Sample size and statistical power of randomised, controlled trials in orthopaedics

K. B. Freedman, S. Back, J. Bernstein

From the University of Pennsylvania School of Medicine, Philadelphia, USA

We reviewed all 717 manuscripts published in the 1997 issues of the British and American volumes of the *Journal of Bone and Joint Surgery* and in *Clinical Orthopaedics and Related Research*, from which 33 randomised, controlled trials were identified. The results and sample sizes were used to calculate the statistical power of the study to distinguish small (0.2 of standard deviation), medium (0.5 of standard deviation), and large (0.8 of standard deviation) effect sizes.

Of the 33 manuscripts analysed, only three studies (9%) described calculations of sample size. To perform post-hoc power assessments and estimations of deficiencies of sample size, the standard effect sizes of Cohen (small, medium and large) were calculated. Of the 25 studies which reported negative results, none had adequate power ($\beta < 0.2$) to detect a small effect size and 12 (48%) lacked the power necessary to detect a large effect size. Of the 25 studies which did not have an adequate size of sample to detect small differences, the average used was only 10% of the required number

Our findings suggest that randomised, controlled trials in clinical orthopaedic research utilise sample sizes which are too small to ensure statistical significance for what may be clinically important results.

J Bone Joint Surg [Br] 2001;83-B:397-402.
Received 16 September 1999; Accepted after revision 15 June 2000

Clinical research assessing treatment and outcome relies on statistical evidence. Since it is rarely feasible to assess the entire patient population, representative samples are studied. These are examined in detail, the attributes of interest are measured and differences between samples are sought. If differences are found, the question then posed is whether the samples truly represent distinct groups, or if the observed difference was caused by chance. Statistical tests are used to help to decide which is the case.

A primary purpose of using statistical tests is to minimise the probability of a type-I error in which it is concluded that there are differences between groups when no such disparity exists. This probability (p) is given an α value, typically set at 0.05, at which level there is a 1 in 20 (5%) risk that the difference detected is entirely owing to chance. When the p value lies above that level, i.e. $p > 0.05$, the results are said to be not statistically significant.

A study can produce results which are not statistically significant for two reasons. There may be no real differences or, alternatively, if present, the study may be insufficiently powerful to detect them. This latter possibility is known as a type-II error. The likelihood of committing a type-II error is quantified as beta (β). The probability of avoiding such an error is the complement of beta ($1-\beta$), and is termed the statistical power of the study. Adequate power in respect of a given effect size has been defined at 80% (i.e., $\beta < 0.2$).[1] To say that a study has 80% power means that if there is indeed a difference of this given magnitude between groups, there is an 80% chance of correctly detecting it as statistically significant.

When the sample size is small,[2] a study is particularly susceptible to a type-II error. Nevertheless, there is evidence that calculations of explicit sample size or power to anticipate this error are rarely performed before the start of a research study.[3-6] The failure to have sufficient subjects in a study is at times unavoidable. For example, a surgeon may present all of his patients who had a given operation, but, although this series may be the largest ever reported, it may still be too small for valid statistical inference. In many instances, however, the failure to have adequate numbers in a study is a preventable error. Power analysis and calculations of sample size performed before the initiation of the study can inform the researchers as to exactly how many subjects are needed. This is especially relevant for prospective, randomised controlled trials.

This type of trial is considered to be the method of

K. B. Freedman, MD, MSCE, Resident
S. Back, BA, Medical Student
J. Bernstein, MD, MS
Department of Orthopaedic Surgery, Leonard Davis Institute of Health Economics, University of Pennsylvania School of Medicine, Veteran's Hospital, Philadelphia, Pennsylvania, USA.

Correspondence should be sent to Dr J. Bernstein at 424 Stemmler Hall, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6081, USA (email: orthodoc@post.harvard.edu).

0301-620X/01/310582 $2.00

choice for establishing an outcome of treatment. The efficacy of a new intervention is most readily accepted when the results are from such an investigation,[7] since it removes many of the flaws, such as bias and confounding, which may be present in other designs of study. Given their prospective nature, calculations of sample size should be considered an essential preliminary in the design of these trials. Moreover, since the performance of a randomised, controlled trial is difficult in surgery, it would be disappointing to expend the effort only to find that the study lacks adequate statistical power.

We have investigated the role of sample size and statistical power in randomised, controlled trials in clinical orthopaedic research. As a sample group, all papers from the 1997 volumes of three major orthopaedic journals were examined. Our aims were to determine how often calculations of sample size were performed, to measure the statistical power of studies which failed to detect differences between sample groups and those which did so, and, finally, to quantify any deficiencies of sample size.

