

---

## CURRENT TOPICS

---

### Biostatistics: How to Detect, Correct and Prevent Errors in the Medical Literature

STANTON A. GLANTZ, PH.D.

**SUMMARY** Approximately half the articles published in medical journals that use statistical methods use them incorrectly. These errors are so widespread that the present system of peer review has not been able to control them. This article presents a few rules of thumb to help readers identify questionable statistical analysis and estimate what the authors would have concluded had they used appropriate statistical methods. To prevent such errors from appearing, journals should secure review by someone knowledgeable in statistics before accepting a manuscript. In addition, human research committees should insist that an investigator define an appropriate strategy for data analysis before approving a protocol. Such policies should quickly and effectively increase the reliability of the clinical and scientific literature.

FEW clinical or biomedical researchers have had formal training in biostatistics. As a result, most readers assume that when an article appears in a journal the reviewers and editors have scrutinized every aspect of the manuscript, including the statistical methods. Unfortunately, this is not so. Critical reviewers of the biomedical literature have consistently found that about half the articles that used statistical methods did so incorrectly.<sup>1-6</sup> Articles published in *Circulation* follow this pattern. Figure 1 summarizes the results of my analysis of the use of statistical procedures in all the original articles published in *Circulation* from July to December 1977. Twenty-seven percent used statistical methods incorrectly and of those articles that used statistics at all, 44% had errors. Almost all of these errors involved inappropriate use of the *t* test in a way that often leads the authors to assert that a treatment produced an effect when the data do not support such a conclusion. When confronted with this observation (or the seemingly conflicting results that arise when comparable articles arrive at different conclusions), readers often conclude that statistical

analyses are maneuverable to one's needs, meaningless, or too difficult to understand.

Except when the statistical test merely puts a *p* value on an obvious effect (or when the article includes the raw data), a reader cannot tell whether the data do in fact support the conclusions. These errors may mislead other investigators.<sup>7-10</sup> In clinical studies, physicians may rely on the erroneous conclusions and expose patients to the risk and expense associated with useless treatments and unnecessary delay in the use of helpful treatments. Ironically, these errors rarely involve sophisticated issues that provoke debate among professional statisticians, but are simple mistakes, such as neglecting to include a control group, not allocating treatments to subjects at random, or misusing elementary tests of hypotheses. This article presents a few basic ideas and rules of thumb that can be used to evaluate the use of statistics in a published article. The real solution to this problem, however, is more careful review of research before publication.

#### The Problem

In 1951, Ross<sup>1</sup> published an analysis of 100 randomly selected articles, published between January and June 1950 in five American medical journals,\* that recommended or criticized some therapy or procedure. Sixty-three percent had an inadequate control group or none at all. Badgley<sup>2</sup> examined 103 articles published in 1960 in two Canadian journals and

---

From the Cardiovascular Research Institute and the Department of Medicine, University of California, San Francisco, San Francisco, California.

This work was supported by NHLBI Program Project grant HL-06285.

Dr. Glantz is the recipient of an NIH Research Career Development Award.

Address for correspondence: Dr. Stanton A. Glantz, Division of Cardiology, Room 1186-M, University of California, San Francisco, California 94143.

Received December 7, 1978; revision accepted July 13, 1979.

*Circulation* 61, No. 1, 1980.

\*Journal of the American Medical Association, American Journal of Medicine, Annals of Internal Medicine, Archives of Neurology and Psychiatry, American Journal of Medical Science.

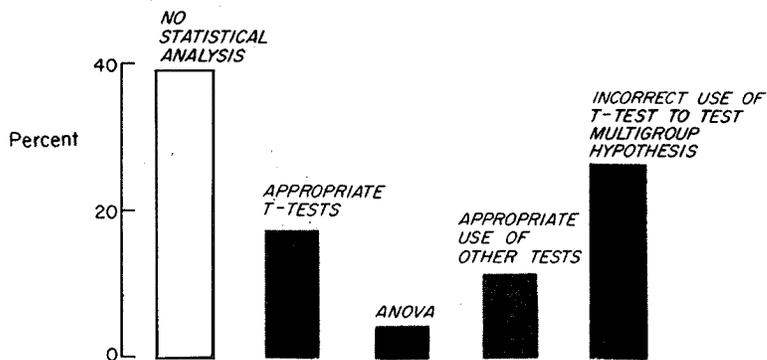


FIGURE 1. Of 142 original articles published in volume 56 of *Circulation* (excluding *Radiology*, *Clinicopathologic Correlations*, and *Case Reports*), 39% did not use statistics, 34% used the *t* test correctly to compare two group means, analysis of variance (ANOVA) or other methods, and 27% used the *t* test incorrectly to compare more than two groups.

