MEASUREMENT IN CLINICAL TRIALS

D.M. CHAPUT DE SAINTONGE & D.W. VERE

Department of Pharmacology and Therapeutics,
The London Hospital Medical College, Turner Street, London E1 2AD

Measurement is the delimitation and fixation of our ideas upon things. It appears to be necessary for answering questions about the order of our environment and particularly for that form of questioning we call experiment. Not only is it necessary to know what is to be measured but also why it is to be measured so that the instruments may be appropriately chosen and the results presented in a useful form. Measurement is not a process of discovery but a tool for sorting observations according to preconceived rules. It is apparent that not everyone operates the same rules, so there will be disagreement about what has been observed. This is an accepted and generally enjoyable fact of individual existence. However much human action is organised, concerted and applied to common tasks. Conflicting individual practices must therefore be resolved if these tasks are to be accomplished efficiently. This paper aims to identify the major issues concerning measurement in clinical trials and indicate the direction in which solutions may be found.

To avoid confusion two words will be given defined

and different meanings. These are *measure*, which will be restricted to apply to a technique for measuring some biological variable, and the *measurements* which that measure yields; the numerical data which are collected when it is used. It is helpful to discriminate between measures and measurements in this way because errors can arise in connection with either, and it is useful to discuss them separately (Figure 1). The word *parameter* (literally 'a measure laid alongside' something) should refer only to measures (a ruler is a parameter), but in strict convention it refers only to a mathematical class of measure. It should never, as in clinical slang, be used to refer to measurements.

In Figure 1 some questions which should be answered in relation to measures and to measurements are listed.

Choosing valid outcome measures

The clinical significance of a trial is judged on the difference in outcome between the treatment groups.

Figure 1 The relationship between variables, measures and measurements and some important questions which require answers in relation to each of them.

But whether the effect is useful depends not only just on its numerical size but in which outcome variables it occurred. An antirheumatic drug which produced a 50% reduction in joint pain might seem to be more useful than one causing a 50% fall in ESR, all else being equal. It is clearly important to choose measures of outcome which are accepted as appropriate and valid. Unfortunately many of the clinicians for whom clinical trials will provide data have only vague notions of what constitutes a treatment 'success'. In general practice a successful outcome has been cynically described as one in which the patient does not bother you again with that problem. In the treatment of diabetes it may mean that the patient shows occasional post-prandial hyperglycaemia or alternatively that no microvascular complications occur, depending on your point of view. In the treatment of malignant disease it may mean the patient lives longer than the median time expected for his particular cancer or that the quality of his life was improved though its duration was not extended. Even those doctors who can define their expectations of a successful treatment more closely than this frequently disagree. In a recent study on judgements in rheumatoid arthritis 80 rheumatologists were asked to score 17 symptoms, signs and laboratory results according to their usefulness in measuring 'success' of an anti-rheumatoid treatment (Kirwan, Chaput de Saintonge & Joyce, unpublished observations, 1982). Not only were there large differences between the policies between individuals but the variables appeared to behave as two distinct groups. 'Patient-orientated' variables such as morning stiffness and patients' assessment of pain tended to correlate directly with each other and inversely with 'laboratory-based' measures like the C-reactive protein and the ESR. There seemed to be groups of opinion which differed in the relative importance of these two groups as measures of success. How can we now decide which of these variables should be used in a clinical trial? Is there any 'correct' answer? The answer must depend, in part, on the standpoint from which the disease is viewed. From one view point disappearance of diagnostic features might constitute a success. When this means disappearance of an organism from infected urine or recovery of strength after a stroke the position seems sensible enough. However that is only because those diagnostic variables correlated well with functional disturbance. Titres of rheumatoid factor in excess of 320 may assist in the diagnosis of rheumatoid arthritis but this is not usually considered a useful measure of treatment success because it correlates poorly with other outcome variables. Since diseases affect patients it seems reasonable to consider outcome from the patients' point of view. In simple terms the patient is concerned with his ability to function, and continue to function, at a chosen level within his chosen

environment. Treatments which fulfil his expectations in this direction are usually judged successful, providing unwanted effects are not important.

