Analyzing Retrospective Data: The Clinical Implications of Statistical Methods

Lee A. Fleisher, MD, FACC, FAHA

The decision to perform a clinical intervention is dependent upon the clinician's prior experience, clinical findings and reading (or interpretation) of the literature, and the assessment of whether the probability of disease reaches some threshold (Fig. 1).

To interpret the literature and apply the principles of evidence-based medicine, it is important to understand the strengths and limitations of the different study designs and their applicability to a given clinical situation. The gold standard for evidence of causation and justification for action is the prospective randomized clinical trial (RCT). RCTs have defined inclusion and exclusion criteria, treatment protocols, and outcomes of interest. They are usually either single of double-blind (both patient and physician) and are designed to test the effect of a drug or intervention.

Randomized clinical trials derive their strength from an evidence-based perspective because of their high degree of internal validity i.e., the randomization scheme and use of placebo (or accepted alternative treatments) provide strong evidence that the results are related to the intervention. If performed properly and with a sufficient sample size, the randomization should ensure that all important variables are distributed evenly, even unidentified confounders. Importantly, these trials have a lower degree of external validity because the intervention may not behave in the same manner as when it is diffused into a more heterogeneous population in whom treatment is not defined. Therefore, it is important to determine if the results of the study can be applied to the specific clinical situation of interest.

In many instances, there is insufficient evidence to justify a randomized clinical trial or it is important to determine how an intervention works in a different population than previously studied in an RCT. In these situations, cohort studies can be utilized to study the question of interest. Although the topic of this lecture is the analysis of retrospective studies, the approach can also be applied to any cohort study, These types of studies are frequently "hypothesis generating" rather than designed to prove or disprove a hypothesis (the goal of the RCT). Prospective cohort studies involve the identification of a group of subjects who are followed over time for the occurrence of an outcome. The goal is to determine those patients who develop the outcome, and those factors which are associated with the development of morbidity or mortality can be discerned. An example of a prospective cohort study identifying factors associated with perioperative cardiac morbidity and mortality is that of Goldman and colleagues, which led to the development of the cardiac risk index (1).

Another example of a prospective cohort study is one in which patients with a known disease are studied for the development of predefined outcomes. Such studies provide the natural history of patients with the disease. An example would be studies of patients who have sustained a myocardial infarction, the importance of which is that the optimal time between the infarct and surgery can be determined (2–4).

An important strength of any prospectively collected data is that the method of surveillance for an outcome of interest can be defined *a priori*, whereas this may not be true in retrospective studies. For example, studies which focus on the incidence of perioperative myocardial infarction are dependent both on the definition of an event and the frequency with which surveillance laboratories are obtained to detect that event.

Although prospective cohort studies have great value in identifying risk factors for the outcome of interest, there are significant limitations. The selection of the cohort of interest can significantly impact the results obtained. The larger the cohort, the more the results can be generalized. A second bias is that many patients may be lost to follow-up. In perioperative studies, this may not be an important issue for short-term outcomes. Finally, the importance of a risk factor depends upon the completeness of the data. For example, if the presence of severe angina was not included in the database, then it could not be a risk factor and other factors may appear to be more important (1).

Evidence gathered from retrospective trials is considered weaker than prospective studies, but may offer an excellent means of further generating hypothesis without collecting new data (5). All of the previously discussed limitations of the prospective cohort study are also applicable to the retrospective study. Unlike prospective studies, retrospective studies are also totally dependent on the data collected in the medical record or billing (discharge) data. In many instances, identification of the outcome of interest was

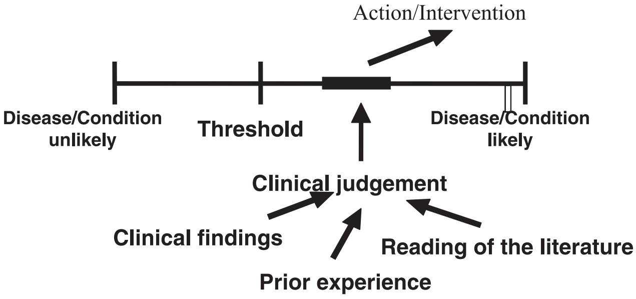

Figure 1. Factors incorporated in medical decision making regarding the appropriateness of performing an action or intervention on a given patient.

not performed in a systematic method. For example, the frequency of obtaining electrocardiograms and serial biomarkers was not consistent in a study with perioperative myocardial infarction as an outcome.

