Editorial

Observational Studies and Predictive Models

L. Richard Smith, PhD

Key Words: STATISTICS.

The paper by Shah and colleagues (hereafter referred to as Shah), "Angina and Other Risk Factors in Patients With Cardiac Diseases Undergoing Noncardiac Operations" (1), is an important study because of the prevalence of cardiac patients that require noncardiac surgery. It is also important because it is a paradigm of a number of studies that have appeared in the literature recently (e.g., 2,3). Statistical analysis plays a central role in these studies. The purpose of this editorial is to use Shah's study as a point of departure to expand on the statistical methodology and suggest some other methodologies for future investigations. This editorial will concentrate on the design of these studies, patient selection, variable selection, statistical model selection and evaluation, and interpretation of results. The principles and pitfalls of the statistical methodology are crucial in evaluating these epidemiologic studies.

There are two classes of observational studies in the clinical setting, retrospective and prospective. In retrospective studies, past patient records are reviewed to collect the relevant prognostic and outcome variables. In prospective studies the relevant prognostic and outcome variables are collected from patients as they are treated and followed.

Retrospective studies are generally quicker and cheaper to conduct because patients have already been treated and the data collected. There are, however, several problems with retrospective studies. Among them are the following: patient records may be lost; definition of the variables may have changed over the time period of the study; and data required for assessing outcome may be missing. Furthermore, these problems may be related to the type of surgery or seriousness of the patient's illness, thus creating serious biases in the results. If a patient has any one of the prognostic variables missing, that patient drops out of any analysis in which the variable is missing. This reduces the sample size and hence reduces the power of the model to predict outcomes.

Prospective studies, on the other hand, take longer to conduct and thus are more expensive. But with careful design, prospective studies permit unbiased patient selection (reducing the likelihood of overlooking patients) and consistent observation and data collection for the length of the study.

Shah chose to study prospectively 688 consecutive cardiac or elderly (>70 yr) patients undergoing various noncardiac surgeries. Twenty-four prognostic variables and two adverse outcome variables were collected. Complete data were collected on all patients. The prognostic variables were analyzed to determine which subset of them was predictive of an adverse outcome, perioperative myocardial infarction (PMI) and/or cardiac death during the hospitalization.

When one goal of the study is to determine the relative risk of adverse outcomes in two or more groups of patients, it is important that patients be stratified into groups that are relatively homogeneous with respect to procedure. Shah classified patients into three groups based on surgical procedure: thoracic and abdominal, aortic or other peripheral vascular disease, and an "all other" group. An "all other" group may be so heterogeneous with regard to risk that any effect will be diluted. If the "all other" group is not at significantly higher or lower risk than the well-defined groups, it can be difficult to generalize from the results. This can be circumvented by

Received from the Division of Biometry and Medical Informatics, Department of Community and Family Medicine, Duke University Medical Center, Durham, North Carolina. Accepted for publication November 15, 1989.

Address correspondence to Dr. Smith, Box 3391, Duke University Medical Center, Durham, NC 27710.

including in the analysis only those groups that are well defined.

In any clinical study, variable selection is very important. Two sets of variables are required to be defined: the outcome variables and those variables known as prognostic variables which may predict outcome variables. All variables should have commonly understood definitions with conventional meanings, with minimal room for interpretation by the observers or data collectors. The greater the accuracy in assessing the variables, the greater the predictive power of the model, and the fewer the number of patients required to achieve meaningful results. For example, Shah has defined PMI explicitly and narrowly as an increase of serum CK-MB isoenzyme levels with the presence of at least one of two other criteria: chest pain or electrocardiographic evidence of subendocardial or transmural myocardial infarction. An independent confirmation of PMI by a cardiologist not directly involved with the patient or the study was required. Cardiac death was defined as death after PMI, after documented cardiac dysrhythmia or cardiogenic shock, or when death was sudden and unexplained. Either PMI or cardiac death, or both, constituted the adverse outcome for the study. Thus, the outcome variables are well defined and easily understood. The prognostic variables are similarly well defined.