It is remarkable to see in rheumatological trials how measures are used, presumably just because they happen to be there, when they may not really be relevant to the main issue of the trial. For example, one of the authors of this paper has suffered from arthritis in both knees, the other from arthritis in both thumb joints. A 'joint count' would yield identical results for both, changes in joint circumferences would no doubt yield statistically solid results, but neither of these simple, available measures would reveal that supplication would be a problem for one of us, whilst writing this paper is a problem for the other. Hence measures need to be relevant and to reflect disease activity in terms which mean something to the patients under study, even if this means sacrificing a simple, robust and statistically analysable measure, or combining it with other less straightforward but more relevant measures like walking distance or handwriting analysis.

The problem is further complicated by the ability of the doctor to modify the patient's therapeutic response and frequency of unwanted effects independently of the treatments he gives (Joyce, 1968). One factor is likely to be his degree of expressed satisfaction with the improvement in response variables he considers most important.

This problem is sometimes, mistakenly, viewed as a conflict between the importance of subjective and objective effects of treatment. Until recently it was fashionable to look for the most objective measure that could reasonably be related to the therapeutic response. Such measures were often biochemical and were claimed to be 'more accurate' or 'more precise' than corresponding subjective measures. In fact neither term has any meaning unless the correlation between the subjective and the objective measure has been established, which is uncommon. Change in haemoglobin concentration is frequently used as an objective measure of response to haematinics yet few who use these drugs can describe the correlation with improvements in tiredness, breathlessness or any other symptom of anaemia. The clinical need for powerful trials able to detect statistically significant but small changes in haemoglobin is therefore doubtful. Similarly if symptomatic measures alone are used a useful effect could be missed if the patient simply titrated his symptoms against an improved ability to do work.

Remarkably often in clinical trials a measure is used because it happens to exist, not because it reflects the variable under study in any really useful way. This may happen because no other measure is available; it may then represent the best that can be done till better times, but it raises the questions

whether the work is premature and whether it might not have been wiser to spend the efforts improving measures. At other times a poor measure is used because someone else used it. It makes it possible to bury all the hard work of validating methods under (Snooks & Bloggs, 1893). This can be wise only if the work of Snooks & Bloggs is repeatable, is relevant to the present study, and if the system used by those distinguished authors operates similarly in the hands of the present investigators. These are no mean assumptions, and temper enthusiasm for the otherwise very proper scientific practice of building on the work of others which is no substitute for my checking my own work to see that it is truly analogous to theirs. This example is given to illustrate the importance of reliability, repeatability, availability and relevance of measures.

The validity of outcome measures depends on the extent to which they correlate with the 'true' outcome. This has two components: the immediate restoration of the patients' function and the alteration in the long-term natural history of the disease or the prognosis. Although 'function' appears to be concerned with patient-centred and often subjective variables of the type mentioned earlier, much of it is homeostatic and only called into action in response to environmental changes. Thus an improvement in granulocytopenia would only be accompanied by subjective improvement in patients challenged by infection. It would be unethical to do such a thing deliberately in the course of a trial, and since the correlation between granulocyte count and liability to bacterial infection is well established it could be used as an indirect and objective measure of patient defence function. Scales of general functional ability have found some application in geriatrics and cancer chemotherapy but often suffer from lack of specificity and rigidity of concepts. It is likely that functional measures of outcome chosen by the patient himself are more sensitive measures of drug effect than those chosen by the doctor (Scott & Huskisson, 1979). Where death or a specified severe disability is chosen as the target to be avoided, there is likely to be little disagreement. Otherwise there seems to be a general reluctance to consider using specific functional impairments as measures of immediate outcome and to favour instead laboratory-centred variables whose correlation with subjective variables is often unknown.

