MARKETING OF THE NEW DRUG

Once sufficient clinical and preclinical data have been collected on the safety and efficacy of a new chemical, submission is made to the Committee on Safety of Medicines. If the information provided is satisfactory a product licence can then be issued. By this time ten or more years may have elapsed since the taking out of a patent on the "new product candidate," and a total of £20m expended. Only a few hundred patients, however, will have been exposed to the drug. Thus the full benefits and problems associated with its use may not become apparent until months or years after it has been marketed. Because of the latter possibility I would suggest that practitioners should adopt a cautious attitude towards new medicines. In my view there is no justification for switching automatically to the 15th "β-blocker" from one that has been in use for ten years or more, since it is unlikely that the new compound will have a measurable advantage for most patients. Possibly, however, it may have advantages for some patients—for example, a cardioselective β-adrenoceptor antagonist would be preferable to propranolol if the patient had coexistent airways obstruction or diabetes mellitus. Similarly, a tetracyclic antidepressant might be preferable for a patient with coexistent cardiac disease.

In addition my personal rule is to insist that there are at least two good clinical studies (with similar results) showing that the drug has therapeutic (rather than just statistically significant) advantages in a particular condition before prescribing it.

References

1 Dollery CT, Davies DS. Conduct of initial drug studies in man. *Br Med Bull* 1970;**26**:233-6.

2 Downie CC. Clinical pharmacology. In: Harris EL, Fitzgerald JD, eds. *The principles and practice of clinical trials.* Edinburgh and London: Livingstone, 1970;15-22.

3 George CF. The investigation of new drugs in man. *Br J Hosp Med* 1974; **12**:780-9.

4 Clark CJ, Downie CC. A method for the rapid determination of the number of patients to include in a controlled clinical trial. *Lancet* 1966;ii:1357-8.

Medicine and Mathematics

Statistics and ethics in medical research

Collecting and screening data

DOUGLAS G ALTMAN

Even with an impeccable design there are many ways in which a study can go wrong when the data are being collected. In general, the more complicated the design the more chance there is of the study not being carried out properly. As an example, consider this historic study. The story was related by "Student" (he of *t*-test fame):

"In the Spring of 1930 a nutritional experiment on a very large scale was carried out in the schools of Lanarkshire. For four months 10 000 schoolchildren received three-quarters of a pint of milk per day; 5000 of these got raw milk and 5000 pasteurised milk; another 10 000 children were selected as controls, and the whole 20 000 children were weighed and their height was measured at the beginning and end of the experiment."[1]

There was no power problem here. The study found that children getting extra milk gained more weight in the period than did the controls. But did the extra milk cause the extra gain? The figure is a simplified chart showing the weight changes for girls during the study. Since the two milk groups are very similar, only one is shown here. There are two striking features of this graph. The first is that the controls were in all cases heavier than those getting extra milk (they were taller too). This can be easily explained by the discovery that some of the teachers who allocated children to groups had juggled the randomisation to enable the poorer children to get the extra milk.

The second curious feature is that the observed growth rate in each group was much less than would be expected by looking at the next age group. The explanation for this is also very simple. The study began in February and ended in June, and the children were weighed on both occasions with their clothes on. The shortfall in weight increase is thus largely due to a different amount of clothing, and the smaller effect in the milk feeding group can be explained by the poorer children wearing relatively fewer clothes in winter.

It may be thought that errors such as these are really obvious, and nobody would make such mistakes nowadays. Two points may be made about the altruistic adjustment of the randomisation. Firstly, this procedure is not unknown in more recent times. Carleton *et al*[2] reported that strongly motivated doctors may upset trials by transilluminating envelopes containing the names of drugs in order to find the desired treatment. However well-intentioned, such underhand activities are by their nature likely to go undetected and can invalidate a whole study. Doctors should not agree to participate in a randomised controlled trial if they have a prior preference for one treatment. Equally, the study sample should not include subjects for which one treatment is clearly medically preferable. A trial where either of these conditions was broken would be unethical.[3]