found that 58% of the experiments failed to include an adequate control group (table 1). The lack of a control group biases the study on behalf of the treatment.<sup>10</sup> Badgley also observed that 57% of the papers used statistical tests of hypotheses in inappropriate ways and that these and other errors left the conclusions open to question in 42% of the papers. Schor and Karten<sup>3</sup> published a detailed analysis of the use of statistical methods in 295 papers from 1964 issues of 10 leading medical journals: 53% of these papers were acceptable and 47% were not (table 2). Gore, Jones and Rytter<sup>6</sup> analyzed papers published in the *British Medical Journal* during 3 months in 1976. Their results are similar to those of Badgley and Schor and Karten: 42% of the papers had at least one error (table 3).

Table 4 and figure 1 present my analysis of the use of one statistical procedure, the *t* test, in one volume each of *Circulation Research* and *Circulation*. The *t* test is the most popular statistical test in biomedical research.<sup>11</sup> I did not examine the experimental design or explore whether the use of some statistical test would have been appropriate in articles that did not use statistics. Forty-six percent of the articles published in *Circulation Research* and 27% of the articles published in *Circulation* used the *t* test when they should have used an analysis of variance or multiple comparison test.<sup>12</sup> The lower absolute percentage of inappropriate *t* tests in *Circulation* reflects the fact

TABLE 1. Canadian Medical Association Journal and Canadian Journal of Public Health (1960)

	Percentage of articles
No or inadequate control group	58
Inadequate sampling procedure	42
Inappropriate or incomplete use of statistical tests	57
Questionable use of statistical inference in drawing conclusions	42

Based on all 103 original research papers published between January 2, 1960 and July 2, 1960. Case reports, reviews, descriptive papers, and literature surveys were excluded. The categories are not mutually exclusive and some articles have more than one error, so the percentages do not sum to 100%. Source: Badgley.<sup>2</sup>

TABLE 2. Sample of 10 Journals (1964)

Type of article	Acceptable		Percent
	n	n	
Analytical	149	41	28
Case descriptions	146	114	78
Total	295	155	53

Based on random sample of articles drawn from January to March 1974 issues of the following journals: *Annals of Medicine*, *New England Journal of Medicine*, *Archives of Surgery*, *American Journal of Medicine*, *Journal of Clinical Investigation*, *American Archives of Neurology*, *Archives of Pathology*, *Archives of Internal Medicine*. Source: Schor and Karten.<sup>3</sup>

TABLE 3. British Medical Journal (1976)

	n	Percent of total	Percent of papers that used statistics
No statistical analysis	15	19	—
Acceptable use of statistical methods	30	39	48
At least one error	32	42	52

Based on all papers and originals (excluding descriptive papers and short reports in the 13 issues in January to March 1976, 8% of the papers that used statistics made some claim in the summary that was not supported by the data presented. Source: Gore et al.<sup>6</sup>

TABLE 4. Circulation Research (1977)

	n	Percent of total	Percent of papers that used statistics
No statistical analysis	20	25	—
Appropriate use of <i>t</i> test	16	20	27
Inappropriate use of <i>t</i> test	36	46	61
Analysis of variance	7	9	12

Based on all 79 original research papers published between January and June 1977.

that a smaller fraction of the articles in *Circulation* used any statistical test; if one considers only those articles that used statistical techniques, about half the articles in both journals (61% and 44%, respectively) contained errors.

#### What Difference Does It Make?

Errors in experimental design and misuse of elementary statistical techniques, such as the *t* test, in a large number of published papers is especially important in clinical studies. These errors may lead investigators to report a treatment or diagnostic test to be of statistically demonstrated value when, in fact, the data do not support this conclusion. Physicians who believe, on the basis of publication in a reputable journal, that a treatment has been proved effective may use it for their patients. Because all medical procedures involve some risk, discomfort or cost, people treated on the basis of erroneous research reports gain no benefit and may be harmed. Scientific studies that document the effectiveness of medical procedures will become even more important as efforts grow to control medical costs without sacrificing quality.<sup>13</sup> Such studies must be designed and interpreted correctly. In addition to indirect costs, there are significant direct costs associated with errors: money is spent, animals may be sacrificed, and human subjects may be put at risk to collect data that are not correctly interpreted.