The situation is rather different when the aim of treatment is to alter the deferred outcome as in hypertension or the hyperlipidaemias. Immediate subjective benefit is uncommon and although the ultimate aim is to prevent future functional impairment measures of treatment effect often depend on indirect measures such as reduction in blood pressure or serum cholesterol. The validity of these measures depends, of course, not just on their existence as risk

factors but the relation between changes in their value and changes in the outcome.

Validity of outcome measures thus depend on how well they are chosen in relation to the treatment aims. The effects of analgesics and anti-inflammatory agents in rheumatoid arthritis will be measured in terms of changes in the symptoms and signs of inflammation most of which have a strong patient-centred component. Antirheumatoid drugs, on the other hand, may have relatively little immediate symptomatic effect yet reduce the rate of joint destruction, the number of erosions and possibly the ultimate degree of disability. As in most aspects of trial design the correct procedure depends on the questions the trial aims to answer.

In many cases there is no single measure of outcome which can be accepted as of overriding importance. Usually several criteria of success are measured, each chosen to represent a different aspect of the drugs' action and the patients' response. Changes in symptoms and signs are usually combined with changes in biochemical, haematological or radiological values to give a composite picture of the effects of treatment. When trials are planned there is a tendency to include large numbers of outcome measures in the assessments either because some of them are thought to represent advances over established measures (which nevertheless 'ought to be included') or because of uncertainty as to which is the most sensitive measure of drug effect in the context. Plausible as these arguments may seem the use of many measures of outcome can cause a number of serious problems.

- (a) Trial organisation becomes more complex in proportion to the number of measures, assessors, and centres. Unless correspondingly more time (and money) is spent on quality control, loss of specimens and data will increase. Such losses will not just reduce the power of the trial and complicate the analysis but may occur in such a biased way that simple (but not essential) variables are recorded at the expense of more time-consuming, but more important measures of effect.
- (b) The greater the number of variables studied and the greater the number of occasions on which comparisons are made, the greater is the chance that spuriously significant results will be found. Although corrections are available for multiple comparisons they are infrequently used because it is usually impossible to say how many informal looks at the data preceded each statistical analysis.
- (c) When sequential designs are used decisions to terminate the trial are usually made on the results of changes in one variable. This has the advantage of forcing the investigators to decide which of the outcome variables they consider

most important. In situations where they refuse to be forced or excess toxicity is an outcome variable, then several sequential plans may have to be used simultaneously.

- (d) The number of patients needed for the trial must be calculated on the basis of a specified difference in a specified variable with a known variance. It is likely that calculations from several variables will give different estimates, so here again the trialists must decide which of the variables is most important.

Indices of response which involve combining several variables have been proposed as a solution to some of these problems. This is usually a simple addition of values weighted according to the relative 'importance' of the variable included. These indices which are mathematical models of disease activity, are of two main sorts. In the so-called 'explicated' models experts 'explain' the model they habitually use in terms of choice of variables and the relative importance of each variable. The functional form of the model is then decided and the model tested in terms of its ability to predict the judgement of the expert when presented with a clinical data set. In statistical models, on the other hand, the model is generated from a series of observations of the effects of treatment. A well-defined outcome is expressed in terms of several independent measures of that outcome and regression analysis used to determine the form of the relationship.

An explicated model of disease activity is capable of representing the judgements of experts and applying them consistently to appropriate data sets. Thus any novice can make similar judgements to an expert given the same data set. The use of judgement analysis has been proposed as a way of standardising the judgement of outcomes in international or multi-centre clinical trials (Joyce *et al.*, 1977). However the ability of a model to describe completely the judgement processes of an expert is no guarantee of its validity. He still has to choose which variables estimate disease activity and may disagree with other clinicians in this respect. Where there is some universally accepted criterion of disease activity such as death or confinement to a wheel-chair this may serve as an independent test of the validity of a model. Where there is not, validity may only be assessed in terms of content or construct. Content validity refers to the extent to which the individual variables in the model represent the components of disease activity that they are supposed to represent. Construct validity concerns coherence of the model with other independent measures of activity and its divergence from measures of other features of the disease (e.g. diagnostic, aetiological). It is sometimes suggested that the validity of outcome measures is related to their ability to detect treatment differences. Thus a

