The evaluation of clinical trials

R.R.P. Jackson

Clinical Operational Research Unit, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK.

Introduction

One of the earliest talking pictures 'Arrowsmith', circa 1931, brought the ethical problem of clinical trials to the attention of the general public. In it, Ronald Colman portrayed a gifted doctor, who after the usual trials and tribulations of general practice in a small American town eventually succumbed to the call of research and went to work in a prestigious laboratory in New York, under his old Professor. After much work, but still by accident, he discovered a wonder drug which appeared to obliterate most known germs (*in vitro* of course). Time passed and eventually bubonic plague broke out in the West Indies. Our medical hero went to see what could be achieved with his serum. His Professor's words – not to forget science for humanity – were ringing in his ears, with instructions to put alternately, one patient on the serum and one not. Although starting in this vein, his wife's death of the plague, back at the home base (apparently, he did not give her his protective serum!), unbalanced our good doctor, who in his hour of grief said, 'Give all of them the serum,' and this cleared the plague.

By current conventions, his initial trial would have been criticized as being not properly randomized, and therefore open to bias. His 'treat everybody' behaviour, despite the evidently excellent results, would have been viewed with caution. Here the problem is presented as a dilemma, the doctor's ethical considerations versus the need for 'scientifically based' trial results.

The real world is seldom so clear cut and evaluating treatments remains difficult and controversial. The problem of evaluation, however, is not one peculiar to medicine but occurs and has been widely discussed in all fields involving human behaviour, such as education, social services and health care, and wherever major decisions are made. In all of these contexts, it is the variability of individual behaviour, and a related factor – often referred to as 'system response' – which make the task difficult. To illustrate the two effects, consider a clinical trial in which many patients are being treated with the same regimen. Considerable variation in clinical response (i.e. individual be-

haviour) between patients is likely to be observed by the medical staff, who cannot help but adapt their own behaviour to it (i.e. system response). The whole process is complicated further by the fact that the evaluation approach adopted can itself influence the outcome: different methods may yield different results (e.g. ultra-sound and computerized axial tomography as ways of measuring tumour burden).

A recent review of evaluation methodology (Jackson, 1975) contains many points of relevance to the problem of clinical trials evaluation. In particular, attention is drawn to the distinction between 'Evaluation Before' the taking of a decision, and 'Evaluation After'. The former asks, on the basis of what is known, whether a particular decision (e.g. whether to employ a particular experimental treatment regimen) should be taken, and it involves the analysis of existing data (e.g. historical clinical trials data) to establish the likely benefits. The latter seeks to establish the consequences of having taken the decision (perhaps through the analysis of a prospectively planned Phase III clinical trial).

Several methods of evaluation are reviewed in the paper and it is concluded that the development of appropriate mathematical models (equations which describe the behaviour of those system elements of interest) and their manipulation, provide a rational and effective means of obtaining insights during both evaluation phases. The approach will be described in more detail in relation to clinical trials evaluation later.

Sir Austin Bradford Hill (1962) was the first to attempt a comprehensive discussion of the philosophy and the procedures of controlled clinical experiments. The sound sense and moderate line of his work sets the tenor of good clinical research. For instance: '... personal observations of a handful of patients, acutely made and accurately recorded by the masters of clinical medicine have been, and will continue to be, fundamental to the progress of medicine. Of that, however statistically minded this age may be or may become, there can be no doubt whatever'. This is tempered by the caution that 'On the other hand, one has to remember that in many respects the reactions of

human beings to most diseases are, under any circumstances, extremely variable. They do not all behave uniformly and decisively. They vary, and that is where the trouble begins. 'What the doctor saw' with one, two or three patients may be both acutely noted and accurately recorded: but what he saw is not necessarily related to what he did. The assumption that it is so related, with a handful of patients, perhaps mostly recovering, perhaps mostly dying, must, not infrequently, give credit where no credit is due, or condemn when condemnation is unjust'.

How, then, are we best to proceed?

One suggested way, which has become the current vogue, is by means of the randomized controlled trial (RCT).