Among the prognostic variables, two types are of interest: those that cannot be altered but can be used for improved patient selection (e.g., age, prior myocardial infarction), or those that can be altered to reduce risk (e.g., hematocrit, blood pressure). For surgical studies, prognostic variables represent the patient's condition before surgery or represent those factors that are determined before surgery, such as anesthetic or surgical technique. This precludes the use of variables occurring after the beginning of surgery. Variables such as length of time in surgery or episodes of hypotension during surgery are outcome variables. For example, even though perioperative hypotension may be strongly associated with adverse outcomes (indeed, it is an adverse outcome itself), it cannot be used as a selection criterion for surgery nor can it be altered until it occurs—after surgery begins. Whatever underlying physiologic process causes perioperative hypotension may also predispose the patient to PMI or cardiac death. Shah has clearly used the preoperative variables only, although he alludes to measuring and analyzing the correlations of some intraoperative factors with adverse outcome.

There is a tendency to collect as many variables as possible, analyzing all of them with regard to the outcome variables, hoping not to miss any important prognostic factor. There is a limit, however, to the number of variables that can be effectively analyzed jointly. Harrell et al. (4) suggested approximately one variable for every 10 observations in the least frequent category of outcome variable. With the 40 adverse outcomes in Shah's study, this rule of thumb would suggest that only four variables could be jointly evaluated effectively. To analyze 24 variables and keep the 10-to-1 ratio would require 240 adverse outcomes for the predictive model to be trustworthy. By fitting many variables to few outcomes, there is a likelihood of an unreliable model—a model that results in inaccurate predictions on an independent patient sample.

There are several ways to avoid the problem of too many variables. One common method is to analyze each prognostic variable separately for its influence on outcome, then use only those variables that are "significant" at some predetermined level. This technique raises a multiple comparison problem. The likelihood of finding spuriously significant factors is increased when many variables are examined. In addition, very often several of the significant variables will be highly correlated, and, when they all are included in the model, only one or two will be significant. Another serious problem with this technique is that some important prognostic variables may be overlooked. Variables that are not significant by themselves may become significant in combination with other variables.

Another approach centers on collapsing groups of like variables into one index. Gersh et al. (5), in analyzing the effect of coronary bypass surgery on patients over the age of 65 yr, collapsed several associated disease variables into one variable by simply counting the number of associated diseases. Similarly, Hickey et al. (6) summed the number of each patient's comorbid disorders in a group of patients undergoing treatment for ischemic mitral regurgitation. In both studies, these count variables proved to be important prognostic factors. Califf et al. (7) produced a strong predictor of adverse outcomes by grouping anginal characteristics into a single angina score, and Harrell et al. (8) showed how 30 original variables could be reduced to 10 entities by linear combinations of like variables. Increased use of such indices that *do not use the outcome data* can reduce the number of prognostic variables and avoid the problem of model instability.

How an individual variable is represented in the model is also of importance. Continuous variables, such as age, are often dichotomized into two ranges for ease of explication. This often masks important

effects. In Shah's study, age is dichotomized into two groups: <70 yr and ≥70 yr. This assumes that the risk is constant for every age up to the age of 70 yr, whereupon the risk jumps to a higher level and is constant thereafter. By including age as a continuous variable, an incremental risk could be estimated for any age increment, thus improving the predictive power of the model. Even this may not be sufficient. There is evidence that increasing age increases risk of adverse outcomes at a fixed rate up to (approximately) age 65 yr, after which there is a dramatic increase in the incremental risk for an increase in age. New techniques (9) have been recently developed to model such nonlinear effects of the prognostic variables using splines or segmented polynomials.

Model selection for analysis of binary outcomes is of primary importance. There are two candidate statistical models for this analysis: the linear discriminant function (LDF) and the logistic regression model (LRM). Both the LDF and the LRM classify patients based on a linear combination of the prognostic factors, $a_0 + a_1x_1 + \ldots + a_nx_n$. The a_i are coefficients to be estimated, and the x_i are the prognostic variables. If an individual a_i is significantly different from zero, we conclude that the corresponding variable is a significant factor. The LDF, which is a linear regression model for classification, is a very powerful tool when the assumption of jointly normally distributed prognostic variables holds. When this multivariate normality does not hold, such as when the prognostic variables are dichotomized, the results can be biased. The LRM considers P to be the probability of being classified in the adverse outcome group and assumes that $\log(P/1 - P)$ is a linear combination of the covariates. The LRM model makes no assumptions about joint multivariate normality. It produces a probability from 0 to 1 of the outcome. Harrell and Lee (10) made a study of the LDF and LRM and concluded that the LRM was very nearly as good as the LDF when multivariate normality holds and that the LRM is superior to the LDF when multivariate normality is not achieved. A further reason for use of the logistic regression model is the ease of interpreting the coefficients. Shah gives a nice explication of this in Table 3. Shah has chosen the logistic regression for the correct reasons: that multivariate normality is violated and would introduce unknown biases into the LDF analysis, and that the LRM lends itself to ease of interpretation. Hosmer and Lemeshow (11) provide an excellent overview of this subject. The LRM is the model of choice here, as it is in most studies involving acute outcomes.

