WHAT CAN WE DO WITH THE DATA WE THROW AWAY?

S.J.W. EVANS

Department of Biometry and Medical Statistics, The London Hospital Medical College,
Turner Street, London E1 2AD

Introduction

At the design stage of a clinical trial or experiment the elementary questions are posed—what is the effect of this treatment and does its effect differ importantly from that of another treatment or no treatment at all? Whatever sophisticated techniques may add when more analyses are carried out it is important not to lose sight of these simple questions. The data we throw away may be thrown away correctly but it may cause us to answer the simple questions differently. Real life is rarely simple and our desire for simple answers may lead us to prefer straightforward inaccuracies to complex realities. Examining more of the data may change our estimate of the effect of the treatment generally or it may enable us to say that its effect is different in different stated circumstances. We will examine the general problem of loss of information and how we can retrieve it under various headings. It may not be possible to arrive at a single answer as to the correct method of retrieval, but it is important to realise the dangers of too great a confidence in analyses which do neglect this information. If it is concluded that the data should be thrown away the wisdom of having collected it should be questioned: whilst some wastage of data may be inevitable, some may be a result of woolly thinking in the planning stage.

Lost significance

The origins of statistical significance testing were largely determined before the advent of pocket calculators by the need to limit the number and size of the tables of distributions, such as the Normal, *t*, *F* and chi-squared. Certain 'critical' values of these distributions were chosen totally arbitrarily and have since acquired an almost magical significance such as the much loved $P < 0.05$! With modern calculating facilities *P* values may be quoted exactly, although it is even better to quote a confidence interval, either 95% or 99% again by arbitrary convention. This changes the emphasis to estimation of the value of some difference, which tends to encourage appreciation of the biological as opposed to the statistical significance of the results.

Thus the use of the two letters NS (not significant) which appear in many published reports should be discouraged. The most important valid use of 'NS' is in testing and selecting a mathematical model of a set of data but this is not the usual use of significance testing as quoted in the (medical) literature.

A common situation occurs when a decision is made regarding throwing away the data gathered so far because it does not yield statistically significant results. It may be dangerous to go on and collect material under rather different conditions and then merge the data without adjustment in the analysis. It is possible to combine data, see for example Lewis & Ellis (1982) or Mantel & Haenszel (1959), and careful use of, say, two way analysis of variance, can lead to insight into what are the clinically important factors in a problem. Journal editors should encourage publication of carefully conducted research on important issues whether statistical significance is achieved or not. Criteria for acceptance should return towards care of conduct, relevance, importance and away from slavish adherence to *P* values.

It is of more help to give the actual probability level of a statistical significance test. This is frequently done if the value is 0.06 and the phrase 'nearly significant' is sometimes quoted (!) but it does no harm to give the actual value of probability computed (to one or at most two, significant figures). From the quoted '*P*-value' an approximate idea of the confidence interval for the value of the mean (or other statistic such as a difference in means) may be obtained provided the actual value of the mean is given. For example if $P = 0.025$ for a mean change in blood pressure of 12 mm Hg, this means that a 97.5% confidence interval (C.I.) just touches 0 mm Hg so the 97.5% C.I. is from 0 to 24 mm Hg. If these measurements had been made on 21 individuals (a paired *t*-test with 20 degrees of freedom), then to obtain a 95% C.I., we look up the table of the *t* distribution with 20 df for both 0.05 and 0.025. These values are 2.0860 and 2.4231 respectively. So 2.4231 times the standard error (s.e.) equals 12, so the s.e. = 4.95 and the 95% C.I. is the mean $\pm 2.0860 \times \text{SE}$ i.e. 12 ± 10.33 (1.67 to 22.33). A confidence interval itself can and in many statisticians' views, should, be presented

whether a statistically significant difference is found or not. This enables the reader with clear views on what is clinically significant to decide whether the lack of statistical significance reflects the lack of clinical importance or not. This is vital when the power of the significance test ($1-\beta$, where β is the probability of a type 2 error) is low (and hence the confidence interval is wide), such as in whether adverse reactions are different in two groups, or whether an interaction is present in a two way analysis of variance. A full discussion of the failures of 'NS' where clinically important differences may well be present is given by Freiman *et al.* (1978). The 'pragmatic' approach to clinical trials as expounded by Schwartz *et al.* (1980) does not have a problem in this area since the concept of 'not-significant' does not apply. The criterion is simply 'which treatment is better?'; not 'is this treatment significantly better than the other one?'. It may be worth considering the use of 'support limits' which provide a similar view of the likely range from a very different logical approach (see Edwards (1972) and Vere (1982)).

