LOGIC AND INFERENCE IN CLINICAL TRIALS

D.W. VERE

Department of Pharmacology and Therapeutics, The London Hospital Medical College, Whitechapel, London E1

1. Trial design

Treatments are given in a wide variety of spontaneous circumstances, but a *trial* takes place when treatment is taken in events which are partly constrained or patterned, a deliberate constraint made in ways which allow *comparisons* to be drawn without environmental *bias* as far as this can be achieved. The object is still to treat patients, but to do it in such a way that reliable measurement and comparison is possible. The measurement may be of the simplest kind, 'A is better than B', or something very complex indeed. But without constraint, upon the events of which treatment is a part, it is seldom possible to interpret even the simplest measurements or to draw the simplest conclusions from what is observed. All trials involve some kind of *comparison*; measures made without this are simply observations. The constraints imposed upon spontaneous events constitute the experimental aspect of a trial.

A *clinical trial* is one which involves patients (there can be healthy volunteer trials), but a *controlled clinical trial* is one in which unforeseeable bias errors are dealt with in a particular way. *Errors* arise in several ways, each of which can be minimised by rigorous trial design (Table 1). As the table shows, there are many kinds of error variation, but all can be minimised using one, or both, of two methods. Sources of systematic variation can be dealt with by anticipating them and removing them wherever possible. For example, a trial of remedies against thyrotoxicosis would be certain to yield a 'result' if treatment A was given to women during the winter, and treatment B to men during the summer. The 'result' would *confound* variation due to a genuine difference between the treatments with bias, both in 'space' (between groups) and in time (seasonal weather differences). To take a real example, a trial was made in severe angina by stopping all previous therapy suddenly, rejecting subjects who developed crescendo angina, giving the rest a β -adrenoceptor blocking drug after their angina frequency had exceeded n attacks per week, and then attributing the 'benefit' which followed to the drug. This 'trial' (which came before a regulatory authority!), (a) was uncontrolled, (b) lacked all inter-therapy comparisons, being comparative only in time, (c) biased the selec-

tion of subjects and (d) confounded spontaneous improvement with drug effects. It was, of course, also unethical.

In these two examples, it would not be difficult to remove bias by anticipating its sources (weather, gender, changes in angina following abrupt treatment withdrawal) and finding ways to remove their effects (e.g. synchronous comparisons between treatments, waiting for a stable response before making measurements) or to make them equal (e.g. allocating each treatment to equal numbers of men and women, using a placebo group).

Having done this sort of thing to the best of the researcher's ability, it will be recognised that much variation cannot be anticipated. The only way to deal with that is to allocate patients to treatments in such a way that every patient's measured response runs the same risk of incorporating an error element. The ways to do this are outside the scope of this article, but include random allocation schemes, often within some sort of stratified design, the minimisation procedure (Taves, 1974), part ordered, part random schemes such as Latin squares and so on (Cochran & Cox, 1957; Li, 1964). Schemes which tend to induce bias include alternate allocation, schemes which depend on initials, on birth date, or on hospital number. The result of proper allocation should be that error variation, which will continue to occur, is reduced to random components alone in so far as they affect the treatment comparisons. The essential element of randomisation is its impersonal, objective character, which frees treatment allocation from any limit of personal choice however subtly it may intrude.

A final element in the logic of trial design is that whether treatments are rotated so that each patient received more than one (within subject allocations), or only one (between subject allocation) or some combination of these, the essential thing is *repetition* of treatment exposure. Repetition of treatment makes it possible to measure probabilities (e.g. 'the chance of r positive events out of n tries'). Repeated measurements within a defined, replicable field are the essence of scientific work, and such collected numerical sets are called *statistics*. Single observations, or patterns of these, however interesting they

Table 1 The kinds of error which arise in trial comparisons

Type of error variation	Method used to balance out error variation
A. Systematic errors, 'bias' <ul style="list-style-type: none"> (i) in 'space' e.g. between trial groups (ii) in time – time dependent drift (a) in the environment (b) within the trial subjects (c) within the observer(s) 	Anticipate these where possible and remove all sources of variation that can be eliminated. If not, proceed as for B.
B. Random, or unsystematic error: <ul style="list-style-type: none"> (a) due to random 'noise' events in 'space' (b) due to random 'noise' events in time (c) individual differences between patients 	Allocate treatments according to an appropriate plan e.g. random allocation Latin square etc.

