Current Concepts Review

Sample Size and Statistical Power in Clinical Orthopaedic Research*

BY KEVIN B. FREEDMAN, M.D., M.S.C.E.†, AND JOSEPH BERNSTEIN, M.D., M.S.†, PHILADELPHIA, PENNSYLVANIA

Investigation performed at the Sports Medicine Service, University of Pennsylvania School of Medicine and Veterans' Hospital, Philadelphia

Classic principles of treatment in orthopaedic surgery — the immobilization of fractures or the draining of infected wounds, for example — were not first established in prospective clinical trials or laboratory experiments. Rather, they were derived from perceptive observation: the methods were seen to work in practice, and they were retained. Observation has a noble history in medicine and science. Still, the modern reader is at least intuitively aware of the limitations of mere observation. Imagine if an investigator were to claim that prophylaxis against deep-vein thrombosis after hip replacement is not needed simply because only two thromboses were observed in ten patients who did not receive prophylaxis compared with three thromboses in ten patients who did. Such a study, were it to be published, would be the object of ridicule.

A single clinical observation is actually a sample of the entire set of possible observations, and methods of statistical inference are needed to draw valid conclusions. Almost all clinical research for the assessment of treatments and outcomes relies on statistical sampling — that is, a set of rules that help to ensure that the individuals included in a study are representative of the larger population to which the investigator wishes to generalize the findings. After the data have been gathered, statistical inference is applied. Statistical inference is a mechanism for evaluating an observed finding (for example, a difference between two treatment groups) relative to differences that may have occurred by chance alone, given the observed variability in measurements. This allows statements about an entire population to be made without the necessity of studying every member of that population — which is rarely feasible even if desired. To determine if observed differences represent true differences as opposed to differences that could be expected to occur because of random chance alone, the investigator uses statistical tests. These tests determine the probability of a difference as extreme as that observed or more so being observed under the null hypothesis of no true underlying difference. When this probability is small, it may be concluded that there is a real difference between the two populations that the samples represent. This is the meaning of the term significant.

It is, of course, possible that the testing of samples from two truly distinct groups will not always disprove the null hypothesis and thus will fail to show that the groups are significantly different. Stated another way, failure to prove that two groups are different is not equivalent to proving that they are the same. Accordingly, if no significant differences are found, the reader may ask whether the investigator failed to demonstrate significant differences because the samples were not unique or because they were unique but were not proved to be so.

Many readers and researchers are aware of the need to consider the possibility that two samples that seem to differ actually come from a single distribution and therefore do not really differ. Thus, they are familiar with p values and their associated alpha threshold (typically, $p < 0.05$). A p value of less than 0.05 means that there is less than a one in twenty chance that the two samples emanate from a single underlying population, and we therefore reject the null hypothesis and accept the two samples as coming from different populations. A second possibility, that two seemingly similar samples are indeed different, has been historically less publicized. Nonetheless, it is no less important when an investigator seeks to determine the validity of a clinical or experimental observation.

The purpose of the current article is to review the concepts of statistical sampling and hypothesis testing, with particular emphasis on study results that are not statistically significant. We also will consider the related issues of statistical power and the need for adequate sample size in clinical orthopaedic research. Although our goal is to familiarize the reader with the topic, it would be neither practical nor desirable for this review to replace a comprehensive textbook; we will favor concepts over equations. We especially encourage the use of statistical consultants when needed.

*No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article. No funds were received in support of this study.

†University of Pennsylvania, 3400 Spruce Street, Philadelphia, Pennsylvania 19104. The e-mail address for Dr. Bernstein is: orthodoc@mail.med.upenn.edu.

Study Results	Truth: Infection rates different	Truth: Infection rates not different
Infection rates observed in the sample different	Cell 1 Correct Conclusion	Cell 2 Type-I Error (alpha)
Infection rates observed in the sample not different	Cell 3 Type-II Error (beta)	Cell 4 Correct Conclusion

FIG. 1

Grid showing the possible outcomes of hypothesis testing. In this hypothetical study, the rates of infection associated with use of external fixation and intramedullary nailing for the treatment of open tibial fractures were compared.

