Estimating Treatment Effects Using Observational Data

Ralph B. D'Agostino, Jr, PhD
Ralph B. D'Agostino, Sr, PhD

THE RANDOMIZED CLINICAL TRIAL (RCT) IS THE ideal method for measuring treatment effects. Participants in clinical trials are randomly assigned to a treatment or control group. Randomization reduces biases by making treatment and control groups "equal with respect to all features," except the treatment assignment. When randomization is performed correctly, differences in efficacy found by statistical comparisons can be attributed to the difference between the treatment and control.[1]

However, the RCT does not necessarily provide the final answer to treatment effectiveness, as there are many restrictions that limit generalizability. For example, RCTs are often restricted to patients with limited disease, comorbidity, and concomitant medications. Thus, RCTs generally demonstrate efficacy rather than effectiveness, where efficacy is the treatment effect under the restricted conditions of the RCT and effectiveness is the treatment effect under the conditions of usual practice.[1]

Observational, nonrandomized studies have a role when RCTs are not available, and, even when RCTs are available, to quantify effectiveness and other real world experiences. A contemporary example of this is the evaluation of drug-eluting stents, for which RCTs have demonstrated short-term efficacy for relatively healthy patients and observational studies are beginning to address long-term effectiveness and safety problems and use of clopidogrel in a broader array of patients.[2]

There are many approaches for making statistical inferences from observational data. Some approaches focus on study design, others on statistical techniques.[1] However, even with the best of designs, observational studies, unlike the RCTs, do not automatically control for selection biases. Therefore, statistical methods involving matching, stratification, and/or covariance adjustment are needed.

Lack of randomization in observational studies may result in large differences on the observed (and unobserved) participant characteristics between the treatment and control groups. These differences can lead to biased estimates of treatment effects. The goal of the statistical techniques that focus on observational data is to create an analysis that resembles what would occur had the treatment been randomly assigned.

In RCTs, the balance is achieved on participant characteristics that occur before the treatment is administered. The success of randomization in creating balance can be assessed before any outcome measurements are taken. Therefore, in observational studies, the first goal of a statistical technique is to create balance between treatments on characteristics that are assessed before the actual treatment is administered. Once this balance has been achieved, outcome measurements can be ascertained and compared between groups. In practice, this goal is often difficult to achieve because the data available for observational studies usually contain measured patient characteristics that are obtained before, during, and after treatment administration, and it is often difficult to determine exactly which patient characteristics are pretreatment or not. Furthermore, there frequently are unmeasured characteristics that are not available, inadequately measured, or unknown. Thus, the statistical methods in observational studies need first to be judged based on their performance in creating a balance on background characteristics between treated and control groups and the impact of outcome data should play no role in this assessment.[3]

To adjust for pretreatment imbalances, 2 statistical approaches often used are analysis of covariance methods and propensity score methods. These 2 methods complement each other and generally should be used together rather than choosing between one or the other.

Analysis of covariance refers here to standard statistical analyses that produce estimates of treatment effects adjusted for background characteristics (covariates), which are included explicitly in a statistical (regression) model. With observational studies, this technique can produce biased estimates of treatment effects if there is extreme imbalance of the background characteristics and/or the treatment effect

See also p 278.

Author Affiliations: Department of Biostatistical Sciences, Wake Forest University School of Medicine, Winston Salem, NC (Dr D'Agostino, Jr); and Department of Mathematics and Statistics, Boston University, and Harvard Clinical Research Institute, Boston, Mass (Dr D'Agostino, Sr).
Corresponding Author: Ralph B. D'Agostino, Sr, PhD, Department of Mathematics and Statistics, Boston University, 111 Cummington St, Boston, MA 02215 (ralph@bu.edu).

is not constant across values of the background characteristics.[1]