measure which could detect the difference between two doses of propranolol in the treatment of hypertension would be considered more valid than a measure which could not. Adopting this attitude the optimal index of activity would be that which discriminated best between the treatment groups. It is hard to see the merit of this approach, no more information is gained than by examining the treatment allocation code which will also discriminate effectively between the treatment groups. The fact that a variable is treatment-sensitive may be useful in bio-assay but is usually irrelevant to the main trial question.

The validity of outcome measures is not agreed in many non-lethal diseases. The choice of variables in clinical trials is thus inevitably a compromise not only between the views of experts but also with those of the less informed. If the aim of trials is to alter practice they must convince the bulk of clinicians in practice or else their influence will be limited. If outcome variables are chosen on the basis of their novelty or research interest, the results risk rejection as irrelevant or uninterpretable by those they were designed to convince. Although this lack of agreement on the validity of outcome measures may seem surprising it is less so when seen against the disagreement as to what is a disease itself (Campbell *et al.*, 1979).

A problem which frequently arises next is through whose mind a subjective measure should be analysed. The choice is between an 'objective' measure, such as walking distance, which is observable by an investigator, and a 'subjective' measure, such as a symptom rating scale or visual analogue scale (Aitken & Zeally, 1970). The difference is by no means so great as might appear; the patients' subjective responses have major influences upon outward measurable responses, and analogue scales seem less open to subjective bias than might be expected. In any case, in any measure of human function the outcome is really interpreted through *two* minds (the subject's and the observer's), though their influences clearly vary among times and places. Figure 2 shows an example of such effects.

Precision, which must be distinguished from accuracy, indicates the degree of fineness of a measurement. A stopwatch is capable of greater precision than a wall-clock at measuring elapsed time but either may be inaccurate however if they are running fast or slow and do not truly measure what they are supposed to. As we have seen it is obviously a waste of time to choose variables which are invalid but which can be measured accurately and precisely, but assuming a variable is valid, with what accuracy and with what precision should it be estimated?

Tests involving a coarse scale of well-being which consists only of the categories 'well' and 'ill' may lack

Figure 2 The effects upon measure and measurement of having one or two observers in the chain from process to reading.

the power to detect small but consistent degrees of improvement. Whether this loss in power is serious will depend on the clinical importance of the effects that may be missed, threshold value of which must be determined at the outset. High degrees of precision have the disadvantages that they often require proportionately greater amounts of skill, time, money and inconvenience. Other aspects of the trial organisation may thus be weakened by over-enthusiasm for inappropriately precise measurements.

Few clinical investigators seem to check the precision of their measures by making replicate measurements in a situation where no change would be expected in what is measured. The results can be most illuminating when this is done, and in any case provide an estimate of the *minimal* variance which may be expected from the system under study. This can be extremely useful if and when the expected effect of some intervention matches the minimal variance in size. Clinical research seldom shows understanding of this type of problem, which is really one of signal-to-noise ratio detection.

Inaccuracy may be random or non-random. Any man or machine making repeated observations of the same event may show random variation in the results. This simply represents variable information loss during data processing. The loss can be minimised by making further observations of the same event or reducing the opportunities for information to be lost. Wherever possible the subjective components of decisions should be replaced by an objective statement since there is clear evidence that judges apply their own decision policies much more accurately

under these circumstances. The accuracy of non-experts may be improved by providing them with the criteria for trial decisions formulated by experts. This is especially useful when many trial assessors of differing expertise are involved. For example, the use of complex rating scales and symptom inventories in trials of antidepressants may allow relatively inexperienced personnel to measure changes in mood accurately.

