Special Article

Experimental design and statistical analysis

Patricia F. Moodie BSC HONS MSC,
Douglas B. Craig MD FRCPC

The failure of authors to recognize certain basic concepts of experimental design and statistical analysis continues to be a cause of concern to Editorial Boards, including the Board of this Journal. These failures lead in many cases to rejection of manuscripts, after considerable expenditure of time and resources in the conduct of the research study and its written description. Other studies, with less serious flaws are published only following major revision.

We hope that the following observations and recommendations will stimulate potential contributors to the *Journal* to consider design and statistical factors and/or consult a statistician early, ideally when they are planning research studies. The common goal of authors and editors must be better research and improved manuscripts.

What follows is not an exhaustive or in depth presentation of statistical maxims and design. We have addressed some of the major problems which we feel require more attention from authors. These are: generalization of the research, sample size, power, the issue of negative conclusions, and nonrandomized trials.

As a first step, all researchers should be concerned about the generalization of their results, i.e., they must ensure that the conclusions of their research will be capable of extension to the study population (the population from which the study sample is drawn).[1] It is rarely of interest to know, for example, how a given treatment affected the ten patients in a study, if no further generalization of these observations is possible. We need, therefore, a mechanism which will assure us that the study sample chosen (the ten patients) is representative of the population from which it is drawn. This requires that we first define the population in precise detail in terms of who, where, when – for example all patients over 35 years of age admitted to maternity ward X in hospital Y and undergoing Caesarean section during time period Z. Upon having specified the study population, a sampling technique (random, stratified random, etc.) appropriate to the study should be employed. The choice of sampling technique could require consultation with a statistician.

The reality of the situation, however, may be such that it is practical to do research on only those patients in one's own practice. In this less than ideal situation, it is very important to describe all the clinically relevant characteristics of these patients as well as anything else about them (e.g., an unusually high proportion of vegetarians) that might suggest they are not representative of the population of interest. This will help others decide whether the conclusions of the study can be extended to their own patients.

Estimation of sample size is also an important component of the design phase. Some research projects should never be initiated because one can predict that the numbers of subjects available for study will inevitably be inadequate. In order to estimate the required number of subjects, we need certain items of information:

- *the level of significance criterion (alpha) to be used.* This may be defined as the level (extent) of risk one is willing to take in incorrectly declaring a result statistically significant. The level of significance is usually 0.05.[2]
- *whether a one-sided or two-sided statistical test will be applied.* Often in anaesthesia a two-sided

From the Manitoba Cancer Treatment and Research Foundation, 100 Olivia Street, Winnipeg, R3E 0V9, and the Department of Anesthesia, University of Manitoba, Winnipeg, Manitoba, R3E 0Z3

test is appropriate, because if there is evidence to suggest that a difference exists, two possibilities (two sides) should be considered (e.g., drug A better or worse than drug B).

- *the desired power*. Power can be thought of as the likelihood of detecting statistical significance in a particular study, when a true difference exists. There is no consensus as to a cutoff value for power, unlike the magical P value of 0.05. This lack of consensus may well be a good thing. As a rule of thumb, however, it seems that achieving a power greater than or equal to 0.80 is considered acceptable. If the power of a study is 0.80, that means it has 0.80 probability (or 80 per cent chance) of detecting statistical significance when a real difference of a certain size exists.
- *the difference to be detected*. This may be estimated from a pilot study or from the literature. If a pilot study is not possible and one does not know what difference to expect, one could specify the minimum difference between treatments that it is clinically worthwhile to detect. For example, if a difference of 5 mmHg in systolic blood pressure between a standard and a new drug is of clinical importance but any difference less than five is not, then five could be specified as the value for the difference to be detected in this case. However, if the true difference between the treatments is larger than five, say it was ten, then the sample size estimate based on a difference of five would be larger than was actually needed for the study. The larger the true difference between treatments, the smaller the number of subjects required to detect statistical significance.
- *estimated standard deviation*.[3] This is required in the case of estimating sample size needed to detect a given true difference between means. The standard deviation may be estimated from a pilot study or the literature. (Note: it is necessary to know only the ratio of the standard deviation to the mean difference to be detected, not their actual values. However, it is not always easy for researchers to think in terms of this ratio.)

Often researchers do not realize that these items of information are required for sample size estimation and are disappointed (sometimes indignant) when a statistician cannot specify the number of subjects required, in the absence of this input from them. A very useful handbook for researchers on sample size estimation is that of Cohen.[4] His approach is practical throughout and he provides tables from which sample size estimates can be obtained.

