The Royal College of Surgeons of England

ORTHOPAEDICS

Ann R Coll Surg Engl 2008; **90**: 58–61
doi 10.1308/003588408X242312

The misuse of 'no significant difference' in British orthopaedic literature

SA SEXTON, N FERGUSON, C PEARCE, DM RICKETTS

Department of Orthopaedic Surgery, Princess Royal Hospital, Haywards Heath, West Sussex, UK

ABSTRACT

INTRODUCTION Many studies published in medical journals do not consider the statistical power required to detect a meaningful difference between study groups. As a result, these studies are often underpowered: the sample size may not be large enough to pick up a statistically significant difference (or other effect of interest) of a given size between the study groups. Therefore, the conclusion that there is no statistically significant difference between groups cannot be made unless a study has been shown to have sufficient power. The aim of this study was to establish the prevalence of negative studies with inadequate statistical power in British journals to which orthopaedic surgeons regularly submit.

MATERIALS AND METHODS We assessed all papers in the last consecutive six issues prior to the start of the study (April 2005) in *The Journal of Bone and Joint Surgery* (British), *Injury*, and *Annals of the Royal College of Surgeons of England.* We sought published evidence that a power analysis had been performed in association with the main hypothesis of the paper.

RESULTS There were a total of 170 papers in which a statistical comparison of two or more groups was undertaken. Of these 170 papers, 49 (28.8%) stated as their primary conclusion that there was no statistically significant difference between the groups studied. Of these 49 papers, only 3 (6.1%) had performed a power analysis demonstrating adequate sample size.

CONCLUSIONS These results demonstrate that the majority of negative studies in the British orthopaedic literature that we have looked at have not performed the statistical analysis necessary to reach their stated conclusions. In order to remedy this, we recommend that the journals sampled include the following guidance in their instructions to authors: the statement 'no statistically significant difference was found between study groups' should be accompanied by the results of a power analysis.

KEYWORDS
Sample size – Data interpretation – Orthopaedic

CORRESPONDENCE TO
SA Sexton, Vachery Cottage, Hook Lane, Shere, Surrey GU5 9QQ, UK
E: shaunsexton@doctors.org.uk

Throughout the orthopaedic literature, we have observed that the statement 'no significant difference' is often not accompanied by the results of a power analysis. In this context, 'no significant difference' is difficult to interpret and the reader has no easy way of determining the probability of the statement being incorrect.[1–4]

This study aimed to examine the British journals to which orthopaedic surgeons commonly submit papers in order to measure the extent to which 'no significant difference' was misused.

Rationale for study

Any clinical trial designed to investigate the efficacy of treatment (surgical procedure, drug, prosthesis or other factor) aims to test a hypothesis. It requires a null hypothesis (the statement that there is no difference between the population values of the parameter of interest (*e.g.* the mean) in the study and control groups) and, therefore, an alternate hypothesis, usually the statement that there is a difference between the population values of the parameter of interest (*e.g.* the mean) in the study and control groups.[5–7]

A clinical trial involves extrapolating the trial sample results to a larger population. Making this generalisation about a population based on the results of a trial sample introduces the possibility of the wrong conclusion being drawn as a result of sampling variability. So, even if the trial has been designed and carried out to a very high standard, it is still possible that the trial sample could, purely by chance, be too small to detect an effect of interest (*e.g.* a difference in the means) between the study and control groups. In other words, a trial can result in a Type I or Type II error.

Type I error

Incorrectly rejecting the null hypothesis when it is in fact true. In other words, a false positive – falsely concluding there is a treatment effect when none is present. The probability of committing a Type I error is known as α (alpha) and is denoted in trial results as a P value (*e.g.* $P < 0.05$).

Type II error

Failure of a trial to reject the null hypothesis when it is false. In other words, a false negative – falsely concluding there is no treatment effect when there is. The probability of committing a Type II error is known as β (beta).

The ideal trial would have a low value of α to avoid the chance of judging a new technique to be effective when it is not. The ideal trial would also be able to detect a real effect, minimising the risks of a Type II error or, in other words, a low value of β. Therefore, the ideal trial aims to minimise both α and β.