Lost variables

Some variables may not be reported on because they seem to show little of interest when analysed. The issues here are, firstly, why were they measured? Unnecessary measurements are expensive in time and money and very often in human resources as well. There is a tendency in Clinical Trials and other prospective investigations to attempt to collect far too much data. Some variables are measured to check that the individual patients do not show any significant pathology, and this is clearly justified. Such variables should also be analysed to check that they do not change importantly on average for the whole group, as well as for the individuals. Experience shows that the quality of data recording is in inverse proportion to the square of the number of variables recorded on each individual. Many variables are measured 'just in case' they show something of interest. Presumably at the beginning of the trial there was an intention to analyse those variables and an expectation that there would be something of importance to report on, otherwise the data would hardly have been collected. The report may point out that no clinically significant findings occurred but this should be mentioned. Secondly, even if the variables concerned cannot contribute to important measures of outcome of the trial, could they be used to obtain some information on the characteristics of patients who do respond well? It will probably require techniques of multivariate analysis (discriminant analysis particularly; for further details see Mardia *et al.* (1979) for theory, Pettit *et al.* (1980) for an example, though not in a clinical trial) to be able to identify those variables which distinguish between patients

who are likely to do well on one treatment or likely to do well on another treatment. The emphasis in trials is to make an overall comparison of treatments whereas clinical practice attempts to use the optimal treatment for a particular patient. Unless the trial has been planned to include analysis of treatment differences within separate strata as part of the protocol, the idea of attempting to find, *post hoc*, strata where the treatment difference is homogeneous within a stratum but heterogeneous between strata is frowned on. This technique does not seem to have been tried as often as it might when two active treatments are compared. It needs to be applied with caution but could be used fruitfully to generate hypotheses to be tested in further work.

Alternatively can these variables which have not been analysed as outcome variables be added to an equation to help understand what are the things that actually affect the outcome? To take a specific example it may be true that there is no difference in the mean age between two treatment groups, but it may also be true that the age of any individual has an effect on the outcome variable being analysed. In this case it may be helpful to do an analysis of co-variance taking into account the age (called the COVARIATE) of patients which will enable an improved precision (that is a reduction in the standard error) for the treatment comparison, even though ignoring the effect of age will not bias the treatment comparison that is made (Anderson *et al.*, 1980). It should be noted that if there is a difference between the two groups in, say, age and age is correlated with the outcome variable then analysis of co-variance (ANCOVA) is essential to avoid bias. There are two criteria to decide on the need for ANCOVA. The first is that there are good reasons for believing there is a relationship with a covariate and the data shows a noticeable difference between the treatment groups in the mean value of that covariate. The second is that a statistically significant difference between the groups in the mean of the covariate is found and a statistically significant association of the covariate with the outcome variable is found. The level of significance may need to be $P = 0.1$ rather than $P = 0.05$ if this is the criterion used. Strict rules do not need to be applied to decide if ANCOVA is necessary since if the relevant association and difference are not statistically significant then the adjustment will be small.

The 'covariate' argument can also apply to a variable whose measurement scale is categorical when two way analysis of variance rather than ANCOVA is necessary. Care needs to be taken in interpreting such analyses when the design of the trial has not allowed for balance in the additional categorical variables.

Residuals and transformations

In many instances in reporting a clinical trial and in the process of its analysis, a treatment effect is looked

for and if found the investigator(s) will be content to report it as it stands, without further investigation. In the event that they do not find a significant treatment effect, they may well, if they believe there should be a treatment effect, put rather more energy into looking for other ways of analysing the data in order to find some difference between the two treatments under consideration. The effort put into analysis should be independent of whether one has been successful in finding a treatment effect or not. For example, further examination of the data may show that bias in a covariate, as discussed above, explains the apparent effect of the treatment. Having found some variables in the data which explain a great deal of the variation in it we can then use the equations that we have developed, what the statistician calls the 'model', in order to obtain a predicted value for the outcome variable(s) of interest. The difference between this predicted value and the observed value which was measured in a particular instance is in general terms a residual. Thus we can have a residual when we have done a simple between groups *t*-test. Here the residual is the difference between the mean for the relevant treatment group and the value for each of the individuals in that treatment group.