may be, cannot be subjected to statistical analysis. They may be studied legally, archaeologically or clinically, but they cannot be expressed in a generalisable form of the type 'if Z happened, then in other situations the chance of X will probably be Y'. Hence they lack even local predictive power. It must be admitted that this process, moving as it always must from the particular to the general, is *induction* which is usually a weak, logical mode. But it is the best that we can do. However, it explains the scientist's reluctance to accept unrepeatable or inconsistent work. Even so the aim of clinical trials is to collect knowledge in such an orderly way that it helps to change genuine doubt into confidence, in the minds of informed people. It provides, literally, information. This is to say that it takes a chaotic, noisy region and puts in its place *formation*, or order. The philosophy of induction and deduction in trial designs is discussed at length by Barnett (1975).

Because *crossover designs* are so important in clinical trials it seems important to digress briefly on their logic. The attraction of these designs is that the variance attributable to individual differences (Table 1) is shared out with scrupulous equality among the treatment blocks, or so it seems in the model. This reduction and confinement of a large, unpredictable, and seemingly irrelevant piece of variance would have major advantages. However, they often melt away in practice. For one thing, time-associated variation often steps in to take the place of that tiresome inter-individual variance; responses to therapy differ, are unpredictable and may be prolonged. Hence 'washover' effects occur. The environment of the trial may vary in time. This matters less if it can be done without ambiguity in a crossover design where treatment effects confound with others in time. Also, may it not be the case that it is inter-individual differences which are the most interesting outcome of a trial? It is not always realised that within-subject comparison alters the logic of statistical tests, so that as with the 'related *t*-test', so with crossover designs

one needs a 'repeated measures within subjects' version of the analysis of variance design. All these points and more are discussed by Hills & Armitage (1979).

To summarise the argument so far, it has been shown that the aim of controlled trials is to induce a body of ordered information about treatments. This is done by providing therapy in a milieu of *repetition*, of *induced replication* of observations, of error reduced to its *random component* by the control of bias, and of *comparison*. Such a design will generate statistics. It will always be capable of analysis. It will provide information. Readers will notice that nothing has been said so far about statistical tests, 'P values', distributions or the like, because these are irrelevant to the essentials of good trial *design*. A well designed trial can always be analysed. A badly designed trial will tend to mislead, however much mathematical ingenuity (or deviousness) is used in attempts to retrieve its lost cause. As a conceptual process, trial analysis coheres with trial design, hence incongruity of data is congruent with loss of power in the final analysis.

Even so, the logic is readily overstated. No one can distinguish between *true* randomness (a series of discrete events where the chance of getting any one is entirely independent of any other) and patterned events which *simulate* a random series. This is because the only way we have to discriminate between these two is a statistical test unless the pattern of ordered events is so obvious that we suspect it directly. The outcome of such test is bound to be an expression of probability; a test cannot say that a series of events is random, only that it cannot be distinguished from a random series. And the smaller the sample the weaker the power of the test to discriminate between random and structured series of events. But more will be needed about that problem in the next section. The fact remains that objective randomisation in an experimental design has very well established and recognisable properties; the fact

that the random element is introduced *through the design* mitigates the logical problem of apparent randomness. That is a problem of interpreting the natural world, not of the use of random designs in experiments.

2. Trial analysis

There are four main inferential approaches to analysis, each with its proper place and each open to controversy. There is no *uniformly* satisfactory way to analyse biological data; despite this fact many research workers force data into one of these approaches, usually into a classical hypothesis test, even when the necessary logical basis is quite absent.

These four inferential emphases are:

- (i) the classical hypothesis test
- (ii) point estimation, with confidence intervals
- (iii) Bayesian inference
- (iv) pure likelihood methods ('pure' because elements of likelihood inference are an essential part of all the other methods).

These four inferential modes* are not logical alternatives; they each answer a different question. (What is confusing is that the mathematics appropriate to one can at times be pressed into the service of another). The question to which the four modes provide answers are these:

- (i) Were the samples drawn from the same (or from different) parent distributions? This is answered by an hypothesis test.
- (ii) Given considerable knowledge about one sample (a) what is the range within which X% of parameter estimates will fall from other samples, and (b) what is the best estimate of the control value parameter based upon a sample?

These questions are answered by point estimates with confidence intervals.

