Propensity scores: Methods, considerations, and applications in the
Journal of Thoracic and Cardiovascular Surgery

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

ABSTRACT

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

Methods: We survey the published statistical literature and make recommenda-
tions for a set of standard criteria for studies that use propensity matching. We
evaluated adherence to these criteria in recent publications in the Journal of
Thoracic and Cardiovascular Surgery and determined how well the standards
were applied.

Results: We found that studies that use propensity matching are rarely docu-
dmented 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 suggests standard criteria for using this method in Journal
publications. (J Thorac Cardiovasc Surg 2015;150:14-9)

Propensity scoring is a powerful tool to strengthen causal
inferences drawn from observational studies. The motiva-
tion is simple: To compare the effects of 2 treatment
options, which we generically refer to as “A” and “B,”
with B being the more common one, we want to compare
the outcomes of similar groups of patients receiving each
treatment. Propensity scoring helps in selecting similar
patient groups for comparison.

Propensity scoring is common in the literature, and
the methodology is widely discussed.1-8 Despite the
popularity of propensity scoring, we are concerned that its
use is conceptually more intricate than many investigators
realize. The consequence can be results that are misleading,
or difficult for readers, referees, and investigators to
evaluate objectively. These concerns persist, despite the fact
that they have been raised previously in the cardiothoracic
surgery literature.9 The problem is compounded by inconsis-
tent recommendations in the methodologic literature (see
later section: Analysis of the Matched Data).

In the present article, we review the basics of propensity
scoring, highlight areas of general agreement and
disagreement on practical recommendations, discuss
choices available to the investigator, examine how this
family of techniques has been applied in recent Journal of
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 1 treatment group, compared with what would have happened had they received the other treatment. Such comparisons need to account for differences in patients that may have contributed to their allocation to 1 of the treatments.

The randomized controlled trial, in which 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 (eg, age, gender, comorbidities, mental disposition) tend to be equally distributed across groups. Thus, significant differences in outcomes can be attributed to the treatment.

However, many questions cannot be addressed by a randomized trial, owing to cost, ethical or practical considerations, and timeliness. For example, in a study of smoking-related healthcare costs, patients cannot ethically (or practically) be randomly assigned to smoke or not smoke. Moreover, a randomized study would take decades to run. Other variables of interest cannot be assigned; Koch and colleagues considered whether men and women fare differently after coronary artery bypass grafting. 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 comparison of outcomes heavily confounded. For example, patients who receive surgery are deemed well enough before treatment to survive the surgery, whereas extremely frail patients may be deemed inoperable. Comparing outcomes between such dissimilar patients is not fair or appropriate.

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 treatment B can be an unreliable extrapolation. Therefore, limiting comparison to only those patients who are legitimate candidates for either procedure is imperative.

An intuitively appealing approach would be to match patients from 1 treatment group with patients from the other on several important characteristics, and 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.

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 that 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 treatment A or B. As a result, a sufficient analysis can be a comparison of outcomes, across treatment groups, of pairs or pools of patients with similar propensity scores.

STEPS IN A PROPENSITY-SCORE ANALYSIS

An analysis using propensity scores has 4 main steps. First, the propensity scores must be estimated (later section: Constructing the Propensity-Score Model). Second, the data need to be matched or grouped based on the estimated propensity scores (later section: Grouping the Data). Third, balance must be assessed, to ensure that the grouping produced similar pools of patients receiving treatment A versus B (later section: Assessment of Covariate Balance in Matched Groups). Finally, data can be analyzed to estimate the treatment effect size and its clinical and statistical significance (later section: Analysis of the Matched Data).

The first 3 of these 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. However, before tackling these technical issues, 2 crucial assumptions must be met for propensity matching to provide useful results.

