹

Reporting the Balance of Baseline Variables Between Treated and Untreated Subjects in the Matched Sample

The CONSORT statement recommends that baseline demographic and clinical data be reported for each arm in a randomized controlled trial.⁴² In a randomized controlled trial, reporting baseline characteristics in each arm of the study allows the reader to assess whether there was potentially a breakdown in randomization, since randomization should, on average, result in similar distributions of baseline variables between the different arms of the study. Although observational studies are, by definition, nonrandomized, matching on the propensity score allows one to balance measured baseline variables between treated and untreated subjects. Describing the balance in measured variables between treated and untreated subjects in the matched sample allows both the researcher and the reader to assess whether the propensity model was adequately specified.

Appropriate statistical methods should be used to compare the distribution of the baseline covariates between treated and untreated subjects. Statistical methods for assessing balance in baseline variables should not be affected by sample size and should reflect the fact that balance is a property of a sample and not refer to a super-population.²⁰ I encourage researchers to use standardized differences to compare distributions between treated and untreated subjects.^{6,7,21} Unlike *P* 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. It can also be used to compare the relative balance of variables measured in different units.

Some studies that use propensity score matching compare characteristics of matched treated subjects with those of unmatched treated subjects. This comparison can provide useful information on differences between treated patients who were used in estimating the treatment effect and the treated patients who were excluded from these analyses. These comparisons can provide useful clinical information and information on generalizability of the results. However, I do not consider the decision concerning their inclusion or exclusion to be a statistical one. Furthermore, their inclusion or exclusion does not affect the quality of statistical methods for assessing balance in measured variables between treated and untreated subjects in the matched sample and for estimating the significance of the treatment effect.

Statistical Methods for Estimating the Effect of Treatment on Outcomes

Researchers should explicitly report that methods appropriate for the analysis of matched data were used. In the section titled “Statistical Methods for Estimating the Treatment Effect,” appropriate methods are described for estimating the treatment effect in propensity score–matched samples. Some researchers have used conditional logistic regression for the analysis of propensity score–matched samples. Whereas conditional logistic regression is appropriate for matched-pairs data, it has been shown to result in biased estimation of odds ratios when used in propensity score–matched samples⁴³ since propensity score methods allow one to estimate marginal treatment effects and not conditional treatment effects.⁴⁴ A marginal treatment effect is the average effect at the population level, whereas the conditional treatment effect is the average effect at the subject level. The odds ratio is not collapsible, meaning that the marginal treatment effect is different from the conditional treatment effect.⁴⁵ However, risk differences are collapsible, whereas the relative risk is collapsible under certain circumstances.⁴⁵ The use of conditional logistic regression is thus discouraged. Because the matched sample has reduced or eliminated systematic differences in measured variables between treated and untreated subjects, regression adjustment should rarely be needed. Researchers are encouraged to report risk differences or relative risks, rather than odds ratios. Agresti and Min²⁸ describe statistical methods for estimating relative risks and associated confidence intervals in matched data. When researchers want to adjust for possible residual baseline differences between treated and untreated subjects, regression methods that account for the matched nature of the sample should be incorporated. Appropriate statistical methods for different types of outcomes are summarized in Table E2, which appears online only.

Comparison of Propensity Score Matching With Other Propensity Score Methods

In the introduction, I stated that there were three commonly used propensity score methods: propensity score matching, stratification on the propensity score, and covariate adjustment using the propensity score. In this review, I have focused on propensity score matching. Propensity score matching may require more analytic steps than competing propensity score methods. However, there are several arguments for the use of propensity score matching. First, propensity score matching allows for the direct comparability of treated and untreated subjects in the matched sample. Both researchers and readers of the published research can assess the degree to which matching on the propensity score resulted in a matched sample in which systematic differences between treated and untreated subjects were reduced or eliminated. When stratification on the propensity score

