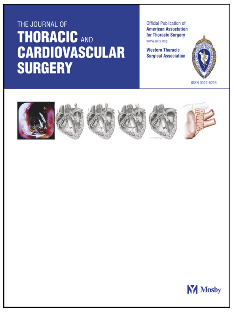

Propensity Scores: Methods, Considerations, and Applications in the Journal of Thoracic and Cardiovascular Surgery

Timothy L. McMurry, PhD, Yinin Hu, MD, Eugene H. Blackstone, MD, Benjamin D. Kozower, MD, MPH

PII: S0022-5223(15)00508-5

DOI: 10.1016/j.jtcvs.2015.03.057

Reference: YMTC 9484

To appear in: *The Journal of Thoracic and Cardiovascular Surgery*

Received Date: 16 March 2015

Accepted Date: 20 March 2015

Please cite this article as: McMurry TL, Hu Y, Blackstone EH, Kozower BD, Propensity Scores: Methods, Considerations, and Applications in the Journal of Thoracic and Cardiovascular Surgery, *The Journal of Thoracic and Cardiovascular Surgery* (2015), doi: 10.1016/j.jtcvs.2015.03.057.

Title: Propensity Scores: Methods, Considerations, and Applications in the Journal of Thoracic and Cardiovascular Surgery

Date: March 12, 2015

Authors:

Timothy L. McMurry, PhD[1]; Yinin Hu, MD[2]; Eugene H. Blackstone, MD[3]; Benjamin D. Kozower, MD, MPH[1,2]

Affiliations:

[1]Department of Public Heath Sciences, University of Virginia, Charlottesville, VA 22908.

[2]Department of Surgery/Division of Thoracic Surgery, University of Virginia Health System, Charlottesville, VA 22908.

[3]Department of Thoracic and Cardiovascular Surgery and Department of Quantitative Health Sciences, Research Institute, Cleveland Clinic, Cleveland, OH 44195.

Keywords: Propensity score; matching; causal inference; observational studies; statistical methods

Word count: 3570

Funding: None

Conflicts of Interest: The authors have no conflict of interest.

Corresponding author:

Benjamin D. Kozower, MD, MPH
University of Virginia Health System
General Thoracic Surgery
P.O. Box 800679
Charlottesville, VA 22908-0679
E: bdk8g@virginia.edu
P: 1(434) 924-2145
F: 1(434) 244-9429

Structured Abstract

Objective: We review the published literature using propensity scoring, describe shortcomings in the use of this technique, and provide conceptual background for understanding and correctly implementing propensity matched studies.

Methods: We survey the published statistical literature and make recommendations for a set of standard criteria for propensity matching studies. We then applied these criterion to recent publications in the Journal of Thoracic and Cardiovascular Surgery and determined how well the standards were applied.

Results: We found that propensity matched studies are rarely documented well enough to be convincing in their results. When documentation is available, statistical shortcomings are common.

Conclusions: Improved statistical practice is needed when using propensity scoring. This article creates standard criterion for using this method in JTCVS publications.

Central Picture

Timothy L. McMurry, PhD

Central Message

We provide conceptual background and good practice recommendations for the use of propensity matching in surgical outcomes research.

Clinical Perspective

The frequent use of propensity matching requires that it be understood by clinicians and researchers. Retrospective studies comparing two treatments are hampered by selection bias; propensity scoring reduces the bias by selecting similar groups for comparison. Review of recent JTCVS publications suggests that the methods are not consistently applied. This article highlights the important concepts.

1. Introduction

Propensity scoring is a powerful tool to strengthen causal inferences drawn from observational studies. The motivation is simple: in order to compare the effects of two treatment options, which we generically refer to as “A” and “B” with B being the more common, we want to compare the outcomes of similar groups of patients receiving each treatment. Propensity scoring helps select similar patient groups for comparison.

Propensity scoring is common in the literature and the methodology is widely discussed[1-8]. Despite propensity scoring's popularity, we are concerned that its use is conceptually more intricate than many investigators realize. The consequence can results that are misleading or difficult for readers, referees, and investigators to evaluate objectively. These concerns persist despite having been raised previously in the cardiothoracic surgery literature.[9] The problem is compounded because recommendations in the methodological literature are not consistent (see Section 8).

