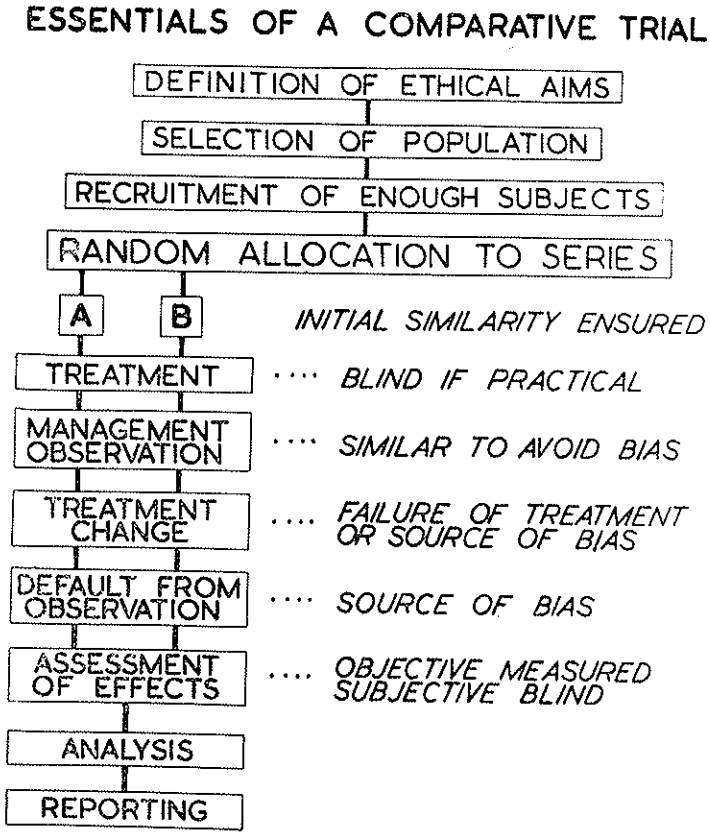
STATISTICAL ASPECTS OF CLINICAL TRIALS

A paper on 'Statistical Aspects of Clinical Trials' was presented by I. Sutherland, M.A., D.Phil., of the Statistical Research Unit, Medical Research Council, London.

DR. SUTHERLAND: Statistics, or perhaps I should say quantitative thinking, permeates clinical trials to such an extent that I am uncertain where this contribution should begin and end—unless it begins with 'aims and ethics' and ends with 'finance'. I propose therefore to illustrate first how statistical concepts enter to a greater or less degree into each stage of the planning and execution of what I may call the classical clinical trial, and then to describe some of the variants and developments which have been increasingly used during the last few years.

The clinical trial has developed from two basic principles—that the effects of one treatment can be evaluated only by comparing them with the effects of another treatment, and that, because human beings differ in their responses, several patients must be observed on each treatment if the comparison is to be adequate and the conclusions convincing. Because the idea of comparison is fundamental, I shall refer to comparative trials rather than controlled trials, as this avoids the misinterpretations



of the word 'controlled' which were referred to by Professor Crofton. The aim of a good clinical trial is simply to make such a comparison, say between an 'old' and a 'new' treatment, used on different series of patients, as precise and as informative as possible.

This diagram shows the various stages, and indicates essential points to be observed, in the course of a comparative trial. I should like to comment upon each of these from the statistical standpoint. The first two stages, which are concerned with specifying the purpose of the trial, and deciding what treatments are to be compared, in what type of individual, and in what type of disease, may seem at first sight to be purely medical, with no place for statistics. I shall refer to these again, however, since statistical requirements later in the trial often have important repercussions on these preliminaries.

The next point, the recruitment of enough subjects, raises the question which the statistician is most frequently asked in connection with any clinical trial, and to which it is almost impossible to give a simple answer. The only response to the enquiry 'How many patients shall we need?' is another question: 'What would you regard as the minimum benefit of the new treatment over the old which would justify its widespread introduction?' If the clinician can be pinned down to some numerical indication of what he would regard as a result of practical importance, the statistician has some hope of ensuring that this result, if it is attained (or exceeded), will also be of statistical significance (that is, unlikely to be due to chance). The trouble is that neither the clinician nor the statistician can really answer their respective questions properly until the results of the trial are known, and by then the answers are unimportant. The statistician, however, if he errs, must endeavour to do so on the side of larger numbers. A trial which does not give a decisive answer is often worse than no trial at all. Impressions of the value or otherwise of a new treatment will inevitably be formed from the equivocal results, and an ethical situation may well be created in which any further comparative trial involving the treatment is quite impracticable. As far as possible, one must avoid the situation in which either a poor treatment gains currency, or a good one is condemned, on insufficient evidence; indeed, it is for this very reason that the trial is undertaken.