(usually the quintiles) is employed, balance must be assessed within each of the strata, requiring a more complex assessment of balance. Second, prior empirical and theoretical research that my colleagues and I^{6,7} have published has demonstrated that propensity score matching tends to result in the elimination of a greater degree of the systematic differences between treated and untreated subjects than does stratification on the propensity score. Third, when covariate adjustment using the propensity score is employed, it is unclear how to assess whether the propensity score model has been correctly specified. With propensity score matching, one knows that the propensity score model has been adequately specified when matching on the estimated propensity score results in treated and untreated subjects having similar distributions of measured baseline variables. Fourth, with propensity score matching and, to a lesser extent, with stratification on the propensity score, one can directly assess the degree of overlap between treated and untreated subjects. This is less apparent when covariate adjustment using the propensity score is employed. Fifth, covariate adjustment using the propensity score is a model-based approach and thus requires the assumption that the outcomes model is correctly specified. Sixth, with covariate adjustment using the propensity score, one loses the ability to compute measures of effect such as risk differences or relative risks when examining the effect of exposures on binary outcomes. Seventh, sensitivity analyses for assessing the impact of potential unmeasured confounders on the treatment effect have been proposed for propensity score-matching methods.¹²

Limitations

There is a limitation to this current systematic review. First, the quality of articles employing propensity score matching was assessed using published articles in the cardiovascular surgery literature. It is possible that authors had provided greater details on the analyses and results, but that these were removed during the revision and editorial process. However, this limitation is tempered by the fact that 65% of the reviewed articles explicitly used statistical methods inappropriate for matched-pairs data when estimating the effect of treatment on the outcome. Furthermore, 47 (78%) of articles used statistical significance testing to compare differences in baseline variables between treated and untreated subjects in the matched sample, despite this not being appropriate. Most of the errors that were highlighted were errors of commission, rather than errors of omission.

Conclusions

In conclusion, propensity score matching tended to be poorly implemented in the cardiovascular surgery literature. The majority of studies ignored the matched nature of the propensity score-matched sample in the subsequent analyses. I have provided suggestions for improving the analysis

of propensity score-matched samples and for improving the reporting of these analyses.

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. *J Am Stat Assoc*. 1984;79:516-24.
3. Grunkmeier 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.
4. Blackstone EH. Comparing apples and oranges. *J Thorac Cardiovasc Surg*. 2002;123:8-15.
5. Austin PC, Mamdani MM, Stukel TA, Anderson GM, Tu JV. The use of the propensity score for estimating treatment effects: administrative versus clinical data. *Stat Med*. 2005;24:1563-78.
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. *Stat Med*. 2007;26:734-53.
7. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Stat Med*. 2006;25:2084-2106.
8. 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*. *Am Stat*. 2007;61:47-55.
9. Austin PC. A critical appraisal of propensity-score matching in the medical literature from 1996 to 2003. *Stat Med*. Accepted.
10. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Anal*. In press.
11. Parson LS. Reducing bias in a propensity score matched-pair sample using greedy matching techniques. In: Proceedings of the Twenty-sixth Annual SAS Users Group International Conference. Cary (NC): SAS Institute Inc; 2001. p 214-6.
12. Rosenbaum PR. *Observational studies*. Springer-Verlag: New York; 1995.
13. Altman DT, Dore CJ. Randomisation and baseline comparisons in clinical trials. *Lancet*. 1990;335:149-53.
14. Altman DG. Comparability of randomised groups. *Statistician*. 1985; 34:125-36.
15. Senn S. Testing for baseline balance in clinical trials. *Stat Med*. 1994;13:1715-26.
16. Senn S. Baseline comparisons in randomized clinical trials. *Stat Med*. 1991;10:1157-60.
17. Senn SJ. Covariate imbalance and random allocation in clinical trials. *Stat Med*. 1989;8:467-75.
18. Begg CB. Significance tests of covariate imbalance in clinical trials. *Control Clin Trials*. 1990;11:223-5.
19. Rothman KJ. Epidemiologic methods in clinical trials. *Cancer*. 1977; 39:1771-5.
20. Imai K, King G, Stuart EA. Misunderstandings among experimentalists and observationalists: balance test fallacies in causal inference. Technical report. Princeton University. Available from: <http://imai.princeton.edu/research/balance.html>.
21. Flury BK, Riedwyl H. Standard distance in univariate and multivariate analysis. *Am Stat*. 1986;40:249-51.
22. Normand ST, Landrum MB, Guadagnoli E, Ayanian JZ, Ryan TJ, Cleary PD, et al. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *J Clin Epidemiol*. 2001;54:387-98.
23. Rothman KJ, Greenland S. *Modern epidemiology*. Philadelphia: Lippincott Williams & Wilkins; 1998.
24. Breslow NE, Day NE. *Statistical methods in cancer research. Vol I. The analysis of case-control studies*. Lyon: International Agency for Research on Cancer; 1980.

