ISAKOS Scientific Committee Report

Considerations on Sample Size and Power Calculations in Randomized Clinical Trials

Jon Karlsson, M.D., Ph.D., Lars Engebretsen, M.D., Ph.D., and Katie Dainty, M.Sc., C.R.P.C.

Abstract: Many studies in orthopaedics and sports medicine have not considered sample size or statistical power as important issues in study design. This article addresses the importance of a sample size calculation in randomized clinical trials and the components of the calculations that researchers must consider in their preliminary planning of an investigation. The types of data being collected, level of significance, types I and II errors, and power are also addressed. **Key Words:** Sample size—Power—Type I error—Type II error—Level of significance—Effect size.

Traditionally, many studies in orthopaedics and sports medicine have not considered sample size or statistical power as important issues in study design. It is therefore not surprising that many investigations have failed to answer the question asked for a variety of related reasons, i.e., due to low power (most often too few subjects included), a traditional type I statistical error, or simply too few numbers. Good sampling practice is essential in any investigation, and includes concern for the size of the sample used. Whole populations can rarely be measured and, therefore, decisions as to an adequate size of sample have to be made. Although practical and ethical issues must be considered, one's initial consideration when determining trial size should include the scientific requirements of the study. Common sense will suggest that the more subjects that are studied, the more likely one is to obtain a representative sample. The layman's formula that describes this relationship is basically:

The margin of error in a sample = 1 divided by the square root of the number of people in the sample.[1]

EFFECT SIZE OR TREATMENT SIZE

The sample size of a clinical trial depends on the size of the difference to be detected between the 2 groups. The larger the difference one wishes to detect, the more patients will be required for the study. It is usually best to require that there is a sufficient number of subjects to detect the minimal "clinically" important difference to avoid the need for astronomically large numbers of patients. Applying the above rule, it makes sense that studies in which strong treatment effects and large differences between groups are expected, require fewer subjects.

ALPHA ERROR (A) = TYPE I ERROR

The sample size calculation also involves consideration of the risk of an α error (accepting the treatment is effective when it is not). The acceptable size for this risk is a value judgment, and can be as large as 1 or as small as 0. When calculating the number of patients required, the number will vary according to the size of the chance the investigator is willing to take of

From Sahlgrenska University Hospital/Ostra, Gothenburg, Sweden (J.K.); Bergslia 22, Oslo, Norway (L.E.); and Fowler Kennedy Sport Medicine Clinic, London, Ontario, Canada (K.D.).

Address correspondence and reprint requests to Katie Dainty, M.Sc., C.R.P.C., Fowler Kennedy Sport Medicine Clinic, 3M Centre, University of Western Ontario, London, Ontario N6A 3K7, Canada. E-mail: kdainty@uwo.ca

0749-8063/03/1909-3889$30.00/0

doi:10.1016/S0749-8063(03)00837-5

falsely concluding that a given intervention is valuable. The larger the chance accepted, the fewer the patients needed. As easy as this sounds, it does not do much for the credibility of a trial.

Alpha (p_α) is usually set at 0.05 (1 in 20) or 0.01 (1 in 100).

BETA ERROR (B) = TYPE II ERROR

Type II or β error is similar to α error, but it is the chance of missing true differences in a particular study, or in other words, accepting a null hypothesis even though it is false. β errors are conventionally much larger than α errors, reflecting the higher value placed on being sure that an effect is really present when one reports that it is.[3]

LEVEL OF SIGNIFICANCE

The chosen level of significance sets the likelihood of detecting a treatment effect when none exists (leading to a so-called false-positive result), and defines the threshold *P* value. Results with a *P* value above the threshold lead to the conclusion that an observed difference may be due to chance alone, while those with a *P* value below the threshold lead to rejecting chance and concluding that the intervention has a real effect. The level of significance is most commonly set at 5% (i.e., $P = .05$) or 1% ($P = .01$). This means that the investigator is prepared to accept a 5% (or 1%) chance of erroneously reporting a significant effect.

POWER (1-B)

The power of a study is defined as its ability to detect a true difference in outcome between the standard or control arm and the intervention arm. It is the power to detect a statistical difference between 2 different treatments, when the difference in fact exists. This can also be described as the possibility that the study shows a statistical significance (at a level decided on beforehand) provided there is an effect. This is correlated to the possibility to find a clinically relevant effect for any given drug or treatment. This may be termed "clinical effect," which is not always the same as statistical effect. Before deciding on the null hypothesis, a minimal clinically important difference should be discussed. One must consider this carefully, since "clinically significant difference" can be a matter of great debate.

