

Reducing bias using propensity score matching

Charity J. Morgan, PhD^a

^a Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL

Received Jun 1, 2017; accepted Jun 2, 2017
doi:10.1007/s12350-017-1012-y

See related article, pp. 394–403

Randomized controlled trials are considered by many to provide the one of the strongest forms of evidence.¹ The goal of randomization is to balance treatment groups on any confounding factors (whether observed or unobserved), eliminating treatment selection bias and ensuring that the groups are comparable. But when randomization is impractical, unethical, or impossible, nonrandomized observational studies may be useful. However, these studies are subject to treatment selection bias as patient covariates often confound treatment selection. For example, when determining a course of treatment, an oncologist may choose to pursue more aggressive therapies only in patients with advanced tumors; a nonrandomized study comparing the effectiveness of cancer therapies is thus going to be confounded by the fact that patients receiving aggressive therapies are likely to have worse prognoses and are therefore not be comparable to those receiving less aggressive therapies. Propensity score matching is a statistical procedure for reducing this bias by assembling a sample in which confounding factors are balanced between treatment groups. The paper by Nappi et al.²

published in this issue provides an example of this approach.¹

In a simple randomized trial, subjects in different treatment groups are comparable because all subjects have the same probability of being assigned to a particular treatment condition. However, in a nonrandomized study, a subject's probability of receiving a treatment is not known and will depend both his or her observed and unobserved covariates. The propensity score was first proposed by Rosenbaum and Rubin as an estimate of a subject's probability of receiving a treatment given that subject's *observed* baseline covariates.^{3,4} The key assumption underlying propensity score analysis is that, because the propensity score is estimated using observed baseline covariates, subjects whose propensity scores are equal will have similar baseline covariates values and thus be comparable. Another important assumption necessary for propensity score analysis is that there are no unmeasured confounders. That is, we assume that all factors that might affect treatment assignment and/or the outcome of interest have been observed. The presence of an unmeasured confounder can lead to biased results (see below).

Logistic regression is the most commonly used method for estimating the propensity score⁵, although more sophisticated data analysis methods are gaining popularity (see Westreich et al.⁶ for a discussion of alternatives). In the propensity score model, the dependent variable is the (logit) probability of receiving a particular treatment; baseline covariates, particularly any that may be confounders for both treatment selection and the outcome of interest, are included as independent variables.⁷ A propensity score for each subject in the study is then found by using the fitted model to estimate the probability of receiving the treatment given that subject's baseline covariates. Once a propensity score for each subject has been estimated,

Reprint requests: Charity J. Morgan, PhD, Department of Biostatistics,
University of Alabama at Birmingham, Birmingham, AL 35294-0022;
cjmorgan@uab.edu

J Nucl Cardiol 2018;25:404–6.
1071-3581/\$34.00

Copyright © 2017 American Society of Nuclear Cardiology.

¹ Note that, the “treatment” groups in this paper are diabetic and nondiabetic patients.

subjects are matched using the propensity scores in order to create a balanced sample.

As a simple example, suppose that an observational study has been conducted comparing survival times for subjects receiving either a new treatment or control (i.e., standard of care). To estimate a propensity score for each subject, we would first fit a logistic regression model to estimate the effect of selected baseline covariates on the probability of receiving the new treatment. Next, propensity scores for each subject would be calculated by plugging that subject's covariate values into the estimated regression equation to find the subject's estimated probability of receiving the new treatment. The propensity score-matched sample would then be constructed. For each subject receiving the new treatment, one (for a 1-to-1 match) or multiple (for a many-to-1 match) control subject(s) whose propensity score(s) were equal or close to the propensity score of the treated subject would be chosen as matches for that subject.

Other approaches to propensity score matching approaches are available; for example, instead of matching on the propensity score itself, the propensity score may simply be used to narrow down the pool of potential matches. That is, only control subjects whose propensity scores are within a pre-specified range (or "caliper") of the propensity score of the treated subject are considered as possible matches. From this subset, the control subject whose covariate values are "closest" to that of the treated subject (according to some measure of distance) is matched with that subject.⁸ Nappi et al.² followed this approach, using a caliper of 0.25 times the standard deviation of the propensity score and the Mahalanobis distance⁹ as their measure of closeness for the propensity scores. See D'Agostino for a thorough review of propensity score matching methods.¹⁰