- (iii) If there exists a body of information about a process from the past, which is exposed in the form of a probability, how can new information be incorporated with it, and what revisions does it provide of the parameter estimates for the combined samples? This question is answered by Bayesian logic. (If the previous information is in the form of data, the problem is then one of data combination).
- (iv) If all discussion of distributions seems artificial, and if the null hypothesis test seems absurd, what are the best comparisons between the population parameters based solely upon information provided by the samples? If I treat the

central values as random variates, what is their most likely ratio? These questions are answered by likelihood methods, with or without Bayesian additions.

These arguments will be shown in more detail now by detailed discussion of each method in turn.

(i) Classical hypothesis tests

The basic notions are to imagine that samples are drawn without bias from a distribution or at least that there are null expectations for differences between their values. If the parameters of the samples are used as *estimates* of the (otherwise unknown) parameters of the parent distribution the test is a 'parametric test'. There are also non-parametric, or even distribution free, tests of null expectations, though the samples are sometimes assumed to be derived from a normal distribution as in the χ^2 test. Sometimes a parameter of the parent distribution is *believed* to have some necessary value, e.g. the mean of deviations from a mean is assumed to be zero. Sometimes a structured hypothesis (H_1) is tested instead of the null hypothesis (H_0). But whatever the test, the common idea is to compare the differences between the samples with the differences which would be expected from a plausible theoretical *model*.

The next step is to use the sample estimates (or sometimes the expected values) of the parent distribution parameters to calculate areas under the (imagined) parent distribution curve.

The next step is to erect *a priori* 'rejection regions' under the distribution, the size of which depends solely on arbitrary judgments by the investigator about the risks he is willing to accept of being wrong were he to assert falsely that two samples were drawn from the same parent distribution when in fact they were not (rejecting a true null hypothesis, H_0 , and making a Type I error; the α probability of error), or were he to assert falsely that two samples were drawn at random from the same distribution when in fact they were not (accepting a false null hypothesis, and making a Type II error: the β probability of error).

The final step is to compare two actual samples, or one real and one theoretical measure, with the areas under the curve, and to find the α and β probabilities which are associable with the decision which the investigator made. Usually, the logic of this step is inverted, the investigator making a decision which depends upon whether α exceeds a presupposed limit or not, often without regard for β . The pattern of reasoning is shown in Figure 1.

This chain of 'reasoning' has become intolerably slipshod in routine use; most investigators seem unaware of the glaring inconsistencies in the routine application of statistical tests. They may be listed like this:

- (a) All tests should be made against a random

*Decision theory, which includes elements of (iii) and (iv) above, will not be discussed in this paper. It will be the subject of a later chapter in this series.

Figure 1 The general pattern of inference from a set of results (R_i) through a hypothesis (H_i), indicating points to which particular attention should be directed when considering each component of reasoning.

expectation. There is no reason to expect randomness in many clinical situations, both with regard to how samples are drawn, or to how the imagined 'parent' distributions are compared in circumstances of human disease. Tests for non-randomness are not strong with small samples, and in practice are seldom used. The only aspect of randomness which can reasonably be assured is the allocation of treatments. The situation is wide open to bias despite the reasonable claim that tests are robust against departures from the normal expectation. Bias can *seem* very 'normal' (in the clinical sense of that word) (Leff, 1973).

(b) The information which we possess is entirely in the samples, in the majority of cases. When this is so, inference about their 'parent' distribution is necessarily fragile. But this is *doubly* so when we infer from the sample to an imagined 'parent', and then back from it to other samples (i.e. the process of Figure 1 applied sequentially to successive samples).

(c) If we have real data, why do we concern ourselves with other imaginary possibilities, such as all the possible values *less* than the one we actually found? (Sometimes they are possible only in statistical models, not in the biological world. For example, if height is normally distributed and the mean is 70 inches, do I need to consider all possible lesser heights under a Gaussian model, including the vanishingly small but mathematically real possibility of people one inch high?)

(d) Why should one be concerned with an H_0 which cannot possibly be true? If two groups of patients received two quite different treatments, the chance of H_0 being true is literally vanishingly small. What we

are really deciding in a statistical test is not whether H_0 is true or false, but (knowing it to be false) whether we have enough data to show the falsehood. It is more a matter of having enough evidence to catch the thief, rather than believing that the thief was, after all, an honest man!