Power is affected by many factors, namely the significance level, the effect size, and the sample size. The lower the significance level, the lower the power. If the researcher wishes to reduce the risk of concluding that an ineffective medication (or any other treatment) is effective, then they must understand that they, in turn, run the risk of missing a treatment which is, in fact, effective. The larger effect a treatment has, the larger is the possibility to discover it within the trial. A generally accepted power is 80% or 0.8. In many recent studies, the power limit is set to 0.9 or even 0.95. This is more often the case in pharmacologic studies, i.e., testing new drugs, than is the case in investigation of surgical techniques. A power of 0.8 means that the researcher is correct 8 times in 10 when accepting his conclusion. Often, a higher power is requested in studies for dramatic new treatments such as for diabetes or cancer. A power of 0.8 also means that the researcher may be incorrect 2 times out of 10. The clinically relevant difference and the risk of committing a type II error should be taken into consideration when defining the power limit desired.

METHODOLOGY

Sample size calculations should be performed during the design phase of the study. There are several computerized methods that can be easily accessed through the Internet if statistical support is not at hand. However, the formula must be appropriate for the design of the trial and the subsequent analysis. Each sample size calculation is unique, and there is no global solution applicable to all trials.

Sample Size Equation for 2 independent groups means:

$$n/\text{group} = 2\{(Z_\alpha + Z_\beta)\,\sigma/\Delta\}^2$$

Z_α = z-value for α found in a Z table (i.e., Z for $\alpha = 0.05$ is 1.96).

Z_β = z-value for β found in a Z table.

σ = standard deviation in mean response.

Δ = minimal difference desired in outcome.

It is strongly recommended that one enlist the aid of a biostatistitian at the beginning to help determine the correct sample size calculations. However, it is always a good idea to understand the calculation and why the statistician did what he or she did.

Things to consider when determining sample size:

1. The investigation or analytical procedure that one wishes to follow as part of the initial experimental design stage of the investigation.

2. The statistical procedure that one may wish to use for analysis (e.g., 1-tailed or 2-tailed *t* test, linear regression).
3. The percentage change and/or level of significance at which one would wish to detect the change.
4. The appropriate sample size formula for the study design in question.

One can see why it is extremely useful to have the help of a statistician from the outset. Finally, one should explicitly choose between 2-sided or 1-sided statistical testing. Unfortunately, the importance of this decision is often neglected. In the 2-sided option, one is able to detect the predefined relevant difference, if present, in both directions. That is, if the intervention is better than the reference of a given amount, this will probably be found, but also a similar advantage of the reference over the principal intervention can be detected. The 1-sided option would only allow a 1-sided evaluation (e.g., whether or not the intervention is better than the reference treatment of a given amount), without testing whether the reference is superior to the principal intervention.[4-6]

CONCLUSIONS

It can be concluded that the larger the sample size, the more precision one will have when stating the findings. With increased precision, there is an increased possibility of establishing an effect (if it truly exists). In other words, it is possible to find even very small effects (e.g., differences between groups) if the sample size is large enough. However, it can be questioned whether it is correct, from an ethical perspective, to perform a very large investigation ("the larger the sample, the higher the power") to identify an effect that may be small and not clinically relevant. This may, in fact, be both incorrect and dangerous, and one must always remember that any new treatment may be potentially harmful. Sample size calculation, including power analysis, should always be included in the methodologic design of any clinical trial. Especially when no statistical difference is found, the possibility of either type I or type II error should be discussed. In fact, a study showing no statistical difference is often more the result of low power than to true lack of difference between the study groups.

REFERENCES

1. Friedman LM, Furberg CD, Demets DL. Sample size. In: Fundamentals of clinical epidemiology. New York: Mosby, 1996;94-125.
2. Pocock SJ. *Clinical trials, a practical approach.* Chichester: Wiley, 1983.
3. Bland JM, Altman DG. One and two sided tests of significance. *BMJ* 1994;309:248.
4. Dunnett CW, Gent M. An alternative to the use of two-sided tests in clinical trials. *Stat Med* 1996;15:1729-1738.
5. Koch GG. One-sided and two-sided tests and *P* values. *J Biopharm Stat* 1991;1:161-170.
6. Peace KE. The alternative hypothesis: One-sided or two-sided? *J Clin Epidemiol* 1989;42:473-476.

RECOMMENDED READING

1. Greenfield ML, Kuhn JE, Wojtys EM. A statistics primer. Power analysis and sample size determination. *Am J Sports Med* 1997;25:138-140.
2. Lachin JM. Introduction to sample size determination and power analysis for clinical use. *Control Clin Trials* 1981;2:93-113.