In the present article, we review the basics of propensity scoring, highlight areas where practical recommendations are in general agreement or disagreement, discuss choices available to the investigator, examine how this family of techniques has been applied in recent Journal of Thoracic and Cardiovascular Surgery articles, and establish guidelines for future articles. Sections 2–9 provide conceptual background, Section 10 presents general guidelines, and Section 11 examines recent articles published in the Journal of Thoracic and Cardiovascular Surgery (JTCVS).

2. The Conceptual Framework for Propensity Matching

Many medical studies are designed to evaluate the effectiveness of a treatment in comparison to an alternative. The goal is to estimate how much better (or worse) the outcomes are for patients in one treatment group as compared with what would have happened had they received the other treatment. Importantly, these comparisons need to account for differences in patients that may have contributed to their allocation to one of the treatments.

The randomized controlled trial, where patients are randomly assigned to treatment groups, is the gold standard for this comparison. The importance of random assignment is that the patient characteristics affecting outcomes (e.g., age, sex, comorbidities, mental disposition) tend to be equally distributed across groups. Thus, significant differences in outcomes can be attributed to the treatment.

However, there are many questions that a randomized trial cannot address due to cost, ethical or practical considerations, or timeliness. For example, in a study of smoking related healthcare costs, patients cannot ethically (or practically) be randomly assigned to smoke or not smoke. [6] Moreover, a randomized study would take decades to run. Other variables of interest cannot be assigned; Koch et al. considered whether men and women fare differently after coronary artery bypass grafting. [10] Many of these clinical questions can be addressed with robust observational data.

Unfortunately, in observational data, patients receiving treatment A typically differ systematically from those receiving B, leaving direct comparisons of outcomes heavily confounded. For example, patients who receive surgery are deemed well enough before treatment to survive the surgery, while extremely frail patients may be deemed inoperable. It is not fair or appropriate to compare outcomes between such dissimilar patients.

Historically, investigators tried to account for differences between groups by using multiple regression to adjust for confounding characteristics. However, because the groups of patients may differ systematically, using a regression to estimate the potential effect of treatment A on a patient who received B can be an unreliable extrapolation. Therefore, it is imperative to only compare patients who are legitimate candidates for either procedure.

An intuitively appealing approach would be to match patients from one treatment group with patients from the other on a number of important characteristics, and then compare their outcomes. Unfortunately, matching on even a modest number of criteria often leaves a large majority of patients unmatched and unavailable for analysis, making the results less reliable. [11]

Propensity scores solve the problem of matching on multiple covariates by reducing them to a single quantity, the propensity score. A patient's propensity score is defined as the probability the patient receives treatment A (instead of B) given all relevant conditions, comorbidities, and other characteristics at the time the treatment decision is made. What makes propensity scores so powerful is that, under some conditions, patients with the same propensity score have the same probabilistic distribution of other covariates regardless of whether they received A or B. [12] As a result, it can be sufficient to compare the outcomes across treatment groups of pairs or pools of patients with similar propensity scores.

3. Steps in a propensity score analysis

There are four main steps in an analysis using propensity scores. First, the propensity scores must be estimated (Section 5). Second, the data need to be matched or grouped based on the estimated propensity scores (Section 6). Third, balance must assessed to ensure that the grouping produced similar pools of patients receiving treatments A and B (Section 7). Finally, data can be analyzed to estimate the treatment effect size and its clinical and statistical significance (Section 8). The first three of these steps are "design" steps used to frame a comparison around similar groups of patients; they must be performed without looking at the outcomes data. None of the steps can be adequately performed by following a simple recipe. Importantly, even before tackling these technical issues, two crucial assumptions must be met in order for propensity matching to provide useful results.

4. Crucial assumptions

There are two conditions on the data which must be met for analyses based on propensity scores to provide valid results. The most important condition is known as "strong ignorability," which technically says a patient's treatment assignment (A or B) is independent of his/her potential outcomes under the two treatment scenarios given the covariates. In other words, the observed covariates contain all the information about the patient's condition that is relevant to the patient's potential outcomes. Strong ignorability makes intuitive sense. If the goal is to compare similar groups of patients receiving different treatments, we need to know all the factors that determine whether patients are comparable at the time of treatment allocation.