Non-random inaccuracy occurs when a subset of those making measurements show a consistent trend in the values of their observations unrelated to treatment effects. An example is seen in multicentre assessments of hearing loss by audiology. Differences between centres in the calibration of instruments and in the ambient noise levels will result in systematic differences between centres in addition to random inaccuracy. Non-random variation is most serious when it is unequally expressed in the treatment groups and is thereby capable of biasing the results. This would be the case if one treatment was disproportionately represented in a centre which tended to record greater hearing losses than the others. While it is not always possible to standardise every aspect of the measuring procedure (the physician's personal approach for example), non-random inaccuracy which could bias the outcome must be anticipated in advance and neutralised appropriately. How much trouble is taken to reduce other forms of inaccuracy depends on the likely pay-offs. Since measurements are one of the sources of variance within the trial improved accuracy will increase power. To what extent depends on what proportion of the error variance can be attributed to the current method of

measurement, if it is small then even perfect accuracy will little improve the power of the trial. The usual solution in small and moderate size trials is to attempt the greatest accuracy that can reasonably be achieved within the constraints of the trial organisation. It seems likely that the use of trial simulations may allow these goals to be set more rationally but as yet they do not seem to be widely used.

Alteration of measurements by the measuring instrument

A frequent cause of systematic, or bias error affecting measurements is the influence of the instrument; this was a familiar problem in the early days of catheter pressure transduction, where the damping and smoothing effects of instruments upon rapidly varying pressures was a major concern. Modern equipment has largely removed that problem, but it reappears in new forms. Chip technology has made a new range of instruments available, often in full microprocessor form, so that measurements appear in some convenient LED display, or even on paper. But it is essential to remember that every instrument placed between measure and measures can distort the measurements. For example, automatic blood pressure recorders often drift with time; it is essential to check their constancy against another method at intervals, at least before and after each set of recordings. Now, clinical trials are in a sense 'instruments' and can falsify interpretation in analogous ways. For example, it may be supposed that the prolonged effects of a drug last for up to five days, and a washout period lasting five days is arranged between treatment blocks. Now, if the range of drug remanence is in fact from three to ten days in various subjects a systematic bias will appear in the trial which has been induced entirely by the measuring 'instrument'. Another familiar example is where fixed times are chosen for measurements which would better be made at variable times which coincide with peak response. This problem appears in a most seductive and dangerous form with acute disorders which vary rapidly near their onset. The Coronary Care Unit trials are a good example; apparent 'benefits' of coronary care depend so much upon the time of arrival at hospital in relation to the acute stage survival curve for the disease, that some trials are less measures of the efficacy of a coronary care unit than they are of the speed of the ambulance service.

Compliance

An important principle of trial design is that every treatment should be given an equal chance of success. It is frequently assumed that all trial patients are perfectly compliant and take all the treatment prescribed for them in spite of overwhelming evidence to

the contrary. Without some measure of how much of the trial medication was actually taken it may be impossible to attribute effects to treatments. An effective treatment which tastes unpleasant may be taken less reliably than one that is less effective but more acceptable. The therapeutic effects of the less effective treatment may be falsely judged superior if the differences in compliance are not taken into account. Biased conclusions of this sort may be avoided if the consumption of trial medication is recorded throughout the trial. This may be accomplished by return tablet counts, patients' records of doses taken, tests for the drug in plasma, saliva, or urine or by a number of more sophisticated techniques. Of course the trial will be designed and the treatment so presented as to maximise compliance throughout. If differences appear between the treatment groups in the amount of trial medication consumed it may be possible to allow for this in the analysis. An analysis of covariance using the amount of trial medication consumed as a co-variate will allow a comparison of the treatment effects to be made which is independent of treatment compliance (Joyce, 1962). Such methods are open to the objection that non-compliant patients may have a different outcome for reasons unrelated to their non-compliance. This suggestion gets some support from the observation that the mortality amongst patients taking an inactive placebo in the Coronary Drug Project was 5% amongst compliant patients and 10% amongst the non-compliant (Canner, 1979).