Randomized clinical trial

Randomization is advocated so as to remove personal bias from allocation procedures (Hill, 1952), and at the same time eliminate systematic errors caused by unknown factors (Chatfield, 1975), thus creating a sample space which can confidently form the basis for tests of statistical significance (Cornfield, 1976).

'Controlled' implies the use of *concurrent* patients to be randomized into say two arms of a clinical trial so as to avoid differences in the composition of patient groups and the risk of patterns of disease changing over time. Usually, the 'control group' is given the standard treatment and the other group the experimental treatment being evaluated. A strict protocol for the administration of each treatment will usually be agreed, and blind or double blind (clinician and patient) allocation may be used to further avoid subjective bias.

To summarize, an idealized RCT should (Armitage, 1960) avoid changes in diagnostic and patient management procedures over time; should guard against bias and selectivity when allocating patients to each arm of the trial, and should avoid subjective bias in measuring response variables.

The analysis of all RCT focusses on the application of statistical significance tests to observed treatment differences, and a cautionary discussion of the popular application of such tests will be presented later. The tests do, however, have a bearing on RCT design, since their discrimination is highly dependent on the number of patients in the trial (Mould, 1979). Given that ethical considerations demand little or no suspected treatment difference to justify an RCT, calls for recruitment of large numbers of patients to trials are commonplace. As an example, statisticians at a conference in Oxford (Peto, 1978) were asked to go as far as to sign a declaration supporting the inclusion of many more patients into trials with larger numbers in each arm, for example: 'Stage III breast cancer trials

should be set up annually and recruit at least 1000 girls to each arm'.

Let us reflect that this is all in the cause of achieving statistical significance with that level of confidence which allows the statistician to pronounce, without due regard to the individual patients. Notwithstanding the evidence obtained from many thousands of patients treated similarly over the years, the message every time seems to be 'let us start again'!

The RCT was given a great boost by the MRC espousing its cause, and by the publication of the papers by Peto *et al.* (1976–77) on the design and analysis of clinical trials. This was further carried forward in a British Medical Journal editorial (1977) in which the doctor's faith in the RCT and the associated statistical approach was accepted as the best way forward, with the reminder of several, apparently wrong, decisions made without the benefit of RCTs: 'Let us remember the number of drugs that have been roman candles, making a bright and beautiful flash for a short time and then burning out. And how many new drugs (such as lithium for the treatment of manic depression) which were useful have been discarded or ignored for many years because their flash was not bright enough?'

This is, no doubt, very poetic, but RCTs have themselves produced conflicting results, obtained from apparently similar or identical trials, and have come out in favour of some treatments which have later been found to be of no benefit such as the use of immunotherapy in acute myelogenous leukaemia.

There are, in fact, many situations in which, with the best will in the world, RCTs cannot be used in practice. These are notably where the number of patients being treated is too small for such a trial to be conducted over a reasonable time span; where all standard therapies are known to be ineffective; where treatments are believed to be substantially different in effect and where costs (patient lives, money, lost opportunity) cannot, or will not, be met. Indeed, Sir Douglas Black (1979), while describing the RCT as having become a 'medico-social belief', expresses reservations 'about its over-riding efficacy'. He gives examples where clear cut benefits (such as treatment with insulin) would not require a trial, and he describes situations where to take account of all possible known variations, would require a trial of inordinate length. Nevertheless, he ends, 'In spite of these reservations, I would agree that if a controlled trial is practicable and can produce a result, it is a most valuable contribution to progress; but it seems to me unrealistic to suggest that nothing should be done without a controlled trial and that any issue can be settled for all time, even by an RCT'.

Another BMJ Editorial (1979) begins to show a lessening of total commitment to RCTs. Some of their practical problems and misuses are given, and atten-

tion is drawn to results in the literature of the effect of drug A being greater than B, being greater than C, being greater than A! A most important part of the editorial is the recognition of Cranberg's (1979) argument that the use of retrospective (as opposed to concurrent) controls is now worth considering, since data is collected more objectively and is more easily retrievable than in the past. The BMJ editorial finished with arguably the quote of the decade that: 'the controlled trial has been placed on too high a pedestal and needs to be brought back down to earth'. The question is: How? There is clearly a need for a more realistic and less restrictive scientific evaluation, bearing in mind Cranberg's remarks (1979) '... no one method is best in every case – the choice of method rests on consideration of the options and their consequences'.