The second important condition states that, given the covariates, the patient needs to have a positive probability of receiving either of the treatments. Intuitively, there is no gain in asking what the potential benefits of surgery are for a patient whose comorbidities preclude surviving an operation. The only interest is in comparing patients for whom both treatments are realistic.

5. Constructing the propensity score model

The first step is to estimate the propensity scores for each patient. The most common approach is to use logistic regression, but there are other regression models that can estimate classification probabilities.[13,14]

Which variables to include

Guidelines for constructing and evaluating a regression model depend on its intended application. [15] For propensity scores, the "strong ignorability" condition necessitates inclusion of covariates that predict potential outcomes under either treatment scenario as well as any covariates that predict treatment assignment, although these two criteria are typically related. From a practical point of view, the second of these requirements deserves particular emphasis: if the data do not contain the information used to make the treatment decision (or are systematically missing in one of the two patient groups), the propensity model will be inadequate, and all subsequent analyses will be suspect.

The consensus is that if the sample size is too small for the propensity score model to include all variables of interest, it is most important to include the variables that are strongly related to outcome. [16] These should be selected *a priori* based on scientific understanding and previous literature, and without reference to the outcomes within the data set. [17] It is probably better to err on the side of too many predictors rather than too few, [7] and when data sizes are large, good propensity models can contain many predictors.[6]

Logistic regression model diagnostics

The goal of the propensity score model is to create balanced groups of patients receiving each treatment. Therefore, some model evaluation tools, such as those evaluating discriminative ability (e.g., the *c*-statistic), multicollinearity, and model selection are only of secondary importance. [18] The crucial diagnostic step is to compare covariate balance between the resulting groups of patients receiving the two treatments; see Section 7.

A model which accurately estimates the likelihood of treatment allocation is the key to achieving this balance. Nonetheless, some metrics which are common in many regression applications are of diminished importance in the present context. For example, multicollinearity occurs when highly correlated predictors produce instability in their corresponding coefficients. Fortunately, multicollinearity does not affect the resulting fitted values, in this case, the propensity scores. However, if the sample size is limited, it may still be advantageous to remove highly correlated variables in order to include less correlated covariates.

There is also concern about traditional model selection strategies such as stepwise variable selection. These approaches are designed for prediction rather than covariate balance. The concern is that these selection methods might remove variables that are weakly related to treatment assignment but strongly related to outcome[16], while variables related to outcome are thought to be at least as important.

The commonly used *c*-statistic also requires nuanced interpretation in this setting. In most applications, a predictive model with a low *c*-statistic is useless. A propensity model with a low *c*-statistic could be caused by poor construction. However, it could also be indicative of differences in practice that are not related to patient condition. The first of these problems invalidates subsequent analyses, while the latter can be beneficial. To illustrate: imagine trying to estimate propensity scores for a randomized trial. A well constructed model accounting for all relevant clinical covariates will have a *c*-statistic around 0.5, and all patients should have similar propensity scores. Nonetheless, a randomized trial is ideally suited for causal inference. At the other extreme, a *c*-statistic close to 1 indicates that the regression model is able to differentiate patients receiving A from those receiving B. This may be an indication that the these two groups are so different that their outcomes will be difficult to meaningfully compare.

6. Grouping the data

Once propensity scores have been estimated, the data are typically grouped by either subclassification (sometimes called stratification) or matching. Both of these methods prune the original data set down to groups or sets of patients with similar propensity scores. While there are other approaches, we focus on these two for their simplicity and frequency of use.

It may be reasonable to remove patients receiving one treatment who have propensity scores either much larger or much smaller than any patient receiving the other treatment, the “oranges” as discussed in Blackstone. [2] The rationale for exclusion is that these patients do not appear to have been candidates for the alternative treatment. Nonetheless, excluded patients should be examined carefully. If there are many unmatchable patients, the propensity score model may include a variable that is a strong surrogate for treatment assignment, which may be removed. [2] Evaluation of the “oranges” will also help reveal the limits within which a valid comparison of the two treatments is possible.