The actual sample size used in any study has an important bearing on what can be concluded from the research when no statistically significant differences are found. **Lack of statistical significance should not be equated with negative proof.** When a difference is not found to be statistically significant, one of the following possibilities may be true – either there is no real difference between the treatment study populations or there is a real difference but the study was not able to detect it because of one or any combination of the following reasons: (a) the sample size was inadequate; (b) the variability within the samples was too great; (c) one (or more) of the assumptions of the statistical test applied was not met by the data.

Too often we have encountered authors who claim that their study demonstrates that two treatments do not differ in their effects. A study may tentatively suggest this. However, to give credence to even a tentative suggestion of no difference in effects, one should, if possible, calculate the power (capability) which that particular study had of detecting a statistically significant difference. For example, if the real difference between two treatment means is 70, the common standard deviation is 100, the sample size in each group is ten and one applied a two-tailed t test at the five per cent level of significance, the power (likelihood), one would have of detecting a statistically significant difference would be 0.31. In this case, the likelihood of detecting statistical significance is not very high and one would not be on very solid ground in postulating there really was no difference between the population means. Cohen[4] also discusses calculation of power and supplies tables which furnish appropriate power values for many types of analyses and situations.

As an alternative or supplement to power, a confidence interval[5,6] for the true difference between the study populations can help one decide whether lack of statistical significance really means absence of treatment differences of medical importance. Most medical research is conducted on samples rather than entire populations. Hence any measurements obtained on these samples only estimate true population values. If, for example, we observe in an experiment that there is a difference of 5 mmHg in systolic blood pressure between the means of the samples of patients receiving therapy A and those receiving therapy B, we cannot conclude there is the same 5 mmHg difference between the means of the *entire* study populations. Neither can we conclude that there is a zero difference between the study populations' means, if we apply a statistical test to these data and find no statistical significance. However, a confidence interval for the difference between population means provides us with a range of values which, with a specified level of confidence, will include the true value of this difference. If we calculate a 95 per cent confidence interval for the true difference in population means for this example and find it to be 0 to 60 mmHg, this tells us that the real difference between means of the study population could be a value from 0 to 60 mmHg. So, on the basis of this study, one could not very convincingly affirm that therapy A and therapy B have the same effect, although no statistical significance was found, for the 95 per cent confidence interval indicates there is a possibility that the real difference between the means of the study populations could be as great as 60 mmHg.

Ideally in a clinical trial, all factors that affect response to treatment should be the same in the study groups so that valid comparison of the treatments can be made. Random allocation (randomization) of the subjects to the treatments helps to achieve this objective, as it tends to balance the different types of subjects among the treatment groups. Although randomization is seldom totally effective, it ensures that any imbalance of subject types that may occur is due to chance and hence will be validly taken into account in the statistical analysis.[7,8] We realize that it is not always ethically possible to implement a study as a randomized clinical trial. However, when it is not possible to assign subjects to the treatment groups at random, it is essential for researchers to realize they cannot conclude that response differences observed on the subjects under different treatment regimens were caused by the treatment received. Without randomization, researchers cannot formally determine the probability that some characteristics of the subjects themselves are responsible for the differences observed in the treatments. The researcher should therefore take this weakness of nonrandomized trials into account in the discussion of the results. In the case of a randomized trial, especially if the sample sizes are small, the researcher should examine the characteristics of the subjects (age, sex, race, etc.) in each treatment group to see how effectively randomization has balanced these characteristics and discuss the results in the light of these findings.

Some are of the opinion that perhaps researchers should not use statistical techniques at all if they do not understand them. Our response to this is that one does not need to know the mathematical foundation or arithmetic calculations involved in a statistical technique but it is essential to understand, on an intuitive basis at least, how that technique is testing one's scientific hypothesis. In short, find a statistician with a good "dataside" manner who can translate statistical jargon. The reward to science will be great!

References

1 *Lavori PW, Louis TA, Bailar III JC, Polansky M*. Statistics in practice. Designs for experiments – parallel comparisons of treatment. N Engl J Med 1983; 309: 1291–9.

2 *Glantz SA*. Biostatistics: how to detect, correct and prevent errors in the medical literature. Circulation 1980; 61: 1–7.

3 *Swinscow TDV*. Standard deviation. Chap 2 *In*: Statistics at square one. British Medical Association, London, 1977.

4 *Cohen J*. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1977.

5 *Sokol RR, Rohlf FJ*. Biometry Ed 2. W.H. Freeman, San Francisco 1980.

6 *Brown WB, Hollander M*. Statistics. A Biomedical Introduction. John Wiley & Sons, Toronto, 1977.

7 *Gilbert JP*. Randomization of human subjects. N Engl J Med 1974; 291: 1305–6.

8 *Byar DP, Simon RM, Friedwald WT et al*. Randomized clinical trials. N Engl J Med 1976; 295: 74–80.

9 *Lebacqz K*. Controlled clinical trials: some ethical issues. Controlled Clin Trials 1980; 1: 29–36.