The situation is more difficult where non-compliance takes the form of taking additional treatments (analgesics, antacids etc.) which are not included in the trial. If this possibility can be foreseen in advance, as it can with many trials involving the relief of symptoms, at least the choice of secondary, 'rescue' treatments can be standardised. An analysis of the frequency of recourse to the rescue treatment will then give some measure of the success of the primary trial treatment. If the rescue treatment differs from patient to patient no satisfactory analysis is possible without making assumptions about the equivalence of the various treatments.

Although it has been stated that no trial is nowadays acceptable without some measure of compliance enthusiasts for pragmatic trials may not agree. If the reasons for non-compliance are entirely linked to the characteristics of the primary trial treatments then, it can be argued, non-compliance is an integral feature of the use of the treatment within the context of the trial. Thus the fact that patients do not take unpleasant treatments should be allowed to alter the outcome and should not be accounted for in the analysis. This analysis by intention to treat has much realism to commend it. It has the disadvantage that potentially useful pharmacological effects may be obscured as may causes of non-compliance which

could be removed by changes in formulation of the trial treatment or its presentation. In general it seems likely that many of the factors which maximise compliance during a clinical trial will operate less strongly when the treatment is prescribed for individuals and benefit may be correspondingly reduced. For some trials patients are recruited according to their ability to comply with the trial regimen during a run-in period, non-compliers are excluded. Not only will the conclusions of such a study be difficult to apply to a population whose compliance is unknown, but bias may arise when non-treatment determinants of outcome such as age and sex are also determinants of non-compliance.

Measurements

It is important, for successful analysis, to consider several factors about measurements. The *scale* on which changes are measured may be quantal (or nominal), ordinal, or interval in type. This distinction partly determines the statistical tests which may be used in analysis (Siegel, 1956). What escapes investigators more often is the grouping of measurements, especially whether they are in part replicates made within each subject, which would influence the form of analysis of variance test used, or whether they are irregularly distributed in time, which may seriously affect regression analysis.

Again, just as investigators often fail to study replicate results of measures to check their reproducibility and precision, so they often fail to collect a small pilot group of measurements to study their distribution. For example, were this bimodal, or highly skewed (Figure 3), the entire test procedure might be invalidated, a point which would be missed were the observations to begin with the actual *experiment*, such that bimodality or skew is attributed falsely to treatment effects.

It is especially important to think about the *timing* of observations in relation to known profiles of disease (Sims, 1965). In a desire for a stringent protocol people may fix observations at a defined time following some intervention, when they would do much better to fix them in relation to clinical response, e.g. by taking the peak response, whenever it may occur. Taking means from data which differ in time distribution between themselves is another splendid way to induce bias (Oldham, 1968). For this reason it is desirable for scientists (but not, it seems, for some journal editors) to publish 'diagrams' or tables for individual as well as grouped measurements in such cases. Another important set of constraints is concerned with the discomforts caused by measurement, be they pain, tiredness, boredom or embarrassment. These may be slight, and well within the limits acceptable to ethics committees, but may still influence compliance and bodily functions.

Lastly, observer bias influences and selects measurement more than most imagine. This has been shown most clearly for digit preference in relation to blood pressure (Wright & Dove, 1970; Rose *et al.*, 1964). However, there is also a time drift which affects successive blood pressure measurements (Dunne, 1969), and similar factors occur in many analogous fields, e.g. slow change in salt, or trace nutrient balances following admission to a ward, or consequent upon relative immobility, slow drifts in laboratory temperature during experiments (Redisch *et al.*, 1952), and even spatial effects of the sites at which tests are made (Ingle, 1962). These all require to be borne in mind by experimenters, who must consider the need for equilibration and washout periods, for replicable arrangements of experimental apparatus and rooms, and for careful random choice of locality for experiments if numerous subjects are to be studied in a large room (e.g. class experiments). This may seem very theoretical, but one author is aware of some experiments which were submitted to a colleague for analysis in which the whole of a very exciting difference 'due to treatment' proved to be such in the disappointingly agricultural sense that the 'treatments' were different operating suites, not different operations!