Subclassification

Subclassification is frequently suggested in the methodological literature but less frequently applied. The idea is simple: propensity scores are grouped into (e.g.) quintiles or deciles (5 to 10 groups is typical). [19] Within each group, the propensity scores are similar, so grouped patients should have similar covariate distributions, and thus can be compared to each other. An analysis is performed in each group, and then results are aggregated. Subclassification has intuitive appeal because it focuses comparisons on pools of patients with similar propensity scores. In contrast, if patients are matched, there may be many suitable matches for a group A patient, with some potential matches arbitrarily excluded from final comparisons.

Matching

The more common approach is to match individual patients receiving one treatment to patients with similar propensity scores receiving the other. While conceptually simple, the details lead to different algorithms which can affect subsequent analyses. These variations include the methods for measuring distances between propensity scores, the threshold for what constitutes matching scores, how one match is chosen from many candidates, the number of patients in group B (the larger group) matched to each patient in A, and whether or not a single patient in group B can be matched to more than one individual in group A.

Intuition suggests that the distance between propensity scores should be measured by the simple difference between estimated probabilities of treatment. This approach is commonly used, however there is evidence that it is more effective to match on the “linear propensity score,” or the difference between propensity scores on the logit scale.[20]

One needs to decide how close two propensity scores need to be before they can be potential matches; this threshold is known as a “caliper.” A narrow caliper can prevent inaccurate matching, but if too many patients go unmatched the results can become uninterpretable. The appropriate caliper size depends on the relative variations of propensity scores in the two treatment groups.[20,21]

Next, the user must decide how many patients in group B should be matched to each patient in A. The most common approach is to match each patient in A to a single patient in B. If group B is much larger than A, it may be advantageous to match a larger number of patients in B, but the benefits are reduced if the extra matches are of poor quality. Finally, it is possible to re-use patients in group B. This makes the matching process independent of the order in which the matches are selected and may improve the overall match quality. However, without adjustment, reused patients have too much weight in the final comparison of outcomes.

Once these decisions have been made, pairing is often done by a “greedy” algorithm. The group A patients are randomly ordered. The first of these randomly ordered patients is then matched to their best group B counterpart. The group B patient is removed from the set of potential future matches, and the process is iterated. An alternative to “greedy” matching is “optimal” matching, [4] which seems to produce better matched pairs, but does not substantially improve the balance of the matched groups as a whole. [22] Stuart maintains a web page describing available software. [23]

7. Assessment of covariate balance in matched groups

Since the goal in using a propensity scores is to create pools of similar patients for comparison, it is extremely important to assess the post-matching similarity across groups before any assessment of outcomes. In particular, all covariates affecting patients’ prognoses prior to treatment and indications for treatment need to be compared. If clinically-relevant differences remain post-match, subsequent analyses are unreliable.

The types of comparisons differ depending on how the data are grouped. If the data are subclassified, then one should perform diagnostics within each subclass. If the data are matched, then typically comparison is between the matched pools of patients.

Most investigators assess covariate balance using hypothesis tests. For example, an investigator might test the hypothesis that the average age of patients in group A is the same as their matched counterparts in B. Unfortunately, hypothesis tests answer the wrong question. The *p*-value from a hypothesis test depends on the difference between the two groups and their sample sizes. However, only the difference between the two groups is relevant to covariate balance.[24]

A better metric for continuous covariates would be to use a measure that does not depend on sample size, such as the standardized difference in means: $(\bar{X}_A - \bar{X}_B)/\sigma_A$[20], which expresses the difference between the two groups in standard deviations. The improvement in balance achieved by matching can be demonstrated by comparing standardized differences in means before and after matching (using the same estimate for σ_A in both quantities). Binary covariates can be compared with a simple difference in proportions or by a similar standardized difference. [25] Alternatively, Rubin suggests a set of powerful but less intuitive diagnostics. [17]

8. Analysis of the matched data

The final step in a propensity score analysis is to estimate the treatment effect size and its clinical and statistical significance. Literature on the proper analysis of matched data is sparse and occasionally in conflict. For example, Rosenbaum recommends analyzing the data with permutation tests in the same way one would analyze an unmatched observational trial. [4] Austin argues that propensity matched data should be analyzed using procedures for matched analyses, such as paired *t*-tests, and McNemar's test. [26] Stuart replies that matched analyses are not necessary, and that the data can be analyzed using a standard regression that includes a treatment indicator and the variables used in the matching. [27] A recent article by Li and Greene suggests that a weighting method is optimal. [28] Many articles make almost no mention of statistical inference.