(e) Gödel's theorem shows that no logical system can be self-authenticating (Körner, 1960). Though biostatistics is a very far cry indeed from the precise auto-denumeration of a formal Gödel proof, the logic of it (if not the content) is similarly self-reflexive, and it may be suspected even more open to doubt than the formally 'undecidable' Gödel question. In short, the more formal falsehood necessarily includes the less.

(f) Most statistical tests are made, in practice, without reference to the β probability. Negative results quoted from such tests are virtually meaningless, unless confidence intervals are given.

Despite these problems, classical statistical tests will continue to play a considerable part in analysing clinical experiments wherever there is reasonable knowledge about the likely parent distribution, and where meticulous attention has been paid to sampling, treatment allocation, and other important aspects of experimental design.

(ii) Point estimates and confidence intervals

These are much neglected; they avoid many of the problems listed for classical hypothesis tests where they can be applied appropriately. The main advantages are that, because they pay no regard to the distribution of samples other than that of primary interest they involve no calculation of rejection

regions; in addition, they allow a range of possible values to be expressed as the 'confidence intervals'. However, the obvious disadvantage of this is that in the very nature of the process they do not permit discussion of other samples quantitatively *in relation to* the first, which is something we want to do whenever our work is comparative—a frequent situation in clinical trials. It is possible to decide that the mean of a second sample lies within boundaries which represent X% of possible parameter values, and no more, which is scarcely informative in many clinical trials.

(iii) Bayesian inference

The enormous clinical attraction is that Bayes' theorem provides a bridge across time, making it possible to compare a succession of similar trials (Levine, 1979) even when the only information they yield is in the form of a rough probability estimate. But the difficulty is whether it is always logical so to do in many of the situations where the theorem has been applied (Feinstein, 1977). As Edwards (1972) points out, the essential detail to remember is that Bayes' theorem is a purely mathematical operation which makes it possible to calculate the probability of an hypothesis being true, given certain results, from the prior probability of that hypothesis and the chance of getting those results under it. In other words it treats the hypothesis as if it were part of a probabilistic random variable. Now that may be possible in many situations, but Bayes' theorem is pressed into duty in many others where there can be no possibility that the condition holds. For example, when prior knowledge is vague, some people have replaced it by some *arbitrary* distribution of probabilities.

It seems feasible so to contrive a series of trials that it will be possible to link them through Bayes' theorem, but that will only be useful if the conditions are reproducible, coherent, and under strict control. But even then time dependent bias will still exist, and may easily prejudice results; checks for it will be an essential part of such studies.

(iv) Likelihood methods

Since some clinicians seem unfamiliar with arguments based on likelihood, it seems wise to reiterate the main points in the logic even though they have seldom been used in clinical trial analysis; an extended account is given by Edwards (1972) and summarised succinctly by Barnett (1975). The main points are:

- (a) The data are regarded as the sole solid source of inference. Instead of looking at the probability of getting certain data given an hypothesis, we look at the likelihood of an hypothesis being supported given certain data.
- (b) The data are often treated as frequencies (r 'successes' out of n 'tries').

- (c) Though distributions are assumed, only their ordinates are considered (i.e. the vertical distance from zero to the distribution curve), not their assumed parameters nor their expected values, nor areas under distribution curves and histograms.
- (d) The ordinates, (c), are used to examine the likelihood of the distribution parameter, treating it as a random variable. Usually, two ordinates are compared as *relative* likelihoods, a likelihood ratio.
- (e) Because (c) and (d) are used, there are no significance tests in the orthodox sense, no α and β probabilities, no power functions of the orthodox type.
- (f) Significance of findings is evaluated as 'support units' (G), it being generally accepted that more than 2 units represents significant support for a structured hypothesis, H_1 . The reason for the value 2 is mathematical (Edwards, 1972, page 182). It is related to the normal deviate of 1.96.
- (g) Likelihood, is, in fact, the firm substructure of both classical (Neyman-Pearson) and Bayesian inference—though much is added to it in both—and it is true that all the *information* in a result is carried forward through likelihood reasoning, it is a 'sufficient' statistic.

