JAMA Guide to Statistics and Methods

The Propensity Score

Jason S. Haukoos, MD, MSc; Roger J. Lewis, MD, PhD

Two recent studies published in *JAMA* involved the analysis of observational data to estimate the effect of a treatment on patient outcomes. In the study by Rozé et al,[1] a large observational data set was analyzed to estimate the relationship between early echocardiographic screening for patent ductus arteriosus and mortality among preterm infants. The authors compared mortality rates of 847 infants who were screened for patent ductus arteriosus and 666 who were not. The 2 infant groups were dissimilar; infants who were screened were younger, more likely female, and less likely to have received corticosteroids. The authors used propensity score matching to create 605 matched infant pairs from the original cohort to adjust for these differences. In the study by Huybrechts et al,[2] the Medicaid Analytic eXtract data set was analyzed to estimate the association between antidepressant use during pregnancy and persistent pulmonary hypertension of the newborn. The authors included 3 789 330 women, of which 128 950 had used antidepressants. Women who used antidepressants were different from those who had not, with differences in age, race/ethnicity, chronic illnesses, obesity, tobacco use, and health care use. The authors adjusted for these differences using, in part, the technique of propensity score stratification.

Use of the Method

Why Were Propensity Methods Used?

Many considerations influence the selection of one therapy over another. In many settings, more than one therapeutic approach is commonly used. In routine clinical practice, patients receiving one treatment will tend to be different from those receiving another, eg, if one treatment is thought to be better tolerated by elderly patients or more effective for patients who are more seriously ill. This results in a correlation—or confounding—between patient characteristics that affect outcomes and the choice of therapy (often called "confounding by indication"). If observational data obtained from routine clinical practice are examined to compare the outcomes of patients treated with different therapies, the observed difference will be the result of both differing patient characteristics and treatment choice, making it difficult to delineate the true effect of one treatment vs another.

The effect of an intervention is best assessed by randomizing treatment assignments so that, on average, the patients are similar in the 2 treatment groups. This allows a direct assessment of the effect of the intervention on outcome. In observational studies, randomization is not possible, so investigators must adjust for differences between groups to obtain valid estimates of the associations between the treatments being compared and the outcomes of interest.[3] Multivariable statistical methods are often used to estimate this association while adjusting for confounding.

Propensity score methods are used to reduce the bias in estimating treatment effects and allow investigators to reduce the likelihood of confounding when analyzing nonrandomized, observational data. The propensity score is the probability that a patient would receive the treatment of interest, based on characteristics of the patient, treating clinician, and clinical environment.[4] Such probabilities can be estimated using multivariable statistical methods (eg, logicistic regression), in which case the treatment of interest is the dependent variable and the characteristics of the patient, prescribing clinician, and clinical setting are the predictors. Investigators estimate these probabilities, ranging from 0 to 1, for each patient in the study population. These probabilities—the propensity scores—are then used to adjust for differences between groups. In biomedical studies, propensity scores are often used to compare treatments, but they can also be used to estimate the relationship between any nonrandomized factor, such as the exposure to a toxin or infectious agent and the outcome of interest.

There are 4 general ways propensity scores are used. The most common is *propensity score matching*, which involves assembling 2 groups of study participants, one group that received the treatment of interest and the other that did not, while matching individuals with similar or identical propensity scores.[1] The analysis of a propensity score–matched sample can then approximate that of a randomized trial by directly comparing outcomes between individuals who received the treatment of interest and those who did not, using methods that account for the paired nature of the data.[5]

The second approach is *stratification* on the propensity score.[4] This technique involves separating study participants into distinct groups or strata based on their propensity scores. Five strata are commonly used, although increasing the number can reduce the likelihood of bias. The association between the treatment of interest and the outcome of interest is estimated within each stratum or pooled across strata to provide an overall estimate of the relationship between treatment and outcome. This technique relies on the notion that individuals within each stratum are more similar to each other than individuals in general; thus, their outcomes can be directly compared.

The third approach is *covariate adjustment* using the propensity score. For this approach, a separate multivariable model is developed, after the propensity score model, in which the study outcome serves as the dependent variable and the treatment group and propensity score serve as predictor variables. This allows the investigator to estimate the outcome associated with the treatment of interest while adjusting for the probability of receiving that treatment, thus reducing confounding.

The fourth approach is *inverse probability of treatment weighting* using the propensity score.[6] In this instance, propensity scores are used to calculate statistical weights for each individual to create a sample in which the distribution of potential confounding factors is independent of exposure, allowing an unbiased estimate of the relationship between treatment and outcome.[7]

Alternative strategies—other than use of propensity scores—for adjusting for baseline differences between groups in observational studies include matching on baseline characteristics, performing stratified analyses, or using multivariable statistical methods to adjust for confounders. Propensity score methods are often more practical or statistically more efficient than these methods, in part

because propensity score methods can substantially limit the number of predictor variables used in the final analysis. Propensity score methods generally allow many more variables to be included in the propensity score model, which increases the ability of these approaches to effectively adjust for confounding, than could be incorporated directly into a multivariable analysis of the study outcome.