Lost measurements Inevitably data which should have been collected will be missing at the time of analysis. Although corrections for missing data are available for many statistical tests the implications may be more serious than simply a loss of power in the trial. Possible causes for loss of data must be considered when the trial is planned and steps taken to minimise the losses. First, the participants must all agree the need for every measurement on each occasion specified by the protocol. This is particularly important when the trial collaborators delegate the hard work of the trial to their juniors. Measurements which are uncomfortable for the patients or irksome to the assessors are particularly likely to be missed. If the trial procedure documents are not clear inappropriate samples may be taken in inadequate amounts into the wrong containers sent to the wrong laboratories with the wrong instructions. If doctors, patients, relatives, nurses, porters, technicians, secretaries and all trial staff understand and accept their role in the collection of data many losses can be avoided. Pilot trials can usefully detect weak points in a trial and are especially valuable where the organisation is complex.

It is unwise to assume that data lost for any of these reasons will be randomly distributed between the treatments. Biased loss may occur when one treatment results in events (adverse reactions, changes in disease severity) and the other does not. There is a tendency to note things that happen and to fail to record things that do not which may result in dispro-

Figure 3 Some visual analogue scale results from a study of pain in preterminal illness (Hunt *et al.*, 1977). The scales reveal examples of unimodal, skewed, and perhaps even bimodal distributions.

portionate data loss in the less eventful treatment periods. Monitoring data returns during the trial is an important function of the trial controller. Increasing data loss over a number of variables may be an early indication of disillusionment or disaffection in staff or patients. Selective loss restricted to measurements which are uncomfortable or hazardous to patients may suggest doubts about the ethics of the trial procedure. In either case prompt action can prevent further loss.

Potentially more serious problems involve the loss of patients from the study. This may occur through drop-out and failure of the patient to attend for assessment or by the clinician's deliberate exclusion of the patient from the analysis. In a trial of azathio-

prine in rheumatoid arthritis, male patients with few symptoms had a significantly higher drop-out rate than any other group. Although the reason was never established it seems likely that for these patients frequent hospital attendances for assessment caused unacceptable interference with work. Death and other serious events are causes of loss to follow-up which are particularly important to detect. The consequences for the trial will be serious when the treatment outcome of the drop-outs differs from the remainder. Unfortunately the outcome in the drop-outs is often never known. Their number can be reduced by imaginative trial organisation to minimise inconvenience to patients and by persistent personal effort to contact defaulters. Earlier articles in this

series have emphasised that a 'pragmatic' trial analysis will include all patients allocated treatments, whether they completed the trial or not. Patients who fail to attend and to continue in the trial are classified as 'trial failures', which is not, of course, the same as a treatment failure. Apart from the difficulty of extrapolating the results of a trial where there are many drop-outs, there is a problem with analyses where the outcome is other than 'success' or 'failure'. Because of the difficulty of defining success already mentioned, trialists frequently estimate the main treatment effect on a continuous scale (morning stiffness, grip strength etc.) and leave the reader to judge its clinical importance. But how can primary treatment failures be brought into this context? In practice we are often presented with several separate analyses. One of those who completed the trial, another of those who were removed because of drug toxicity or failure to respond, and a third of those who were lost to follow-up. It is customary to include in the main analysis data from all drop-outs and withdrawals up to the time of their loss from the trial, after which they no longer contribute. However there is a