Unfortunately, for any given sample, α and β are linked. If α is made smaller, β must get larger and *vice versa*. In order to reduce both α and β (all other things remaining unchanged within the trial) the sample size must be increased. This is where power analysis comes in.

The power of a trial is a measure of its ability to detect an effect if it is real. In other words, it is a measure of the ability of the trial to reject the null hypothesis correctly when it is false. In most orthopaedic trials, β is usually set at either 0.2 or 0.1 which gives power values (1 – β) of 0.8 (80%) and 0.9 (90%), respectively. Therefore, if a treatment effect of a given magnitude is truly present, the trial has an 80% or 90% probability of detecting it. The statement 'there is no significant difference between groups', which is often seen in the orthopaedic literature, may only mean 'there is no statistically detected difference between the groups in our study'. Often, what the reader would like to know is 'is there no clinically significant difference between the groups in the study?' Calculating the power of a trial enables the authors to revise the statement 'there is no significant difference between groups' to one that is far more informative – 'there is no statistically significant difference between groups and our study had an 80% (or 90%) probability of detecting our minimum specified difference in treatment effect'. An example of a power calculation is given in Appendix 1.

Often in clinical trials, the size of the study groups are limited by ethical, financial or other constraints. There are several methods, other than increasing sample size and changing the P-value, that can be utilised to increase the power of a study:

1. *Increase the sample size in the control, low-risk, cheaper or less rare group.*
2. *When comparing means, reduce the standard deviation. Standard deviation can be reduced by improving measurement techniques (less 'messy' data) or by using a more homogeneous group of subjects (although this will limit the trials conclusions to that particular population from which the sample was selected, and so is often not appropriate).*
3. *Increase the size of the effect you would be satisfied to detect. Any study will have a greater power to detect a larger difference. However, the tail must not be allowed to wag the dog and one must be careful to ensure that samples size calculations should not be allowed to change the clinical effect you would be satisfied to detect.*
4. *Change the variables measured so that they are continuous measurements. Simple yes/no outcomes need far higher sample sizes. For example, using a scoring system with a continuous variable measurement will require smaller sample sizes than a binary good/poor outcome measurement.*
5. *Avoid using too many experimental groups. This will result in smaller sample sizes in each group. Consider reducing the scope of the study to focus on the minimum number of groups.*

An ideal high-power trial would have a large sample size, would be looking for a large clinical effect in a population with a low standard deviation. Conversely, if a trial uses a small sample size, looking for a small clinical effect in a population with a high standard deviation, then a conclusion of no statistically significant difference will be of very little use.

The principles described above also apply when calculating confidence intervals. If the sample size is too small, then the trial will produce wide confidence intervals and, as a result, a real effect can be missed (Type II error). The paper by Dorey *et al.*[9] provides a good review of the importance of confidence intervals in the presentation of data.

It is clear that power analysis should be used where possible at the start of prospective trials. However, the use of power analysis retrospectively is more open to debate. If used correctly, power analysis is still very useful; however, if used incorrectly, it can have serious short-comings. For example, if a trial reaches a conclusion of 'no statistical difference between groups', then the trial has a low power to detect the effect actually observed. Using the *observed* data from the study and calculating the power using the *observed* difference is known as a *post-hoc* power analysis and is generally futile.[8] However, what is useful is the calculation of the power of the trial to detect a difference would have been clinically relevant. If this information is provided in a paper, it enables the reader to determine whether the conclusion 'no significant difference between groups' is a reasonable one to make in order to detect this stated clinical effect.