Next comes the crux in the design of a clinical trial—the allocation of patients to the treatment series. Since Professor Crofton has dealt with this aspect very thoroughly, I will remind you only that unless the treatments are used on similar series of patients with similar disease, managed in all other respects in the same way, the comparison will be less precise, and probably very much less informative, than we wish it to be. In particular, although the results may in fact be quite correct, they may be largely unacceptable to other workers because of the possible alternative explanations for the findings. It is for these reasons that, save in exceptional circumstances, the old and new treatments must be studied concurrently, and the choice of which patient is to receive which treatment must not be left to individual preference —requirements which are now accepted by clinicians with varying degrees of enthusiasm.

The effect of a properly handled allocation procedure is to provide two series of patients, ready to start treatment, which are similar in all respects, within chance

limits. It is desirable to compare the series, at the start, in characteristics such as age, or extent of disease, which may affect the response to treatment; there may occasionally be a large chance difference between the series which will have to be taken into account in assessing the results. As Professor Crofton pointed out, disparity of this kind can often be avoided at the allocation stage—for example, by having separate balancing allocation lists for different types of disease, to ensure equal representation of each type in the two treatment series.

Having obtained two similar series, which are to differ only in their treatment, no further differences must be allowed to enter the comparison. The management of the patients, any ancillary treatments which are given and the methods by which the results are assessed, must all be the same for patients in the two series. A detailed routine for the management and observation of patients, both before and during treatment, is usually laid down in advance in a 'protocol' for the trial. Although this degree of specification is regarded as irksome by some clinicians, it is of considerable assistance in avoiding bias in the handling of patients, as well as ensuring the necessary observations on their progress.

Another valuable safeguard against bias, where it is practical, is not to let the patients know which of the treatments they are receiving, by supplying them in units of similar size, colour, appearance and taste—the trial then being described as 'blind'. It is of equal value to keep the clinician also in ignorance of the treatment, this representing a 'double-blind' trial. I once took part myself in what I suppose was a 'triple-blind' trial, and found myself analysing the results in ignorance of whether 'A' or 'B' was the placebo. The exercise was interesting, but not unduly exacting, since the results were practically the same in the two groups. In many trials, however, the identity of the treatments—or at least the fact of a difference—cannot be hidden; one treatment may have to be injected, for example, and the other given by mouth.

It is important to realise that blind treatment does not in itself provide a complete safeguard against differences in management. During the Second World War, the U.S. Navy was anxious to determine the possible value of a new anti-seasickness drug (Tucker, 1954[1]). 'It was decided to put two lifeboats to sea in rough weather, and give the drug to the men in one boat, and in the other placebos. Care was taken to have the two boats of the same kind and size, manned by crews of the same size and weight, under the direction of men of similar skill and experience', and, presumably, sent out in the same storm. 'Preliminary results were encouraging, when it was ascertained that the crew of Boat A, given the drug, observed significantly less seasickness than the crew of Boat B, given placebos. Aware of the possible pitfalls of such a study, however, it was decided to repeat the experiment, this time giving placebos to the crew of Boat *A*, and the drug to the crew of Boat *B*. The results were completely contradictory: the crew of Boat A, given placebos, reported significantly less seasickness than did the crew of Boat B, given the drug. The inference was that there must have been more difference between Boats A and B than between the drug and placebo, and so it proved. On careful 'dissection' it was ascertained that a leading air-tank in Boat A had permitted water to enter, lowering the boat in the water so that it rode in rough seas in a different manner than did Boat B.' This source of bias should clearly have been eliminated in the planning

of the enquiry, by allocating the drug to half the crew of each boat, chosen at random, and the placebo to the other half.