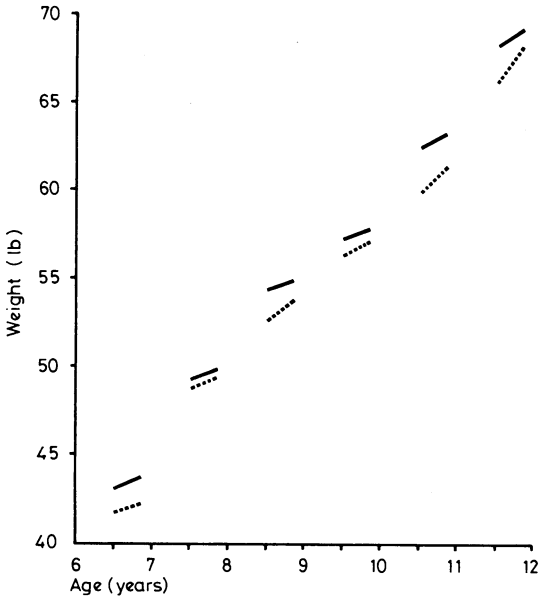
The second point relating to the allocation of subjects to treatments is that a major reason for random allocation is to eliminate the effect of both deliberate and unconscious biases. If the groups are not selected randomly it will be impossible to know whether any observed treatment effect is genuine, as in the

Division of Computing and Statistics, Clinical Research Centre, Harrow, Middx HA1 3UJ

DOUGLAS G ALTMAN, BSC, medical statistician (member of scientific staff)

Lanarkshire milk trial. So what reliability can we place on the results of a study in which patients were allocated to treatments "nearly at random"?[4]

The other error in the Lanarkshire study, that of weighing children with full clothing at different times of the year, would be unlikely to be made in that form now. Errors of this sort, however, are very easy to make, and usually occur when a source of variation is overlooked. For example, in studies looking for small differences it may be important to allow for the fact that height and blood pressure are less in the evening than in the morning, or that lung function is better in summer than in winter. Failure



Lanarkshire milk experiment[1]: comparison of control group (————) and milk feeding group (- - - - -) showing mean weight at beginning and end of study for each yearly age group.

to allow for such things can lead to two effects being "confounded" or inseparable. So, in the milk study we cannot say how much of the difference between the groups was due to the milk, how much to the non-random allocation, and how much to the changes in clothing.

Perhaps to try to insure against this sort of problem, it is quite common for a study to collect information on anything that might possibly be of some value or interest. This seems particularly common in surveys, where one is not always investigating a specific issue but looking at a general situation. If information is being collected by questionnaire, however, then increasing the number of questions may lower the response rate, with the results being less reliable as a consequence. Further, excessive amounts of information may reduce the care given to data collection.

Data screening

Before proceeding to the analysis, some degree of data screening should be carried out. By screening is meant checking so far as is possible that the recorded values are plausible, since one can not usually know if the data are correct. Simple data sets obviously need minimal checking in comparison with studies concerning a large amount of information for each subject.

Screening the data (sometimes called cleaning or validation) entails checking that for each variable all the observations are within reasonable limits. Where feasible, each variable should also be cross-checked against other relevant information. This may show inconsistencies such as an 18-year-old woman with six children. It may also show that values that appeared odd are quite compatible with other data.

Much can be learnt from an initial close examination of the data, taking variables both one and two at a time, using histograms and scatter diagrams.[5] As well as identifying outliers, such screening of the data should disclose whether it will be necessary to transform any of the variables before analysis. It will also help to discover if any observations are missing. All of these aspects merit examination.

WHAT CAN WE DO ABOUT OUTLIERS?

Outliers are observations that are not compatible with the rest of the data. Typically there may be one or two such values in a set of data, but they can have an unduly large influence on the results of an analysis.

The first thing to do with suspicious values is to make sure that they have not been incorrectly transcribed. Any impossible values should be treated as missing data, but defining what is impossible may be very difficult. For example, how large would a value for length of gestation or maternal age be before it was considered impossible?

If an outlying observation appears correct in that the value is possible (although unlikely) and there is no evidence to suggest that it is wrongly recorded, then it should not be excluded from the analyses. It is particularly bad to remove such values purely on the grounds that they are the smallest or largest.

In small samples outlying values may have a very large influence on the results—for example, a regression line will be "pulled towards" outlying values. Ranking methods can be used, but they are generally only useful for testing hypotheses, not for the estimation of means, standard deviations, regression slopes, and so on.