Shah has also taken the necessary step of testing the adequacy of the model. Using the prognostic variables found from the model based on the entire population, the model was refitted on half the data (the observation sample). These new parameter estimates were then used to estimate the probability of adverse outcome for each of the patients in the second half of the data (the holdout sample), and predictions were compared with observed outcomes. The holdout sample is not a truly independent test because the selection of variables was based on the entire study population, which includes the holdout sample. It would be interesting to know how the values of the parameter estimates and their standard deviations changed from the model based on the entire sample of 688 patients to the model based on the observation sample with 336 patients. These changes could shed some light on the stability of the model. For instance, if the coefficient for a prognostic variable changes sign, or its magnitude changes dramatically, this could signal some model instability. Shah plots the probability of adverse outcome versus case number for those patients in the holdout group who had an adverse outcome and for those who did not. The median probability for the adverse-outcome group is 0.09 and for the no-adverse-outcome group is 0.02. The model predicts higher probabilities of adverse outcome for the adverse-outcome group, but more information is needed about the predictive accuracy of the model. Perhaps the simplest information to provide is the sensitivity and specificity for the model. The sensitivity of a test is the proportion of patients with adverse outcomes that are predicted to have adverse outcomes. The specificity is the proportion of patients free of adverse outcomes that are predicted to be free of adverse outcomes. The sensitivity and specificity are dependent on the probability level at which discrimination is to occur. If this probability level is arbitrarily set to 0.50 for testing model adequacy, one infers from Shah's plot that the model has low sensitivity because so few patients in the adverse-outcome group have a probability of adverse outcome greater than 0.50. Sensitivity and specificity, however, have their own problems. Two patients with probabilities of adverse outcomes of 0.49 and 0.51 would be classified into groups 0 and 1 respectively, even though they are very similar in risk.

Of course, the 0.50 discrimination level for model evaluation may not be the point at which a clinician makes a decision about a patient. Depending on the risk to the patient of not having surgery, the clinician may choose a higher or lower discrimination level. A more comprehensive measure of model adequacy is required. If the discrimination level is varied from 0 to 1 and the sensitivity and specificity calculated at each point, the receiver operating characteristic curve (12)

can be constructed by plotting the sensitivity versus 1 − specificity. The area under the receiver operating characteristic curve is a powerful measure of predictive discrimination. It is the probability that in randomly selected pairs of patients where one patient has an adverse outcome and the other does not, the patient with the adverse outcome is the one with the higher predicted probability. This measure provides a single number for the model's discrimination ability. A direct comparison of the discriminatory power of two different models can be made by statistically comparing the areas under the receiver operating characteristic curves.

Finally, the model's predictive reliability should be validated on the scale of absolute predicted risk. One way this can be done is by comparing each quintile of risk with the observed prevalence in the quintile. Large differences between observed and predicted risks would indicate an unreliable model. Because the use of the LDF and LRM has become common in anesthesiology clinical studies, there needs to be a consistent standard for reporting and testing the adequacy of predictive models.

In the model used by Shah, all the prognostic variables are dichotomized into 0 or 1, depending on whether the variable is present in the patient. The base, or reference, model, with all prognostic factors set to zero, represents the probability of an adverse outcome for a patient in the "all other" group who has no other risk factors. Inferences beyond the range of the data should be made with care, however. For example, if a 1 is substituted for each of the definitive surgical groups, with all other factors being 0, this would describe a patient with no risk factors having two different kinds of surgery. In this case, the probability of adverse outcome is 0.03 (3%). There is some danger in too literal an interpretation here. No patients in the study had two or more kinds of surgery, although the model allows for such a circumstance.

