Propensity Scores: Uses and Limitations

Kristin L. Sainani, PhD

INTRODUCTION

Observational treatment studies are limited by the lack of randomization. Factors that influence treatment selection also may influence outcomes, leading to bias. For example, patients with severe disease may be more likely to receive a particular drug and also may be more likely to die, thus creating a spurious link between the drug and mortality. Propensity scores attempt to correct this problem by ensuring that the treatment groups under study are balanced with respect to measured covariates. They are not a magic bullet and cannot circumvent the lack of randomization, but they do offer certain advantages over more traditional methods of controlling for confounding by indication. In particular, their use may be warranted when the number of confounders is large or the number of outcomes is small. This article reviews propensity score methods, including what they are, how they work, and their advantages and limitations.

This article will refer to 2 database studies that illustrate the use of propensity scores. In the first example, Mokhles et al [1] compared late survival rates in patients who underwent aortic valve replacement and received a Ross procedure (an autograft) versus those who received a mechanical valve (with optimal anticoagulation therapy). The authors of previous studies had suggested a survival advantage for patients who underwent the Ross procedure but could not rule out bias caused by patient selection. Indeed, patients selected for the Ross procedure in the study by Mokhles et al tended to be younger and in better physical condition than patients selected for a mechanical valve. In the second example, Murray et al [2,3] examined whether rehabilitation in nursing homes improves community discharge rates and function for patients who have had a stroke. Patients who received rehabilitation services had better outcomes but also were less disabled, better insured, and had more social support at baseline compared with patients who did not receive rehabilitation. Therefore it was unclear whether the benefits were attributable to the rehabilitation per se or to these other factors.

WHAT ARE PROPENSITY SCORES?

In a randomized trial, all participants have an equal chance of being assigned to each treatment. However, in an observational study, participants have variable chances of receiving each treatment. The propensity score estimates these probabilities based on each participant's measured characteristics (covariates). For example, a young patient in good physical condition might have a 70% chance of undergoing the Ross procedure, whereas an older patient in poor physical condition might have only a 30% chance. These 2 patients are not comparable, but patients with the same propensity score are comparable. For example, if a patient with a 70% propensity score underwent the Ross procedure and another with a 70% propensity score received a mechanical valve, then, in theory, any difference in outcome can be attributed to the treatment rather than to patient selection.

Covariates: the clinical, demographic, and social characteristics that are measured for study participants, including potential predictors and confounders.

Different treatment groups that have been matched or grouped by propensity scores should be balanced with respect to measured covariates. The effect is similar to randomization but with one critical distinction: Propensity scores ensure balance only in observed covariates, whereas randomization balances both observed and unobserved covariates. This

K.L.S. Department of Health Research and Policy, Division of Epidemiology, Stanford University, HRP Redwood Bldg, Stanford, CA 94305. Address correspondence to: K.L.S. e-mail: kcobb@stanford.edu
Disclosure: nothing to disclose

Disclosure Key can be found on the Table of Contents and at www.pmrjournal.org

distinction means that if unmeasured confounders are present, propensity score methods will do nothing to correct for them [4].

Besides helping to balance the groups, the propensity score also is a data reduction tool in that it reduces a large number of variables about a patient—all of the collected information about factors that may influence treatment selection—into a single probability value.

HOW ARE PROPENSITY SCORES CALCULATED?

Propensity scores are estimated using a logistic regression with treatment as the outcome (eg, Ross procedure or mechanical valve) and measured covariates as the predictors. Several practical issues need to be addressed, including what predictors to include in the model, how to handle missing data in the covariates, and how to evaluate the model [4].

Logistic regression: the multivariate regression technique that is used when the outcome (in this case, treatment group) is binary. Logistic regression can be used to calculate a predicted probability (in this case, the probability of treatment) for each study participant.

When one builds a propensity score model, many of the typical principles of good model building do not apply. For example, researchers do not need to worry about overfitting because the goal is to fit the model as closely as possible to the data (to balance the observed covariates) rather than to generalize to new data. Researchers typically jam many variables in the model and also may include several interactions and quadratic terms. For example, in the rehabilitation study, the authors included 112 predictors (eg, demographics, use of assistive devices, and nutrition) in their propensity score model. Although researchers need not be parsimonious, they may benefit from discarding variables that turn out to be related to treatment selection but not to outcomes; simulations show that including such variables does not improve balance or reduce bias but may make finding people with matching propensity scores more difficult [5].

Overfitting: When a regression model includes too many predictors relative to the number of events (or overall sample size, for continuous outcomes), then the model may fit the quirks of the particular sample but have no predictive power outside the sample.

