Statistics for clinical trials and audit

Ian Kestin

Abstract

This article covers the application of statistics to clinical trials and audit, including the basic types of study design, bias, power analysis, guides to good clinical practice, the presentation of results and applications in quality assurance.

Keywords bias; crossover; randomization

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

Study designs

Bias is any factor that may alter the results and lead to false conclusions; over 30 different types of bias have been described. Some common types of bias in clinical trials are listed in Table 1.

Bias in medical research may occur at all stages and may be entirely unrelated to the conduct of the researchers. Statistical techniques have been used to demonstrate the effects of bias in clinical trials (Table 2).

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

Retrospective studies are observations of patients who have completed their treatment, and the data obtained after the events, for example, 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 in all respects except the disease in question. 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.

Common types of bias in clinical trials

- **Selection bias** 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 accidental or a deliberate manipulation of the study by the investigators
- **Measurement bias** can arise if the measurements made on the patients have systematic errors that affect some groups more than others. This can occur if equipment is not calibrated uniformly, and is especially likely if different observers are making subjective assessments of the patients. Observers inevitably make different assessments of the same observation
- **Publication bias** is a significant cause for bias in medical knowledge when not all data are submitted for publication*
- **Commercial bias** occurs because studies sponsored by pharmaceutical companies generally show more favourable results for drug therapy than independent studies,[2] and meta-analyses also seem biased by the commercial interests of the authors[3]
- **Attrition bias** may lead to errors in interpretation if patients who have entered the study from analysis are excluded, as the rate or causes of drop-out from the study may not be equal for all groups. For example, when comparing a medical and a surgical treatment, 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 (called **intention to treat**)

*See example in *Anaesthesia and intensive care medicine*, 2006; ***7:*** 135–42.

Table 1

Effects of bias on medical research

- There have been four meta-analyses comparing studies with adequate techniques to ensure that the investigators were unaware of the treatment allocation (i.e. adequate 'blinding') with studies of the same topic in which the investigators probably could have discovered the treatment of the patients. All reported that there was an obvious exaggeration of the benefits of treatment in studies in which masking was inadequate. The effects were similar but less marked if the trials had inadequate randomization of the subjects or if the patients were aware of the treatment allocation[4]

Table 2

Ian Kestin, *FRCA, is Consultant Anaesthetist at the Western Infirmary, Glasgow, UK. He trained in anaesthesia in Southampton, Bristol and the USA. His interests include education in anaesthesia and mathematical applications in medicine. Conflicts of interest: none declared.*

A randomized controlled trial is an experiment in which the eligible patients are randomly allocated to receive one of the treatments. Usually one or more groups receive 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 3.

The purpose of **randomization** is to distribute the confounding factors that may affect the response equally across all treatments. Some of these factors may be known or obvious, for example age, gender and smoking, but, more importantly, there will nearly always 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 someone unconnected with the conduct of 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. It is often easy for the investigators to 'adjust' the randomization if they are doing it themselves.

There are several types of randomization. **Simple randomization** allocates the patients to a one of the treatment groups entirely by chance. If the patients are initially subdivided according to baseline characteristics, for example age or gender, and then these subgroups are allocated randomly to one of the treatments this is called **stratified randomization**. Stratified randomization will reduce the risk that the groups are unbalanced at the end of the study, and is used if there are important baseline characteristics known to affect the outcome of treatment. The disadvantage is that it may be difficult to recruit sufficient patients to all the categories, so delaying the study. **Minimization** is a technique that is particularly useful if the study has a small number of patients. The first patient is allocated randomly, and then the second and subsequent patients are allocated using a weighted randomization. The weighting is adjusted in each patient to increase the chance that the patient is allocated to a treatment group that would 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.

Randomization cannot ensure that the confounding variables are equally distributed across the groups as it is still possible by chance for the groups to end up unequal, for example all the males being allocated to one group. It is common to use statistical tests to check whether the groups are similar in ages, weights, etc. after the study has been completed. These tests will detect only major differences, and 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 and/or patient are unaware of their treatment group; if both are unaware of the treatment, this is called a **double-blind trial**. Masking is important, as the response to treatment is often considerably altered by expectation, either by the patient or by the investigator. If the patients knew they were receiving the placebo, they would not expect to improve, whereas 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 it is not possible, for example 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 (Table 4). A poorly conducted study is unethical (see *Anaesthesia and intensive care medicine,* 2006; **7:** 5–9).

New drugs undergo a series of clinical studies in order to be given a product licence (Table 5).