Retrospective analyses have the advantage of easily analyzing a large number of patients and can include traditional chart reviews and more recently administrative databases. Examples of administrative databases include Medicare claims files, private insurance company claims, and hospital electronic records. These databases include a small number of data points on an extremely large number of subjects. For example, the Medicare database includes both financial data and ICD-9 (disease) and CPT (procedure) codes for each patient. They also include information regarding location of care and provider type. The Medicare claims file is now being extensively utilized to benchmark rates of mortality and major complications after coronary bypass surgery (6). Hospitals can compare their rates with those of neighboring and competing hospitals, and may use this data as markers for quality of hospital care (7,8).

The major issue in evaluating retrospective data is the potential influence of confounders. Bias in retrospective studies is a serious threat to the validity of the findings. Bias can come in a myriad of forms, and it may be impossible for statistical analysis to detect and eliminate bias. The traditional method to statistically adjust for bias is the use of multivariate analysis.

Multivariate analysis provides an estimate of the relationship between the risk factor (or intervention) and the outcome after adjusting for the differences of known or suspected confounders between the risk factor (or treatment) groups. (Fig. 2) Multivariate analysis can involved three different types of analysis: multiple linear regression for continuous outcomes (e.g., total minutes of anesthesia, total cost), multiple logistic regression for categorical or dichotomous outcomes (e.g., myocardial infarction or death), and proportional hazards regression for time to event outcomes (e.g., time to death).

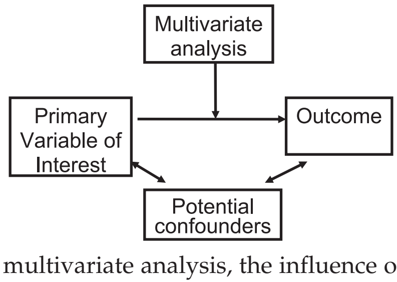

Figure 2. In multivariate analysis, the influence of confounders on both the primary variable of interest (intervention) and outcome is assessed.

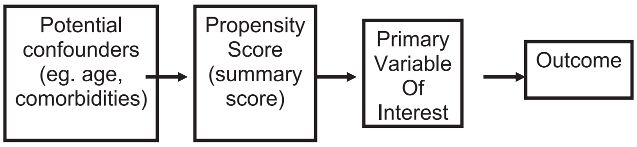
Potential confounders (eg. age, comorbidities)
Propensity Score (summary score)
Primary Variable Of Interest
Outcome

Figure 3. In propensity analysis, the propensity score is calculated from the confounders for the primary variable of interest. The relationship between the propensity score and the primary variable of interest is modeled.

In a multiple regression analysis, a regression equation is developed that relates the outcome of interest to all of the explanatory variables in the study. Because several explanatory variables are involved, each has a "slope" or regression coefficient associated with it. For example,

$$\text{Days in the ICU} = \text{Intercept} + (\text{age} \times \text{coefficient 1})$$
$$+ (\text{presence of heart disease} \times \text{coefficient 2})$$
$$+ (\text{presence of diabetes} \times \text{coefficient 2}) + \ldots\ \ldots$$

This multivariate linear regression is able to look at the influence of each of these potential variable on the outcome of ICU days with the strength of the relationship incorporated into the coefficients.

Recently, propensity scores are increasingly being used to help determine whether the outcome differences seen are true effects of the treatment or just a sign that the risk factors for the outcome were not evenly distributed between the groups (9). Propensity scores are different from regression analyses because they take into account the variable's influence on the likelihood for the subject to receive treatment, the variable's impact on outcome, and the variable's impact on the relationship between treatment (or nontreatment) and outcome. Such analyses increase in importance in studies where randomization is impossible or impractical.