This confusion has resulted because statistical understanding is still evolving, and assumptions made about the data and matching process can alter the estimates' derived properties. In most cases relevant to surgical outcomes, regression is defensible and even recommendable. Propensity scores provide an objective way to restrict the domain of analysis to patients who are legitimate candidates for either procedure. Outcomes in the two groups are then compared using a regression model that controls for all covariates used in matching, plus a treatment indicator variable. The coefficient associated with this indicator is interpreted as the treatment effect. An advantage of regression is that it provides some level of "double robustness" by adjusting for any remaining small covariate imbalances. [29] It is for this reason that even randomized trials are sometimes analyzed with regression models.

Regression is more important following subclassification because, within subclasses, meaningful covariate imbalances may still remain. In this setting, recommendations for the regression remain similar. If enough data are available, a regression model containing the treatment indicator and all covariates can be fit in each subclass, and the results combined. If data are more limited, a single regression model may be fit containing subclass indicators and subclass by treatment interactions along with the other covariates. This keeps the covariate relationships fixed but allows different size treatment effects across the subclasses. After regression modeling, subclass-specific treatment effects are then combined by a weighted average of the treatment effects in each subclass, where the effects are typically weighted by the number of group A individuals in the subclass.

9. Interpretation

For matched data, patients receiving A have been grouped with a probabilistically similar pool of group B patients. Therefore, the estimated effect size represents the average improvement of the group A patients relative to similar patients in group B. This quantity is traditionally described in the literature as the Average Treatment Effect in the Treated (ATT). ATT is not the same as the average effect of treatment across the entire population, referred to in the literature as Average Treatment Effect (ATE). In most cases, we suspect ATT is the desired quantity, as it describes the benefits/risks of A relative to similar patients receiving B, rather than as a potential benefit averaged across all patients.

Both ATT and ATE assume that all group A patients in the initial data set were included in the final analyzed groups. If many patients have been excluded, the interpretation may change or results may become uninterpretable; see Section 6.

10. Recommendations for published literature

While we recognize the importance of brevity, it is important that propensity scoring methods be described well enough that results can be evaluated and replicated. Most of our recommendations can be implemented with one or two paragraphs. In some cases, additional tables may be provided in on-line appendices.

While different analyses are appropriate for different data sets and clinical questions, we propose that articles utilizing propensity matching should include the following:

1. The original sample sizes for the pools of patients in each group.
2. The sample sizes available after matching.
3. The type of regression model used to estimate the propensity scores.
4. The variables considered for inclusion in the propensity model, the variables included in the final model, and the inclusion criteria.
5. The type of matching algorithm used.
6. Diagnostics demonstrating the quality of the resulting matches.
7. Characterization of the unmatched patients.
8. An indication of the statistical procedures used for analyses.

11. JTCVS literature review

We reviewed all publications in JTCVS from 2013 and 2014 using propensity score matching. We found 25 such articles in 2013 and 64 in 2014. While many of these papers were well done, some exhibited substantial statistical shortcomings, and many did not provide enough detail for objective evaluation. Our results are summarized in Table 1. Notably, many papers showed evidence of inadequate covariate balance after matching, and no paper carefully evaluated the excluded patients.

12. Conclusions

Propensity matching is a powerful tool for observational data analyses because it facilitates the comparison of outcomes between similar groups of patients. Although propensity matching has become a popular technique, the methodology is acqtually quite complex. This review is intended to help surgeons understand the concepts behind propensity matching which may influence their own research and/or help them critically evaluate the published literature. We identified eight criteria which we feel should be reported in any manuscript using propensity matching. When we applied these criteria to the publications in JTCVS from 2013 and 2014, concerns were raised about the use of this methodology and appropriateness of the applications. We recommend that the Journal adopt these criteria to create a standard for future articles submitted to JTCVS using propensity matching.