The propensity score for an individual is the probability of being treated conditional on the individual's background (pretreatment) characteristics.[4,5] Intuitively, the propensity score is a measure of the likelihood that an individual would have been treated based on his or her background characteristics. Mathematically, the propensity score is the probability (between 0 and 1) that a participant is in the "treated" group given his or her background (pretreatment) characteristics. This score is frequently estimated by using logistic regression, in which the treatment variable (treated yes or no) is the outcome and the background characteristics, not the study outcomes, are the predictor variables in the model. Matching, stratification, or regression (covariance) adjustment using the propensity score can be used to produce unbiased estimates of the treatment effects and create participant characteristic balance between groups. In some of these methods, the propensity score is used in the analyses as a weight or factor (regression adjustment), whereas in others it is used to construct the appropriate comparisons (stratification or matching) but not in the analyses directly.

In practice, the success of propensity score modeling is judged by whether balance on pretreatment characteristics is achieved between the treatment and control groups after its use. Because of this, the analysis can be more liberal with inclusion of covariates in the model than in most traditional settings. For instance, covariates with $P>.05$ can be included in the propensity score model. In addition, in the same way that randomization in a clinical trial will create balance on all patient characteristics, both those related to outcomes to be assessed later and those unrelated to outcomes, the focus should be on including variables in propensity score models that are unbalanced between the treated and control groups, and not necessarily be concerned specifically with whether they are related to the outcomes of interest.

Propensity score modeling should be assessed, as would randomization in an RCT, on its performance in creating balance and not on whether the eventual treatment-effect estimates are larger or smaller than expected. The decision about whether a propensity score model "worked" should be made based only on examining the characteristics measured on the participants before the collection of any outcome measures. The only way to assess whether unmeasured characteristics are balanced is to examine the balance on measured covariates to which they are related.

Once a propensity score model has been selected and is successful in creating a balance between groups on the observed characteristics (eg, by showing that within strata defined by the propensity score, the background characteristics are balanced when they were not balanced based on the overall data, or by creating matched pairs of individuals based on their propensity score and showing that the background characteristics on the matched pairs are balanced), the treatment effect can be estimated. This is where analysis of covariance can still be useful.

As in RCTs, it may be useful to include a set of important covariates in the final models to estimate the treatment effect to increase precision of the treatment effect estimate or reduce bias if the randomization process did not create perfect balance. After using the propensity score to create strata or matched pairs for analysis, a reasonable approach is to consider fitting a model to estimate the treatment effect that includes a subset of patient characteristics that are thought to be the most important known potential confounders. In this way, the investigator should be able to add precision to the treatment-effect estimate by taking advantage of the propensity score modeling to create balance and the analysis of covariance modeling to create precision (and adjust for any residual imbalances that may exist after the propensity score modeling).

In an observational study, the participants and their physicians self-select for either receiving the treatment or not receiving it, which may limit interpretation. However, even in RCTs, there is self-selection by the patient.[1] For example, patients who decide to allow themselves to be randomized to receive a treatment (such as cardiac catheterization) are not necessarily a random sample of all potential patients. In fact, in many cases randomized trials have inclusion and exclusion criteria that restrict participation, such as by the participant's age, health status, and medication use, so that in fact the participants in a trial may not resemble closely the actual individuals who may take the treatment once available. In general, observational study data resemble more closely the real world; that is, they include all individuals who are eligible to have the procedure/treatment, not only the subset of individuals who are comfortable with being randomized to receive a treatment, and who fit into the particular inclusion/exclusion criteria of a trial. Evaluation of a propensity score analysis of observational data because it does not match perfectly RCT results may miss the mark completely. Focus should be on understanding the differences (patient population, less experienced health care professionals and organizations) between the different studies.

In this issue of *JAMA*, Stukel and colleagues[6] compare statistical methods for addressing selection biases in observational studies by using Medicare data on the use and survival outcomes of management by cardiac catheterization after acute myocardial infarction. In addition to the propensity scores methods, the authors also use an instrumental variable analysis. Using these approaches, the authors obtain different estimates of treatment effects. As the instrumental variable method gives a treatment effect

closer to what RCTs produce, the authors imply the instrumental variable method is better than the propensity score method because it eliminates the bias due to unobserved variables.