25. Snedecor GW, Cochran WG. Statistical methods, 8th ed. Ames (IA): Iowa State University Press; 1989.
26. Conover WJ. Practical nonparametric statistics. 3rd ed. New York: John Wiley; 1999.
27. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. 3rd ed. New York: John Wiley; 2003.
28. Agresti A, Min Y. Effects and non-effects of paired identical observations in comparing proportions with binary matched-pairs data. *Stat Med*. 2004;23:65-75.
29. Klein JP, Moeschberger ML. Survival analysis: techniques for censored and truncated data. New York: Springer-Verlag; 1997.
30. Harrington D. Linear rank tests in survival analysis. In: Armitage P, Colton T, editors. Encyclopedia of biostatistics. 2nd ed. New York: John Wiley; 2005. p. 2802-12.
31. Cox DR, Oakes K. Analysis of survival data. London: Chapman & Hall; 1984.
32. Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. New York: Springer-Verlag; 2000.
33. Lin DY, Wei LJ. The robust inference for the proportional hazards model. *J Am Stat Assoc*. 1989;84:1074-8.
34. Cox DR, Snell EJ. Analysis of binary data. London: Chapman & Hall; 1989.
35. Diggle PJ, Liang KY, Zeger SL. Analysis of longitudinal data. Oxford: Oxford University Press; 1994.
36. 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. *J Clin Epidemiol*. 2006;59:437-47.
37. Rosenbaum PR, Rubin DB. The bias due to incomplete matching. *Biometrics*. 1985;41:103-16.
38. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*. 1998;17:2265-81.
39. Joffe MM, Rosenbaum PR. Invited commentary: Propensity scores. *Am J Med*. 1999;150:327-33.
40. Rubin DB. Estimating causal treatment effects from large datasets using propensity scores. *Ann Intern Med*. 1997;127:757-63.
41. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. *Sankhya Series A*. 1973;35:417-46.
42. Moher D, Schulz KF, Altman D, for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA*. 2001;285:1987-91.
43. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med*. 2007;26:3078-94.
44. Austin PC, Grootendorst P, Normand SLT, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Stat Med*. 2007;26:754-68.
45. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*. 1987;125:761-8.

E-References (systematic review articles)

