Statistics for clinical trials and audit

Ian Kestin

The importance of scientific evidence has only recently been recognized in clinical medicine and most of this evidence has been obtained from clinical trials. We need to be sceptical when reading trial reports in journals because there are commercial, institutional and personal influences on the 'impartial' use of scientific methods in medicine; one-third of all major original clinical research is later proved wrong.[1]

Study designs

Case studies describe the outcomes of an intervention in one or more patients. Case studies lack any control patients for comparison and methods to avoid bias.

Bias is any factor that may alter the results and lead to false conclusions. More than 30 different types of bias have been described. Bias in medical research may occur at all stages, and may be entirely unrelated to the conduct of the researchers (Table 1).

Retrospective studies involve observations of patients who have completed their treatment, and the data are obtained from written records. A common type of retrospective study is the case-control study in which patients who have the disease or condition of interest are compared with control patients who do not. These control patients are selected to match the patients as closely as possible. This selection inevitably has a risk of introducing hidden bias, the effect of which cannot be assessed. Missing data is another common problem of retrospective studies.

Prospective studies are those in which the patients are selected in advance and then studied in a structured format according to the study protocol.

A randomized controlled trial is a study in which the eligible patients are randomly allocated to receive a treatment. Usually one (or more) group receives the drug of interest, and one group, the control group, is used for comparison. Depending on the purpose of the study, the control group may receive an inert substance, a placebo, or a standard treatment for the disease studied. In some studies the patients receive all the treatments in sequence, and thus serve as their own controls. These are called crossover studies, and confounding factors will be equal across all treatments. Not all trials can be done using a crossover design, and the limitations are given in Table 2. The data can be tested statistically for the presence of these limitations after the study has been completed, but these tests will detect only major effects. It is best to ensure that a crossover trial is the appropriate design and is well conducted.

The purpose of randomization is to distribute the confounding factors, which may affect the response, equally across all treatments. Some of these factors may be known or obvious (e.g. age, gender, smoking) but, more importantly, there will be unknown factors (e.g. genetic) that may affect outcome. Recruiting an adequate number of patients and randomly allocating them to the different treatments is the only method of minimizing the effect of confounding variables. The method of randomization is important, and a recognized method, such as random-number tables, should be used by a researcher not connected with the study. Allocation by days of the week, hospital number or birthday is not random, and may introduce bias into the characteristics of the groups. There are several types of randomization.

Simple randomization allocates the patients to one of the treatment groups entirely by chance.

Stratified randomization occurs when the patients are initially subdivided according to baseline characteristics (e.g. age, gender). The subgroups are then allocated randomly to one of the treatments. Stratified randomization will reduce the risk of

Some common types of bias in clinical trials

Selection bias
This occurs when the patients are selected in a manner that introduces systematic differences between the groups. This can occur in many ways (e.g. poor methods of randomization). It may be an accidental or a deliberate manipulation of the study by the investigators

Measurement bias
This can arise if the measurements taken from the patients have systematic errors that affect some groups more than others. This can occur if the equipment is not calibrated uniformly, and is especially likely if observers are making subjective assessments of the patients

Publication bias
Failure to submit data for publication is a significant cause for bias in medical knowledge (see *Anaesthesia and Intensive Care Medicine* **7:4**: 135)

Attrition bias
The exclusion of patients, who have entered the study, from analysis may lead to errors in interpretation, as the rate or causes of drop-out from the study may not be equal for all groups. For example, when a medical and a surgical treatment are compared, if those who died as a direct or indirect result of surgery are excluded from analysis, a bias towards the surgical treatment is introduced. In general, all patients should be analysed in the original groups to which they were allocated (intention to treat)

Table 1

Ian Kestin, *FRCA, is Consultant Anaesthetist at the Western Infirmary, Glasgow. He trained in anaesthesia in Southampton, Bristol and the USA. His interests include education in anaesthesia and mathematical applications in medicine.*

Limitations of crossover design

Period effects
There should not be any significant change of disease severity during the study period. If the disease significantly worsened or improved between the first and second treatments, the two treatments were not studied under similar conditions. For example, transient diseases like the common cold cannot be studied using a crossover design. The order of the two treatments under investigation is usually varied among the patients to avoid bias from period effects