Hypothesis Testing

Clinical research is designed to determine if different treatments produce different outcomes. For example, in a hypothetical study of patients with open tibial fractures in which treatment with external fixation is compared with treatment with intramedullary fixation, the investigator seeks to answer the question of whether the outcome achieved with one treatment is, in some meaningful way, distinct from that attained with the other. To answer this question, representative samples of patients managed with each method are examined in detail, the attributes of interest are measured, and differences between samples are sought.

If every patient in the world who received (or will receive) either type of treatment were studied, there would be no question for the statistician; such a study would be a complete survey, not a study based on a sample. If the results were different for the two groups, then, clearly, the groups would be different. This would not be an inference but a demonstrated fact, and there would be no role for statistical testing. More typically, however, samples are used and inferences are drawn from them. The question for the statistician then becomes whether these inferences are valid — that is, whether the observations of the study can be extrapolated to a larger context. (It should be pointed out that there are many reasons other than statistical issues that may render an inference invalid. Merely meeting statistical criteria does not validate a study.)

When the results of any study are considered in the context of the actual (unknown) truth, one of four outcomes is possible (Fig. 1). These outcomes can be illustrated with use of the hypothetical example of a study measuring the rates of infection associated with two different methods of treatment. In outcome 1, the rates of infection in the study sample are found to be different and are truly different (correct conclusion). In outcome 2, the rates of infection are found to be different but are truly not different (type-I error). In outcome 3, the rates of infection are found to be not different but are truly different (type-II error). In outcome 4, the rates of infection are found to be not different and are truly not different (correct conclusion).

A type-I error is made when the results of a study indicate a difference between groups even though, in reality, there is no difference. The likelihood of a type-I error being committed is reported as the p value. When this value is less than a given alpha (α) threshold, typically set at 0.05, a result is said to be statistically significant. When the p value is more than the chosen alpha threshold, a result is said to be not significant. A type-II error results when the p value fails to reach statistical significance even though the underlying groups are truly distinct. This usually occurs when the sample size is too small relative to the variability among subjects and the difference between the groups. The probability of committing a type-II error is given by beta (β), and the likelihood of avoiding it, the complement of beta ($1-\beta$), is termed the statistical power of the study. Adequate power traditionally has been defined at 80 percent (that is, $\beta \leq 0.20$)[4]. Saying that a study has a power of 80 percent means that, if a difference of a particular stated magnitude exists between groups, there is an 80 percent chance of correctly detecting it.

Maximizing Statistical Power

Statistical power is not an arbitrary feature of a study; rather, it can be controlled by the design of the study. The one concrete step that a researcher can take to ensure that a study will have adequate power to detect the desired difference (effect size) between groups is to enroll adequate numbers of subjects. When

TABLE I

NUMBER OF STUDIES HAVING ADEQUATE POWER TO DETECT SMALL, MEDIUM, AND LARGE EFFECT SIZES[4] FROM A SAMPLE OF FIFTY-NINE STUDIES HAVING NEGATIVE RESULTS

	Effect Size		
Power	Small (N = 59)	Medium (N = 59)	Large (N = 59)
≥0.80	2	17	38
$0.60 \leq x < 0.80$	0	12	13
$0.40 \leq x < 0.60$	3	13	4
$0.20 \leq x < 0.40$	15	13	4
<0.20	39	4	0

the sample size is small[13], a study is particularly susceptible to type-II error. Nevertheless, there is evidence in the medical literature that explicit sample-size and power calculations to minimize the risk of this error are infrequently performed before the start of a research study[1,5,9,11]. Furthermore, even *post hoc* analyses of power, which can help the reader to interpret negative results, are frequently omitted.

We studied the prevalence of studies, in the orthopaedic literature, in which an inadequate sample size had been used. A review of the 1997 American and British volumes of *The Journal of Bone and Joint Surgery* and the 1997 volumes of *Clinical Orthopaedics and Related Research* revealed eighty-six clinical studies in which the scientific methods of hypothesis testing had been used. For each article, we determined whether significant differences had been found for the primary outcomes of the study and whether a sample-size calculation or a power estimation had been performed. The parameters of the study then were used to determine the study's statistical power to detect a small, medium, or large treatment effect, as defined by Cohen[4] and as used by several other authors[7,10-12], and to determine the sample size necessary to detect a small, medium, or large treatment effect. Cohen defined the means to calculate small, medium, and large effects for a range of statistical comparisons. These effects are derived according to mathematical formulae, but, according to Cohen, they also can be defined loosely. A small effect is one that is typically of interest in clinical studies; a medium effect, one that is visible to the naked eye; and a large effect, one that is so stark that the study is probably unnecessary[4]. It must be emphasized that standardized effect sizes are used only in retrospective, external calculations of power, in the absence of assertions by the authors of the study.