The modelling approach

The modelling approach to evaluation consists of considering an environment as it is found; of establishing its important features; of producing an idealization of the relationships between them, and then constructing mathematical equations to describe them. The mathematical equations when calibrated on a given set of data, form the basis for predictions of the future, or evaluations of the past or present.

The first point to notice is that no longer is a strict set of design constraints declared, as for the RCT. It follows that data emerging from randomized controlled trials may themselves be accessible to some form of modelling evaluation, but more significantly, the body of data encompassed by the modelling approach is much larger than that small subset to be found in RCTs. That is not to say that scientific principles need to be discarded. In broadening the clinical trials 'data base' to include retrospective data for example (and situations have been identified (Gehan Freireich, 1974; Fleming, 1983 personal communication), in which the use of historical controls is in fact ethically preferable), care needs to be taken that comparability between patient groups is, as far as possible, maintained, and appropriate methodologies exist (Kay, 1977), to improve reliability in this respect, by taking account of known prognostic information. In the absence of known prognostic factors, modelling results may actually lead to the identification of candidate variables.

Because of the nature of the problem under consideration, or the data available, modelling evaluation may often address slightly different questions to those addressed by the RCT, and may use any of a range of comparison methods (other than the statistical test). Where data is scant (as for example in Phase I or II studies), formal use of the modelling approach may

establish the most pertinent question(s) to be addressed in larger scale prospective studies – which may in turn also be evaluated by means of mathematical models.

With the post-war development of the branch of science known as Operational Research, has come a new theoretical repertoire of techniques which lend themselves to the evaluation of some aspects of clinical trials.

A number of mathematical models have been developed covering various aspects of oncology (Tautu, 1978). However, of those constructed for investigating the results of clinical trials (for example, Fix & Neyman, 1951; Le Cam & Neyman, 1982), disappointingly few seem to have been applied in actual clinical studies. A more recent model (Jackson & Aspden, 1979), however, based on an idealization of patient progress (Figure 1) has been used for the analysis of leukaemia trials at St Bartholomew's Hospital, and it will serve to exemplify the modelling approach. Application of the model leads to a method of comparing trials which yields hypotheses for more formal testing, and using clear graphical forms of presentation (an important aspect of all analytical methods (Altman, 1980; Gore, 1981), it provides sequential monitoring of new trials.

From the basic schematic (Figure 1) the mathematics of the system are written down, calibrated on the results of a particular trial and then used to monitor and compare a new ongoing trial (Jackson *et al.*, 1982). An example of such a comparison is given in Figure 2. This represents just one possible form of analysis, given the model.

The individual graphs correspond to the boxes in the schematic. Each one shows the number of patients in the vertical axis, against time from the start of the new trial along the horizontal. The solid central line on each graph is the prediction of the numbers of patients in the new trial expected to be there at particular times, given patients in both trials behave in the same way. The dotted lines on either side represent the upper and lower limits between which numbers are likely to occur on 90% of occasions, and represent some measure of confidence in the predictions obtained from the mathematical model. From these graphs two very important hypotheses can be drawn. First, by considering the dead non-remitters graph, it is seen that from the very start of the new trial there are more patients dying without remission than would be expected, all asterisks lying on or above the prediction curve. This could be due to the remission proportion being substantially lower than in the previous trial, and/or the time to death for non-remitters being shorter. The second important inference was at that time really no more than speculation, but it was noted that none of the patients lucky enough to remit had relapsed.

Was this then a kill or cure treatment? The detailed,

Figure 1 Schematic of acute myelogenous leukaemia.

up-dated analysis is presented elsewhere (Jackson *et al.*, 1982), and all the above inferences were confirmed with further follow-up.

