Rare Outcomes, Common Treatments: Analytic Strategies Using Propensity Scores

When treated patients are compared to controls, differing outcomes may reflect either effects caused by the treatment or differences in prognosis before treatment. Random assignment of patients to treatment or control, as in a randomized, controlled clinical trial (1), ensures that the groups were comparable before treatment and the prognosis in treated and control groups was nearly the same, so that differing outcomes indicate treatment effects. Somewhat more precisely, random assignment ensures that the only differences in prognosis between groups are due to chance, the flip of a coin in assigning treatments. In an ideal randomized trial, if a common statistical test rejects the hypothesis that the difference in outcomes is due to chance, a treatment effect is demonstrated. Notice that randomization does nothing to make patients have individually similar prognoses; rather, it ensures that assignment to treatment or control is unrelated to prognosis.

When random assignment is not used—that is, in an observational study—treated and control groups may differ in prognosis, and differing outcomes may not be effects of the treatment. Measured and recorded differences in prognosis—overt biases—can often be controlled by analytical adjustments (2), whereas unmeasured differences—hidden biases—may exist and must be addressed by other means (2–4). A prognostic variable or covariate is a variable describing the condition of patients before treatment. *Bias* refers to systematic differences between treated and control groups with respect to one or more prognostic variables; the bias is overt if the variable is measured and hidden if it is not.

Analytical adjustments for overt biases are of two kinds: 1) those that focus on the relationship between prognostic variables and outcomes and 2) those that focus on the relationship between prognostic variables and assignment of patients to treatment or control. The first strategy models the response directly, for example, through use of regression or logistic regression. The second strategy, which uses propensity scores, is an attempt to reconstruct, after the fact, a situation similar to random assignment, albeit only with respect to observed prognostic variables. In principle, either strategy (separately or in combination), properly used, can control overt biases. Neither strategy does much to control hidden biases. In practice, the second strategy has advantages over the first when the outcome is rare, the treatment is common, and there are many prognostic variables. Here, the terms *rare* and *common* refer to the available data: A rare outcome is seen in a small fraction of the patients under study, and thus there are limited data with which to model the outcome and its relationship to prognostic variables; however, a large fraction of patients received each of the two treatments under study, so there are plenty of data with which to model the relationship between treatment assignment and prognostic variables. An example of this can be seen in studies of relatively rare adverse side effects of relatively common treatments. One such study is by Jasmer and colleagues (5) in this issue: The authors examined 18 cases of hepatotoxicity among 411 patients given one of two competing treatments for latent tuberculosis infection.

Propensity Scores: What They Are, Why They Work, What They Can't Do

The *propensity score* is the chance of receiving the treatment rather than the control for a patient with given observed prognostic variables (6–8). In the simplest randomized experiment, a coin is flipped to assign patients to treatment or control, so the propensity score is 1/2 for every patient. In contrast, in an observational study, without random assignment, the chance of being assigned to one treatment or another may vary from patient to patient depending on prognostic variables (for example, frail patients may be less likely to be treated surgically, or, as in Jasmer and colleagues' study, patients with other risk factors for liver injury may be less likely to receive a particular tuberculosis treatment). In that case, certain types of patients will be overrepresented in the treated group, and other types will be overrepresented in the control group; as a result, the groups will not be comparable.

An adjustment using propensity scores attempts to undo the problem created by unequal chances of receiving treatment. It does this by comparing patients who had the same chance of receiving treatment. A treated patient who had a 3/4 chance of receiving treatment is compared to a control who also had a 3/4 chance of receiving treatment, while a treated patient who had a 1/4 chance of receiving treatment is compared to a control who also had a 1/4 chance of receiving treatment. If two patients both have a 3/4 chance of receiving treatment on the basis of their observed prognostic variables, these variables will not help to predict which one receives treatment; thus, the comparison is expected to be balanced with respect to these prognostic variables (2, 6). For instance, in a report by Rosenbaum and Rubin (6), patients with good left ventricular function and substantial occlusion of several coronary arteries were more likely to be treated with coronary artery bypass graft surgery, while patients with poor left ventricular function or fewer occlusions of the arteries were more likely to be treated with drugs; as a result, the surgical and drug groups were different in terms of prognostic variables. Indeed, they differed significantly on 74 observed prognostic variables. To control for this, the authors grouped patients into five strata of the same size by using their esti-

mated chance of receiving surgery based on the 74 prognostic variables: that is, using the quintiles of their estimated propensity scores. Within these five strata, all 74 prognostic variables were balanced. Within each of the five strata, the patients in the surgery and drug groups had similar distributions of the 74 prognostic variables. In fact, the balance on observed variables was slightly better than would be expected from random assignment of treatments. Of course, propensity scores balance just observed covariates used to construct the score, but randomization balances both observed and unobserved covariates. Propensity scores can remove overt biases, but unlike randomization, they cannot be expected to remove hidden biases.