We found sample-size calculations described in the Materials and Methods section of only five (6 percent) of the eighty-six studies, and only two other studies included *post hoc* power analyses. In fifty-nine studies (69 percent), results that were not significant for at least one primary-outcome variable were reported. Of these fifty-nine studies with negative results, only two (3 percent) had adequate power ($\beta \leq 0.20$) to detect a small effect size, and twenty-one (36 percent) lacked the power necessary to detect even a large effect size (Table I). Among the fifty-seven studies that lacked an adequate sample size to detect small differences, the average sample-size deficiency was 85 percent of the required number. The lack of adequate power and the failure to report a beta statistic make it difficult for the reader to know whether there really was no difference between study groups. In such situations, the authors were forced to conclude that, on the basis of the numbers available, no differences were seen. Had there been adequate power for the effect size that was hypothesized, a more forceful statement could have been made. If, for example, a 30 percent difference in the rates of infection had been hypothesized, the investigators could have made a statement such as: "There is at least an 80 percent likelihood that, had there been a 30 percent difference between groups, we would have found that difference with a value of p of less than 0.05." The results of our review indicate that the elite orthopaedic literature may include many studies with inadequate power due to inadequate sample size.

The factors that affect the power of a study are the sample size, the level of statistical significance (α), the variability of the samples, and the effect size chosen by the investigator. Clearly, all study designs should have adequate power as their goal. Nevertheless, there is a countervailing pressure not to have too many subjects, as enrolling subjects entails time, effort, and expense. The investigator should enroll the correct number of subjects, minimizing costs but optimizing the chance to find differences between groups if they really exist. Accordingly, to paraphrase Einstein, the study should be made as simple as possible but no simpler. The minimum number of subjects needed to provide adequate power can be derived mathematically. Both free and commercial software programs can be used to perform these sample-size calculations[6]. To ascertain the number of subjects needed, the investigator supplies the level of alpha at which significance is achieved, the desired degree of statistical power, the variance expected in the sample data, and the size of the effect that is clinically important. The program then returns the value for the requisite number of subjects. In the sections that follow, we review the role that these parameters play in determining sample-size requirements as well as their effect on the power of the study.

The Role of Alpha

Statistical tests do not provide binary yes-or-no results; rather, they produce a p value, which is a measure of the likelihood of a type-I error. P values range from zero to one. These continuous values are converted into binary results by the investigator applying an alpha threshold, or a point of statistical significance. When the value of p is less than alpha, the null hypothesis is rejected and the two samples are considered distinct,

TABLE II

POWER RESULTING FROM N VALUES AND EFFECT SIZE FOR A HYPOTHETICAL STUDY ON THE RATES OF INFECTION IN PATIENTS WITH OPEN TIBIAL FRACTURE*

No. of Subjects per Group	Effect Size			
	30% in Control Group vs. 10% in Study Group	30% in Control Group vs. 15% in Study Group	30% in Control Group vs. 20% in Study Group	30% in Control Group vs. 29% in Study Group
10	0.19	0.12	0.08	0.05
20	0.35	0.20	0.11	0.05
80	**0.89**	0.62	0.30	0.05
100	**0.95**	0.72	0.37	0.05
300	**0.99**	**0.99**	**0.81**	0.06
500	**0.99**	**0.99**	**0.95**	0.42

*$\alpha < 0.05$. Adequate statistical power (≥0.80) is indicated in bold type.

and vice versa. Although the alpha threshold is used to minimize the chance of a type-I error, this criterion also increases the chance of a type-II error. To illustrate this, let us assume, for example, that our hypothetical study of open tibial fractures reveals a rate of infection of 20 percent in eighty patients managed with an external fixator and a rate of 35 percent in eighty managed with an intramedullary nail. This sample size and these observed rates of infection result in a p value of 0.03. Let us also stipulate that the unknown truth is that the rates of infection are different. If an alpha threshold of 0.05 is used, then the results are significant and the correct conclusion is reached. However, if a more stringent alpha criterion of 0.01 is used, then the results are not significant and the implication that the rates are not different is incorrect — a type-II error.