Limitations of crossover trials

- **Period effects**: there should not be any significant change with time in the condition of the disease during the study period. If the disease significantly worsened or improved between the first and second treatments, the two treatments would not be studied under similar conditions. For example, transient diseases such as the common cold cannot be studied using a crossover design. The order of the two treatments under investigation is usually varied between the patients to avoid bias from period effects
- **Treatment–period interactions:** one of the treatments may work differently if given in one of the study periods. The treatment may work more effectively earlier or later in the disease process, or its effects may be modified by the other treatment
- **Carryover 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 are not being measured
- The data can be tested statistically for the presence of each of these effects after the study has been completed, but these tests will detect only major effects. It is better to ensure that a crossover trial is the appropriate design and is well conducted

Table 3

Power analysis

An essential part of the design of any clinical trial is a power analysis, which is a statistical technique to estimate the number of patients required to reduce the risk of a type II error (see *Anaesthesia and intensive care medicine,* 2006; **7:** 135–42) to an acceptable value. The actual calculations in a power analysis depend on the type of data, that is, categorical, ordinal or continuously variable data. The probability of a type II error is denoted as β, and should be 20% or less, and the power of the study is defined as $(1-\beta)$. For example, if β has been chosen to be 10%,

Guides to good research practice: trial quality

- **Good clinical practice:** the Association of the British Pharmaceutical Industry, the General Medical Council, the Department of Health, the Medical Research Council, the European Parliament and the British Medical Association have all published guidance on the conduct of clinical trials
- **Governance:** the NHS document *Research Governance Framework* outlines the standards expected from NHS and university staff conducting research within the NHS
- **National research register:** the NHS maintains a database of all UK research of interest to the NHS. There are several other databases of ongoing clinical trials
- **The Consolidated Standards of Reporting Trials (CONSORT):** documents produced by this organization are guides for authors to write papers such that readers can judge the internal and external validity of the trial. Internal validity is the quality of the trial, and external validity is the degree to which the conclusions can be extended to other patients in different contexts

Table 4

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 information 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 and β.

Clinical studies of new drugs

- **Phase 1 studies:** healthy volunteers are given low doses to study the pharmacokinetics of the drug
- **Phase 2 studies:** observational studies in patients with the disease for which the drug is thought to be useful. The effects of a range of doses are studied with further pharmacokinetic analyses
- **Phase 3 studies:** formal randomized controlled blinded clinical trials in patients. The control groups are usually given a placebo, as the company has to demonstrate only efficacy and safety, not superiority over existing treatments. These studies will be used by the medicines licensing authorities to assess whether the drug will be given a product licence
- **Phase 4 studies:** carried out after the drug has a product licence, and are initiated by doctors or the company to compare the drug with other treatments or to extend the product licence. It is rare for new drugs released onto the market to have been studied in children or during pregnancy
- **Post-marketing surveillance:** most countries have some method of reporting and collating information on adverse effects of new drugs. Many adverse effects are sufficiently rare not to have been detected in phase 2 and phase 3 studies

Table 5

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 better 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 that the treatment may be ineffective or even harmful.

Confidence intervals are particularly useful to indicate the possible true incidence of uncommon complications after a trial in which there were few adverse events. Many authors report that there were no complications in their study and recommend their technique as safe, without giving an estimate of what the true incidence could be (Table 6).

The number needed to treat (NNT) (the reciprocal of the absolute risk reduction of a treatment) is another useful statistic for reporting clinical trials (Table 7).

Diagnostic tests

It is very rare for a single diagnostic test to clearly and reliably distinguish 'normals' from those with the condition. Nearly always, there is an overlap of test results that could occur in either 'patients' or 'normals'. The properties of a diagnostic test are commonly given by the sensitivity and specificity (Table 8), but knowledge of these two values is not useful in practice.

We are usually more interested in the probability of the disease being present if the test is positive; this is called the positive predictive value (PPV). (The sensitivity is the probability of a positive result if the disease is present and is related to the PPV

Upper 95% incidence for the true population incidence of a complication after a trial with *n* patients

Number of complications reported in a study of *n* patients	Upper 95% confidence limit for the true incidence of the complication in the population
0	3/*n*
1	5/*n*
2	7/*n*
3	9/*n*
4	10/*n*

For example, if nothing goes wrong in a study of 50 patients, then the upper 95% confidence limit for the true incidence of adverse events is 6%, normally an unacceptable incidence of serious complications. A technique could not be recommended as safe from a small trial such as this.