Materials and Methods

Sampling of orthopaedic research. All papers published in the 1997 issues of the British and American volumes of the *Journal of Bone and Joint Surgery* and of *Clinical Orthopaedics and Related Research* were studied. All basic science and review articles, as well as 31 other articles which could not be clearly categorised as clinical research, were excluded from further analysis (Table I). Any reports on animal models, cadavers or histological and cellular analyses only, were considered as basic science. The clinical manuscripts were then further classified as a case report, case series, case-control studies, retrospective or prospective cohort studies, or randomised, controlled trials[8] (Table I).

Data collection. All of the randomised, controlled trials were examined in detail. First, we determined whether an appropriate calculation of sample size was documented in the 'Materials and Methods' section of the article. The primary outcomes of each study were then identified and their values tabulated. In studies that did not identify which outcomes were primary, we established this by applying appropriate criteria.[5] The sample size of each group in the study was then recorded. Outcomes were listed as either 'positive', when significant differences were found, or 'negative', with no statistically significant differences. Finally, the manuscript was scanned to determine whether a subsequent power analysis had been performed. A database containing these parameters was established.

Power calculations. To measure statistical power and the requirements of sample size, the magnitude of the difference between groups, the so-called 'effect size', which is both plausible and worth detecting is established. Larger samples are required to elicit smaller differences. Therefore, for a given sample, the power decreases as smaller effect sizes are examined, or, for a given level of power, the requirement of a sample size increases as the effect size of interest decreases. Unfortunately, researchers rarely state what they consider to be the minimal clinically significant difference of interest. Accordingly, it is not possible to calculate power retrospectively, unless some proxy value is used. Such proxies should be employed in retrospective studies such as ours, and then only if the authors of the manuscript under review did not name an effect size of interest. They are not to be used in the calculation of sample sizes at the outset of a clinical study. They are second-class substitutes for the preferred method, namely a statement of values by the researchers themselves. Such values can also be obtained from previous trials using the same or similar treatments, or be suggested by peer review.

The surrogate values chosen for our analysis were the standardised effects as defined by Cohen[1] and used previously by several authors.[6,9-11] Cohen has defined a means to calculate the effects of 'small', 'medium', or 'large'

Table I. Survey of original orthopaedic research in 1997

Type of study	*Journal of Bone and Joint Surgery [Am]*	*Journal of Bone and Joint Surgery [Br]*	*Clinical Orthopaedics and Related Research*	Total
Clinical research	138	134	225	497
Case reports	42	17	31	90
Case series	71	95	155	321
Case-control	1	2	5	8
Retrospective cohort	8	7	25	40
Prospective cohort	1	1	3	5
Randomised, controlled trial	15	12	6	33
Excluded	56	50	114	220
Basic science	45	41	103	189
Other				
Diagnostic testing	4	4	4	12
Health policy research	2	1	2	5
Classification or scoring systems	4	3	2	9
Cost studies	0	1	3	4
Meta-analysis	1	0	0	1

Table II. Definitions of effect size and formulae according to Cohen[1]

Test	Small	Medium	Large	Formulae*
t-test	0.2	0.5	0.8	$d = u_a\text{-}u_b/\sigma$
chi-squared	0.2	0.5	0.8	$h = \phi_1\text{-}\phi_2$
F test	0.1	0.25	0.4	σ of means/pooled σ

* u_a, mean of sample a; u_b, mean of sample b; σ standard deviation; $\phi = 2 \arcsin \sqrt{P}$

sizes, for a range of statistical comparisons. These effects are derived according to mathematical formulae, considering both the absolute and the relative magnitude of the changes detected. The formulae used to derive small, medium and large effect size are given in Table II. According to Cohen, a small effect can be loosely defined as one typically of interest in clinical studies, a medium effect as one visible to the naked eye, and a large effect as so stark that the study is probably unnecessary.

Using the information obtained from each manuscript, we determined the power to detect a small, medium and large effect size for each primary outcome. The power for each outcome was averaged for studies with multiple outcomes to determine an overall power for each study to detect the primary outcomes. In addition, the sample size necessary to detect a small, medium and large effect size for each primary outcome was determined, using $\alpha = 0.05$ and power = 0.80 ($\beta = 0.20$). By convention, 80% power is considered the lowest acceptable value since the risk of a type-II error is too high with less.[1] The arithmetical difference between the sample size needed and the actual sample size used yielded the sample size deficiency for each study. No adjustment was made for multiple comparisons, since only the primary outcomes were considered.