#### Why Has the Problem Persisted?

Because so many people are making these errors, there is little peer pressure on academic investigators to use statistical techniques carefully. Quite the contrary, some investigators fear that their colleagues — and especially reviewers — will view a correct analysis as unnecessarily theoretical and complicated.

The journals are the major force for quality control in scientific work. Some journals have recognized that the regular reviewers often are not competent to review statistical methods in papers submitted for publication and these journals have modified their review process accordingly. Generally, they have someone familiar with statistical methods review manuscripts before they are accepted for publication.<sup>14-17</sup> Most editors, however, apparently assume that the reviewers will examine the statistical methods as carefully as they examine the clinical protocol or experimental preparation. If this assumption were correct, one would expect all papers to describe, in detail as explicit as the description of the protocol or preparation, how the authors have analyzed their data. Yet, Feinstein<sup>11</sup> reported that one-third (128 of 389) of the procedures used to test for statistical significance in six leading medical journals during January to June 1973 were not identified. Schor and Karten<sup>9</sup> reported similar findings, based on their review of papers published during 1964. My reading of *Circulation Research* and *Circulation* showed a similar pattern. It is hard to believe that the

reviewers examined the methods of data analysis with the same diligence with which they evaluated the experiment used to collect the data. Thus, the only manuscripts routinely sent to biostatisticians for comment, according to Schor and Karten,

are usually heavily documented with statistical jargon, and often a statistician is either co-author or co-worker on the study. Thus, the manuscripts which are usually sent out for statistical review are frequently not in need of evaluation. Those with only a few probability values mentioned and those lacking statistical jargon appeared in our study to be the ones most in need of a statistical review. These, however, are the ones which under the present system are not reviewed by the biostatistician.<sup>9</sup>

#### What Can a Reader Do?

The best solution to this problem is to improve the quality of statistical analysis in biomedical research. In the meantime, there are a few rules of thumb the reader can use to spot potential errors and estimate what the author would have concluded had statistical techniques been correctly applied to the data. They are: 1) the difference between the standard deviation and standard error of the mean; 2) the meaning of *p*; and 3) common errors in the use of the *t* test and how to compensate for them.

#### The Difference Between the Standard Deviation and the Standard Error of the Mean

*Experimental data are often summarized as mean ± SD, SE or SEM. SD stands for "standard deviation" and SE and SEM both stand for standard error of the mean. These two quantities are not equivalent; they quantify different things.*

When the variable being observed behaves so that any given observation is equally likely to be above or below the mean and more likely to be near the mean than far from it, it makes sense to quantify the spread of values using the standard deviation. Under these conditions, the standard deviation has the useful property that roughly 68% of the observations will be within 1 standard deviation of the mean and roughly 95% of the observations will be within 2 standard deviations of the mean. This property makes the standard deviation a good way to summarize the variability in data with a single number.

For example, an article reporting that diastolic blood pressure in healthy adults is  $78 \pm 6$  mm Hg (mean ± SD), implies that roughly 95% of all healthy adults have diastolic blood pressures within  $2 \times 6$  mm Hg = 12 mm Hg of 78 mm Hg, i.e., 66–90 mm Hg. The "2 standard deviations rule" is a good rule of thumb: *When observations are (or can be assumed to be) equally likely to be above or below the mean and more likely to be near the mean than far away, about 95% of them will be within 2 standard deviations on either side of the mean.*

Most authors, however, fail to summarize their data with the standard deviation; they use the standard error of the mean. Unlike the standard deviation, the standard error of the mean does not summarize the

variability in the observations or give the reader insight into the range of the observations. Why do most authors use the standard error of the mean to summarize their data? First, tradition; second, the standard error of the mean is always smaller than the standard deviation. If an author reports the standard error of the mean and the sample size, a reader can compute the standard deviation using the simple formula:

$$SD = SEM \times \sqrt{\text{sample size}}$$

Confusing the standard error of the mean with the standard deviation can be misleading. For example, suppose an article reported that diastolic blood pressure in nine healthy adults was  $78 \pm 2$  mm Hg (mean  $\pm$  SEM). What is the range of diastolic pressures that should include roughly 95% of the observations? The standard error of the mean is 2 and the sample size is 9, so the standard deviation is  $2 \text{ mm Hg} \times \sqrt{9} = 2 \text{ mm Hg} \times 3 = 6 \text{ mm Hg}$ . The answer is 66–90 mm Hg, as before. In contrast, applying the “2 standard deviations rule” with the standard error of the mean would estimate this range to be 74–82 mm Hg, which is 16 mm Hg too narrow.

