Anesthesiology 2008; 108:3–4

Liars, Damn Liars, and Propensity Scores

PROPENSITY score methods are being used increasingly to reduce the impact of treatment-selection bias when using observational data to estimate causal treatment effects.[1] In the article by Vincent *et al.*,[2] 821 pairs of patients were matched according to a propensity score. The data in the study of Vincent *et al.* were derived from a previous study called the Sepsis Occurrence in Acutely Ill Patients Study, which was a multicenter, observational study that included all adult patients admitted to 198 European intensive care units.[3] The authors demonstrated that the 30-day survival rate was higher in those patients who received a transfusion compared with those who did not. These results contradict those of Hebert *et al.*,[4] who demonstrated that a restrictive strategy of erythrocyte transfusion was at least as effective as and possibly superior to a liberal transfusion strategy in critically ill patients in a multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. The use of propensity score analysis is not without controversy, because occasionally its use has resulted in some disputed conclusions.[5] So what is propensity score analysis, and what strengths, weaknesses, and biases are inherent to the analysis?

The accepted standard for demonstrating that a treatment produces a certain outcome is a prospective, randomized, blinded (controlled if appropriate) trial. This is the case because random assignment of patients to treatment groups balances both known and unknown patient characteristics that may affect outcome and reduces the likelihood that there will be differences in the patient characteristics between study arms. Unfortunately, many therapies cannot be randomized for ethical, economical, or practical reasons, and on these occasions, observational studies can provide valuable information about treatment effectiveness. However, because of the very nature of the study design, the interpretation of observational studies is fraught with difficulty.

An inherent problem in the methodology of observational studies is that the investigators do not have control over the treatments given to participants. As in the study of Vincent *et al.*,[2] patients are often "assigned" to a treatment condition based on a conglomeration of characteristics that make it very likely/unlikely that they will experience the outcome under study. Unlike random assignment, where the groups systematically differ on only the treatment intervention, the treatment groups in observational studies are likely to differ on both the treatment intervention and also a myriad of other variables that may independently affect outcome (called covariates). As a result of differences in treatment groups, investigators must rely on statistical adjustments to control for the confounding effect of the observed covariates when estimating the unique effect due to treatment. Often, there are large amounts of data on potential cofounders available for analysis, but the large volume and complexity of this data does not guarantee reliable and accurate analysis. It is for the improvement of such analyses that propensity methodology was created.

Propensity methodology was first proposed in 1983 as a novel strategy for statistical control in observational studies.[6] The method first focuses on the relation between baseline patient characteristics and the primary treatment variable of interest, such as receiving erythrocyte transfusion *versus* not. Conceptually, the propensity score is the conditional probability that an individual study participant would have been treated based on that individual's observed pretreatment variables. Statistically, propensity score methods require a two-step process in which a logistic regression model is first built to predict the probability ("propensity") of exposure to treatment condition (treatment model). A second model incorporating the information on the propensity score is then constructed to evaluate the exposure–outcome association (outcome model). Statistical adjustments using the estimated propensity score have the advantageous property of balancing observed covariates that were used to construct the score, thus producing a situation closer to actual randomization. The propensity score can also be used outside of a model-based approach to compare patients with similar characteristics. The three most common methods for using the estimated propensity score are matching,[7] regression adjustment,[8] and weighting (stratification).[9] Regardless of the technique, the propensity score is calculated the same way.

Patient matching by propensity score is one technique for addressing baseline characteristics. In this method, a propensity score summarizes all measured confounders in a single score, and subjects are then matched by the propensity score. This greatly simplifies the matching of subjects, because patients would otherwise have to be independently matched on all of the covariates, an endeavor whose complexity increases with each considered covariate. In regression adjustment based on propensity scores, the propensity score is entered into the

This Editorial View accompanies the following article: Vincent J-L, Sakr Y, Sprung C, Harboe S, Damas P: Are blood transfusions associated with greater mortality rates? Results of the Sepsis Occurrence in Acutely Ill Patients Study. ANESTHESIOLOGY 2008; 108:31–9.