Materials and Methods

The most recent six issues of the *Journal of Bone and Joint Surgery* (British), *Injury*, and *Annals of the Royal College of Surgeons of England* prior to April 2005 were reviewed by two of the authors (SS and NF) and their results compared to ensure agreement. Inclusion criteria for a paper were that

Table 1 Results broken down by journal

	Number of papers included in study	Papers with no significant difference as primary conclusion	Adequate power analysis done
J Bone Joint Surg Br	102	18 (17.6%)	3 (16.7%)
Ann R Coll Surg Engl	26	7 (26.9%)	0
Injury	42	24 (57.1%)	0
Total	170	49 (28.8%)	3 (6.1%)

the study had as its primary conclusion a statistical comparison between two or more groups. Both prospective and retrospective studies were included for the reasons described above. Papers using either numerical or binary data were included. Studies where no statistical comparison was made between groups (*e.g.* descriptive data) were excluded. We noted whether a power analysis had been performed or, failing that, whether enough information was provided in the text of the paper to enable us to perform our own power analysis.

Results

A total of 170 papers met the criteria described in Materials and Methods. Of these 170 papers, 49 had a primary conclusion of no significant difference between two or more groups. Only 3 (6.1%) of these 49 papers, where there was a primary conclusion of no significant difference, reported a power analysis demonstrating adequate sample size. One further paper performed a power analysis, and the sample size used was confirmed as inadequate (Table 1).

Discussion

It appears from these results that power analysis is underutilised in the three British journals studied. Without power analysis, a reader of these papers is unable to assess the validity of the statement 'no significant difference' and, therefore, does not know the likelihood of a Type II error. This study demonstrates that, for the three journals investigated, a conclusion of no significant difference perhaps should be treated with scepticism, as 93.9% of these conclusions are not backed up by an adequate power analysis.

This paper has concentrated on the importance of power analysis in relation to a 'non-significant' result. However, it must be emphasised that a power analysis should be performed for all studies where groups are compared. The sample size should be large enough to detect a given difference as significant, but should not be too large to be wasteful of patients. Therefore, investigators should always justify their sample size at the outset of a study in the form of a power statement. This is a statement that gives values for the significance level sought, power, clinical treatment effect sought, and standard deviation of the observations.

Many of the papers we have reviewed which have a conclusion of no significant difference add important information to medical knowledge. However, there is a danger that, without an adequate power analysis, the prevalence of Type II errors may be high and, as a result, beneficial new interventions may be discarded or follow up studies with adequate sample size may not be performed.

One argument against the use of power analysis is the difficulty in performing the calculations required to determine sample size. We have found the statistical program nQuery[10] to be a useful aid. However, real difficulty can be faced in sample size estimation when dealing with more complex analyses rather than the more simple two sample comparisons.

Post hoc power calculation has been deprecated, as it can not change the outcome. Used incorrectly we would agree. A *post hoc* power calculation using the *observed* treatment difference in order to calculate sample size is futile. However, a power analysis calculated retrospectively, as described in the introduction utilising the *clinically important* treatment difference (rather than the observed treatment difference), will enable the reader to assess the risk of a Type II error having occurred, and also what clinical difference the study would be likely to pick up. Ideally, a power analysis should be performed at the start of a trial; failing that, a correctly applied retrospective power analysis is better than no power analysis.

Conclusions

We recommend that the journals studied provide additional guidance in their instructions to authors to include the following: the statement 'no significant difference was found between study groups' should be accompanied by the results of a power analysis.

References

1. Chung KC, Kalliainen LK, Spilson SV, Walters MR, Kim, HM. The prevalence of negative studies with inadequate statistical power: an analysis of the plastic surgery literature. *Plast Reconstr Surg* 2002; **109**: 1–6.