What, then, does the standard error of the mean measure? In an experiment, an investigator rarely studies all possible members of a population, but only a small, representative sample. The mean value computed from such a sample is an estimate of the true mean that would be computed if it were possible to observe all members of the population.\* Because the sample used to compute the mean consists of individuals drawn at random from the population being studied, there is nothing special about this sample or its mean. In particular, had the luck of the draw been different, the investigator would have drawn a sample containing different individuals and computed a different mean value. Likewise, chance could have led to yet a third collection of observations and a third associated mean. Each of these three samples has a mean and each of these sample means is an estimate for the true (and unobserved) population mean. In theory, one could compute the means of all possible samples containing the number of observations the investigator chose to examine. In general, each of these sample means will differ from the others, but all will cluster around the true mean value that would be computed if it were possible to observe all members of the population. The standard deviation of all possible sample means is the standard error of the mean.

Thus, the standard error of the mean does not quantify variability in the observations, as the standard deviation does, but rather the precision with which a sample mean estimates the true population mean.

Because the standard error of the mean is the standard deviation of the collection of all possible sample mean values, we can apply the “2 standard deviations rule” to obtain the following rule of thumb: *There is a roughly 95% chance that the true mean of the population from which the sample was drawn lies within two standard errors of the mean of the sample mean.* That is, the standard error of the mean quantifies the certainty with which one can estimate the true population mean from the sample.

Returning to the diastolic blood pressure example, the sample of nine healthy adults permits a reader to be 95% confident that the mean diastolic blood pressure of all healthy adults is 74–82 mm Hg. While this fact is often of interest, it says nothing about variability in the data. The standard deviation contains this information. Thus, the standard deviation, not the standard error of the mean, should be used to summarize data.

### The Meaning of *p*

In addition to summarizing data, statistical techniques permit investigators to test whether their observations are consistent with their hypotheses. The result of such procedures is a so-called significance level or *p* value.

Understanding what *p* means requires understanding the logic of statistical hypothesis testing. For example, suppose an investigator wants to test whether a drug altered body temperature. The obvious experiment is to select two similar groups of people, administer a placebo to one group and the drug to the other, measure body temperature in both groups, then compute the mean and standard deviation of the temperatures measured in each group. The mean responses of the two groups will probably be different, regardless of whether the drug has an effect, for the same reason that different random samples drawn from the same population yield different estimates for the mean. Therefore, the question becomes: Is the observed difference in mean temperatures in the two groups likely to be due to random variation associated with the allocation of individuals to the experimental groups or due to the drug?

To answer this question, statisticians first quantify the observed difference between the two samples with a single number called a test statistic, such as *t*. The greater the difference between the samples, the greater the value of the test statistic. If the drug has no effect, the test statistic will be a small number. But, what is “small”?

To find the boundary between “small” and “large” values of the test statistic, statisticians assume that the drug does not affect temperature. If this assumption is correct, the two groups of people are simply random samples from a single population, all of whom received a placebo (because the drug is, in effect, a placebo). In theory, the statistician repeats the experiment, using all possible samples of people, and computes the test statistic for each hypothetical experi-

\*Likewise, the standard deviation computed from the sample is an estimate of the true standard deviation of the entire population. If it were possible to observe all members of the population there would no longer be any need for statistics, because statistical inference deals with making statements about larger populations from limited samples.

ment. Just as random variation produced different values for means of different samples, this procedure will yield a range of values for the test statistic. Most of these values will be relatively small, but sheer bad luck requires that there be a few samples that are not representative of the entire population. These samples will yield relatively large values of the test statistic, even if the drug had no effect. This exercise produces only a few of the possible values of the test statistic, say 5% of them, above some cutoff point. The test statistic is "large" if it is larger than this cutoff point. There are tables containing these cutoff values in most statistics books.

Having determined this cutoff point, we perform an experiment on a drug with unknown properties and compute the test statistic. It is "large." Therefore, we conclude that there is less than a 5% chance of observing data that led to the computed value of the test statistic if the assumption that the drug had no effect was true. Traditionally, when the chances of observing the computed test statistic when the intervention has no effect are below 5%, one rejects the working assumption that the drug has no effect and asserts that the drug does have an effect. There is, of course, about a 5% chance that this assertion is wrong. This 5% is the "*p* value" or "significance level." Precisely, the *p* value is the probability of obtaining a value of the test statistic as large or larger than the one computed from the data when in reality there is no difference between the different treatments. In other words, the *p* value is the probability of being wrong when asserting that a true difference exists. If one asserts a difference when  $p < 0.05$ , one accepts the fact that, over the long run, one assertion of a difference in 20 will be wrong.