It is important to note, conversely, that if, in truth, the rates were not different, the alpha threshold of 0.05 would result in a type-I error, which would be avoided with use of the stricter value. Accordingly, one cannot make a blanket statement that a high or low threshold is appropriate. Setting the acceptable probability of error to reject the null hypothesis is one of the key judgments made by investigators, editors, and readers. Although a p value of less than 0.05 has some historical acceptance, it is an arbitrary value. We do not believe that there should be a universal cutoff; rather, the p value should be reported and the reader should be allowed to draw his or her own inferences. If a study were to demonstrate that a safe and inexpensive treatment for cancer was 50 percent better than standard treatments but the p value was 0.06, that study still would have value to the medical community. Alternatively, if the treatment was costly or risky, a much stricter p value might be appropriate. Thus, use of the standard of $p < 0.05$ is a judgment, not a rule.

The investigator must realize that decreasing the risk of a type-I error comes at the expense of increasing the risk of a type-II error. As alpha is decreased, the statistical power for a particular sample size and effect size decreases as well. For simplicity, it may be reasonable to base initial power calculations on the conventional threshold of $p < 0.05$; however, if the researcher subsequently decides that a more stringent alpha threshold is in order, the enrollment of additional study subjects will be required to maintain an acceptable risk of a type-II error.

The Role of Effect Size

When two modes of treatment are compared, a finding of a significant difference does not mean that the difference is clinically important. For example, if the hypothetical study of open tibial fractures had a sufficiently large number of patients and the observed rates of infection were 50 percent for the group managed with the external fixator and 51 percent for the group managed with the intramedullary nail, a p value of 0.01 could be derived; however, a clinician might still consider the rates of infection to be clinically equivalent. This 1 percent absolute difference is therefore significant statistically but irrelevant clinically. In statistical terms, this effect size (the magnitude of the difference demonstrated between groups) is trivial. It should be understood that, given enough study subjects, any true difference between study groups can be detected at a chosen p value, even if the effect size is clinically unimportant.

In attempting to determine the number of study subjects needed, it is important to determine the minimum clinically relevant effect size. It is preferable to have sufficient numbers of subjects so that any clinically meaningful differences are also statistically significant. If the rates of infection were 29 and 30 percent with a p value of 0.06, it is doubtful that the investigator would

be disappointed that statistical significance was not attained. However, if a 10 percent difference in the rate of infection is clinically relevant, then the finding that a difference between rates of 20 and 30 percent is not statistically significant would be disappointing and inconclusive.

The mechanics of statistical tests are such that, for fixed sample sizes, as the effect size increases the p value decreases. This makes intuitive sense; it is less likely that two samples come from the same underlying population if the differences between them are large. Thus, for a given number of subjects in each group, a rate of infection of 25 percent compared with a rate of 30 percent might have a p value of 0.06, for example, whereas a rate of 24 percent compared with a rate of 30 percent might have a p value of 0.04. If the number of subjects were to be doubled, a rate of 25 percent compared with a rate of 30 percent might have a p value of 0.01. The clinical question is whether it would be worthwhile to enroll these additional subjects in order to attain statistical significance if the difference between the two rates is not clinically important. Accordingly, it is of major importance for the investigator to stipulate the minimum effect size of interest when planning the study. The smaller the effect size that is clinically important, the greater the number of subjects needed to establish significance.

Effect size is based on clinical judgment, not statistical inference. There are times when a 1 percent difference is irrelevant, as in the case of a 90 percent rate of success compared with a 91 percent rate; there also are times when a 1 percent difference is crucially important, as in the case of a 1 percent rate compared with a 2 percent rate of fatal pulmonary embolism. In each study, the investigator should assert and defend the minimum effect size of interest that is chosen.