These types of models have been used to analyse data from a number of trials in Hodgkin's disease (Aspden *et al.*, 1981), Wilms' tumour (Jackson *et al.*, 1981), small cell bronchogenic cancer (Hughes, 1981) and early breast cancer (Metasa Ltd, 1979). An

important feature of such models is that they 'provide illuminating analyses of data that cannot be studied by conventional statistical methods, either because the data are too heterogeneous or because it is not clear which are the alternative hypotheses between which the statistical analysis is to adjudicate' (Beale, 1979 personal communication).

The form of the graphs allows comparisons between

Figure 2 Use of calibrated model to monitor an ongoing trial (for explanation see text).

the results at any time (K.D. Tocher, 1981, personal communication). Indeed, if in the hypothesis testing phase, likelihood methods (Edwards, 1972) are used, it is possible to overcome the problems normally associated with sequential testing (Armitage, 1975). The carefully elicited structure of the model allows for the presentation of results in a form consistent with the clinicians' understanding of the course of trials whilst highlighting the effects of different phases.

One phase of patient progress (box in Figure 1), namely the 'in-remission' box, has been modelled in more detail (Birkhead, 1984), to try to account for the shape of observed remission duration curves. The model was based on *a priori* assumptions about cell growth and the effects of treatment, and developed from application and discussion of the more general model.

Models of tumour response to chemotherapy (Birkhead & Gregory, 1984) and of the effects of drug resistance (Birkhead, 1984) have also recently been constructed with the aim of helping the more rational design of cancer treatment strategies. They exemplify another aspect of the 'Evaluation before' approach as it applies to clinical trials, which is featuring increasingly in the medical research literature (Goldie & Coldman, 1979; Goldie *et al.*, 1982).

Routine statistical analysis – a word of warning

It is unlikely that any clinician will have the resources available for him to develop models for his own particular problem. At the present time none of the type of models discussed above in the references is available in sufficiently clear form for him to use alone. Generally, they are developed by operational researchers in close collaboration with clinicians – which is one of the strengths of the approach.

What the clinician often does currently, and probably will continue to do for the foreseeable future, is to analyse clinical trials survival times using either the log-rank test (Peto & Peto, 1972) or, more popular in America, the generalized Wilcoxon statistic (Gehan, 1965). Often small trials are effected by randomizing only 40 or 50 patients to two or more treatments, but analysis proceeds regardless of design faults, via survival curves, and *P*-values are reported. '*P*-values or no publication' seems to be the current editorial dictum, when the question 'why *P*-values?' might be usefully asked. Research clinicians however, are going to continue to analyse data on small trials addressing questions which are not always well-defined. One or two notes of caution about life-table analysis therefore need to be sounded.

Clinicians often misunderstand the meaning of *P*-values, assuming that a number less than 0.05 is sufficient. Interim analyses are often performed (and

are sometimes necessary on ethical grounds), yet many clinical analysts fail to appreciate that a small *P*-value obtained at any one of a series of interim analyses is not as strong as one obtained at a single pre-determined point. Frequent testing at the 0.05 level will lead to many false positive results in early tests (Armitage *et al.*, 1969). Instructive examples have been constructed to demonstrate precisely this (Fleming *et al.*, 1981, unpublished).

In addition to the danger of false positives, there is a probability with every test procedure of obtaining false negative results. That is, of failing to identify a significant difference between survival distributions when one actually does exist. A statistical analysis (Freiman, 1978) of 71 published RCTs which produced non-significant treatment differences, indicated that half of these trials which were otherwise well-conducted had a 40% chance of missing even a 50% reduction in mortality! As has been remarked earlier, to increase the power of such tests and thus the sensitivity of the trials, the number of patients must be increased.

Finally, it is often not realized that the log-rank and similar tests as described in layman's language by Peto *et al.* (1976–77), are based on sets of assumptions, which are almost never checked but which in many cases may be violated.

Despite the many pitfalls awaiting them, clinicians will continue to undertake analysis of small trials using their limited understanding of conventional statistics. Such analysis, therefore, should be made available to them in as comprehensible and comprehensive a way as possible, but should contain the necessary strictures to avoid the common errors. Similar precautions will eventually be necessary with interactive mathematical modelling methods as they become available on desktop computers.