As discussed above, the goal of propensity score matching is to create a sample in which treatment groups are balanced on baseline covariates. Thus, an assessment of covariate balance in the matched sample is a crucial step of the analysis. Rather than conducting statistical tests comparing the covariates values in the two groups, it is recommended that the absolute standardized differences between groups for each covariate be examined.⁵ The absolute standardized difference (ASD) is defined as

$$ASD = \frac{|\bar{x}_T - \bar{x}_C|}{\sqrt{\frac{S_T^2}{2} + \frac{S_C^2}{2}}}$$

where \bar{x}_T and S_T are the sample mean and standard deviation, respectively, for subjects in the treated group and \bar{x}_C and S_C are the sample mean and standard deviation for the control group, respectively. Larger values

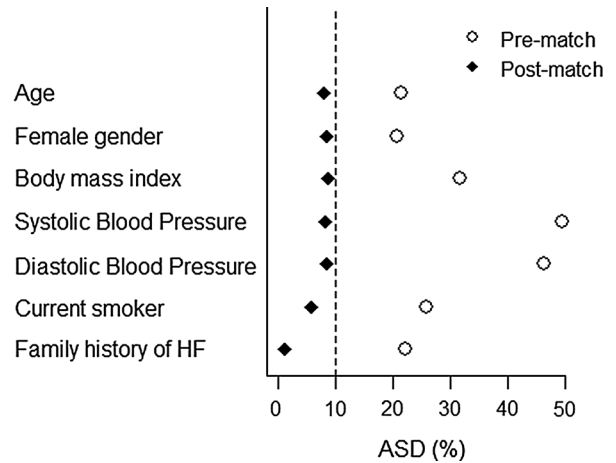


Figure 1. Absolute standardized differences before and after propensity score matching.

of ASD indicate greater imbalance in covariate values. Covariate balance may be assessed by comparing the ASD to a pre-specified threshold (<10% is a common choice).

An example of absolute standardized differences before and after propensity score matching is shown in Figure 1. We see that the ASDs for all covariates are smaller after propensity score matching and all below the threshold of 10%, suggesting that the propensity score matching has balanced the treatment and control groups on these covariates.

If covariate imbalance remains after the propensity score matching, the propensity score model should be revised, for example by adding interaction terms and/or other covariates. Once covariates are sufficiently balanced, statistical analysis is conducted using the matched sample.

While on average randomization will balance both observed and unobserved confounding factors, it is important to remember that propensity scores can only balance *observed* covariates. As a result, statistical inferences may still be subject to bias from unmeasured confounding variables.¹¹ Sensitivity analyses should be conducted to assess how robust study conclusions are to the presence of an unmeasured confounder; see Liu et al. for an introduction to some of the available sensitivity analysis methods.¹²

Although we have limited our discussion here to propensity score matching, propensity scores may be used in other ways to adjust for covariate imbalance. Instead of using the propensity scores to create a balanced sample, analyses may be conducted on the full sample but with either weighting or stratifying by the propensity score. Another approach is to treat the propensity score as a covariate in regression analyses.

For discussion and comparison of other propensity score analyses, see.^{5,10}

In their paper, Nappi et al.² use a propensity score-matched sample to compare left ventricular shape in diabetic and nondiabetic subjects. This approach allowed the authors to adjust for any observed baseline differences between diabetic and nondiabetic patients that may have confounded analyses of left ventricular shape. Although it cannot replace a true randomized trial, propensity score matching is a powerful tool for adjusting for confounding variables and reducing treatment selection bias.

Disclosure

The author has no conflicts of interest to disclose.

References

1. Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. *Plast Reconstr Surg.* 2011;128:305–10.
2. Nappi C, Gaudieri V, Acampa W, Assante R, Zampella E, Mainolfi C, et al. Comparison of left ventricular shape by gated spect imaging in diabetic and nondiabetic patients with normal myocardial perfusion: a propensity score analysis. *J Nuclear Cardiol.* 2017. doi:10.1007/s12350-017-1009-6.
3. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70:41–55.
4. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc.* 1984;79:516–24.
5. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res.* 2011;46:399–424.
6. Westreich D, Lessler J, Funk MJ. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *J Clin Epidemiol.* 2010;63:826–33.
7. Shadish WR, Steiner PM. A primer on propensity score analysis. *Newborn Infant Nurs Rev.* 2010;10:19–26.
8. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat.* 1985;39:33–8.
9. Mahalanobis PC. On the generalised distance in statistics. *Proc Natl Inst Sci India.* 1936;2:49–55.
10. D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med.* 1998;17:2265–81.
11. Joffe MM, Rosenbaum PR. Invited commentary: Propensity scores. *Am J Epidemiol.* 1999;150:327–33.
12. Liu W, Kuramoto SJ, Stuart EA. An introduction to sensitivity analysis for unobserved confounding in non-experimental prevention research. *Prev Sci.* 2013;14:570–80.