- E1. Rice TW, Khuntia D, Rybicki LA, Adelstein DJ, Vogelbaum MA, Mason DP, et al. Brain metastases from esophageal cancer: a phenomenon of adjuvant therapy? *Ann Thorac Surg.* 2006;82:2042-9, 2049. e1-2.
- E2. Koch CG, Li L, Van Wagoner DR, Duncan AI, Gillinov AM, Blackstone EH. Red cell transfusion is associated with an increased risk for postoperative atrial fibrillation. *Ann Thorac Surg.* 2006;82:1747-56.
- E3. Kunihiro T, Schmidt K, Glombitza P, Dzindzibadze V, Lausberg H, Schafers HJ. Root replacement using stentless valves in the small aortic root: a propensity score analysis. *Ann Thorac Surg.* 2006;82:1379-84.
- E4. Mishra M, Malhotra R, Karlekar A, Mishra Y, Trehan N. Propensity case-matched analysis of off-pump versus on-pump coronary artery bypass grafting in patients with atheromatous aorta. *Ann Thorac Surg.* 2006;82:608-14.
- E5. DeCamp MM, Blackstone EH, Naunheim KS, Krasna MJ, Wood DE, Meli YM, et al; NETT Research Group. Patient and surgical factors influencing air leak after lung volume reduction surgery: lessons learned from the National Emphysema Treatment Trial. *Ann Thorac Surg.* 2006;82:197-206.
- E6. Di Mauro M, Di Giammarco G, Vitolla G, Contini M, Iaco AL, Bivona A, et al. Impact of no-to-moderate mitral regurgitation on late results after isolated coronary artery bypass grafting in patients with ischemic cardiomyopathy. *Ann Thorac Surg.* 2006;81:2128-34.
- E7. Brunelli A, Xiume' F, Al Refai M, Salati M, Marasco R, Sabbatini A. Gemcitabine-cisplatin chemotherapy before lung resection: a case-matched analysis of early outcome. *Ann Thorac Surg.* 2006;81:1963-8.
- E8. Baskett RJ, Cafferty FH, Powell SJ, Kinsman R, Keogh BE, Nashef SA. Total arterial revascularization is safe: multicenter ten-year analysis of 71,470 coronary procedures. *Ann Thorac Surg.* 2006;81:1243-8.
- E9. Olsson C, Thelin S. Antegrade cerebral perfusion with a simplified technique: unilateral versus bilateral perfusion. *Ann Thorac Surg.* 2006;81:868-74.
- E10. Toumpoulis IK, Anagnostopoulos CE, Balaram S, Swistel DG, Ashton RC Jr, DeRose JJ Jr. Does bilateral internal thoracic artery grafting increase long-term survival of diabetic patients? *Ann Thorac Surg.* 2006;81:599-606.
- E11. Chukwuemeka A, Weisel A, Maganti M, Nette AF, Wijesundera DN, Beattie WS, et al. Renal dysfunction in high-risk patients after on-pump and off-pump coronary artery bypass surgery: a propensity score analysis. *Ann Thorac Surg.* 2005;80:2148-53.
- E12. Sabik JF 3rd, Blackstone EH, Houghtaling PL, Walts PA, Lytle BW. Is reoperation still a risk factor in coronary artery bypass surgery? *Ann Thorac Surg.* 2005;80:1719-27.
- E13. Calafiore AM, Di Mauro M, Di Giammarco G, Teodori G, Iaco AL, Mazzei V, et al. Single versus bilateral internal mammary artery for isolated first myocardial revascularization in multivessel disease: long-term clinical results in medically treated diabetic patients. *Ann Thorac Surg.* 2005;80:888-95.
- E14. Gillinov AM, Blackstone EH, Rajeswaran J, Mawad M, McCarthy PM, Sabik JF 3rd, et al. Ischemic versus degenerative mitral regurgitation: does etiology affect survival? *Ann Thorac Surg.* 2005;80:811-9.
- E15. Lawton JS, Barner HB, Bailey MS, Guthrie TJ, Moazami N, Pasque MK, et al. Radial artery grafts in women: utilization and results. *Ann Thorac Surg.* 2005;80:559-63.
- E16. Louagie YA, Jamart J, Gruslin A. Do coronary bypass graft flows differ between on-pump and off-pump operations? *Ann Thorac Surg.* 2005;79:2004-12.
- E17. Habib RH, Zacharias A, Schwann TA, Riordan CJ, Durham SJ, Shah A. Effects of obesity and small body size on operative and long-term outcomes of coronary artery bypass surgery: a propensity-matched analysis. *Ann Thorac Surg.* 2005;79:1976-86.
- E18. van Domburg RT, Takkenberg JJ, Noordzij LJ, Saia F, van Herwerden LA, Serruys PW, et al. Late outcome after stenting or coronary artery bypass surgery for the treatment of multivessel disease: a single-center matched-propensity controlled cohort study. *Ann Thorac Surg.* 2005;79:1563-9.
