JACC Vol. 21, No. 3
March 1, 1993:835–7

EDITORIAL

It Is All in the Numbers

STANTON A. GLANTZ, PHD, FACC, ASSOCIATE EDITOR, *JACC*
San Francisco, California

The new editors of the *Journal of the American College of Cardiology* have decided that one of the editors should be responsible for ensuring the statistical accuracy of studies published in the Journal. We have implemented this quality control by ensuring that every manuscript that is returned to an author for revisions is first reviewed for statistical issues. Although I have not done a formal study of the kinds of problems appearing in the manuscripts, several recurring areas of difficulty have emerged. About one third of the papers have no statistical problems when they are initially accepted. Occasionally, a paper that would have been acceptable for other reasons is rejected on purely statistical grounds. Most of the time, the papers have some methodologic difficulties that the authors are expected to fix in revising the manuscript; sometimes correcting the statistical methods actually leads to alterations in the paper's conclusions.

The problems identified in manuscripts submitted to JACC—inappropriate use of the *t* test, use of parametric statistical tests when the underlying assumptions are badly violated, problems with the study design that introduce biases toward treatments and drawing negative conclusions on the basis of small sample sizes—are typical problems that have been identified in several reviews of the medical literature over the years (1). Each of these problems is briefly reviewed here, together with some guidelines on avoiding them.

Inappropriate use of the *t* test. The Student *t* test is the most commonly used statistical method in biomedical research. The unpaired *t* test is used appropriately to compare the mean responses of two treatments or two groups of different individuals, as in a comparison of mean blood pressure in men and women; the paired *t* test is used to compare mean responses in the same individuals before and after an intervention, as in a comparison of blood pressures measured in the same patients before and after they received a drug.

In addition to these appropriate uses, the *t* test is widely misapplied to compare treatments or responses when there are three or more treatment conditions by comparing each pair of treatments two at a time. For example, if one is comparing a placebo and two drugs it is common to use *t* tests to compare placebo with drug A, placebo with drug B, and drug A with drug B. Because these comparisons are not independent, this procedure violates the assumptions underlying the *t* test. If a statistical comparison is considered to reveal a significant difference when $p < 0.05$ for each individual test, then in making the three comparisons just mentioned the overall risk of erroneously concluding that one of the treatments had an effect is approximately $3 \times 0.05 = 15\%$ (because there is a 5% false positive error rate in each of the individual tests). Thus, by using multiple *t* tests to compare more than two groups, the risk of a false positive conclusion—that is, asserting that a treatment had an effect when in fact it did not—is higher than the nominal 5% risk used for the individual tests. Similar problems arise when comparing interventions over time with multiple *t* tests.

The appropriate statistical method to use in this situation is not the *t* test but analysis of variance, the multigroup generalization of the *t* test. There are two broad categories of analysis of variance. Factorial analysis of variance is the generalization of the unpaired *t* test, which involves comparing three or more treatment groups consisting of different individuals. Repeated measures analysis of variance is the generalization of the paired *t* test, which is used for analyzing data collected in the same individuals over multiple interventions. If an analysis of variance, which examines *all* of the data at once, detects a significant difference, it is then reasonable to use one of many multiple comparison procedures to isolate which of the groups are different.

Another situation in which people often misuse *t* tests is to compare two relationships over time—say, the response over time of blood pressure after giving two different drugs. It is common to compare the responses to the two drugs at each time period or the responses over time to each drug by using multiple *t* tests. A more accurate and sensitive method would be to use a linear regression model to describe the temporal relationship and then compare the linear regressions obtained over time for the different conditions (2). It is possible to collapse the entire analysis into a single multiple regression analysis using dummy variables (2) as well as to account for repeated observations in the same individuals.

From the Cardiology Division, University of California, San Francisco, Moffitt Hospital, San Francisco, California.
Address for correspondence: Stanton A. Glantz, PhD, Cardiology Division, University of California, San Francisco, 1186 Moffitt Hospital, 505 Parnassus, San Francisco, California 94143.

0735-1097/93/$6.00

JACC Vol. 21, No. 3
March 1, 1993:835–7

Use of nonparametric tests when the underlying assumptions are badly violated. The *t* test and analysis of variance are so-called parametric statistical techniques because the theory underlying these methods assumes that the populations from which the data were drawn follow a normal (bell-shaped) curve and that the only differences associated with the different treatments is a change in the means of these populations, but not in the standard deviation parameters. While the parametric methods are robust, the results will still be reasonably accurate even if there are moderate deviations from these assumptions. These methods produce unreliable results when there are serious deviations from the assumptions of normality and equal variance (that is, equal standard deviations). If the data appear not to meet these assumptions it is more appropriate to use nonparametric methods such as the Mann-Whitney rank sum test (instead of an unpaired *t* test), the Kruskal-Wallis analysis of variance on ranks (instead of parametric analysis of variance), the Wilcoxon signed rank test (instead of the paired *t* test) or the Friedman one-way analysis of variance based on ranks (instead of repeated measures analysis of variance). Using these nonparametric methods when appropriate produces more reliable results than using parametric methods. Some authors (3) believe that, because the nonparametric methods are generally only slightly less sensitive than the parametric methods even when the data meet the assumptions of the parametric methods, nonparametric methods should be used most of the time.