Rare Outcomes, Common Treatments

If the outcome is rare but the treatment is common, there may be little data with which to estimate the relationship between outcome and prognostic variables but plenty of data with which to estimate the relationship between treatment assignment and prognostic variables, that is, to estimate the propensity score. In this case, adjustments using the propensity score may be practical, whereas adjustments based on modeling the outcome may not. Logistic regression is a method used to model a binary (two-category) outcome. If several hundred patients are assigned to each of two treatments and 20 binary outcome events occur, then a logistic regression model for the propensity score may incorporate 30 prognostic variables, but a logistic regression model for the outcome cannot. This situation is not uncommon in studies of relatively rare side effects of standard drugs. (Standard logistic regression is not even possible unless the number of events is greater than the number of prognostic variables—otherwise, the maximum likelihood estimate does not exist—and many more events are needed for a stable model. If a small category of patients in a study, say 10 patients who incidentally have asthma, contains no side effects, then logistic regression will say that patients with asthma will never have side effects, even though there are very few data to warrant such a claim.)

In their study of treatments for latent tuberculosis infection, Jasmer and colleagues compare hepatotoxicity after treatment with isoniazid or rifampin plus pyrazinamide. Of the 411 patients for whom liver enzyme tests at 1 or 3 months were available (about half of whom received isoniazid), only 18 cases of grade 3 or 4 hepatotoxicity occurred; as result, there is little hope of building a good model relating grade 3 or 4 hepatotoxicity to prognostic variables. Patients were allocated to isoniazid or rifampin plus pyrazinamide in alternate weeks. Thus, although the study is not randomized, large biases would not be anticipated under normal circumstances.

Jasmer and colleagues estimated the propensity score—the probability of receiving isoniazid given observed prognostic variables—using a logistic regression model, and then stratified patients into five groups using the quintiles of the estimated score (6). Within these five strata, they found more cases of grade 3 or 4 hepatotoxicity in patients receiving rifampin plus pyrazinamide than in those receiving isoniazid.

Propensity scores can be used in several other ways. One application forms matched pairs of a treated patient and a control with similar propensity scores (7, 9). Another uses a logistic regression model to predict the treatment, in this case isoniazid or rifampin plus pyrazinamide, from prognostic variables plus the outcome, in this case hepatotoxicity; the model rejects the hypothesis that treatment has no effect on the outcome if the coefficient of outcome is a statistically significant predictor of treatment (2, 10). Provided that the treatment is common, this technique is practical even if the outcome is rare. All three approaches can be used in conjunction with models for the outcome; for an example, see Rosenbaum and Rubin's paper (6). When the comparison is not between treatment and control but rather between several doses of treatment, propensity score methods may still be used if the doses can be predicted from the prognostic variables by an ordinal logistic regression model (8, 11). A user of propensity scores should be aware of certain technical issues that are summarized briefly with limited technical detail in encyclopedia entries (10, 12, 13), and in greater technical detail in a textbook (2).

Addressing Hidden Biases

Propensity scores remove overt biases but do little or nothing to address hidden biases due to unobserved or unrecorded differences between treated and control patients before treatment. In an ideal clinical trial, randomization prevents hidden biases, although even experiments may need to address some hidden biases from protocol violations, such as frequent withdrawals of patients from treatment or extensive nonadherence by patients. In contrast, in an observational study, hidden bias is a serious and central problem that could undermine the study's conclusions.

The problem of hidden bias cannot be eliminated from an observational study, although it can often be reduced in magnitude, and it can always be discussed with candor (2, 13). A sensitivity analysis indicates the magnitude of hidden bias that would need to be present to materially alter the conclusions of a study (2–4, 13). For instance, although hidden bias could, in principle, explain away the association between heavy smoking and lung cancer, the magnitude of such a hidden bias would have to be enormous; thus, that association is highly insensitive to hidden bias (3). In contrast, the association between coffee consumption and myocardial infarction (14), although potentially a serious public health concern, is sensitive to comparatively small hidden biases (2). A finding that is sensitive to small hidden biases needs to be viewed with

greater caution; however, a sensitivity analysis does not demonstrate that the postulated hidden bias is present. As a result, a conclusion that is sensitive to small biases may nonetheless be correct and should not be dismissed on that basis. A sensitivity analysis is simply a device for candidly discussing and measuring the possible impact of hidden bias, which varies markedly from one study to the next. For a recent example, see Normand and colleagues' analysis (15). Familiar devices intended to shed light on hidden biases, such as Sir Austin Bradford Hill's famous criteria for causality (16) (for example, coherence, specificity, and a dose–response relationship), may be appraised in terms of their ability to reduce sensitivity to hidden bias (2, 13, 17).