Authors must carefully address missing data because the logistic regression will exclude any participant for whom even one data point for one covariate is missing. To avoid drastically shrinking the sample size, researchers must impute these missing values. For example, in the rehabilitation study, 7 of the 112 predictors were missing values for 0.5%-5.5% of the sample. Thus at least 5.5% of the sample—and likely a much greater percentage—were missing at least one covariate value. The authors appropriately replaced these missing values with the mean values from the sample that were not missing data.

Finally, authors must assess the resulting balance of covariates. If the balance is poor, they may need to refit the logistic regression by including additional covariates (or higher order terms). Balance is best assessed with standardized differences rather than *P* value tests (which are highly dependent on sample size). The standardized difference for a covariate is the mean difference between the groups divided by the standard deviation, and then converted to a percent. For example, patients who underwent the Ross procedure were on average 7.9 years younger than patients who received a mechanical valve, and the SD for age was approximately 10.5; thus the standardized difference was 75% of an SD. Standardized differences of less than 10% indicate good balance. Figure 1 shows the standardized differences in covariates before and after treatment groups were matched by propensity scores. The treatment groups after matching appear balanced because all the measured covariates have standardized differences below 10%.

HOW ARE PROPENSITY SCORES USED?

One key use of the propensity score is to reveal when it is simply impossible to compare groups. Researchers should always plot the distributions of propensity scores in the treatment groups. If the groups have little overlap in propensity scores, they are inherently incomparable, and no statistical tricks can overcome this problem. Traditional methods for controlling for confounding by indication may fail to reveal this irreconcilable limitation in the data, leading to erroneous conclusions. If the groups do overlap sufficiently in their propensity scores, then the propensity scores can be used in 3 ways to evaluate treatment effects: stratification, matching, or statistical adjustment.

Stratification

A simple and intuitive method for controlling for confounders is to divide (stratify) participants into groups according to various covariates and to calculate treatment effects within these groups. Of course, this method is inherently limited because one can only control for a few covariates at once; stratifying on too many covariates will result in groups that are too sparse to estimate effects. The propensity score solves this issue because it contains information about all the measured covariates summarized into a single variable that researchers can use to stratify patients.

For example, in the rehabilitation study, researchers stratified patients who had a stroke into quintiles by propensity scores (Figure 2). Then they calculated the relative rate of community discharge within each quintile. Averaging over the 5 quintiles gave an overall relative rate of 1.58, suggesting

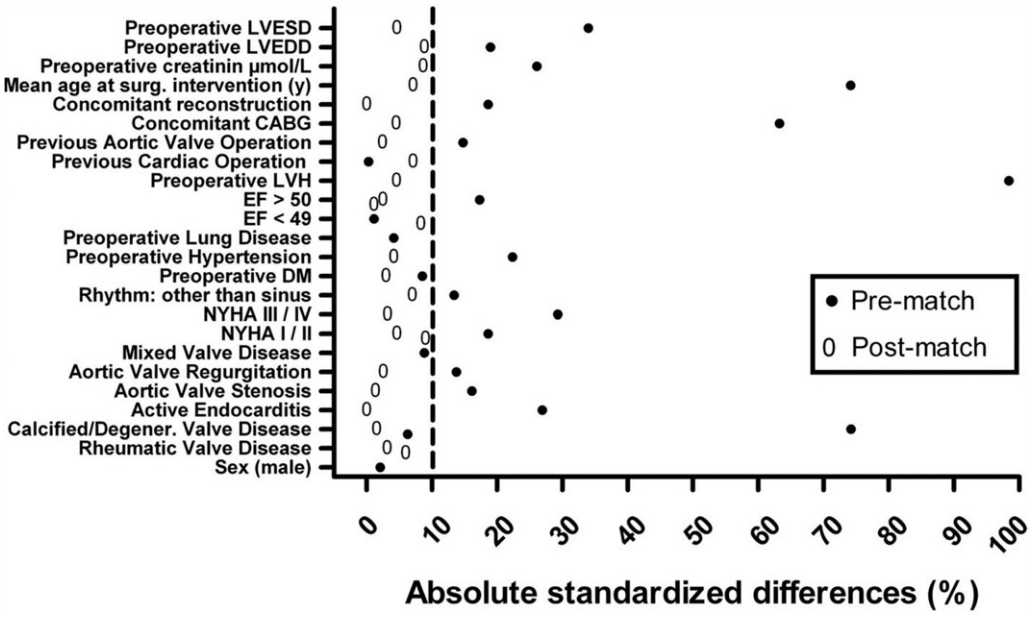

Figure 1. Plots of the standardized differences for baseline covariates between patients who received the mechanical valve and patients who underwent the Ross procedure, before and after propensity score matching. CABG = coronary artery bypass grafting; EF = ejection fraction; LVEDD = left ventricular end-diastolic dimension; LVESD = left ventricular end-systolic dimension; NYHA = New York Heart Association. (Reproduced with permission from Mokhles MM, Körtke H, Stierle U, et al. Survival comparison of the Ross procedure and mechanical valve replacement with optimal self-management anticoagulation therapy: Propensity-matched cohort study. Circulation 2011;123:31-38.)

that rehabilitation improves community discharge after one controls for patient selection. However, the authors also discovered that significant differences existed in the treatment effect across propensity quintiles: Those who were most likely to receive rehabilitation derived the least benefit from it. Thus the authors concluded that doctors may be selecting the wrong patients for rehabilitation. This insight could only be revealed with the use of a propensity score approach.