The attraction of likelihood methods is not only in their secure logic; in many situations, though not all, they are far easier to compute than either classical or Bayesian statistics. For example, consider a situation which is relevant to clinical trial work, a four fold table of the kind usually analysed by Fisher's exact test (which is, by the way, a method based on likelihood reasoning or the hypergeometric distribution). Tables of conditional (Bayesian) values for the predicted number of patients in a trial were given by Bennett & Hsu (1960) and Cassegrande *et al.* (1978) after several years of work using major computing facilities. When we computed a binomial approximation to their results using the University of London computing facilities (Holmes *et al.*, 1982), even this exercise, though more brief, cost several hundred pounds. But almost identical numerical results were produced in one hour on a pocket calculator using the likelihood calculation. Likelihood quanta are simply additive in a very attractive way. For example, for the four fold table with cells labelled a and b in the top row, c and d in the lower row, and $(a + b) = (c + d) = n/2$, $G = n \ln 2 - a \ln (1 + c/a) - b \ln (1 + d/b) - c \ln (1 + a/c) - d \ln (1 + b/d)$. It will be seen from the ratios within the brackets that, whatever the value of n, any G value will be a simple n multiplier of any other, given that the conditional probabilities in the cells remain the same. This is a remarkable simplification when compared with any of the calculations required for the classical test.

Unfortunately complex analysis and comparisons cannot be pursued as yet by likelihood methods. The

main problem with likelihood is how to use the results. People still feel ill at ease with 'support units', having had no experience of relating them to real events so that their meaning can be appreciated. Presumably, most clinical journal editors would still invite likelihood users to take their results elsewhere, probably to a decent burial in the literature of applied statistics. Perhaps the remarkable economies in assumption, work and money which likelihood methods could provide in some cases may slowly grow in the current economic climate. But not all clinical trial data fit comfortably within a likelihood framework.

In summary, of the inferential methods to hand clinical trialists have two (hypothesis test, Bayesian logic) which seem to answer the questions they wish to ask, but which involve several highly dubious leaps of inference which are often unjustified. They have

two others (point estimates, likelihood) which are logically conservative and often impeccable, but which do not answer all the questions which doctors wish to ask. These include, for example, the power of a test when the result is not statistically significant, or the estimates of central value and dispersion of an assumed sampling distribution. Should we be content to ask fewer, more basic questions, or should we continue to delude ourselves that we have found answers to all that we would wish to know? To go back to the start, a well designed study carries conviction through its balance, its symmetry of comparison, its coherence, its completeness, its small variance and its unambiguous message. These qualities of fairness and informativeness transcend all the subtleties of logical frame and statistical testing, indeed they usually render such discussions unnecessary.

References

- BARNETT, V. (1975). *Comparative statistical inference*. London: Wiley.
- BENNETT, B.M. & HSU, P. (1960). On the power function of the exact test for the 2×2 contingency table. *Biometrika*, **47**, 393–398.
- CASSEGRANDE, J.T., PYKE, M.C. & SMITH, P.G. (1978). The power function of the 'exact' test for comparing two binomial distributions. *Appl. Statist.*, **27**, 176–801.
- COCHRAN, W.G. & COX, G.M. (1957). *Experimental designs* (2nd edn). London: Wiley.
- EDWARDS, A.W.F. (1972). *Likelihood*. Cambridge: Cambridge University Press.
- FEINSTEIN, A.R. (1977). The haze of Bayes, the aerial palaces of decision analysis, and the computerised Ouija board. *Clin. Pharmac. Ther.*, **21**, 482–496.
- HILLS, M. & ARMITAGE, P. (1979). The two period cross over clinical trial. *Br. J. clin. Pharmac.*, **8**, 7–20.
- HOLMES, J., STEVENS, A.J. & VERE, D.W. (1982). Anticipating the numbers of subjects in a clinical trial. *Br. J. clin. Pharmac.*, **13**, 274P–275P.
- KÖRNER, S. (1960). *The philosophy of mathematics*, pp. 90–96, especially p. 95. London: University Library, Hutchinson.
- LEFF, J.P. (1973). Influence of selection of patients on results of clinical trials. *Br. med. J.*, **2**, 156–158.
- LEVINE, J. (1979). *Co-ordinating clinical trials in psychopharmacology*. DHEW publication No. (ADM) 79–803, Washington D.C. 20402: U.S. Government Printing Office.
- LI, C.C. (1964). *Introduction to experimental statistics*. London: McGraw-Hill.
- TAVES, D.R. (1974). Minimisation: A new method of assigning patients to treatment and control groups. *Clin. Pharmac. Ther.*, **15**, 443–453.