Drawing negative conclusions on the basis of small sample sizes. One of the primary goals of clinical research is to identify new treatments or diagnostic tests and to clarify their safety and efficacy. Such studies are often expensive and difficult to conduct, especially when investigators are studying relatively rare conditions for which recruiting suitable patients is a problem. During the last year several manuscripts submitted to *JACC* reported new therapies as "safe and effective" when the investigators found no failures or complications in the small number of patients studied. Although finding no failures in a small number of patients is encouraging, it does not prove safety or efficacy. For example, if there are no failures in six patients, the actual failure rate could be as high as 50% (i.e., the 95% confidence interval for the true failure rate extends up to about 50%). With no failure in 10 patients, the 95% confidence interval for the failure rate extends all the way to 30%. Because drawing negative conclusions based on small sample size is sometimes necessary, investigators (and readers) need to be aware of the severe limitations in this procedure (4). In particular, for such a negative conclusion, it is imperative that the authors not only present the observed failure rate, but also compute the confidence intervals for the results, and discuss them in their report.

A related issue is that of concluding that there is no difference between two treatments or conditions. Although most investigators have focused on the false positive error rate in statistical tests (known as the type I, or α, error), they also need to be concerned about the false negative error, that is, the possibility of failing to conclude that a difference exists when there really is one. This latter form of error is known as the type II, or β, error. The sensitivity of a test to detect a specified difference, which equals 1 minus the probability of a type II error, is known as the power of the test. Many people have come to accept a 5% risk of a false positive result (i.e., $p < 0.05$) as an appropriate level for concluding there is a significant difference between two treatments. Similarly, most people prefer that experiments yielding negative results have a power of at least 80% (i.e., an 80% chance of concluding that there is a significant difference between the different groups under study when one really exists). Computing the power of a test depends on the level of confidence you wish to have in drawing a positive conclusion (the p value you consider statistically significant), the sample size and the size of the effect deemed worth detecting. These factors often combine to make it difficult to reach the desired 80% power.

In any event, studies reaching negative conclusions should present the power to detect a clinically meaningful effect and discuss the power in the overall interpretation of the study. It is usually difficult to pin down investigators as to what is "clinically meaningful," but this is a crucial part of presenting the results of a study, particularly if the results are negative. (This is not to say that *JACC* refuses to publish negative studies; such studies can often be as important as positive studies in terms of management of patients or basic understanding of the cardiovascular system.) Because power calculations often reveal low power to back up negative conclusions, it is often necessary to enter more patients or other subjects into the study to support a negative conclusion with a reasonable level of confidence.

Uncontrolled and nonrandomized studies. When evaluating a new treatment, investigators often wish to report the results in terms of the overall success rate. However, it is important to have a reasonable control group to compare the results with, particularly because of the placebo effect in which the simple provision of treatment itself leads some people to feel better regardless of the actual efficacy of the treatment. These control subjects can be untreated persons or patients receiving a conventional therapy. It is well established that presentation of results without an adequate control almost always overstates the value and efficacy of the therapy under study.

A related issue is the need to randomize patients into the various treatment groups. If treatments are not randomly assigned, there are many opportunities for bias to be introduced into the study through the processes of patient selection and treatment allocation. Although it is not always possible to randomize patients in a study, this issue needs to be carefully considered and discussed as a limitation when randomization is not possible. Other approaches, such as matching the study group with patients of similar character-

istics who did not receive the treatment in question, can be used to compensate for some of the limitations of nonrandomized studies.

Conclusions. Although the kinds of statistical errors discussed here rarely lead to rejection of a paper that otherwise would have been accepted for publication, correcting these difficulties can change the conclusions reached. Careful consideration of the statistical and methodologic aspects of research in its design, conduct and reporting will improve the quality of publications in the *Journal of the American College of Cardiology* and, if authors deal with these problems before manuscripts are submitted, will speed the review and publication process.

References

1. Glantz S. Primer of Biostatistics. 3rd ed. New York: McGraw-Hill, 1992.
2. Glantz S, Slinker B. Primer of Applied Regression and Analysis of Variance. New York: McGraw-Hill, 1990.
3. Madansky A. Prescriptions for Working Statisticians. New York: Springer-Verlag, 1988.
4. Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. JAMA 1983;249:1743–5.