In our review of the eighty-six studies in the 1997 literature, we used Cohen's standards of small, medium, and large effect sizes[4] only because the authors of the original studies did not state an effect size of interest. In such *post hoc* analyses, these standards are reasonable proxies if the investigator omitted the important step of stipulating the preferred effect size of interest; however, they are not appropriate for an investigator who is planning a study — that is, the investigator should not estimate the number of subjects needed to attain significance for a small effect but, rather, should enroll a sufficient number of subjects to attain statistical significance for what the investigator considers to be clinically important differences.

The Role of Variance

In our hypothetical study comparing external fixation with intramedullary fixation, dichotomous (binary) outcomes, such as infected versus not infected, as well as continuous data, such as the time to union and the functional score, can be measured. Continuous data are those that can be any value within a range. In a study measuring continuous variables, the variance of the sample data must be taken into consideration when the investigator assesses statistical power or plans the number of subjects needed. Variance is a statistical measure of how much the typical member of the group deviates from the mean; low variance implies that most values are centered around the mean, whereas high variance is found when the data are more spread out. (The well known standard deviation is simply the square root of the variance.)

The importance of variance in determining sample-size requirements is that, if variance is low, a given sample of a group is more likely to be representative of the entire group. Accordingly, with lower variance, fewer subjects are needed to reflect the underlying population accurately and fewer subjects are needed to demonstrate significant differences if real differences exist. Although variance cannot be known in advance, it can be estimated from several sources, including previous studies, pilot data, and educated guesses.

The Role of Sample Size

Since the alpha value and the effect size usually are chosen on the basis of the nature of the particular study, sample size is commonly the only variable in the power equation that is controlled by the investigator. For given values of alpha and effect size, increasing the sample size increases the chance of attaining statistical significance if real differences exist.

The effect of sample size on power is demonstrated in the following example. Suppose that a researcher wants to ascertain if a new treatment for open fracture is associated with a lower rate of infection than is the conventional treatment, which is associated with a 30 percent rate of infection. A table listing the results, sample size, and effect size could be constructed for this hypothetical experiment (Table II). The power to detect a significant difference (set at $p < 0.05$) between the 30 percent rate of infection in the control group and a second rate, ranging from 10 to 29 percent, in the study group can be determined as a function of the sample size. If variance is assumed to be constant for all cases, the power is more than 0.80 only for an n value of eighty or more and even then, only if the effect size is 20 percent. (In this example, alpha is set at the conventional $p < 0.05$. A lower threshold would require still greater numbers of subjects.) Thus, if the investigator decides, before the study, that a 10 percent decrease in the rate of infection is the effect size of clinical interest, then 300 patients must be enrolled in the study in order for it to have adequate power. With fewer than 300 patients, a 10 percent difference will not have a p value of less than 0.05 and will be reported as not statistically significant. If the study is performed with 300 patients and the rate of infection in the group that received the new treatment group is 29 percent, the

investigator will conclude that there is no significant difference. This would be neither disappointing nor inconclusive because an effect size this small has been determined to be clinically unimportant. Adequate power implies that, if differences of clinical interest are detected, they will be statistically significant.

Estimation of Sample-Size Requirements

The calculation of sample size for a study in which dichotomous variables (such as the presence of infection or a nonunion) are assessed requires the definition of three parameters: alpha, or the threshold below which p must be in order for a result to be deemed significant; beta, or the acceptable risk of a type-II error; and the effect size of interest. For continuous variables, such as the Knee Society score and the duration of hospitalization, the information needed to calculate sample-size requirements includes these three parameters as well as an estimation of the variance in the sample data.

The calculation of sample size can be understood by considering the example of a randomized controlled trial in which use of a cast and electrical stimulation is compared with use of a cast alone for the treatment of closed tibial fractures, with the outcome of interest defined as the time to union (continuous data). The sample-size calculation is performed by entering the appropriate information into a standard formula. Here, we use the conventional values of 0.05 for alpha and 0.20 for beta (a power of 0.80); however, we stress that these are simply conventions. In addition, we assert that the effect size of interest is a seven-day difference in the time to union. (In our report of the study, we would, of course, be required to defend this choice.) We estimate that the variance of the data is fourteen days. (Again, in our report, we would cite previous studies to support this.) On the basis of these values, the sample size needed for our study would be sixty-three patients in each group; this does not account for patients who are lost to follow-up or are withdrawn from the study.