CRUCIAL ASSUMPTIONS

Two conditions on the data must be met for analyses based on propensity scores to provide valid results. The most important condition is known as “strong ignorability,” which means that the treatment assignment (A or B) is independent of a patient’s potential outcomes under the 2 treatment scenarios, given the covariates. In other words, the observed covariates contain all the information about the patient’s condition that is relevant to 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 condition is that, given the covariates, the patient needs to have a positive probability of receiving both treatments. Intuitively, this condition can be understood as the idea that there is no gain in asking what the potential benefits of surgery are for a patient whose comorbidities preclude survival of an operation. The only interest is in comparing patients for whom either treatment is realistic.

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 other regression models 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 2 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 1 of the 2 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, the most important to include are 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 dataset.17 Having too many predictors is probably better than having too few,5 and when sample 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 (eg, the c-statistic), multicollinearity, and model selection, are of only secondary importance.18 The crucial diagnostic step is to compare covariate balance between the resulting 2 groups of patients, each receiving 1 of the 2 treatments (see later section: Assessment of Covariate Balance in Matched Groups).

A model that accurately estimates the likelihood of treatment allocation is the key to achieving this balance. Nonetheless, some common metrics in many regression applications are less important 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, 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.

In addition, concern has been raised 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 even though variables related to outcome are considered at least as important.

The commonly used c-statistic 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 former problem invalidates subsequent analyses, whereas the latter can be beneficial. For example, imagine trying to estimate propensity scores for a randomized trial. A well constructed model accounting for all relevant clinical covariates has a c-statistic of approximately 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 treatment A from those receiving treatment B, indicating that the 2 groups may be so different that their outcomes are difficult to compare meaningfully.

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 dataset down to groups or sets of patients with similar propensity scores. While there are other approaches, we focus on these 2 for their simplicity and frequency of use.

Before grouping, it may be reasonable to remove patients receiving 1 treatment who have propensity scores that are either much larger or much smaller than any patient receiving the other treatment—the “oranges,” as discussed in Blackstone.5 The rationale for exclusion is that these patients do not seem to have been candidates for the alternative treatment. Nonetheless, excluded patients should be examined carefully. If many patients are unmatchable, the propensity-score model may include a variable that is a strong surrogate for treatment assignment,
which may be removed. In addition, evaluation of the “oranges” will help reveal the limits within which a valid comparison of the 2 treatments is possible.

**Subclassification**

Subclassification is frequently suggested in the methodologic literature but less frequently applied. The idea is simple: propensity scores are grouped, eg, into quintiles or deciles (5-10 groups is typical). Within each group, the propensity scores are similar, so grouped patients should have similar covariate distributions, and thus can be compared. An analysis is performed in each group, and 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, many matches may be suitable 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 1 treatment to patients with similar propensity scores receiving the other. Although 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 1 match is chosen from many candidates, the number of patients in group B (the larger group) matched to each patient in group A, and whether a single patient in group B can be matched to more than 1 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, evidence indicates that it is more effective to match on the “linear propensity score,” or the difference between propensity scores on the logit scale.

A decision must be made on how close 2 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 variances in the 2 treatment groups.

Next, the user must decide how many patients in group B should be matched to each patient in group A. The most common approach is to match each patient in group A to a single patient in group B. If group B is much larger than group A, matching a larger number of patients in group B may be advantageous, but the benefits are reduced if the extra matches are of poor quality. Finally, reuse of patients in group B is a possibility. This method 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 using a “greedy” algorithm. The group A patients are randomly ordered. The first of these randomly ordered patients is matched to their best group-B counterpart. The group-B patient is removed from the set of potential future matches, and the process is repeated. An alternative to “greedy” matching is “optimal” matching, which seems to produce better matched pairs, but does not substantially improve the balance of the matched groups as a whole. Stuart maintains a web page describing available software.

**ASSESSMENT OF COVARIATE BALANCE IN MATCHED GROUPS**

Given that the goal in using propensity scores is to create pools of similar patients for comparison, assessment of the postmatching similarity across groups is extremely important to complete, before any assessment of outcomes. In particular, all covariates affecting patients’ prognoses before treatment, and indications for treatment, need to be compared. If clinically relevant differences remain after matching, 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 group B. Unfortunately, hypothesis tests answer the wrong question. The P value from a hypothesis test depends on the difference between the 2 groups and their sample sizes. However, only the difference between the 2 groups is relevant to covariate balance.