It is commonly believed that the *p* value is the probability of making a mistake. There are obviously two ways an investigator can reach a mistaken conclusion based on the data: He or she can report that the treatment had an effect when in reality it did not, or can report that the treatment did not have an effect when in reality it did. The *p* value only quantifies the probability of making the first kind of error (the type I error) — erroneously concluding that the treatment had an effect when in reality it did not. It gives no information about the probability of making the second kind of error (the type II error) — concluding that the treatment had no effect when in reality it did.

#### Common Errors in the Use of the *t* Test and How to Compensate for Them

The *t* test is used to compute the probability of being wrong (the *p* value) when asserting that the mean values of two treatment groups are different. It is appropriate to test the hypothesis that the drug discussed above had no effect on body temperatures.

The *t* test is also widely but erroneously used to test for differences among more than two groups by comparing all possible pairs of means with *t* tests. For example, suppose an investigator measured cardiac output under control condition, in the presence of drug A

and in the presence of drug B. It is common to perform three *t* tests on these data: one to compare control vs drug A, one to compare control vs drug B, and one to compare drug A vs drug B. This practice is incorrect because the true probability of erroneously concluding that the drug affected cardiac output is actually higher than the nominal level, say 5%, used when looking up the "large" cutoff value of the *t* statistic in a table.

To understand this, reconsider the experiment described in the last paragraph. If the value of the *t* statistic computed in one of the three comparisons just described is in the most extreme 5% of the values that would occur if the drugs really had no effect, we will reject that assumption and assert that the drugs changed cardiac output. We will be satisfied if  $p < 0.05$ , and we are willing to accept the fact that one statement in 20 will be wrong. Therefore, when we test control vs drug A, we can expect to erroneously assert a difference 5% of the time. Similarly, when testing control vs drug B, we expect to erroneously assert a difference 5% of the time, and when testing drug A vs drug B, we expect to erroneously assert a difference 5% of the time. Therefore, when considering the three tests as a group, we expect to conclude that at least one pair of groups differ about  $5\% + 5\% + 5\% = 15\%$  of the time, even if the drugs did not affect cardiac output. (The *p* value is actually 13%.<sup>12</sup>) In general, simply adding the *p* values obtained in multiple tests produces a realistic and conservative estimate of the true *p* value for the set of comparisons.

We end our discussion of the *t* test with three rules of thumb:

- 1) The *t* test should be used to test the hypothesis that two group means are not different.
- 2) When the experimental design involves multiple groups, other tests, such as analysis of variance or the multigroup generalization of the *t* test should be used.
- 3) When *t* tests are used to test for differences among multiple groups, the reader can estimate the true *p* value by multiplying the reported *p* value times the number of possible *t* tests.

In the example above, there were three *t* tests, so the effective *p* value was about  $3 \times 0.05 = 0.15$ , or 15%. When comparing four groups, there are six possible *t* tests (1 vs 2, 1 vs 3, 1 vs 4, 2 vs 3, 2 vs 4 and 3 vs 4); so, if the author concludes there is a difference and reports  $p < 0.05$ , the effective *p* value is about  $6 \times 0.05 = 0.30$ ; there is about a 30% chance of making at least one incorrect statement if he or she concludes that the treatments had an effect.

These rules of thumb can help readers spot and correct for erroneous use of statistics. Obviously, it would be better to keep such errors out of print or, better yet, correct them early in the research.

#### How Can These Errors Be Prevented?

First, journal editors should insist that statistical methods be used correctly. Second, human research committees should not approve experiments on humans if the proposed study is poorly designed or the

resulting data will not be analyzed correctly. These two actions will force medical investigators to learn enough elementary statistics to design their experiments and analyze their data correctly and to recognize cases that require help from a professional statistician. This is not a call to turn clinicians into statisticians. Virtually all the errors in question deal with misuse of material discussed in most introductory statistics textbooks. Rejection of a paper or a protocol should motivate investigators to acquire an understanding of elementary statistics.