We could also choose a different primary outcome, such as the rate of union. Since healing is a dichotomous variable, there is no need to estimate the variance, but we do have to estimate a baseline rate of union in the control group and the effect size of clinical interest. If we estimate the rate of nonunion in the group managed with a cast alone to be 10 percent and the clinically relevant decrease in the rate of nonunion (the effect size) in the study group to be 5 percent, then a sample size of 474 patients is needed in each group. Our choice of this effect size means that the new treatment (electrical stimulation) would be accepted only if the rate of nonunion is decreased to 5 percent. A difference of less than 5 percent would be considered clinically unimportant, and the treatments would be considered equivalent.

The difference in the sample size that is required for these two different calculations emphasizes a final point: the investigator who is planning a study must declare not only the effect size of interest but also the primary outcomes of interest. These primary outcomes will be the subjects of the sample-size calculations that are performed before the investigation, and the correct sample size will be the largest value that is derived. In the present example, if both outcomes are considered primary (that is, important to both the investigator and the reader), then 474 rather than sixty-three is the appropriate number of subjects per group. Although a study that is based on only sixty-three patients may lead to conclusions about the time to union for the two treatment groups that are both clinically important and statistically significant, no valid conclusions can be made with so few patients with regard to the rates of nonunion.

Post Hoc Analyses

The best time to perform power calculations is before a study is initiated. However, it is possible to perform meaningful calculations (so-called *post hoc* analyses) after the completion of a study, even if they were omitted at the outset. The investigator should be aware that not all readers accept the value of *post hoc* analyses. When no significant differences are found, power calculations can be performed after the fact to determine the likelihood of a type-II error. This is analogous to reporting the exact p value rather than simply stating that p is greater than 0.05. Both after-the-fact power calculations and explicit p values provide the reader with additional information for interpreting the results.

In the example in which use of a cast alone is compared with use of a cast and electrical stimulation, it was established that 474 patients per group were needed to detect a 5 percent difference in the rates of nonunion. If only 100 patients per group were enrolled, the power that such a sample yields, which is obviously less than the desired 80 percent, could be calculated. With use of this sample size and 5 percent as the minimum effect size of interest, the power to detect a difference in the rates of nonunion is 0.23. Therefore, if the rates of nonunion were 10 and 5 percent, statistical significance would not be achieved at $p < 0.05$ and the likelihood of a type-II error would be great. Accordingly, no strong conclusion could be derived from this study.

The investigator can also determine the minimum detectable difference in the rates of nonunion that would be significant ($p < 0.05$), given adequate power (≥ 0.80) and the sample size. In this example, if 100 subjects (rather than the requisite 474) were enrolled, the rate of nonunion in the treatment group would have to be 1 percent (a tenfold risk reduction) in order for a significant difference to be detected. In other words, with only 100 subjects, any rate of nonunion of more than 1 percent might lead to the conclusion that no significant

differences were found, despite the fact that a clinically meaningful difference of 5 percent might exist.

Overview

All investigators who perform clinical studies should aim for adequate power. If clinically meaningful differences are found, it is hoped that they will be statistically significant as well. The problem of inadequate power to detect clinically meaningful differences between study groups in medical research has been demonstrated previously in several fields, including emergency medicine[2], cardiovascular research[14], nursing[11], internal medicine[8,9,12], general practice[7], rehabilitation[10], and hand surgery[3]. We found that clinical orthopaedic research is similarly affected, with sample sizes that are too small to ensure statistical significance for what may be clinically important results.

A study with negative results but adequate power to detect clinically meaningful differences is a valuable contribution to the literature. Properly presented, it informs the reader that the treatments under study are comparable. A negative study with inadequate power is inconclusive at best. A negative study with inadequate power is not appropriate for the practice of evidence-based medicine.