Treatment–period interactions
The effect of a treatment may vary according to the study period. The treatment may work more effectively earlier or later in the disease process than when it was given, or its effects may be modified by the other treatment

Carry-over effects
There must be adequate time for the effects of the first treatment to disappear before starting the second treatment, otherwise the true effects of the second treatment will not be measured

Table 2

unbalanced groups at the end of the study, and is used if there are important baseline characteristics known to affect the outcome of treatment. A disadvantage of this method is that a sufficient number of patients may be difficult to recruit to all the categories and the study may be delayed.

Minimization is a technique that is particularly useful if patients are difficult to obtain for the study. The first patient is allocated randomly, the second and subsequent patients are then allocated using a weighted randomization. The weighting is adjusted in each patient and is used to increase the chance that the patient is allocated to a treatment group that will minimize the differences in baseline characteristics already present between the groups. The principle of randomization is maintained while minimizing the chance of unequal groups at the end of the study.

Baseline characteristics: randomization cannot ensure that the confounding variables are equally distributed across the groups; the groups may still be unequal (e.g. all the males are allocated to one group). It is common to use statistical tests to check if the groups are similar in ages, weights, etc., after the study has been completed. These tests will detect major differences only, and, of course, the unknown confounding variables remain exactly that, unknown. If the two groups are found to differ in important characteristics after the study has been completed, all is not lost. The results can still be analysed using statistical techniques that compensate for differences in baseline characteristics, such as analysis of covariance or multiple regression analysis.

Blinding or masking means that the investigator, patient or both are unaware of the treatment group. In a double-blind trial both are unaware of the treatment group. Masking is important, because the response to treatment is often considerably altered by expectation, either by the patient or the investigator. If the patients know they are receiving the placebo, they do not expect to improve. However, in practice there can often be a considerable response to placebo. A randomized double-blind controlled trial is the gold standard for obtaining medical evidence. Sometimes this design is not possible for all studies (e.g. in studies comparing a surgical with a medical treatment) and these are known as open studies; however, they should still be randomized.

There are a number of guides to good practice in the conduct of research, for example, from the Association of British Pharmaceutical Industry, the General Medical Council, the Department of Health, the Medical Research Council, the European Parliament and the BMA.

Power analysis

An essential part of the design of any clinical trial is a power analysis, a statistical technique to estimate the number of patients required to reduce the risk of a Type II error (*Anaesthesia and intensive care medicine* **7:4:**135) to an acceptable value. The actual calculations in a power analysis depend on the type of data (i.e. categorical, ordinal or continuously variable data). The probability of a Type II error is denoted as β, and should be 20% or less. The power of the study is defined as $(1 - \beta)$. For example, if β has been chosen to be 10%, the power of the study is 90%; with this number of patients, the study has a 90% probability of demonstrating a treatment difference, if such a difference exists. The main determinants of the power of a study are:

- the magnitude of the difference between the treatment and control groups and the variability of the data. This magnitude is often unknown and is usually the purpose for doing the study, but estimates can be obtained from pilot studies, previously published work or chosen by the investigators to be the minimum difference of clinical importance to detect
- the values chosen for α (the level of statistical significance) and β.

Presentation of results

Confidence intervals are generally better than p-values for reporting the results of clinical trials; both contain the same mathematical information, but confidence intervals best show the implications of the analysis. For example, if the 95% confidence interval for the difference between the treatment and control group includes zero, then the reader immediately knows the treatment may be ineffective or even harmful. Confidence intervals are a particularly useful indication of the possible true incidence of uncommon complications after a study in which there were few adverse events. Many authors report no complications in their study and recommend their technique as safe, without giving an estimate of what the true incidence may be (Table 3).

The number needed to treat (NNT) is the reciprocal of the absolute risk reduction of a treatment, and is another useful statistic for reporting clinical trials (Table 4).

Diagnostic tests

It is very rare for a single diagnostic test to clearly and reliably distinguish healthy individuals or 'controls' from patients with the condition. Nearly always there is an overlap of test results that could occur in either 'patients' or 'controls'. The properties