risk that the exclusion of these patients from subsequent time points may bias the results if the pattern of loss was non-random. An alternative is to allow each patient who was lost to continue to contribute data, values for which have been extrapolated from the pre-dropout trends. This makes sometimes unjustifiable assumptions about the natural history of the disease and it is simpler to insert the last recorded data in all subsequent time points. Another proposed solution scores all patients who entered the trial on the basis of observed effects at the end of the trial or the relative values of the cause of withdrawal. Thus 'withdrawn early—cured' scores very high, while 'withdrawn early—drug toxicity' scores low (Gould, 1980). Most of these solutions are untested and their acceptability to the trial-consuming public has not been established. A conservative approach to the problem is to perform all possible analyses from the most conservative to the most liberal. If they all give the same answer the most conservative can be published. If they differ, the conclusions are clearly dependent on the assumptions made about missing data, and these should be discussed.

References

- AITKEN, R.C.B. & ZEALLEY, A.K. (1970). Measurement of moods. *Br. J. hosp. Med.*, **4**, 215–224.
- CAMPBELL, E.J.M., SCADDING, J.G. & ROBERTS, R.S. (1979). The concept of disease. *Br. med. J.*, **2**, 757–762.
- CANNER, P.L. (1979). Monitoring clinical trial data for evidence of adverse or beneficial treatment effects. In *Multicenter Controlled Trials: Principles and Problems*, eds Boissel, J.P. & Klimt, C.R. Paris: INSERM.
- DUNNE, J.F. (1969). Variation of blood pressure in untreated hypertensive outpatients. *Lancet*, **i**, 391–392.
- GOULD, A.L. (1980). A new approach to the analysis of clinical drug trials with withdrawals. *Biometrics*, **36**, 721–727.
- HUNT, J.M., STOLLAR, T.D., LITTLEJOHNS, D.W., TWY-CROSS, R.C., VERE, D.W. (1977). Patients with protracted pain: a survey conducted at the London Hospital. *J. med. Ethics.*, **3**, 61–73.
- INGLE, D.J. (1962). Control and exploration of animal variables in drug research. *Clin. Pharmac. Ther.*, **3**, 242–244.
- JOYCE, C.R.B. (1962). Patient co-operation and the sensitivity of clinical trials. *J. chron. Dis.*, **15**, 1025–1035.
- JOYCE, C.R.B. (1968). Psychological factors in the controlled evaluation of therapy. In *Psychopharmacology, dimensions and perspectives*. Mind and medicine monographs No. 17, Lippincott
- JOYCE, C.R.B., BERRY, H., CHAPUT DE SAINTONGE, D.M., DOMENET, J., FOWLER, P. & MASON, R.M. (1977). *Judgement analysis of investigators' assessments: A way to reduce one important source of error in multicentre trials*. Eular bulletin Monograph 1/1977,
- LYNN, P.A. (1978). Frequency spectra and the frequency range of biological signals. *Br. J. clin. Equipment*, July.
- OLDHAM, P.D. (1968). *Measurement in medicine*, pp. 191–193. London: English University Press.
- REDISCH, W., SHECKMAN, E. & MURRAY STEELE, J. (1952). Skin temperature response of normal human subjects to various conditions. *Circulation*, **6**, 682–687.
- ROSE, G.A., HOLLAND, W.W. & CROWLEY, E.A. (1964). A sphygmomanometer for epidemiologists. *Lancet*, **i**, 296–300.
- SCOTT, P.J. & HUSKISSON, E.C. (1979). Measurement of functional capacity with visual analogue scales. *Rheumatol. Rehabil.*, **16**, 257–259.
- SIEGEL, S. (1956). *Nonparametric statistics*. London: McGraw-Hill.
- SIMS, R.S. (1965). Temperature recording in a teaching hospital. *Lancet*, **ii**, 535–536.
- SNOOKS, A. & BLOGGS, V. (1895). The standard work on things which never happen again. *J. Inaccess. Obscure Meth.*, **1**, 12–579.
- WRIGHT, B.M. & DOVE, C. (1970). A random-zero sphygmomanometer. *Lancet*, **i**, 337–338.