2. Maggard MA, O'Connell JB, Liu JH, Etzioni DA, Ko CY. Sample size calculations in surgery: are they done correctly? *Surgery* 2003; **134**: 275–9.
3. Lochiner HV, Bhandari M, Tornetta P. Type-II error rates (beta errors) of randomized trials in orthopaedic trauma. *J Bone Joint Surg Am* 2001; **83**: 1650–5.
4. Bhandari M, Richards RR, Sprague S, Schemitsch EH. The quality of reporting of randomized trials in the *Journal of Bone and Joint Surgery* from 1988 through 2000. *J Bone Joint Surg Am* 2002; **84**: 388–96.
5. Machin D, Campbell M, Fayers P, Pinol A. *Sample size tables for clinical studies.* Oxford: Blackwell Science, 1997.
6. Campbell MJ, Machin D. *Medical statistics: a common sense approach.* Chichester: John Wiley, 1999.
7. Everitt BS. *Statistical methods for medical investigations.* Hodder Arnold, 1994.
8. Clark VA. Discussion: the prevalence of negative studies with inadequate statistical power: an analysis of the plastic surgery literature. *Plast Reconstr Surg* 2002; **109**: 7–8.
9. Dorey F, Nasser S, Amstutz H. The need for confidence intervals in the presentation of orthopaedic data. *J Bone Joint Surg Am* 1993; **75**: 1844–52.
10. nQuery Advisor® 6.0. Cork, Ireland: Statistical Solutions, 2004.
11. Dallal GE. The 17/10 rule for sample size determination. *Am Stat* 1992; **46**: 70.

Appendix 1

USING POWER ANALYSIS TO CALCULATE SAMPLE SIZE

For many of the commonly used tests comparing two groups, sample size can be estimated using quick formulae, an example of which is given below. Alternatively, tables such as by Machin *et al.*[5] can be used to calculate the sample size, or commercially available computer programs can be used.

Quick formula for calculating samples size when comparing two means where the outcome data are normally distributed (two sample *t*-test)

In order to calculate the sample size required for a trial, five variables must first be determined.

1. α	the significance level or P value: this is usually .05 or 0.01.
2. Power $(1 - \beta)$	this is usually 80% or 90%.
3. E	the minimum effect size that is clinically relevant.
4. σ	the standard deviation of the sample.
5. Δ	E/σ

Using the formula of Dallal:[11] $N = 17/\Delta^{1.9}$

This quick formula assumes a two-sided significance of 0.05 and a power of 80%.

EXAMPLE

A study compares the maximum flexion obtained after implanting one of two types of total knee replacement (knee A and knee B). If the study aims to detect a difference between groups of 10° and there is a between subject standard deviation of 20°, how many patients should be recruited for the trial?

In this study, $E = 10$ and $\sigma = 20$; therefore, $\Delta = 10/20 = 0.5$.

Using Dallal's formula to estimate sample size:

$$N = 17/0.5^{1.9} = 63.4.$$

This is rounded up to give a sample size of 64 patients in each group, a total of 128.

In the above example, the sample size required is calculated assuming continuous normally distributed outcome data. We will consider how altering these and other factors will affect the power of the trial.

1. Deviations from a normal distribution usually result in lower power and, therefore, a greater sample size is required to detect a given effect.
2. If the study is one-tailed (in other words you expect a greater flexion with knee B, for example, based on the results of previous studies), then the study will increase in power and the sample size required will decrease. However, this is true only when the effect is in the expected direction. If it is in the other direction, then the power is zero.
3. Within-subject studies (paired tests) have higher power than unpaired tests.
4. If the outcome data in the example above are replaced by ordered categorical data, the power of the study will usually decrease and increased sample size will be required. In the above example, if the outcome data of flexion in degrees is replaced by an outcome score of 1 (very poor flexion) to 5 (very good flexion), then the data have become ordered categorical.
5. In the example above, a difference of 10° was the clinical effect that the trial was designed to detect. However, if this was increased to 20°, the power of the study would increase and the sample size required would decrease. Care must be taken not to raise the clinical effect (E) such that a clinically important difference may be missed.
6. Changing the significance level (α) will alter the power of the study. Increasing the P value above 0.05 should usually be avoided. However, in some cases, reducing the P value may be appropriate at the start of a study. As a result, the sample size required would increase.
7. If the measurement of maximum flexion is accurate only to the nearest 5°, then the standard deviation of the outcome data will be higher than if the flexion measurement is accurate to 2°. A higher standard deviation will require the study to increase sample size in order to maintain its power. Therefore, accurate and sensitive measurement of data are important in helping to reduce the sample size required.