Table 6

Incidence of stroke after 5 years in treated and untreated hypertensive patients

	Untreated hypertension; absolute risk of stroke	Treated hypertension; absolute risk of stroke	Relative risk	Relative risk reduction	NNT
Mild hypertension	0.015	0.009	0.6	0.4	167
Moderate hypertension	0.2	0.12	0.6	0.4	13

The number needed to treat (NNT) is the number of patients it is necessary to treat to prevent one stroke in that group. In both treated groups the incidence of stroke is reduced to 60% of that in the control group, but the baseline incidence of stroke in the patients with mild hypertension is much less. Because so few patients with mild hypertension have strokes, many more must be treated unnecessarily to prevent one stroke than in patients with moderate hypertension. In these patients, the cost and side effects of treatment would be relatively more important.

Table 7

by Bayes theorem.) The PPV depends on both the qualities of the test and the prevalence of the disease. The PPV usually decreases as the disease becomes rarer because the number of 'normals' with abnormal tests far exceeds the number of 'patients' with abnormal tests. For example, the PPV of most of the single tests to predict difficult intubation is rarely greater than 20%, and is often less than 10%.[5] The prevalence of difficult intubation is actually very low, and the normal patients predicted to be difficult but who are actually easy greatly outnumber those with abnormal predictors who really are difficult to intubate.

Odds ratios are sometimes used instead of probabilities (Table 8). A high likelihood ratio means that the test is a good one, but the usefulness of a positive test result again depends on the prevalence of the disease.

Statistical applications in quality assurance

During the Second World War, statistical methods were developed to monitor the production of shells on factory assembly lines. This work developed into the theory of statistical process control, now widely used in industry to monitor industrial processes. These techniques have considerable applications in medicine.

All clinicians sometimes fail at practical procedures and have patients with complications, adverse effects and poor outcomes. Sometimes these adverse events come in clusters, and can unfairly give rise to suspicion of poor performance. Statistics can help distinguish the good clinician who has had a run of 'bad luck' from a poor clinician who has intermittent runs of 'good luck'.

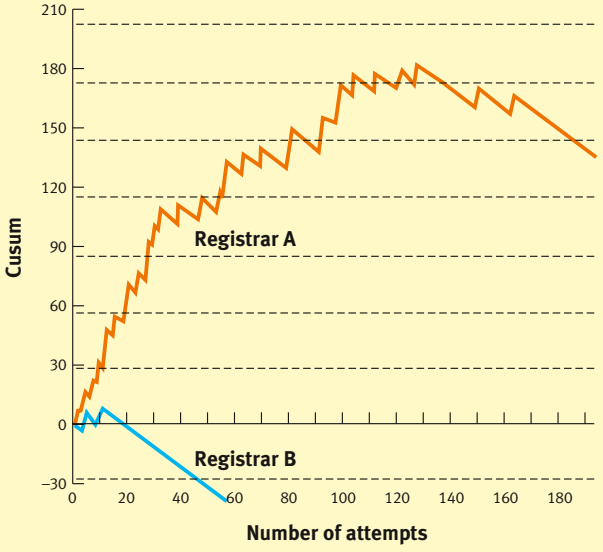
There are several methods available that measure the frequency of a binary outcome, that is, success or failure at a procedure or the presence or absence of a complication. Cusum analysis (*cu*mulative *sum*mation) is one of the better-known methods. There are four variables that must be defined in advance: the acceptable or expected failure rate (i.e. the norm); the unacceptable failure rate, at which some action should be

Diagnostic tests

Sensitivity = (number of patients with the disease AND a positive test result)/(number of patients with the disease)
Specificity = (number of patients without the disease AND a negative test result)/(number of patients without the disease)
Positive predictive value = (number of patients with a positive test result AND the disease)/(number of patients with a positive test result)
Likelihood ratio = probability (positive test with disease present)/probability (positive test without disease)
Pre-test odds = prevalence/(1–prevalence)
Post-test odds = (pre-test odds) × (likelihood ratio)

Table 8

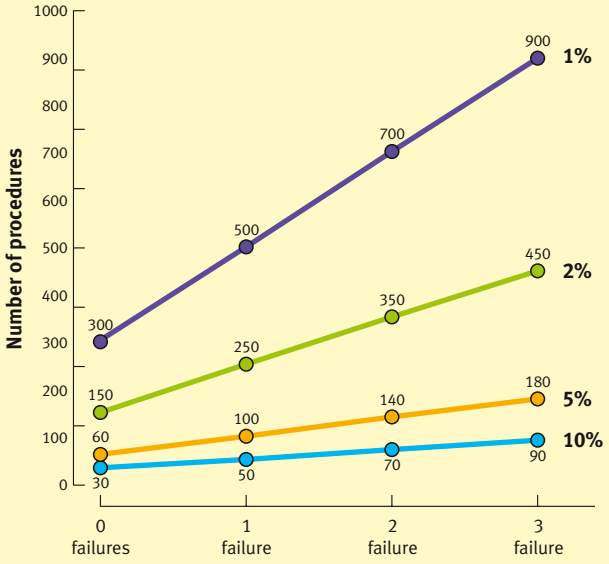
Cusum plots of two trainee anaesthetists learning epidural anaesthesia in obstetric practice