The attempt by Stukel et al to highlight the instrumental variable method is useful and when properly performed offers great hope. Their suggestions that this approach is "better" than the propensity score analysis for this particular data example and can deal better with selection biases from unmeasured variables is theoretically possible. However, 4 issues must be carefully considered.

First, selection of the instrumental variable for this study creates some problems. Instrumental variable methods are 2-stage regression methods in which an instrumental variable related to the decision to have the treatment but not related to the outcome of the study is used to reduce or remove the bias due to the unobserved baseline characteristics. The ability to identify a real instrumental variable is controversial, difficult to justify and understand, and may create confusion in interpretation.[7] The instrumental variable used by Stukel et al, "regional cardiac catheterization rate," is a useful variable in that it correlates to the decision to treat. It also relates, however, to a successful outcome because it is a system variable indicative, as the authors state, to "more high-volume hospitals with specialized staff and equipment and coronary care units."[6] The status of this variable as an instrumental variable is not completely obvious. Its role, however, in a propensity score analysis to attain balance was not considered by the authors. It would be helpful to have a clear explanation as to why the regional characteristic used as the instrumental variable (or some set of variables closely correlated with this variable) was not included in the propensity score analysis, and if it were included what impact would it have.

Second, it is difficult to determine which mortality rate estimates and which analyses are most believable. The propensity analysis gives a treatment-related mortality reduction rate of 0.55. The propensity analysis did balance the groups on measured characteristics (as shown in columns 5, 6, and 7 of Table 1 in the article by Stukel et al[6]). The 0.55 effect estimated from these data is either correct or there must be some unmeasured characteristics that are strongly imbalanced between the 2 groups and not being properly taken into account in the propensity score analysis. Does "regional cardiac catheterization rate" account for this? Would its addition, or the addition of a set of variables closely related to it, to the propensity score analysis adjust the mortality rate?

Third, the comparison with existing RCTs data may not necessarily be a gold standard. The question of comparable populations of the RCT and observational studies reflecting efficacy vs effectiveness does not seem to be completely resolved by the authors.

Fourth, the authors acknowledge, "instrumental variable analyses . . . are more suited to answer policy questions than to provide insight into a specific clinical question for a specific patient."[6] Treatment effects should deal with effects relevant to patients. Such statements as given by the author cloud the issue.

In conclusion, the article by Stukel et al is an important reminder of the need for careful and rigorous approaches to observational data analyses. Because the final inferences appear different depending on the method chosen, investigators must be cautious when conducting observational data analyses and must ensure that they have available what they consider to be the most important patient characteristics measured before treatment assignment. Furthermore, the analytic method for comparing treatments must be shown to properly balance these characteristics. In addition, sensitivity analyses also should be performed in much the same way as Stukel et al did. Moreover, external validation of results should be attempted, but always with caution. RCTs should not always be considered as the only source of valid scientific information. The data collected from such studies are strong only if it can be shown that in fact a truly random sample of eligible patients participate and complete the protocol as designed. When patients self-select to be included in observational studies, the findings may more accurately reflect "real world" experience, but if and only if optimal, rigorous, and appropriate methods for dealing with selection bias and confounding are part of the analytic plan.

Financial Disclosures: None reported.

REFERENCES

1. D'Agostino RB Sr, Kwan H. Measuring effectiveness: what to expect without a randomized control group. *Med Care*. 1995;33(4 suppl):AS95-AS105.
2. Eisenstein EI, Anstrom KJ, Kong DF, et al. Clopidogrel use and long-term clinical outcomes after drug-eluting stent implantation. *JAMA*. 2007;297:159-168.
3. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized studies. *Stat Med*. 2007;26:20-36.
4. 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;17:2265-2281.
5. Rosenbaum P, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41-55.
6. Stukel TA, Fisher ES, Wennberg DE, et al. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA*. 2007;297:278-285.
7. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? analysis using instrumental variables. *JAMA*. 1994;272:859-866.