A standardised program was used for all our power calculations,[12] and we performed all descriptive statistics with Intercooled STATA 5.0 (Stata Co, College Station, Texas).

Results

Survey of orthopaedic research. We reviewed 717 original research manuscripts (Table I). Of the 497 clinical research articles, 33 were randomised, controlled trials; these provided the sample for the remainder of the analysis.

Positive and negative study results. In the 33 randomised, controlled trials, 75 primary outcomes were identified; 56 (75%) were 'negative' in that no statistically significant differences were found, and 19 (25%) were 'positive' with statistically significant results. Eight of the 33 studies (24%) were positive for all primary outcomes, 19 (58%) failed to reveal statistically significant differences between groups for all primary outcomes, and six (18%) were negative for at least one primary outcome; 25 (75%) were thus negative for at least one primary outcome.

Initial calculations of sample size and subsequent power analysis. Of the 33 randomised, controlled trials, three (9%) described the performance of sample-size calculations in the 'Materials and Methods' section of the manuscript. Of the 25 studies with negative results for one or more primary outcomes, only one provided subsequent power analysis to assist the reader in determining whether the study was able to detect real differences had they existed. Only four of the 33 studies (12%) reported any power analysis.

Power to detect small, medium and large effect size in negative studies. A frequency distribution for the mean power to detect a small, medium or large effect size between groups for the 25 studies with negative results is given in Table III. Of the 25 randomised, controlled trials with 'negative' results, none had adequate samples to detect a small effect size, four were able to detect a medium effect size and 13 (52%) a large effect size. Thus, 12 randomised, controlled trials (36%) did not even have adequate power to detect a large treatment effect.

Power to detect small, medium and large effect size in positive studies. A frequency distribution of the mean power to detect a small, medium and large effect size between groups for the eight studies with positive results, i.e. statistically significant differences, is given in Table IV. Merely because a study detected statistically significant results does not mean that it utilised a sufficient number of subjects; positive studies with inadequate power have their own limitations, as will be discussed later.

Sample size deficiencies. Using formulae similar to those employed to determine power, the sample size needed to

Table III. Power required to detect a small, medium and large effect size for the 25 negative studies

Power	Small	Medium	Large
>0.80	0	4	13
$0.60 \leq x < 0.80$	0	5	7
$0.40 \leq x < 0.60$	1	7	2
$0.20 \leq x < 0.40$	4	6	3
<0.20	20	3	0

Table IV. Power required to detect a small, medium and large effect size for eight positive studies

Power	Small	Medium	Large
≥0.80	0	5	6
$0.60 \leq x < 0.80$	1	1	2
$0.40 < x < 0.60$	2	2	0
$0.20 \leq x < 0.40$	2	0	0
<0.20	3	0	0

detect a difference of a given magnitude can be calculated using established values for the type-I and type-II rates of error, (α and β) respectively. With α set at 0.05 and β at 0.20 (80% power), we calculated the sample sizes required to detect small, medium and large effect size.

In the 25 randomised, controlled trials with inadequate sample size to detect a small treatment effect, the mean number of patients was 80, and the mean number needed to detect a small effect size was 770. Thus, the sample size deficiency was 690 patients (SD = 66), 90% of the required number. In the 21 randomised, controlled trials with inadequate power to detect a medium effect size, the mean deficiency in sample size was 70 patients (SD = 30), only 44% of the required sample size.

Discussion

All clinical studies should aim for adequate power with respect to a given effect size. If clinically significant differences are then found, they are likely to be statistically significant as well. The problem of inadequate power to detect important differences between study groups in medical research has been demonstrated previously in several fields, including emergency medicine,[13] cardiovascular research,[14] nursing,[6] internal medicine,[5,10,15] general practice,[11] rehabilitation,[9] and hand surgery.[16] Our study suggests that randomised, controlled trials in clinical orthopaedic research likewise utilise sample sizes which are too small to ensure statistical significance for what may be clinically important results.

The statistical power of the 'negative' studies in the randomised, controlled trials which we have surveyed was severely compromised. All of the negative studies had less than 60% power to detect a small effect size for the primary outcomes, and 21 of 25 studies (84%) did not have adequate power to detect a medium effect size. This means that small but real differences between groups would probably be reported as 'not statistically significant'. For most studies, a medium effect size is of sufficient magnitude to represent a clinically important difference. For example, if a study was performed to compare the rate of nonunion after reamed *versus* unreamed tibial nailing, a medium effect size would represent a difference of 10% *versus* 29%. Inadequate power to detect this difference would mean that if a difference of 10% *versus* 25% were found in the study, this difference would be reported as 'not statistically significant'. Moreover, many studies did not even have enough power to detect a large effect size.