References

1. Blackstone EH. Breaking down barriers: Helpful breakthrough statistical methods you need to understand better. *J Thorac Cardiovasc Surg*. 2001;122(3):430-439.

2. Blackstone EH. Comparing apples and oranges. *J Thorac Cardiovasc Surg*. 2002;123(1):8-15.

3. Brookhart MA, Wyss R, Layton JB, Sturmer T. Propensity score methods for confounding control in nonexperimental research. *Circ Cardiovasc Qual Outcomes*. 2013;6(5):604-611. doi: 10.1161/CIRCOUTCOMES.113.000359 [doi].

4. Rosenbaum PR. *Observational studies.* 2nd ed. New York: Springer; 2002.

5. Rubin D. *Matched sampling for casual effects.* Vol 10. New York: Cambridge University Press; 2006:12.

6. Rubin DB. The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Stat Med*. 2007;26(1):20-36.

7. Stuart EA. Matching methods for causal inference: A review and a look forward. *Stat Sci*. 2010;25(1):1-21.

8. d'Agostino RB. Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*. 1998;17(19):2265-2281.

9. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: A systematic review and suggestions for improvement. *J Thorac Cardiovasc Surg*. 2007;134(5):1128-1135. e3.

10. Koch CG, Khandwala F, Nussmeier N, Blackstone EH. Gender and outcomes after coronary artery bypass grafting: A propensity-matched comparison. *J Thorac Cardiovasc Surg*. 2003;126(6):2032-2043.

11. Rosenbaum PR, Rubin DB. The bias due to incomplete matching. *Biometrics*. 1985:103-116.

12. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.

13. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010;29(3):337-346.

14. Westreich D, Lessler J, Funk MJ. Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol*. 2010;63(8):826-833.

15. Vittinghoff E, Shiboski S, McCulloch CE. *Regression methods in biostatistics.* 2nd ed. New York: Springer; 2010.

16. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163(12):1149-1156.

17. Rubin DB. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Serv Outcomes Res*. 2001;2(3-4):169-188.

18. Rubin DB. On principles for modeling propensity scores in medical research. *Pharmacoepidemiol Drug Saf*. 2004;13(12):855-857.

19. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*. 1984;79(387):516-524.

20. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*. 1985;39(1):33-38.

21. Cochran WG, Rubin DB. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*. 1973:417-446.

22. Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*. 1993;2(4):405-420.

23. Stuart E. Software for implementing matching methods and propensity scores. Elizabeth Stuart's Propensity Score Software Page Web site. http://www.biostat.jhsph.edu/~estuart/propensityscoresoftware.html. Updated 2012. Accessed 9/5, 2014.

24. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A*. 2008;171(2):481-502.

25. Austin PC. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Communications in Statistics-Simulation and Computation*. 2009;38(6):1228-1234.

26. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med*. 2008;27(12):2037-2049.

27. Stuart EA. Developing practical recommendations for the use of propensity scores: Discussion of ‘A critical appraisal of propensity score matching in the medical literature between 1996 and 2003’by peter austin, statistics in medicine. *Stat Med*. 2008;27(12):2062-2065.

28. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. *The International Journal of Biostatistics*. 2013;9(2):215-234.

29. Rubin DB. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*. 1973:185-203.

Table 1: Characteristics of recent (2013-2014) JTCVS papers using propensity scores.

Criteria	**Number (%) of Papers Providing Information**
Sample size for original data set	87/89 (98%)
Matched sample size	81/89 (91%)
Type of regression model used to estimate the propensity score	79/89 (89%)
Matching algorithm	60/89 (67%)
Analysis of covariate balance	66/89 (74%)
Evidence of inadequate covariate balance	17/66 (26%)
Comparison of matched to unmatched patients	0/89 (0%)
Type of statistical procedure	Number of papers
Univariate, independent samples	59/89 (66%)
Univariate, paired	11/89 (12%)
Regression after matching	31/89 (35%)
Regression including the propensity score as a covariate	10/89 (11%)