- E19. Okereke I, Murthy SC, Alster JM, Blackstone EH, Rice TW. Characterization and importance of air leak after lobectomy. *Ann Thorac Surg.* 2005;79:1167-73.
- E20. Lam BK, Gillinov AM, Blackstone EH, Rajeswaran J, Yuh B, Bhudia SK, et al. Importance of moderate ischemic mitral regurgitation. *Ann Thorac Surg.* 2005;79:462-70.
- E21. Di Mauro M, Iaco AL, Contini M, Teodori G, Vitolla G, Pano M, et al. Reoperative coronary artery bypass grafting: analysis of early and late outcomes. *Ann Thorac Surg.* 2005;79:81-7.
- E22. Lytle BW, Blackstone EH, Sabik JF, Houghtaling P, Loop FD, Cosgrove DM. The effect of bilateral internal thoracic artery grafting on survival during 20 postoperative years. *Ann Thorac Surg.* 2004;78:2005-12.
- E23. Svensson LG, Blackstone EH, Rajeswaran J, Sabik JF 3rd, Lytle BW, Gonzalez-Stawinski G, et al. Does the arterial cannulation site for circulatory arrest influence stroke risk? *Ann Thorac Surg.* 2004;78:1274-84.
- E24. Baskett RJ, Kalavrouziotis D, Buth KJ, Hirsch GM, Sullivan JA. Training residents in mitral valve surgery. *Ann Thorac Surg.* 2004;78:1236-40.
- E25. Diodato MD, Moon MR, Pasque MK, Barner HB, Moazami N, Lawton JS, et al. Repair of ischemic mitral regurgitation does not increase mortality or improve long-term survival in patients undergoing coronary artery revascularization: a propensity analysis. *Ann Thorac Surg.* 2004;78:794-9.
- E26. Karthik S, Grayson AD, McCarron EE, Pullan DM, Desmond MJ. Reexploration for bleeding after coronary artery bypass surgery: risk factors, outcomes, and the effect of time delay. *Ann Thorac Surg.* 2004;78:527-34.
- E27. Quader MA, McCarthy PM, Gillinov AM, Alster JM, Cosgrove DM 3rd, Lytle BW, et al. Does preoperative atrial fibrillation reduce survival after coronary artery bypass grafting? *Ann Thorac Surg.* 2004;77:1514-22.
- E28. Sabik JF, Nemeh H, Lytle BW, Blackstone EH, Gillinov AM, Rajeswaran J, et al. Cannulation of the axillary artery with a side graft reduces morbidity. *Ann Thorac Surg.* 2004;77:1315-20.
- E29. Coselli JS, LeMaire SA, Conklin LD, Adams GJ. Left heart bypass during descending thoracic aortic aneurysm repair does not reduce the incidence of paraplegia. *Ann Thorac Surg.* 2004;77:1298-303.
- E30. Guller U, Anstrom KJ, Holman WL, Allman RM, Sansom M, Peterson ED. Outcomes of early extubation after bypass surgery in the elderly. *Ann Thorac Surg.* 2004;77:781-8.
- E31. Khan JH, Lambert AM, Habib JH, Broce M, Emmett MS, Davis EA. Abdominal complications after heart surgery. *Ann Thorac Surg.* 2006;82:1796-801.
- E32. Reddy SL, Grayson AD, Oo AY, Pullan MD, Poonacha T, Fabri BM. Does off-pump surgery offer benefit in high respiratory risk patients? A respiratory risk stratified analysis in a propensity-matched cohort. *Eur J Cardiothorac Surg.* 2006;30:126-31.
- E33. Reeves BC, Ascione R, Caputo M, Angelini GD. Morbidity and mortality following acute conversion from off-pump to on-pump coronary surgery. *Eur J Cardiothorac Surg.* 2006;29:941-7.
- E34. Pai KR, Ramnarine IR, Grayson AD, Mediratta NK. The effect of chronic steroid therapy on outcomes following cardiac surgery: a propensity-matched analysis. *Eur J Cardiothorac Surg.* 2005;28:138-42.
- E35. Ali IS, Buth KJ. Preoperative statin use and in-hospital outcomes following heart surgery in patients with unstable angina. *Eur J Cardiothorac Surg.* 2005;27:1051-6.
- E36. Frankel TL, Stamou SC, Lowery RC, Kapetanakis EI, Hill PC, Haile E, et al. Risk factors for hemorrhage-related reexploration and blood transfusion after conventional versus coronary revascularization without cardiopulmonary bypass. *Eur J Cardiothorac Surg.* 2005;27:494-500.
- E37. Brunelli A, Sabbatini A, Xiume' F, Borri A, Salati M, Marasco RD, et al. Inability to perform maximal stair climbing test before lung resection: a propensity score analysis on early outcome. *Eur J Cardiothorac Surg.* 2005;27:367-72.
- E38. Pandey R, Grayson AD, Pullan DM, Fabri BM, Dihmis WC. Total