A better metric for continuous covariates is a measure that does not depend on sample size, such as the standardized difference in means: \( \frac{\bar{X}_A - \bar{X}_B}{\sigma_A} \), which expresses the difference between the 2 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 \( \sigma_A \) in both quantities). Binary covariates can be compared with a simple difference in proportions, or by a similar standardized difference. Alternatively, Rubin suggests a set of powerful, but less intuitive, diagnostics.

**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. Austin argues that propensity-matched data should be analyzed using procedures for matched analyses, such as paired t tests, and McNemar's test. 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. A recent article by Li and Greene suggests that a weighting method is optimal. 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 2 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. For this reason, even randomized trials are sometimes analyzed with regression models.

Regression is more important following subclassification because, within subclasses, meaningful covariate imbalances may 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 approach 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, with the effects typically weighted by the number of group-A individuals in the subclass.

**INTERPRETATION**

For matched data, patients receiving treatment 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, which is not the same as the average effect of treatment across the entire population, referred to as the average treatment effect. In most cases, we suspect that the average treatment effect in the treated is the desired quantity, as it describes the benefits and risks of treatment A relative to those for similar patients receiving treatment B, rather than as a potential benefit averaged across all patients.

Both measures assume that all group-A patients in the initial dataset were included in the final analyzed groups. If many patients have been excluded, the interpretation may change, or results may become uninterpretable; see earlier section: Grouping the Data.

**RECOMMENDATIONS FOR PUBLISHED LITERATURE**

Although we recognize the importance of brevity, propensity-scoring methods must be described well enough that results can be evaluated and replicated. Most of our recommendations can be implemented with 1 or 2 paragraphs. In some cases, additional tables may be provided in online appendices.

Although different analyses are appropriate for different datasets and clinical questions, we propose that articles on studies utilizing propensity matching 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.

**JOURNAL OF THORACIC AND CARDIOVASCULAR SURGERY 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. Although many of these articles 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 articles showed evidence of inadequate covariate balance after matching, and no article carefully evaluated the excluded patients.
TABLE 1. Characteristics of 2013–2014 JTCVS papers using propensity scores

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Articles that provided information, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size for original dataset</td>
<td>87 (98)</td>
</tr>
<tr>
<td>Matched sample size</td>
<td>81 (91)</td>
</tr>
<tr>
<td>Type of regression model used to estimate the propensity score</td>
<td>79 (89)</td>
</tr>
<tr>
<td>Matching algorithm</td>
<td>60 (67)</td>
</tr>
<tr>
<td>Analysis of covariate balance</td>
<td>66 (74)</td>
</tr>
<tr>
<td>Evidence of inadequate covariate balance</td>
<td>17 (26)</td>
</tr>
<tr>
<td>(of the 66)</td>
<td></td>
</tr>
<tr>
<td>Comparison of matched to unmatched patients</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Type of statistical procedure</td>
<td></td>
</tr>
<tr>
<td>Univariate, independent samples</td>
<td>59 (66)</td>
</tr>
<tr>
<td>Univariate, paired</td>
<td>11 (12)</td>
</tr>
<tr>
<td>Regression after matching</td>
<td>31 (35)</td>
</tr>
<tr>
<td>Regression including the propensity score as a covariate</td>
<td>10 (11)</td>
</tr>
</tbody>
</table>

Values for n are out of a total of 89 articles, unless otherwise indicated.

CONCLUSIONS

Propensity matching is a powerful tool for the comparison of 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 actually quite complex. This review is intended to help surgeons understand the concepts behind propensity matching that may influence their own research and/or help them critically evaluate the published literature. We identified 8 criteria that we feel should be reported in any article that uses 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 reporting on studies that use propensity matching.

Conflict of Interest Statement

Authors have nothing to disclose with regard to commercial support.

References


Key Words: propensity score, matching, causal inference, observational studies, statistical methods