The *cu*mulative *sum*mation (cusum) starts at 0, and, with successive attempts, increments are added to the cusum if there is a failure and subtracted if there is success. The horizontal lines are the boundary lines for statistical significance (see text). Registrar B has a couple of early failures, and then a steady run of successes. After 49 epidurals a boundary line is crossed from above, and the trainee can be signed off as competent. Registrar A, in contrast, has a prolonged series of difficulties (undetected by the trainers!) for 100 attempts before the cusum flattens out. A decreasing cusum (from attempt 140 to attempt 185) that subsequently crosses a boundary below means the trainee can now be signed off as competent

Figure 1

Number of practical procedures to obtain 95% statistical confidence in competence after various numbers of failures with different maximum acceptable failure rates



If a trainee has one failure/complication, and the maximum acceptable rate is 10%, then the trainee needs to have done 50 procedures with only one failure in order for the trainer to be 95% confident that the true complication rate is 10% or less. If the maximum acceptable rate is 1%, e.g. dural puncture in obstetric anaesthetic practice, the trainee has to do 500 epidurals with only one dural puncture in order for the trainer to be confident that the true rate is less than 1%

Figure 2

taken (e.g. further training); and the α and β errors (see *Anaesthesia and intensive care medicine,* 2006; **7:** 135–42). From these four values, an increment, *s*, and boundary values to the cusum can be calculated. The cusum starts at 0 and then, with successive procedures, *s* is subtracted from the previous value of the cusum in the event of success, and (1 − s) is added in the event of failure. Thus, a falling trend in the numerical value of the cusum is associated with successes, and a rising trend is associated with failures. The calculated boundary values distinguish when there is statistically significant deviation from the acceptable standard and not just random variability. If the cusum is increasing and exceeds the boundary value, then the true failure rate is at, or higher than, the unacceptable failure rate and action needs to be taken; if the cusum is falling and falls below the boundary value, then the clinician's true success rate is not statistically different from the acceptable rate, that is, the clinician can be certified as competent (Figure 1).

The data in Figure 1 can also be transformed to provide some information on the competence of trainees if there are rare but serious side effects of treatment (Figure 2).

For example, if a trainer wants to be sure that the trainee's complication rate is 10% or less, the trainee has to do 30 procedures without a complication in order to be 95% confident the true complication rate is 10% or less. If there is one complication, then the number of procedures increases to 50. For rarer complications, for example dural puncture in obstetric anaesthesia in which the desirable rate would be 1%, then trainees would need to do 300 epidurals without a single dural puncture before they could be certified as competent. If they do one dural puncture, then trainees must do a total of 500 epidurals with only one dural puncture if the trainer is to be confident their true rate is 1% or less. This quantity of experience cannot realistically be provided in training programmes. We could use a lower standard of proof (i.e. less than 95% confidence in our decisions). The other options are either to accept a lower degree of competence for the 'trained beginner'; or to acknowledge that there are rare complications of practical procedures and we cannot know whether the trainees are safe and competent when they finish training. Some method of monitoring the 'trained' practitioner is needed, and that, for some, further training may be needed after they are 'trained'.

These techniques are applicable to monitoring the quality of medical practice of both individuals and teams, and are particularly useful for monitoring new or infrequent procedures. The important choice is to set the acceptable and unacceptable failure rates, and these could be obtained from a variety of sources, for example standards set by national bodies, previously published papers or from a local consensus. ◆

REFERENCES

1 Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 2005; **294:** 218–28.

2 Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *BMJ* 2003; **326:** 1167–70.

3 Fister K. At the frontier of biomedical publication: Chicago 2005. *BMJ* 2005; **331:** 838–40.

4 Jüni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ* 2001; **323:** 42–6.

5 Yentis SM. Predicting difficult intubation: worthwhile exercise or pointless ritual? *Anaesthesia* 2002; **57:** 105–9.

FURTHER READING

Also available at: http://www.consort-statement.org/

Also available at: http://www.dh.gov.uk/en/publicationsandstatistics/publications/publicationspolicyandguidance/dh_4108962