- arterial revascularisation: effect of avoiding cardiopulmonary bypass on in-hospital mortality and morbidity in a propensity-matched cohort. *Eur J Cardiothorac Surg*. 2005;27:94-8.
- E39. Stamou SC, Jablonski KA, Garcia JM, Boyce SW, Bafi AS, Corso PJ. Operative mortality after conventional versus coronary revascularization without cardiopulmonary bypass. *Eur J Cardiothorac Surg*. 2004;26:549-53.
- E40. Calafiore AM, Di Giammarco G, Teodori G, Di Mauro M, Iaco AL, Bivona A, et al. Late results of first myocardial revascularization in multiple vessel disease: single versus bilateral internal mammary artery with or without saphenous vein grafts. *Eur J Cardiothorac Surg*. 2004;26:542-8.
- E41. Ben-Gal Y, Mohr R, Uretzky G, Medalion B, Hendler A, Hansson N, et al. Drug-eluting stents versus arterial myocardial revascularization in patients with diabetes mellitus. *J Thorac Cardiovasc Surg*. 2006;132:861-6.
- E42. Guru V, Fremes SE, Tu JV. How many arterial grafts are enough? A population-based study of midterm outcomes *J Thorac Cardiovasc Surg*. 2006;131:1021-8.
- E43. Clark LL, Ikonomidis JS, Crawford FA Jr, Crumbley A 3rd, Kratz JM, Stroud MR, et al. Preoperative statin treatment is associated with reduced postoperative mortality and morbidity in patients undergoing cardiac surgery: an 8-year retrospective cohort study. *J Thorac Cardiovasc Surg*. 2006;131:679-85.
- E44. Svensson LG, Mumtaz MA, Blackstone EH, Feng J, Banbury MK, Sabik JF 3rd, et al. Does use of a right internal thoracic artery increase deep wound infection and risk after previous use of a left internal thoracic artery? *J Thorac Cardiovasc Surg*. 2006;131:609-13.
- E45. Smedira NG, Blackstone EH, Roselli EE, Laffey CC, Cosgrove DM. Are allografts the biologic valve of choice for aortic valve replacement in nonelderly patients? Comparison of explantation for structural valve deterioration of allograft and pericardial prostheses *J Thorac Cardiovasc Surg*. 2006;131:558-564. e4.
- E46. Stamou SC, Hill PC, Haile E, Prince S, Mack MJ, Corso PJ. Clinical outcomes of nonelective coronary revascularization with and without cardiopulmonary bypass. *J Thorac Cardiovasc Surg*. 2006;131:28-33.
- E47. Rajakaruna C, Rogers CA, Angelini GD, Ascione R. Risk factors for and economic implications of prolonged ventilation after cardiac surgery. *J Thorac Cardiovasc Surg*. 2005;130:1270-7.
- E48. Kuniyara T, Grun T, Aicher D, Langer F, Adam O, Wendler O, et al. Hypothermic circulatory arrest is not a risk factor for neurologic morbidity in aortic surgery: a propensity score analysis. *J Thorac Cardiovasc Surg*. 2005;130:712-8.
- E49. Roselli EE, Murthy SC, Rice TW, Houghtaling PL, Pierce CD, Karchmer DP, et al. Atrial fibrillation complicating lung cancer resection. *J Thorac Cardiovasc Surg*. 2005;130:438-44.
- E50. Calafiore AM, Di Giammarco G, Teodori G, Iaco AL, Pano M, Contini JB, et al. Bilateral internal thoracic artery grafting with and without cardiopulmonary bypass: six-year clinical outcome. *J Thorac Cardiovasc Surg*. 2005;130:340-5.
- E51. Ercan S, Rice TW, Murthy SC, Rybicki LA, Blackstone EH. Does esophagogastric anastomotic technique influence the outcome of patients with esophageal cancer? *J Thorac Cardiovasc Surg*. 2005;129:623-31.
- E52. Kumpati GS, Cook DJ, Blackstone EH, Rajeswaran J, Abdo AS, Young JB, et al. HLA sensitization in ventricular assist device recipients: does type of device make a difference? *J Thorac Cardiovasc Surg*. 2004;127:1800-7.
- E53. Kim K, Rice TW, Murthy SC, DeCamp MM, Pierce CD, Karchmer DP, et al. Combined bronchoscopy, mediastinoscopy, and thoracotomy for lung cancer: who benefits? *J Thorac Cardiovasc Surg*. 2004;127:850-6.
- E54. Wijesundera DN, Beattie WS, Rao V, Ivanov J, Karkouti K. Calcium antagonists are associated with reduced mortality after cardiac surgery: a propensity analysis. *J Thorac Cardiovasc Surg*. 2004;127:755-62.
- E55. Sharony R, Grossi EA, Saunders PC, Galloway AC, Applebaum R, Ribakove GH, et al. Propensity case-matched analysis of off-pump coronary artery bypass grafting in patients with atheromatous aortic disease. *J Thorac Cardiovasc Surg*. 2004;127:406-13.
- E56. Mack MJ, Pfister A, Bachand D, Emery R, Magee MJ, Connolly M, et al. Comparison of coronary bypass surgery with and without cardiopulmonary bypass in patients with multivessel disease. *J Thorac Cardiovasc Surg*. 2004;127:167-73.
- E57. Legare JF, Buth KJ, Sullivan JA, Hirsch GM. Composite arterial grafts versus conventional grafting for coronary artery bypass grafting. *J Thorac Cardiovasc Surg*. 2004;127:160-6.
- E58. Sabik JF, Blackstone EH, Lytle BW, Houghtaling PL, Gillinov AM, Cosgrove DM. Equivalent midterm outcomes after off-pump and on-pump coronary surgery. *J Thorac Cardiovasc Surg*. 2004;127:142-8.
- E59. Koch CG, Khandwala F, Cywinski JB, Ishwaran H, Estafanous FG, Loop FD, et al. Health-related quality of life after coronary artery bypass grafting: a gender analysis using the Duke Activity Status Index. *J Thorac Cardiovasc Surg*. 2004;128:284-95.
- E60. Kaw R, Golish J, Ghamande S, Burgess R, Foldvary N, Walker E. Incremental risk of obstructive sleep apnea on cardiac surgical outcomes. *J Cardiovasc Surg*. 2006;47:683-9.