What Are the Limitations of Propensity Score Methods?

The propensity score for each study participant is based on the available measured patient characteristics, and unadjusted confounding may still exist if unmeasured factors influenced treatment selection. Therefore, using fewer variables in the propensity score model reduces the likelihood of effectively adjusting for confounding.

Although propensity score matching may be used to assemble comparable study groups, the quality of matching depends on the quality of the propensity score model, which in turn depends on the quality and size of the available data and how the model was built. Conventional modeling methods (eg, variable selection, use of interactions, regression diagnostics, etc) are not typically recommended for the development of propensity score models. For example, propensity score models may optimally include a larger number of predictor variables.

Why Did the Authors Use Propensity Methods?

In the reports by Rozé et al[1] and Huybrechts et al,[2] both of whom used propensity score methods because their data were observational, the treatments of interest (ie, screening by echocardiography and use of antidepressants in pregnancy) were not randomly allocated, and important characteristics differed between groups. Direct comparisons of the outcomes between treated and untreated groups would have likely resulted in significantly biased estimates. Instead, use of propensity score matching and stratification enabled the investigators to create study groups that were similar to one another and more accurately measure the relationship between treatment and outcome.

How Should the Findings Be Interpreted?

Given the observational nature of these studies, the fact that individuals in the treated and untreated groups were dissimilar, and the goal of accurately estimating the association between treatment and outcome, the investigators had to adjust for differences in the treatment groups. Use of propensity score methods, whether by matching or stratification, resulted in less biased estimates than if such methods were not used. Even though observational data cannot definitely establish causal relationships or determine treatment effects as rigorously as a randomized clinical trial, assuming propensity score methods are properly used and the sample size is sufficiently large, these methods may provide a useful approximation of the likely effect of a treatment. This approach is particularly valuable for clinical situations in which randomized trials are not feasible or are unlikely to be conducted.

What Caveats Should the Reader Consider When Assessing the Results of Propensity Analyses?

The studies by Rozé et al[1] and Huybrechts et al[2] used propensity score matching and propensity score stratification, respectively. Although both methods are more valid in terms of balancing study groups than simple matching or stratification based on baseline characteristics, they vary in their ability to minimize bias. In general, propensity score matching minimizes bias to a greater extent than propensity score stratification. Assessment of balance between the groups, after use of propensity score methods, is important to allow readers to assess the comparability of patient groups.

Although no single standard approach exists to assess balance, comparing characteristics between treated and untreated patients typically begins with comparing summary statistics (eg, means or proportions) and the entire distributions of observed characteristics. For propensity score–matched samples, standardized differences (ie, differences divided by pooled standard deviations) are often used and, although no threshold is universally accepted, a standard difference less than 0.1 is often considered negligible. Assessing for balance provides a general sense for how well matching or stratification occurred and thus the extent to which the results are likely to be valid. Unfortunately, balance can only be demonstrated for patient characteristics that were measured in the study. Differences could still exist between patient groups that were not measured, resulting in biased results.

ARTICLE INFORMATION

Author Affiliations: Department of Emergency Medicine, University of Colorado School of Medicine, Denver (Haukoos); Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, California (Lewis); David Geffen School of Medicine at UCLA, Los Angeles, California (Lewis).

Corresponding Author: Jason S. Haukoos, MD, MSc, Department of Emergency Medicine, Denver Health Medical Center, 777 Bannock St, Mail Code 0108, Denver, CO 80204 (Jason.Haukoos@dhha.org).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

Funding/Support: Dr Haukoos is supported, in part, by grants R01AI106057 from the National Institute of Allergy and Infectious Diseases (NIAID) and R01HS021749 from the Agency for Healthcare Research and Quality (AHRQ).

Disclaimer: The views expressed herein are those of the authors and do not necessarily represent the views of NIAID, the National Institutes of Health, or AHRQ.

REFERENCES

1. Rozé JC, Cambonie G, Marchand-Martin L, et al; Hemodynamic EPIPAGE 2 Study Group. Association between early screening for patent ductus arteriosus and in-hospital mortality among extremely preterm infants. *JAMA*. 2015;313(24):2441-2448.

2. Huybrechts KF, Bateman BT, Palmsten K, et al. Antidepressant use late in pregnancy and risk of persistent pulmonary hypertension of the newborn. *JAMA*. 2015;313(21):2142-2151.

3. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10(1):37-48.

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

5. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46(3):399-424.

6. Schaffer JM, Singh SK, Reitz BA, Zamanian RT, Mallidi HR. Single- vs double-lung transplantation in patients with chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis since the implementation of lung allocation based on medical need. *JAMA*. 2015;313(9):936-948.

7. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550-560.