In any trial, but particularly in the case of long-term treatment for a protracted disease such as tuberculosis, changes of the prescribed treatment may occur. Such changes may indicate a genuine failure of one of the treatments, perhaps due to a lack of clinical effect, or to excessive toxicity, which makes it essential for the clinician to depart from the protocol in the interests of the patient, but they may also reflect a lack of faith in one of the treatments, which may not really be justified. Substantial losses from the latter cause may disturb the similarity of the residual series of patients, and consequently bias the assessment, or even make it impossible to draw valid conclusions. The same applies to losses from observation, whether these are complete, the patient refusing to co-operate further, or partial, when necessary observations on the progress of the patients have been missed. Both sources of bias are less potent if treatment has been blind. But the risk emphasises the general principle that, once allocated to treatment, every patient must be accounted for in the results, and changes of treatment or losses from observation kept to the unavoidable minimum. This is another example of the importance of careful planning at the outset. When considering aims and ethics, the nature, dosage, and duration of treatment, as well as the types of patients and disease to be included, thought should be given to the need for as complete a series of observations as possible. This should not merely ensure a successful trial, but is likely also to result in the comparison of more practical methods of treatment.

Whenever possible, objective, and preferably measured, assessments of the progress of patients should be used. It is, however, probable that these will have to be supplemented by subjective assessments. For a valid comparison between the treatments, these must be made in ignorance of the treatment. Sometimes the patient is asked for subjective reports on the effects of the treatment. The results were recently reported of a trial of a reputed anti-smoking drug (lobeline) (*Manchester Guardian*, 1958[2]). By the end of a fortnight's course of four tablets a day, one-quarter of the sixty volunteers in the trial had stopped or drastically reduced their smoking and rather more than half were smoking less, the remainder reporting no effect. The volunteers, however, were unaware that only one-third of them were receiving lobeline, the others being given either copper sulphate or a placebo. Moreover, success or failure to stop smoking was not affected by the type of tablet. Clearly one could have placed no reliance on the results if each volunteer had known in advance what treatment he was being given; as he did not, the virtues of the anti-smoking drug were reliably shown to be largely psychological.

A clinician, too, may have to make subjective assessments. If the identity of the treatment is known to the clinician responsible for managing the patient, another clinician should make the assessment; it is, for example, standard practice in trials of tuberculosis treatment to use one or more independent observers to assess changes in radiographic appearances, in ignorance of the treatment of the individual patient.

Of analysis I need say little, as this is usually, and thankfully, regarded as the statistician's private domain. Tabulations and statistical appraisal of the results will also inevitably form an integral part of the final report; and despite their logical basis, the preparation of clear tables and a lucid text is a surprisingly difficult art.

It must be emphasised, however, that should the results show a clear-cut superiority of one treatment over another, it is still incumbent upon the investigators not to accept this result without question. Each step of the trial must be reviewed to see whether there could be any difference between the series, other than the difference of treatment, which could account for the results obtained. As Haldane (1927)[3] has put it, 'a good half of most research work consists in an attempt to prove yourself wrong'.

I should like now to describe briefly some of the main variants of the basic pattern which I have just outlined.

1. Although I have been speaking, and shall continue to do so, in terms of therapeutics, there is no difference in principle in a comparison of prophylactic agents—the same precautions have to be taken at each stage of the investigation.

2. A trial may, of course, compare more than two treatment regimes.

3. Ancillary problems may be studied at the same time as the main comparison between treatments, without increasing the numbers in the trial. For example, the effect of differences in dosage rhythm may be studied by putting half the patients (chosen at random) in each treatment series on one of the dosage rhythms, and the other half on the other. The results can then be analysed, either as a comparison of the treatments, or of the dosage rhythms, and there will also be some information on whether the choice of rhythm is more important for one of the treatments than the other.

4. A similar design may be used to investigate the possibility of synergism or antagonism between two drugs.

5. It is usual to think of treatment in terms of specific drug therapy. Trials are increasingly being undertaken in which the main comparison is of other aspects of treatment. I am at the moment concerned with two investigations of this type in the field of tuberculosis. In the one trial, the effects of different durations of the same drug therapy are being compared, and in the other, treatment in sanatorium is being compared with treatment in the home, again using the same drug therapy. As effective drug therapy becomes established for a particular disease, attention is bound to be directed towards these more practical problems of administration.