Clinical information systems

One of the great advances of recent years has been the development and use of computerized clinical data bases. The more recent discipline for clinicians of helping in the designing of data sheets has been extremely beneficial (Gregory *et al.*, 1981). Early systems were developed for use on large (main-frame) computers which required central staff to operate and often technical help at the clinicians' end. However, as the enthusiasm of the clinician increased it became very clear that a system which he could use, on his desk, was wanted. This meant that a new concept of 'User Friendliness' had to be developed. An example of this work is seen in PEDRO (Patient Endoscopy Data Records Organiser), a micro-computer based system on which patients' data are recorded on a visit-by-visit basis by staff in the medical unit.

This kind of development will eventually lead to data base, analysis and graphics all being available for clinicians, in an entirely understandable form on a micro-computer.

The need for safeguards against wrong analysis and inference is recognized and must be built into any systems to be used by non-professional analysts. However, the importance of this computing development lies in the future availability of accurate data, which can then be used with confidence for trials evaluation.

Discussion

RCTs concentrate on detecting differences in patient response due to, say, two different treatments. To achieve this aim an idealized trial is conceived and the analysis follows to the satisfaction of the medical statistician. However, inherent in the trials is the assumption that it is possible, over a long period, to create comparable patient groups and treat patients within such groups, identically. Some of the limitations and short-comings of the approach have been

discussed. In particular, the use of historical data is nearly always excluded. But the recent development of computerized clinical data bases makes this prohibition redundant in some cases and makes it possible to obtain useful information from a large body of previously discarded trials data. Advances in the theories of evaluation point to a need for other approaches (such as mathematical modelling) which could greatly enhance the design and analysis of clinical trials.

Acknowledgements

The views in this paper (which are the author's own) were developed in the Operational Research Service, Department of Health and Social Security and the Research Centre for the Mathematical Modelling of Clinical Trials, University of Warwick, and latterly, at the Clinical Operational Research Unit, Department of Statistical Science, University College London.

I am very grateful to Mr B. Birkhead for his comments, suggestions, and references.

The Clinical Operational Research Unit is grateful for financial aid given to this study by the Department of Health and Social Security.

References

- ALTMAN, D. (1980). Statistics and ethics in medical research. *British Medical Journal*, **281**, 1267.
- ARMITAGE, P. (1960). *Sequential Medical Trials*, 1st Edition, Blackwell Scientific Publications: London.
- ARMITAGE, P., MCPHERSON, C.K., & ROWE, B.C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, Series A*, **132**, 235.
- ARMITAGE, P. (1975). *Sequential Medical Trials*. Blackwell Scientific Publications: Oxford.
- ASPDEN, P., SUTCLIFFE, S.B.J., TIMOTHY, A.R. & JACKSON, R.R.P. (1981). An approach to the treatment evaluation of Hodgkin's disease. *Omega*, **9**(5), 537.
- BIRKHEAD, B.G. (1984). A mathematical model of remission duration in acute myelogenous leukaemia. *Cancer Treatment Reports* (in press).
- BIRKHEAD, B.G. (1984). Drug resistance and the chemotherapy of residual tumours. (submitted).
- BIRKHEAD, B.G. & GREGORY, W.M. (1984). A mathematical model of the effects of drug resistance in cancer chemotherapy. *Mathematical Biosciences*, **72**, 59.
- BLACK, SIR D. (1979). The paradox of medical care. *Journal of the Royal College of Physicians of London*, **13**, 2, 57.
- CHATFIELD, C. (1975). *Statistics for Technology*, 1st Edition, Chapman and Hall, London.
- CORNFIELD, J. (1976). Recent methodological contributions to clinical trials, *American Journal of Epidemiology*, **104**, 408.
- CRANBERG, L. (1979). Editorial: Do retrospective controls make clinical trials 'inherently fallacious'? *British Medical Journal*, Nov 17, 1265.
- , (1977). Editorial: Randomised clinical trials. *British Medical Journal*, **2**, 1244.
- EDWARDS, A.W.F. (1972). *Likelihood*, 1st Edition. Cambridge University Press.
- FIX, E. & NEYMAN, J. (1951). A simple stochastic model of recovery, relapse, death and loss of patients. *Human Biology*, **23**, 205.
- FLEMING, T.R., GREEN, S.T. & HARRINGTON, D.P. (1981). Performing serial testing of treatment effects. (unpublished).
- FREIMAN, J.A. (1978). The importance of beta, the type II error and sample size in the design and interpretation of a randomized control trial. *New England Journal of Medicine*, **299**, 690.
- GEHAN, E. (1965). A generalised Wilcoxon test for comparing arbitrarily single censored samples. *Biometrika*, **52**, 203.
- GEHAN, E.A. & FREIREICH, E.J. (1974). Non-randomised controls in cancer clinical trials. *New England Journal of Medicine*, **290**, 198.
- GOLDIE, J.H., & COLDMAN, E.J. (1979). A mathematical model relating the drug sensitivity of tumours to their spontaneous mutation rate. *Cancer Treatment Reports*, **63**, 1727.
- GOLDIE, J.H., COLDMAN, A.J. & GUDAUSKAS, G.E. (1982). Rationale for the use of alternating non-cross-resistant chemotherapy. *Cancer Treatment Reports*, **66**, 439.
- GORE, S. (1981). Statistics in question. *British Medical Journal*, **283**, 901.
- GREGORY, W., BELL, R. & JACKSON, R.R.P. (1981). An automated clinical research and data analysis system. *Computer and Biomedical Research*, **14**, 472.
- HILL, A.B. (1952). The philosophy of the clinical trial. *New England Journal of Medicine*, **247**, 113.
- HILL, A.B. (1962). *Statistical Methods in Clinical and Preven-*