Table E1. Components of a propensity score-matched analysis

Step	Analytic component
1	Describe how the propensity score model was specified. <ul style="list-style-type: none">● Describe how variables were selected for consideration for inclusion in the propensity score model.● Describe how the propensity score model was formulated.
2	Explicitly describe how the matched sets were formed. <ul style="list-style-type: none">● Was matching done with or without replacement?● Was greedy or optimal matching used?● What was the width of the calipers for the matching method?
3	Report the distribution of baseline variables in treated and untreated subjects in the matched sample.
4	Compare balance in baseline variables between treated and untreated subjects. <ul style="list-style-type: none">● Do not use statistical significance testing.● Use methods, such as standardized differences, that are not affected by sample size and that are a property of the sample.
5	Explicitly describe statistical methods to estimate the effect of treatment on the outcome. This method must account for the matched nature of the sample.

TABLE E2. Statistical methods for estimating treatment effect for different outcomes

Outcome	Statistical method
Continuous	Paired <i>t</i> test or Wilcoxon signed rank test <ul style="list-style-type: none">● Can adjust for residual imbalance in covariates using linear regression model that accounts for the matched-pair nature of the data using GEE methods.
Binary (dichotomous)	Risk differences: McNemar test Relative risks: methods proposed by Agresti and Min. ²⁸ <ul style="list-style-type: none">● Can adjust for residual imbalance in baseline covariates using logistic regression model estimated using GEE methods to account for matched-pairs design.
Time-to-event (survival)	<ul style="list-style-type: none">● Comparison of Kaplan-Meier survival curves using the test of Klein and Moeschberger.²⁹● Cox proportional hazards model stratified on matched pairs.● Cox proportional hazards model with robust variance estimator to account for matching.

GEE, Generalized estimating equation.