In presenting the results from the analyses, Shah includes an estimate of each effect, standard deviation of the estimate, significance levels, and the estimated probability of adverse outcome for the specific factor when considered alone. Presentation of the parameter estimates (rather than just P values) is important because it allows the reader to judge the magnitude and direction of the effect. It is interesting that Shah has not slavishly adhered to the "$P < 0.05$" rule for including factors in the model. There may well be cases in which factors with P values greater than such an arbitrary cut-off should be included. The groups may have small differences in important prognostic factors, none of which would be individually "significant." The cumulative effect, however, could be important. Furthermore, including a factor known to be important may change the estimate or variance of the coefficient of another factor. Choice of the "nonsignificant" factors for inclusion depends on both the investigator's knowledge of the physiological process and the interrelationships among all the prognostic factors.

This study has provided useful insights into the processes leading to adverse outcomes for cardiac patients undergoing noncardiac surgery. The results of the analysis have identified factors that appear to influence adverse outcome but may not be strongly predictive. Before a model is put to clinical use, however, it should be tested on an independent sample, preferably at another institution. Multiinstitutional studies would provide a mechanism for developing and testing these complex models.

The basic tenets of prospective patient recruitment, variable selection and definition, choice of statistical model, and assessment of model adequacy are necessary requirements for clinical studies. The role of the statistician in these studies is as a coinvestigator who must be involved at every step of the study. Beginning with experimental design, the statistician can provide estimates of sample size based on the estimated adverse outcome proportion, the number of prognostic variables which may be analyzed effectively, and the magnitude of prognostic variable effects. The choice of model is of critical concern. Incorrect model choice may yield biased estimates of the prognostic variables effect, or perhaps incorrectly identify significant prognostic variables. Assessment of model adequacy provides the reader with the information required to judge whether the model is robust enough to be used in other institutions or studies. Finally, because of the important roles of statistics in these studies, consistent editorial policies regarding statistical review of manuscripts need to be developed and applied.

The author thanks J. G. Reves, MD, F. E. Harrell Jr., PhD, W. D. White, MPH, and B. M. Lovell, MA for their critical review of the manuscript.

References

1. Shah KB, Kleinman BS, Rao TLK, Jacobs HK, Mestan K, Schaafsma M. Angina and other risk factors in patients with cardiac diseases undergoing noncardiac operations. Anesth Analg 1990;70:240–7.
2. Tuman KJ, McCarthy RJ, Speiss BD, DaValle M, Dabir R, Ivankovich AD. Does choice of anesthetic agent significantly affect outcome after coronary artery surgery? Anesthesiology 1989;70:189–98.

3. Tuman KJ, McCarthy RJ, Speiss BD, et al. Effect of pulmonary artery catheterization on outcome in patients undergoing coronary artery surgery. Anesthesiology 1989;70:199–206.
4. Harrell FE Jr, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems and suggested solutions. Cancer Treatment Rep 1985;69:1071–7.
5. Gersh BJ, Kronmal RA, Schaff HV, et al. Comparison of coronary artery bypass surgery and medical therapy in patients 65 years of age or older. N Engl J Med 1985;313:217–24.
6. Hickey MS, Smith LR, Muhlbaier LH, et al. Current prognosis of ischemic mitral regurgitation. Circulation 1988;78(Suppl 1):I-51–9.
7. Califf RM, Mark DB, Harrell FE Jr, et al. Importance of clinical measures of ischemia in the prognosis of patients with documented coronary artery disease. J Am Coll Cardiol 1988;11: 20–6.
8. Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved diagnostic prediction. Stat Med 1984;3:143–52.
9. Harrell FE Jr, Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. JNCI 1988;80:1198–202.
10. Harrell FE Jr, Lee KL. A comparison of the *discrimination* of discriminant analysis and logistic regression under multivariate normality. In: Sen PK, ed. Biostatistics: statistics in biomedical, public health, and environmental services. Amsterdam: Elsevier Science Publishers B.V., 1985:333–43.
11. Hosmer DW Jr, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons, 1989.
12. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.