The best way to prevent a type-II error from occurring is to perform a sample-size calculation before the initiation of a research study. Nearly all granting agencies, including the Orthopaedic Research and Education Foundation and the National Institutes of Health, mandate sample-size calculations. The researcher should consider the issues of sample size and power before the study is begun. Foremost among the reasons for this is feasibility: the investigator will want to know in advance whether it is possible to draw valid conclusions on the basis of the population available. It may well be that a study of patients from one surgeon's practice would have to include patients seen over ten years of that practice in order to achieve statistical significance. This information is crucial in the planning stage of a prospective study, and it may stimulate investigators to pool their patients or to engage in multicenter trials. Conversely, an investigator may discover that he or she overestimated the number of subjects needed, thereby saving time, expense, and effort as a result of sample-size planning. In addition, as power calculations demand that the minimum effect size of interest be established, such calculations call attention to this critical issue that might otherwise be overlooked.

Proper study design and appropriate statistical analysis are essential to the validity of all quantitative clinical research. Type-I error is better known and easier to evaluate than is type-II error, and reviewers and readers are more cognizant of p values when authors conclude that there are significant differences between groups. However, there must be equal scrutiny when authors conclude that no significant differences exist. In this age of managed care, treating physicians may be forced to use the cheapest among several methods, provided that the choices are deemed equivalent. It is therefore especially important that investigators do not erroneously label two treatments as equivalent when, in fact, all that was shown was that they were not different. All clinical studies should therefore be based on appropriate sample-size calculations. In addition, exact p values should be stated, and statistical power and the underlying effect size of interest should be explicitly expressed. We must base our treatment on evidence whenever possible, and attention to the details of sample size will enhance the quality of that evidence.

References

1. **Altman, D. G.,** and **Doré, C. J.:** Randomisation and baseline comparisons in clinical trials. *Lancet,* 335: 149-153, 1990.
2. **Brown, C. G.; Kelen, G. D.; Ashton, J. J.;** and **Werman, H. A.:** The beta error and sample size determination in clinical trials in emergency medicine. *Ann. Emerg. Med.,* 16: 183-187, 1987.
3. **Chung, K. C.; Kalliainen, L. K.;** and **Hayward, R. A.:** Type II (beta) errors in the hand literature: the importance of power. *J. Hand Surg.,* 23: 20-25, 1998.
4. **Cohen, J.:** *Statistical Power Analysis for the Behavioral Sciences.* Ed. 2. Hillsdale, New Jersey, Lawrence Erlbaum, 1988.
5. **DerSimonian, R.; Charette, L. J.; McPeek, B.;** and **Mosteller, F.:** Reporting on methods in clinical trials. *New England J. Med.,* 306: 1332-1337, 1982.
6. **Dupont, W. D.,** and **Plummer, W. D., Jr.:** Power and sample size calculations. A review and computer program. *Controlled Clin. Trials,* 11: 116-128, 1990.
7. **Fox, N.,** and **Mathers, N.:** Empowering research: statistical power in general practice research. *Fam. Pract.,* 14: 324-329, 1997.
8. **Freiman, J. A.; Chalmers, T. C.; Smith, H., Jr.;** and **Kuebler, R. R.:** The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. *New England J. Med.,* 299: 690-694, 1978.
9. **Moher, D.; Dulberg, C. S.;** and **Wells, G. A.:** Statistical power, sample size, and their reporting in randomized controlled trials. *J. Am. Med. Assn.,* 272: 122-124, 1994.
10. **Ottenbacher, K. J.,** and **Barrett, K. A.:** Statistical conclusion validity of rehabilitation research. A quantitative analysis. *Am. J. Phys. Med. and Rehab.,* 69: 102-107, 1990.
11. **Polit, D. F.,** and **Sherman, R. E.:** Statistical power in nursing research. *Nursing Res.,* 39: 365-369, 1990.
12. **Reed, J. F., III,** and **Slaichert, W.:** Statistical proof in inconclusive "negative" trials. *Arch. Intern. Med.,* 141: 1307-1310, 1981.
13. **Rosner, B.:** *Fundamentals of Biostatistics.* Ed. 4. Belmont, California, Duxbury Press, 1995.
14. **Williams, J. L.; Hathaway, C. A.; Kloster, K. L.;** and **Layne, B. H.:** Low power, type II errors, and other statistical problems in recent cardiovascular research. *Am. J. Physiol.,* 273(1 Part 2): H487-H493, 1997.