If the number of strata is small, some variability in the propensity scores will exist within each stratum. Thus some residual confounding may occur. And, of course, stratification on the propensity scores does nothing to control for unmeasured confounders.

Residual confounding: "leftover" confounding that remains, which occurs when a confounding variable has been accounted for in an imprecise manner.

Matching

Another intuitive way to minimize confounding is to individually match each participant in one treatment group with a comparable participant in the other treatment group (or possibly to multiple comparable participants). Matching on many covariates simultaneously is complex and tricky, but matching on the propensity score is straightforward and optimizes covariate balance [4].

If researchers cannot find an exact propensity score match for each participant, they usually allow matches within a certain distance, for example, 2%. In the study of the Ross procedure, researchers used a "nearest neighbor" matching strategy: they randomly ordered the patients who received a mechanical valve and then sequentially matched each one to a patient who underwent the Ross procedure with the closest propensity score. If no patients who underwent the Ross procedure had a propensity score within 25%, the patient was left unmatched and was excluded. In fact, matches could be found only for 253 of 406 patients who received a mechanical valve. In the matched cohort, surprisingly, the authors found that patients who underwent the Ross procedure actually had slightly worse mortality than did the patients who received a mechanical valve (hazard ratio for death, 1.86; 95% confidence interval, 0.58-5.91).

Researchers always face a trade-off between inexact matching and incomplete matching. Inexact matching may lead to residual confounding. However, incomplete matching may lead to a lack of generalizability of the results and a loss of statistical power. For example, the results of the Ross study do not apply to the types of patients who were excluded from the analysis (ie, those with extreme propensity scores). Also, the results may have failed to reach statistical significance because of the reduced sample size and accompanying loss of statistical power.

Debate is ongoing as to how to best analyze propensity matched data to assess treatment effects. Many experts recom-