SUMMARY

If many patients receive each of two competing treatments but the outcome under study rarely occurs, then there may not be enough data with which to model the relationship between the outcome and prognostic variables. However, there may be plenty of data with which to model the relationship between treatment assignment and prognostic variables. In this situation, among the several methods of adjustment for overt biases, the propensity score method has the advantage of not requiring modeling of the rare outcome event. The propensity score allows us to address the difficult problem of whether a rare outcome event is attributable to a treatment by simultaneously controlling for many measured covariates, even when there are too many covariates to model their relationships with the rare outcome. The principal limitation of all methods of adjusting for overt biases is that they do not address hidden biases from prognostic variables that were not measured. Hidden biases must be addressed by other means.

Leonard E. Braitman, PhD
Albert Einstein Medical Center
Philadelphia, PA 19141

Paul R. Rosenbaum, PhD
University of Pennsylvania
Philadelphia, PA 19104-6302

Grant Support: Dr. Rosenbaum is supported by grant SES-0004205 from the Methodology, Measurement and Statistics Program and the Statistics and Probability Program of the U.S. National Science Foundation.

Requests for Single Reprints: Leonard E. Braitman, PhD, Office for Research and Technology Development, Albert Einstein Medical Center, 5501 Old York Road, Philadelphia, PA 19141; e-mail, lbraitman@att.net.

Current author addresses are available at www.annals.org.

Ann Intern Med. 2002;137:693-695.

References

1. **Piantadosi S.** Clinical Trials: A Methodologic Perspective. New York: J Wiley; 1997.

2. **Rosenbaum PR.** Observational Studies. New York: Springer-Verlag; 2002.

3. **Cornfield J, Haenszel W, Hammond E, Lilienfeld A, Shimkin M, Wynder E.** Smoking and lung cancer: recent evidence and a discussion of some questions. J Natl Cancer Inst. 1959;22:173-203.

4. **Rosenbaum PR.** Discussing hidden bias in observational studies. Ann Intern Med. 1991;115:901-5. [PMID: 1952480]

5. **Jasmer RM, Saukkonen JJ, Blumberg HM, Daley CL, Bernardo J, Vittinghoff E, et al.** Short-course rifampin and pyrazinamide compared with isoniazid for latent tuberculosis infection: a multicenter clinical trial. The Short-Course Rifampin and Pyrazinamide for Tuberculosis Infection (SCRIPT) Study Investigators. Ann Intern Med. 2002;137:640-7.

6. **Rosenbaum PR, Rubin DB.** Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association. 1984;79:516-24.

7. **Rosenbaum PR, Rubin DB.** Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. American Statistician. 1985;39:33-8.

8. **Joffe MM, Rosenbaum PR.** Invited commentary: propensity scores. Am J Epidemiol. 1999;150:327-33. [PMID: 10453808]

9. **Bergstralh EJ, Kosanke JL, Jacobsen SJ.** Software for optimal matching in observational studies [Letter]. Epidemiology. 1996;7:331-2. [PMID: 8728456]

10. **Rosenbaum PR.** Propensity score. In: Colton T, Armitage P, eds. Encyclopedia of Biostatistics. Volume 5. New York: Wiley; 1998:3551-5.

11. **Lu B, Zanutto E, Hornik R, Rosenbaum PR.** Matching with doses in an observational study of a media campaign against drug abuse. Journal of the American Statistical Association. 2001;96:1245-53.

12. **Rosenbaum PR.** Multivariate matching methods. In: Kotz S, Read CR, Banks D, eds. Encyclopedia of Statistical Sciences, Update Volume 2. New York: J Wiley; 1998:435-8.

13. **Rosenbaum PR.** Observational studies. In: Smelser NJ, Baltes PB. International Encyclopedia of the Social and Behavioral Sciences. New York: Elsevier; 2001:10810-5.

14. **Jick H, Miettinen OS, Neff RK, Shapiro S, Heinonen OP, Slone D.** Coffee and myocardial infarction. N Engl J Med. 1973;289:63-7. [PMID: 4710407]

15. **Normand ST, Landrum MB, Guadagnoli E, Ayanian JZ, Ryan TJ, Cleary PD, et al.** Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. J Clin Epidemiol. 2001;54:387-98. [PMID: 11297888]

16. **Hill AB.** The environment and disease: association or causation? Proc Roy Soc Med. 1965;58:295-300.

17. **Rosenbaum PR.** Does a dose-response relationship reduce sensitivity to hidden bias? Biostatistics. 2002;3 [In press].

Current Author Address: Dr. Braitman: Office for Research and Technology Development, Albert Einstein Medical Center, 5501 Old York Road, Philadelphia, PA 19141.

Dr. Rosenbaum: Department of Statistics, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Locust Walk, Philadelphia, PA 19104.