When we examined the eight positive studies, none had adequate power to detect a small effect size, although six could detect a large one. The lack of power in these positive studies emphasises that investigations which find only statistically significant results ('positive studies') are not by definition of adequate power. It is possible to perform a study with inadequate numbers of subjects but still find results which are statistically significant. The danger of a positive study of inadequate power is that it can leave the reader with a false impression of the true difference between groups. This is because in order for a small study to show statistically significant results, the magnitude of the difference between groups must be large. For example, a study which compares the time to union using internal fixation with casting for fractures of the tibia may show a statistical difference between them even if only very few patients were in each group, but only if the difference was vast, say two weeks for internal fixation and 16 weeks for casting. Readers who look first at the difference between groups, and then separately consider the 'truth value' of that difference by looking at the p value, may interpret such a study by concluding that internal fixation is much better than casting, and that the difference between them is statistically significant. In fact, all that the p value states is that it is unlikely that the two groups have an identical effect; it does not comment on the magnitude of the difference. It may be that the 95% confidence interval for the two groups is one to ten weeks for internal fixation compared with 11 to 20 weeks for casting. It is certainly possible that the difference between the two is actually quite negligible, but in the absence of published confidence intervals the importance of the p value may be overstated.

This example highlights the need to report the confidence interval in all study results. It gives the reader an assessment of where the true mean value for the group lies. Because of the law of large numbers, which states that statistical truths are more likely to be manifest as the number of observations increases, the confidence interval for a given mean value narrows as the sample size increases. Accordingly, deficiencies of sample size and the possibilities of type-II error will be highlighted by wide confidence intervals.

The most appropriate way to prevent a type-II error is to perform a calculation of sample size before starting a study. Only 9% of the randomised, controlled trials which we have examined reported such calculations. It is possible that the investigators performed them, but did not report these calculations in the 'Methods' section of their manuscripts. This, however, seems unlikely, since previous studies have found that only rarely do authors perform calculations of sample size and not include them in the published report.[17] We stress that orthopaedic research is not alone in this. A review in 1994 of three prominent medical journals found that only 33 of 102 trials (32%) reported calculations of sample-size.[5] Given the effort and expense of performing a randomised, controlled trial, such calculations must be an integral part of these studies. Including an adequate number of subjects is the only way to ensure that if there is a clinically significant difference, the results will be statistically significant. The inadequate sample sizes yielding low power in the studies which we have sampled may have contributed to the inconclusive (negative) statistical outcomes in 25 of 33 studies.

A negative study with adequate power to detect clinically significant differences provides a valuable contribution to the literature. It informs the reader that the treatments under study are comparable. A negative study with inadequate power is, at best, inconclusive and may easily be misinterpreted. Such negative studies are not appropriate for the practice of evidence-based medicine.

It is important to distinguish between a 'clinically important' effect size and a statistically significant or scientifically plausible one. In comparing two methods of treatment, finding 'statistically' significant differences does not mean that the differences are 'clinically' significant. For instance, in a very large trial, a difference in the rates of infection between two groups of 20% *versus* 21% may have a p value of 0.01, but a clinician may nevertheless consider the rates of infection to be equivalent. This absolute difference of 1% may be highly significant statistically, but it is probably irrelevant clinically. Likewise, if a difference of 10% in infection is clinically relevant, a study which finds a difference in infection rate of 20% or 30% and was not statistically significant would be disappointing and inconclusive. Therefore, in order to decide on the effect size required to detect statistical significance, its clinical relevance must be determined.

Limitations of the study. Are the papers examined in our study representative of the orthopaedic literature? This can be proven only by examining all the contents of the other orthopaedic journals, which defeats the purpose of sampling. We believe that our sample is reasonable because the journals studied have a high citation index and prestige and recognition within orthopaedic surgery. By studying the 1997 volumes we obtained the best possible sample as regards statistical methodology, since editorials and other guidelines have stressed these issues. We have studied only randomised, controlled trials because these are, by definition, prospectively designed, and are most likely to have adequate statistical power.

We calculated power only for the primary outcomes of the study since these are of major clinical interest, and drive decisions on treatment. They would have been the outcome for which the authors would have undertaken power calculations had they thought to do so.