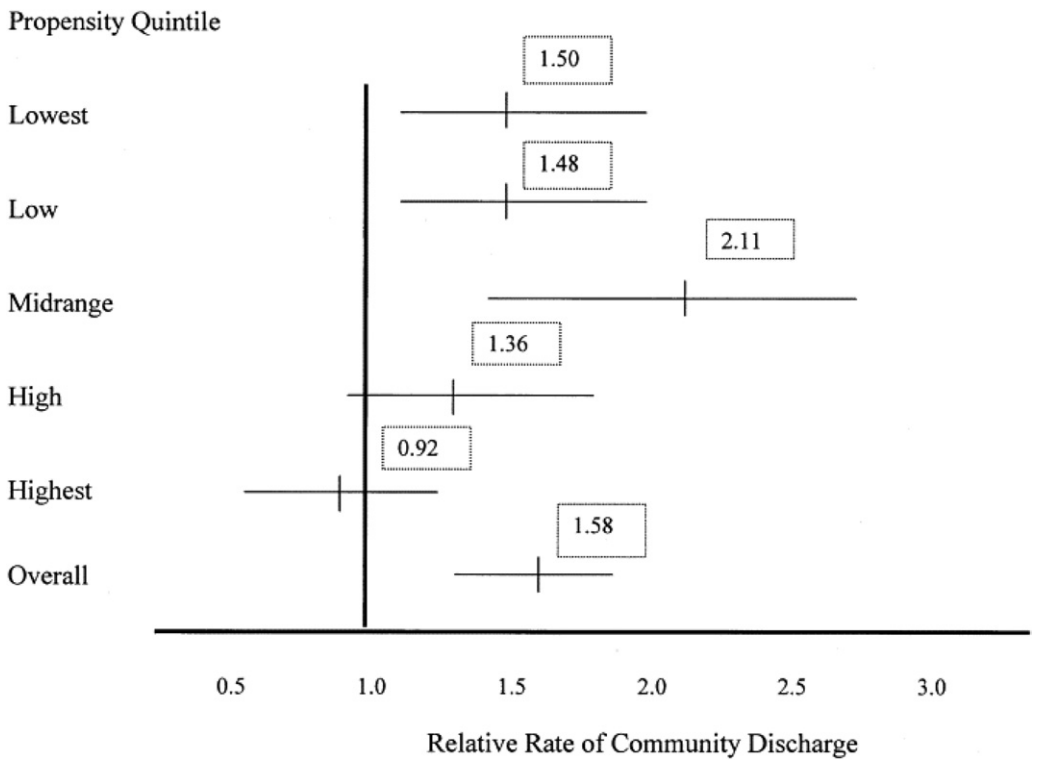

Figure 2. Relative rates of community discharge in the sample, divided into quintile of propensity. Horizontal bars represent the 95% confidence interval. (Reproduced with permission from Murray PK, Dawson NV, Thomas CL, Cebul RD. Are we selecting the right patients for stroke rehabilitation in nursing homes? *Arch Phys Med Rehabil* 2005;86:876-880.)

mend the use of statistical tests that account for the within-pair correlation [6]. For example, in the study of the Ross procedure, the authors used tests that account for the paired nature of the data, including the McNemar test, the paired *t*-test, and the Wilcoxon signed-rank test. However, other investigators recently have argued that it is not necessary to account for the pairing because 2 individuals with the same propensity score may actually be quite different in terms of their underlying characteristics (as the result of divergent indications) and thus may not be more correlated with each other than with others in the dataset [7].

Statistical Adjustment

Researchers also may include the propensity score as a covariate in a regression analysis that evaluates the treatment effect. For example, the authors of the Ross study ran a Cox regression with death as the outcome and treatment as the predictor by using all 1324 participants from the unmatched cohort. The hazard ratio for mortality for patients who underwent the Ross procedure versus patients who received a mechanical valve was 1.33. When the authors added the propensity score as a covariate in the model, the hazard ratio increased to 3.64 (95% confidence interval 1.22-10.88). This result suggests that patients undergoing the Ross procedure have increased mortality after correcting for their younger age and other distinguishing characteristics.

In theory, adjusting for the propensity score is similar to directly adjusting for all the covariates that were used to calculate the propensity score. In fact, some persons have argued that propensity score adjustment offers almost no advantage over directly adjusting for confounders for the majority of cases, and thus that readers should not be overly impressed by propensity scores [8]. However, one situation exists in which there is a clear advantage: when the number of outcome events is small. One risks overfitting if the model contains fewer than 10 events per covariate, and one cannot build the model if there are fewer events than covariates. Thus including the single propensity score in the model in lieu of tens or hundreds of confounders may be statistically advantageous or even necessary. Indeed, in the study of the Ross procedure only 36 total deaths occurred; thus it would have been extremely unwise to fit a Cox regression with 24 covariates in addition to treatment group. Simulations reveal that if the ratio of events to confounders is fewer than 7:1, adjusting for the propensity score rather than all the confounders reduces bias [9].

CONCLUSION

Authors comparing 2 treatments in an observational design should consider the use of propensity scores, particularly if the treatment groups are highly imbalanced, the number of confounders is large, or the number of events is low. Propen-

sity scores reduce an unwieldy set of confounders into a single, intuitive variable. They optimize matching and stratification and make it possible to statistically adjust when the ratio of events to confounders is low. They also may reveal cases in which the patient populations are too divergent to make meaningful comparisons. Readers should keep in mind that propensity scores are not a substitute for randomization. Unlike randomization, propensity score methods only ensure balance in measured, not unmeasured, confounders. When evaluating studies using the propensity score, readers should ask the following 5 questions: How were variables selected for the propensity score model? How were missing data handled? How was balance assessed (preferably with standardized difference scores)? How much did the treatment groups overlap? How much could the results have been influenced by unmeasured or residual confounding?

REFERENCES

1. Mokhles MM, Körtke H, Stierle U, et al. Survival comparison of the Ross procedure and mechanical valve replacement with optimal self-management anticoagulation therapy: Propensity-matched cohort study. Circulation 2011;123:31-38.
2. Murray PK, Dawson NV, Thomas CL, Cebul RD. Are we selecting the right patients for stroke rehabilitation in nursing homes? Arch Phys Med Rehabil 2005;86:876-880.
3. Murray PK, Singer M, Dawson NV, Thomas CL, Cebul RD. Outcomes of rehabilitation services for nursing home residents. Arch Phys Med Rehabil 2003;84:1129-1136.
4. Joffe MM, Rosenbaum PR. Invited commentary: Propensity scores. Am J Epidemiol 1999;150:327-333.
5. 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;26:734-753.
6. Austin P. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med 2008;27:2037-2049.
7. Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: from naive enthusiasm to intuitive understanding. Stat Methods Med Res 2012;21:273-293.
8. Winkelmayer WC, Kurth T. Propensity scores: Help or hype? Am J Epidemiol 2003;158:280-287.
9. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol 2003;158:280-287.