Accepted for publication October 3, 2007. The authors are not supported by, nor maintain any financial interest in, any commercial activity that may be associated with the topic of this article.

model as the only confounding variable, in addition to the exposure to treatment, to better estimate the unique effect of the treatment exposure on outcome. Vincent *et al.*[2] used regression adjustment based on the propensity score in their study. Finally, in weighting or stratification by propensity score, patients are stratified based on their propensity score. Predetermined strata (*e.g.*, quintiles) are used to directly compare treatment and control patients in the same strata. Although each propensity score technique has its own unique advantages, in general they all share the same limitations.

The first and most important limitation of all methods of confounder control such as multivariable logistic regression and propensity score methods is that although they can balance observed baseline covariates between groups, they do nothing to balance *unmeasured* characteristics and confounders. As a result, unlike randomized control trials, propensity score analyses have the limitation that remaining unmeasured confounding variables may still be present, thus leading to biased results. Another limitation of propensity score methods is that the analysis does not "fix" other potential methodologic biases that may exist. For example, in the study by Vincent *et al.*,[2] patients who received a blood transfusion at any time were matched, based on propensity score, with patients who did not receive a blood transfusion. Because blood transfusions could have occurred at any time, the design could have taken into account the fact that transfusions are a time-dependent variable. For example, consider two patients, one who died on postoperative day 1 without receiving transfused blood and one who received an initial blood transfusion on postoperative day 4 and subsequently died within hours after the transfusion. If these patients were selected as a matched pair for the proportional hazards regression analysis, when evaluating this matched set it would seem that transfusing blood improved survival because the patient who received a blood transfusion survived 4 days, whereas the patient who did not only survived 1 day. To overcome this potential bias, the matched control for each case would have necessarily been selected from the pool of nontransfused patients who survived at least until the day at which the transfused patient received his or her first blood transfusion. Other problems with propensity score analysis have been identified, including the performance of the technique under certain conditions, such as when there are seven or fewer events per confounding variable.[10] As a result, it is unclear which adjustment method is most preferable for each given situation. This, coupled with perceived opacity of the statistical process, results in propensity score analysis having a very "black box" feeling about it.

Evidence-based medicine has been established as a cornerstone of good medical practice and as a method to improve patient care. Ideally, evidence-based medicine should be based on prospective, randomized, blinded trials. Frequently, these trials are not available and we must use observational trial data that has been modified by statistical analysis such as propensity analysis. It is imperative that we understand the strengths and weakness of these statistical techniques to improve the care of our patients. Therefore, the limitations discussed above suggest that the results reported by Vincent *et al.*[2] must be interpreted with caution.

Gregory A. Nuttall, M.D.,* **Timothy T. Houle, Ph.D.†**
*Department of Anesthesiology, Mayo Clinic, Rochester, Minnesota. nuttall.gregory@mayo.edu. †Department of Anesthesiology, Wake Forest University School of Medicine, Winston-Salem, North Carolina.

References

1. Austin PC: The performance of different propensity score methods for estimating marginal odds ratios. Stat Med 2007; 26:3078–94
2. Vincent J-L, Sakr Y, Sprung C, Harboe S, Damas P: Are blood transfusions associated with greater mortality rates? Results of the Sepsis Occurrence in Acutely Ill Patients study. Anesthesiology 2008; 108:31–9
3. Vincent JL, Sakr Y, Sprung CL, Ranieri VM, Reinhart K, Gerlach H, Moreno R, Carlet J, Le Gall JR, Payen D: Sepsis in European intensive care units: Results of the SOAP study. Crit Care Med 2006; 34:344–53
4. Hebert PC, Wells G, Blajchman MA, Marshall J, Martin C, Pagliarello G, Tweeddale M, Schweitzer I, Yetisir E: A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. Transfusion Requirements in Critical Care Investigators, Canadian Critical Care Trials Group. N Engl J Med 1999; 340:409–17
5. Mangano D, Tudor I, Dietzel C, Multicenter Study of Perioperative Ischemia Research Group, Ischemia Research Education Foundation: The risk associated with aprotinin in cardiac surgery. N Engl J Med 2006; 354:353–65
6. Rosenbaum P, Rubin B: The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70:41–55
7. Rubin DB, Thomas N: Matching using estimated propensity scores: Relating theory to practice. Biometrics 1996; 52:249–64
8. D'Agostino RB Jr: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998; 17:2265–81
9. Sato T, Matsuyama Y: Marginal structural models as a tool for standardization. Epidemiology 2003; 14:680–6
10. 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; 158:280–7