Objection may be raised to our use of the Cohen standard effect sizes. This standard is less appropriate than would be effect sizes designated by the researchers themselves, but since few of the studies described an effect size, a substitute standard was needed. We chose the Cohen method because it is reproducible, it considers both the relative and absolute magnitude of the differences between groups, it allows evaluation of statistical power based on three effect sizes and it is an approach which is accepted in medicine.[6,9-11]

The use of these standardised effect sizes has been employed only in a retrospective, external calculation of power in the absence of assertions by the authors of the studies. They have no place in the planning of a clinical study. Even in retrospective reviews, the preferred values would be those specified by the authors themselves or by peer consensus. However, typically, authors state an effect size only when they perform calculations of sample size. Thus a review of inadequacies of sample size must rely on some substitute standards. Those of Cohen are imperfect, but are as good as any other.

Conclusions

Sample size and power must be addressed before the start of a randomised, controlled trial. It is necessary to know in advance the likelihood of finding valid conclusions from the population assessed. It may be acceptable to subject a patient to a chance of a less than ideal treatment, or to the psychological stress of being a 'subject', if there is a chance of a valid scientific outcome, but it is not ethical to conduct a study the design of which cannot make valid conclusions.

Knowledge of the requirement of sample size is crucial in the planning of a prospective study, and such information should encourage investigators into engaging in multicentre trials. The performance of power calculations specifically demands that the minimal effect of interest is established and calls attention to important details which may otherwise be overlooked.

A proper design of the study and appropriate statistical analysis are essential to the validity of all quantitative clinical research. A type-I error is better known than a type-II error and reviewers and readers are more cognisant of p values when authors conclude that significant differences between groups are found. Equal scrutiny is required when authors decide that there is no statistically significant difference.

In this age of limited resources and tight budgets, physicians may be forced to employ the cheapest methods, especially if the choices are thought to be similar. It is therefore important that investigators do not erroneously label two treatments as equivalent, when it has merely been shown that the differences were not statistically significant. All clinical studies should be based on appropriate calculations of sample size. The awareness of type-I error and the popularity of p values should be matched by equal cognisance of type-II error and β values. The practice of evidence-based medicine requires no less.

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.

References

1. **Cohen J.** *Statistical power analysis for the behavioral sciences.* Second ed. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc, 1988.
2. **Rosner B.** *Fundamentals of biostatistics.* Fourth ed. Belmont, California: Wadsworth Publishing Company, 1995.
3. **Altman DG, Dore CJ.** Randomization and baseline comparisons in clinical trials. *Lancet* 1990;335:149-53.
4. **DerSimonian R, Charette LJ, McPeek B, Mosteller F.** Reporting on methods in clinical trials. *N Engl J Med* 1982;306:1332-7.

5. **Moher D, Dulberg CS, Wells GA.** Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994;272:122-4.

6. **Polit DF, Sherman RE.** Statistical power in nursing research. *Nurs Res* 1990;39:365-9.

7. **Cook DJ, Guyatt GH, Laupacis A, Sackett DL.** Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest* 1994;105:305-11.

8. **Hennekens CH, Buring JE.** *Epidemiology in medicine.* Boston: Little Brown & Co, 1987.

9. **Ottenbacher KJ, Barrett KA.** Statistical conclusion validity of rehabilitation research: a quantitative analysis. *Am J Phys Med Rehabil* 1990;69:102-7.

10. **Reed JF, Slaichert W.** Statistical proof in inconclusive 'negative' trials. *Arch Intern Med* 1981;141:1307-10.

11. **Fox N, Mathers N.** Empowering research: statistical power in general practice research. *Fam Pract* 1997;14:324-9.

12. **Dupont WD, Plummer WD.** Power and sample size calculations: a review and computer program. *Control Clin Trials* 1990;11:116-28.

13. **Brown CG, Kelen GD, Ashton JJ, Werman HA.** The beta error and sample size determination in clinical trials in emergency medicine. *Ann Emerg Med* 1987;16:183-7.

14. **Williams JL, Hathaway CA, Kloster KL, Layne BH.** Low power, type II errors, and other statistical problems in recent cardiovascular research. *Am J Physiol* 1997;273:487-93.

15. **Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR.** 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. *N Engl J Med* 1978;229:690-4.

16. **Chung KC, Kalliainen LK, Hayward RA.** Type II (beta) errors in the hand literature: the importance of power. *J Hand Surg [Am]* 1998;23:20-5.

17. **Liberati A, Himel HN, Chalmers TC.** A quality assessment of randomized control trials of primary treatment of breast cancer. *J Clin Oncol* 1986;4:942-51.