WHY TRANSFORM DATA?

When analysing continuous variables (height, blood pressure, serum cholesterol, etc) it is usual to make use of a "family" of statistical analyses, including *t* tests, regression, and the analysis of variance, that make important assumptions about the data. Such analyses are not valid if these criteria are not met.

The best known example of this is when data display skewness instead of the required symmetric Normal (Gaussian) distribution. All of the above methods have some sort of Normality assumption. In such cases it is often possible to find a mathematical transformation for the data that will make the analysis valid.[5,6] By far the most common transformation used in medical research is the logarithmic transformation, needed, for example, for various biochemical measurements.[7] It is worth noting that an appropriate transformation may also have the effect of making previously suspicious values become quite reasonable.

Although it is obvious that the more nearly the underlying assumptions are met the more reliable will be the results, it is unfortunately not possible to say how far the raw data can deviate from the ideal before the results become invalid. Because of the subjective nature of this problem expert help can be particularly helpful here.

WHAT CAN WE DO ABOUT MISSING DATA?

An important distinction must be made between data that are missing through random misfortune (if some forms are mislaid, for instance) or for a reason directly or indirectly related to the study itself. Most studies have a few accidentally missing observations. These cases can usually be omitted without greatly affecting the results. It may be thought preferable to include a subject for any analyses for which data exist, only excluding him when the relevant observation is missing. This procedure can cause complications in interpretation, however, as each analysis will be based on different subjects, and is better avoided if possible.

It is also common to have data missing through a subject's refusal to supply information or to participate in a study. The problem here is that refusers are often an atypical subgroup. In a survey it may be possible to study what is known about the refusers to see if and how they do differ from participants, and to try to estimate the effect on the results. Clearly a high refusal rate will mean that little sensible extrapolation from the sample to the population is possible.

In a randomised trial it is essential that refusers (or withdrawals) are considered as part of the group to which they were allocated.[3] A good example is given by a study[8] of the sudden infant death syndrome. High-risk infants were randomly allocated to observed and control groups, where observation consisted of increased health visitor surveillance. In the control group, where active participation did not need to be sought, there were nine unexpected deaths out of 922 infants, a rate of 9·8 per thousand. In those allocated to the "observed" group, there were two unexpected deaths out of 627 who agreed to participate (3·2 per thousand), and three out of 210 among those who refused (14·3 per thousand). This is a good example of the commonly found poor prognosis among refusers.

The purpose of a randomised trial is to be able to make comparisons between randomly allocated groups. Some trials have "observed controls" where one randomly chosen group is offered treatment while the other group is just observed. Any refusing treatment must still be considered with the treated group; otherwise the two groups will no longer be comparable (the control group do not have a chance to refuse), and it will not be possible to draw valid conclusions. Such trials are thus comparisons of different treatment policies. Alternatively trials can have "placebo controls," when only those subjects who give their informed consent to participate are randomised. Such studies give a direct comparison of treatments, although on a less representative group of subjects, but they are not always practical. The two approaches are discussed and illustrated in Meier's fascinating and very readable account of the Salk vaccine trial.[9]

The health visitor surveillance study had observed controls, so that all of those allocated to the observation group should be considered together. This gives five unexpected deaths out of 837, which is a rate of 6·0 per thousand, and is not nearly significantly different from the control group. The authors excluded the refusers from their analysis, giving a much larger apparent effect of observation (although still not statistically significant). In contrast, a recent study[10] comparing treatments for suspected myocardial infarction included withdrawals from the trial when analysing the data.

Another class of missing data is censored data—that is, values that cannot be measured. One common source is in the measurement of substances present in such low concentrations that some of the samples are below the sensitivity of the equipment being used. Another is where records are kept of the length of time for some event to happen (survival data) or the length of duration of some phenomenon, and the experiment is terminated before an answer can be obtained for all subjects. Censored data are clearly very different from missing observations, and must not be excluded from analysis; this would severely affect the results as these are the most extreme observations. Such data sets can be analysed by non-parametric (ranking) methods if only a few observations are censored at the same point. If censoring is at different values (as in survival studies) more rigorous statistical methods are necessary.