- tive Medicine*, 1st Edition, 1–4, Livingstone: Edinburgh.
- HUGHES, D.J. (1981). A review of solid tumour modelling at RCMCT. University of Warwick Working Paper.
- JACKSON, R.R.P. (1975). Evaluation: A review of current thinking. DHSS Internal Report.
- JACKSON, R.R.P. & ASPDEN, P. (1979). Treatment evaluation – A modelling approach with application to acute myeloid leukaemia. *Journal of the Operational Research Society*, **30**, 11.
- JACKSON, R.R.P., McKAY, V.L. & MALPAS, J.S. (1981). Patient progress modelling and results in Wilms' tumour. *Omega*, **9** (5), 527.
- JACKSON, R.R.P., BIRKHEAD, B.G., BELL, R., LISTER, T.A. & GREGORY, W.M. (1982). Application of Jackson, Aspen acute myeloid leukaemia model. *Journal of the Operational Research Society*, **33**, 221.
- KAY, R. (1977). Proportional hazard regression models and the analysis of censored survival data. *Applied Statistics*, **25**, 227.
- LECAM, L. & NEYMAN, J. (1982). *Probability Models and Cancer*. Proceedings of an Interdisciplinary Cancer Study Conference, Berkeley. North-Holland, Amsterdam.
- METASA LTD. (1979). Report on the Breast Cancer Model Development Study. Report prepared for the Department of Health and Social Security, Operational Research Services.
- MOULD, R.F. (1979). Clinical trial design in cancer. *Clinical Radiology*, **30**, 371.
- PETO, R. (1978). Unpublished paper to MRC Cancer Trial Strategy Meeting, Oxford.
- PETO, R. & PETO, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society, Series A*, **135**, 185.
- PETO, R., PIKE, M.C., ARMITAGE, P., BRESLOW, N.E., HOWARD, S.V., MANTEL, N., McPHERSON, K., PETO, J. & SMITH, P.G. (1976–77). Design and analysis of clinical trials requiring prolonged observation of each patient. *British Journal of Cancer*, **34**, 585, and **35**, 1.
- TAUTU, P. (1978). Mathematical Models in oncology: a Bird's-Eye View. *Zeitschrift für Krebsforschung und Klinische Onkologie*, **91**, 223.