In effect, propensity score analysis attempts to reconstruct a situation similar to randomization. The propensity score is calculated by building a regression model with the treatment as the dependent variable (Fig. 3). In cohort studies, the chance of receiving different treatments is frequently a function of different baseline characteristics such as age and comorbidities. Therefore, the different treatment groups may have differing baseline characteristics. As demonstrated above, multivariate analysis evaluates the influence of these characteristics on the outcome. In propensity analysis, patients with a similar chance of undergoing the treatments are compared.

There are three common approaches to propensity analysis. One approach is matching by propensity scores. Propensity scoring summarizes all potential confounders into a single score, allowing a better chance of matching than matching on individual confounders. The advantage of this technique is that once the scores are calculated and matching has been performed, the relationship between the primary variable or intervention of interest and multiple outcomes can easily be calculated. This is unlike multivariate modeling in which new models must be created for each outcome of interest. One example of matching by propensity scoring was utilized in the SUPPORT trial in which patients with a pulmonary artery catheter were matched with patients without a catheter but who had the same aggregate propensity for receiving a catheter (10). Patients who received a pulmonary artery catheter were significantly more likely to die in the intensive care unit than those in the matched, equally ill cohort who did not.

Stratification is another form of analysis. Patients are grouped into different strata by their propensity score. An example of such an analysis was performed by Lindenauer and colleagues to evaluate the effect of treatment with lipid-lowering medications and in-hospital mortality following major noncardiac surgery (11). Propensity matching was utilized to adjust for differences in treatment and after adjusting for quintile of propensity, a significant effect of treatment was observed in which receiving lipid lowering agents on the first two days after surgery was associated with lower mortality.

Finally, the propensity score can actually be utilized in a regression model. In such a manner, the propensity score can be the only confounding variable.

In interpreting the literature, the reader of a propensity score analysis should ask two questions:

1. Matching on the propensity is intended to balance observed prognostic variables. Did it? The authors should include a table showing it did.
2. Propensity score only balance the prognostic variables used to construct the score—the variables in the table in Ref. 1. Did the authors fail to measure some important variable? Are there important variables not in the table?

If these conditions for question 1 are met and there is little likelihood that important variable were not measured, then it is likely that the propensity analysis did prove good evidence of a strong association. Some authors have attempted to use sensitivity analysis to determine if some unmeasured variable would change the results.

REFERENCES

1. Goldman L, Caldera DL, Nussbaum SR. Multifactorial index of cardiac risk in noncardiac surgical procedures. N Engl J Med 1977;297:845–50.
2. Tarhan S, Moffitt EA, Taylor WF, Giuliani ER. Myocardial infarction after general anesthesia. JAMA 1972;220:1451–4.
3. Rao TLK, Jacobs KH, El-Etr AA. Reinfarction following anesthesia in patients with myocardial infarction. Anesthesiology 1983;59:499–505.
4. Shah KB, Kleinman BS, Sami H, et al. Reevaluation of perioperative myocardial infarction in patients with prior myocardial infarction undergoing noncardiac operations. Anesth Analg 1990;71:231–5.
5. Ochroch EA, Fleisher LA. Retrospective analysis: looking backward to point the way forward. Anesthesiology 2006;105:643–4.
6. Ghali WA, Ash AS, Hall RE, Moskowitz MA. Statewide quality improvement initiatives and mortality after cardiac surgery. JAMA 1997;277:379–82.
7. Mukamel DB, Mushlin AI. Quality of care information makes a difference: an analysis of market share and price changes after publication of the New York State Cardiac Surgery Mortality Reports. Med Care 1998;36:945–54.
8. Hannan EL, Racz MJ, Jollis JG, Peterson ED. Using Medicare claims data to assess provider quality for CABG surgery: does it work well enough? Health Serv Res 1997;31:659–78.
9. Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. Am J Epidemiol 1999;150:327–33.
10. Connors AF Jr, Speroff T, Dawson NV, et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators. JAMA 1996;276:889–97.
11. Lindenauer PK, Pekow P, Wang K, et al. Lipid-lowering therapy and in-hospital mortality following major noncardiac surgery. JAMA 2004;291:2092–9.