Conclusions

Problems with data collection are often the result of the failure at the design stage to anticipate unusual circumstances. This is one reason why large studies ought to have a pilot phase to try to spot any major deficiencies. It is because we cannot foresee everything that may be relevant that randomisation is so important, but it must be strictly adhered to.

The wide availability of computers and calculators has made it much easier to carry out statistical analyses. Unfortunately, they have also made it easy to produce results without ever really studying the raw data. Before embarking on analysis there is much that can be learnt from simple inspection of variables both singly and in pairs. Such screening of the data, especially graphically, as well as greatly helping to prepare the data for analysis, can also provide considerable insight into the relationships between variables.

The issues of data screening discussed in this article generally receive scant attention. Yet they concern strategic decisions that can have major implications for the ensuing results, as the criticism[11] of the Anturane study[12] has shown. They directly affect the validity and thus the ethics of research.

This is the fourth in a series of eight articles. No reprints will be available from the author.

References

1 "Student." The Lanarkshire milk experiment. *Biometrika* 1931;**23**: 398-406.
2 Carleton RA, Sanders CA, Burack WR. Heparin administration after acute myocardial infarction. *N Engl J Med* 1960;**263**:1002-5.
3 Peto R, Pike MC, Armitage P, *et al.* Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I Introduction and design. *Br J Cancer* 1976;**34**:585-612.
4 Clarke BF, Campbell IW. Long-term comparative trial of glibenclamide and chlorpropamide in diet-failed, maturity-onset diabetics. *Lancet* 1975;i:245-7.
5 Healy MJR. The disciplining of medical data. *Br Med Bull* 1968;**24**:210-4.
6 Armitage P. *Statistical methods in medical research.* Oxford; Blackwell, 1971:350-9.
7 Flynn FV, Piper KAJ, Garcia-Webb P, McPherson K, Healy MJR. The frequency distributions of commonly determined blood constituents in healthy blood donors. *Clin Chim Acta* 1974;**52**:163-71.
8 Carpenter RG, Emery JL. Final results of study of infants at risk of sudden death. *Nature* 1977;**268**:724-5.
9 Meier P. The biggest health experiment ever: the 1954 field trial of the Salk poliomyelitis vaccine. In: Tanur JM, Mosteller F, Kruskal WH, *et al*, eds. *Statistics: a guide to the study of the biological and health sciences.* San Francisco; Holden-Day, 1977:88-100.
10 Wilcox RG, Roland JM, Banks DC, Hampton JR, Mitchell JRA. Randomised trial comparing propranolol with atenolol in immediate treatment of suspected myocardial infarction. *Br Med J* 1980;**280**:885-8.
11 Kolata GB. FDA says no to Anturane. *Science* 1980;**208**:1130-2.
12 The Anturane Reinfarction Trial Research Group. Sulfinpyrazone in the prevention of sudden death after myocardial infarction. *N Engl J Med* 1980;**302**:250-6.

Which penicillin maintains a prolonged therapeutic concentration in the blood?

Three preparations of penicillin provide low blood concentrations for a prolonged period by slow absorption from the site of injection. Procaine penicillin gives an effective blood concentration against fully sensitive organisms for up to 24 hours after a dose of 600 mg; a single dose of benethamine penicillin gives effective concentrations for four to five days; and benzathine penicillin may last for two to three weeks or more depending on the dose given. These preparations are usually combined with benzylpenicillin to give higher initial blood concentrations, and the duration of effect will be that of the longest-acting constituent. Fortified procaine penicillin contains benzyl penicillin and procaine penicillin; Triplopen contains benethamine, procaine, and benzyl penicillins; and Penidural All Purpose contains benzathine, procaine, and benzyl penicillins. The duration of therapeutic effect will depend also on the sensitivity of the organism, the nature of the infection being treated, and the dose given. Probenecid increases the blood concentrations and prolongs the effect of penicillins. Its main use is in maintaining high concentrations of the short-acting penicillins for longer periods. It may be usefully combined with procaine penicillin to produce higher blood concentrations, but there is little to be gained from using it with preparations containing combinations of short- and long-acting penicillins.

Garrod LP, Lambert HP, O'Grady F. *Antibiotic and chemotherapy.* 4th ed. Edinburgh: Churchill Livingstone, 1973.