The response of the editors of *Circulation Research* to this problem is exemplary. After seeing the article-by-article analysis used to compile table 4, they submitted this analysis to expert reviewers. After considering the reviewers' comments, the editors announced that they were going to revise the review mechanism

so that published papers use statistical tests in the proper way. It is likely that some of our reviewers cannot or will not provide an appropriate critique of statistical evaluation of data or comment on the need for statistical evaluation of certain data. For this reason, we are adopting the following policy. Initial review of all manuscripts will be carried out as in the past. When it appears that a manuscript is likely to be accepted, a statistician will review it to determine whether a statistical method should be employed, or whether its use is appropriate. This may prolong the reviewing process, but we believe the delay will be small. We believe that such a delay will be justified, as it will insure that the information presented in papers published in the Journal has been analyzed and interpreted appropriately.<sup>17</sup>

This is a wise policy for three reasons. First, the key step in the reviewing process remains with the referees who are competent to comment on the scientific question the paper addresses. Once the scientific importance of the question has been established, the manuscript may still be worth publishing even if the statistical review corrects an error that substantially changes the conclusions. Second, if the editors circulate the statistical review to the other referees, it can educate them. As referees learn elementary statistics, it will be possible to return to the current reviewing procedure in which the same referee is responsible for both the methods of data collection and data analysis. Third, and most important, there is empirical evidence that this policy will improve the journal quickly and effectively. Schor<sup>18</sup> reported that in 1964 the *Journal of the American Medical Association* instituted such a statistical reviewing procedure. Only one-third of the papers published before this procedure was instituted were judged acceptable in their use of statistical methods, but three-quarters were judged acceptable 21 months later.

Though their motivations are different, human research committees, like journals, have both the responsibility and the power to force quick improvement, at least for clinical studies. It is unethical to put patients at risk to collect data for a scientifically invalid study. In addition, the number of experimental subjects should be minimized, and well-designed,

carefully analyzed experiments generally require fewer subjects to draw conclusions than poorly designed ones. Members of human research committees are already required to assess the scientific soundness of proposed studies,<sup>19</sup> and the committee should be required to have as a member someone able to comment on the experimental design and data analysis plan. Some institutions already have a statistician on the committee as a matter of policy, while other medical centers have no biostatistics department and, hence, few people to call for expert consultation. In addition to ethical considerations, good scientific practice requires the data analysis plan to be outlined before the data are collected in an investigation directed at testing a specific hypothesis.

These suggestions differ from those of other authors, who have correctly pointed out the need to improve medical education or to have medical investigators consult statisticians.<sup>6, 15, 16</sup> While students and research fellows should receive formal training in applied statistics, if only to increase the skepticism with which they approach the literature, such a step will not, in the short run, improve the literature. Until the indifference toward statistical and other deductive reasoning in much of academic medicine subsides, such courses probably will have little effect. Consulting statisticians may be helpful, but, as noted above, most of the errors are misapplications of elementary techniques that the investigators themselves should learn to use correctly. Most professional (and even amateur) statisticians are not especially interested in grinding out garden-variety statistics for other people. More important, with rare exceptions, no one can do a better job of analyzing the data from a clinical protocol or experimental procedure than the investigator.

#### Acknowledgments

I thank my colleagues for offering suggestions to improve this manuscript: D. Allison, S. Alten, S. Ashley, W. Bengé, B. Brundage, B. Brown, K. Chatterjee, M. Cheitlin, J. Comroe, D. Freedman, H. Gelberg, D. Heilbron, J. Hoffman, A. Jonsen, N. McCall, G. Meier, J. Moore, W. Parmley, C. Peck, N. Phillips, T. Ports, P. Renz, M. Rosen, S. Rubin, S. Schroeder, D. Stanski, C. Zippin. I thank M. Gruen and M. Hurtado for typing the manuscript and M. Briscoe for preparing the illustration.

#### References

1. Ross OB Jr: Use of controls in medical research. *JAMA* 145: 72, 1951
2. Badgley RF: An assessment of research methods reported in 103 scientific articles from two Canadian medical journals. *Can Med Assoc J* 85: 246, 1961
3. Schor S, Karten I: Statistical evaluation of medical journal manuscripts. *JAMA* 195: 1123, 1966
4. Schoolman HM, Becketl JM, Best WR, Johnson AF: Statistics in medical research: principles versus practices. *J Lab Clin Med* 71: 357, 1968
5. Lionel NDW, Herxheimer A: Assessing reports of therapeutic trials. *Br Med J* 3: 637, 1970
6. Gore S, Jones IG, Rytter EC: Misuses of statistical methods: critical assessment of articles in *B.M.J.* from January 10<sup>th</sup> March, 1976. *Br Med J* 1: 85, 1977
7. Weech AA: Statistics: use and misuse. *Aust Pediatr J* 10: 328, 1974