6. Another important trend is the increasing realisation that after the period of prescribed treatment there should be a period of careful follow-up. The immediate results of treatment may not be maintained, and this can only be discovered by a further period of observation. One example arises with the use of A.C.T.H. and cortisone in rheumatic fever; the decrease in erythrocyte sedimentation rate is more rapid with the hormones than with aspirin, but the differences disappear in the course of a period of follow-up without drugs (Medical Research Council and American Heart Association, 1955[4]). Another example is the transient weight gain in tuberculous patients treated with rest in bed, compared with those allowed to continue at work (Tuberculosis Society of Scotland, 1957[5]). As during the period of prescribed treatment, care must be taken during such a follow-up to maintain an absence of bias in treatment, management, and assessment between the series.

7. The precision of assessment can sometimes be enhanced by pairing patients who are similar in initial disease-state and in other factors likely to affect the response to treatment, assigning one of the pair at random to each of the treatments. The

relative progress of the members of each pair then provides an assessment of the benefits of the one treatment over the other. If pairing is not possible at the allocation stage, some form of later matching of patients on their pre-treatment condition is an acceptable alternative. This method has been used to compare the response of tuberculous patients in Britain and in Uganda to similar drug therapy (Fox, *et al.*, 1956[6]).

8. The precision of assessment will be further enhanced if it is possible to apply the two treatments serially, or simultaneously, to the same patient, and to assess their effects separately. This is a logical extension of pairing or matching similar patients—the aim of both approaches being to compare the two treatments on closely similar clinical material. The circumstances in which two separate treatments can be given to the same patient are unfortunately limited; many treatments are systemic, and time, or the first treatment, may cure the disease before the second treatment can be tried. But this valuable approach can be used in the treatment of symptoms (as in asthma), or with local treatments (say for skin diseases), or in studies of local reactions (such as tuberculin tests).

9. Finally, a relatively new technique which appears ideally suited to the clinical trial is the sequential approach, in which each result is scrutinised as it is obtained, and a decision then made on the accumulated results whether to continue the trial, or to accept the treatments, either as equivalent, or as different, in their effects, and to stop the trial (Armitage, 1954[7]). The intake of suitable cases into a trial often proceeds quite slowly, and the prospect of being able to call a halt as soon as a definite result has been obtained is attractive, especially as the average sample size required is less for the sequential than for the corresponding classical procedure. There is unfortunately one main difficulty. The interval between the start of treatment and the assessment of the results for each patient may be much longer than the average interval between the intake of successive patients. If this is so, an unnecessarily large number of patients will have been admitted to the trial by the time it is possible to stop the intake. This may sound like an academic point, but with many diseases, such as tuberculosis, it quite effectively precludes the use of sequential methods for determining the size of the investigation. These methods, however, should be used if it is intended to scrutinise the results as they accumulate. Although this approach has been used only occasionally in the medical field so far (Kilpatrick and Oldham, 1954; Newton and Tanner, 1956; Snell and Armitage, 1957[8,9,10]), in part because of this major difficulty, it will undoubtedly find more use as it becomes better known.

References

1 Tucker, W. B., (1954), Transactions of the 13th conference on the chemotherapy of tuberculosis, p. 28, Veterans Administration, Washington.
2 Manchester Guardian, (1958), 13 January.
3 Haldane, J. B. S., (1927), *Possible Worlds*, p. 174, London.
4 Medical Research Council and American Heart Association, (1955), *Brit. Med. J.*, **1**, 555.
5 Tuberculosis Society of Scotland, (1957), *Tubercle, Lond.*, (1957), **38**, 375.
6 Fox, W., Hutton, P. W., Sutherland, I., and Williams, A. W., (1956), *Tubercle, Lond.*, **37**, 435.
7 Armitage, P., (1954), *Quart. J. Med. N.S.*, **23**, 255.
8 Kilpatrick, G. S., and Oldham, P. D., (1954), *Brit. Med. J.*, **2**, 1388.
9 Newton, D. R. L., and Tanner, J. M., (1956), *Brit. Med. J.*, **2**, 1096.
10 Snell, E. S., and Armitage, P., (1957), *Lancet*, **1**, 860.