If a series of measurements of some variable (e.g. blood pressure), has been at monthly intervals on several patients, a first set of residuals may be obtained by subtracting the median value for each patient from that patient's measurements. (The mean can equally be used but it is easier to use the median in hand calculations and has advantages if the data are prone to occasional errors of measurement). It is enlightening to plot not only the distribution of those residuals, (for example, with a *t*-test the distribution should be approximately Gaussian) but also to plot those residuals against some other variables which might possibly affect them. The shape of the graphs so obtained can lead to further insights into the data, for instance pointing out that the relationship was not linear and that fact on its own may explain why the first analysis did not find a relationship. Transformation of the data (e.g. by taking logs) may then be seen to be necessary. Useful discussions of residuals are found in Daniel & Wood (1981) and Draper & Smith (1980). For a full discussion of transformations see Box & Cox (1964) while for elementary discussion see Gore & Altman (1982). J.W. Tukey has said, 'forget the fit, just look at the residuals'. In other words the interest in scientific terms in the data lies in the unexpected patterns and not those which could have been predicted beforehand.

Statistical tests may be applied to the residuals so obtained, but a very useful technique which does not require a statistical significance test is that of cumulative sums (Cusums). (There are significance tests for cusums but these are unnecessary here, van Dobben der Bruyn (1968)). It is reasonable to obtain the

cusum of residuals and to plot this against the variable of interest. The cusum is obtained simply by adding up all the residuals in logical order (e.g. chronologically). The resulting plot is an extremely sensitive indicator of whether there have been any changes with time which do not always show up in the original plots. The residuals themselves can be plotted on normal probability paper or using the inverse normal distribution to examine whether the assumption of a Gaussian distribution is a reasonable one, and this can be used to check whether any outlying values are reasonable given some assumptions about the data. It can also be a good means of detecting errors that have not been corrected earlier in the analysis. An extensive series of plots of residuals are shown in Draper & Smith (1980). A particularly useful computer program for use with regression models (note that analysis of variance or co-variance and even a *t*-test can be set up as regression models) is BMD-P9R (Dixon, 1981). This program includes a full residual analysis including Cook's distance which indicates how influential a point is in altering the form of the regression equation, Cook (1977). It gives the regression equation with the most distant point omitted (Jackknife), see Tukey (1977), which is often a salutary lesson in small data sets. Some regressions are statistically significant only because of a single data point at an extreme value of the independent variable and occasionally a regression which was statistically 'not significant' becomes 'significant' when a single point is omitted. Similarly the use of transformations may result in biologically more comprehensible models as well as reducing the problem of outliers. *Post-hoc* discoveries of this nature do not demand that the point is omitted but the sensitivity of the analysis to single points warns the alert investigator to be guarded in drawing conclusions based solely on 'P-values'. The use of 'Robust' methods, such as in Huber (1981) could well be employed and deserve fuller recognition in the whole field of analysis of medical data.

The use of sophisticated criteria to decide on the rejection of extreme values (outliers) is not really to be recommended. Such techniques as 'Winsorisation' are preferred, where extreme values are replaced by less extreme ones, but even here there can be concern over what is happening. The use of 'Robust' methods does not reject outliers in general, but assigns smaller weights to them than to 'inliers'. (Rejecting them implies giving a weight of zero to them and a weight of one to all others. Robust methods effectively allow for a graduated scale of weights).

The theory is now relatively well developed but the practice is dependent on more widespread knowledge of their utility and on computer programs to implement them becoming more generally available than at present. This situation is changing rapidly however and work by Marazzi (1980), Barnett & Lewis (1978)

gives good reason to expect that their availability and usage will increase in the next few years.

'Missing' patients

There are two principal reasons why data on individual patients are lost. The first of these is that there is data missing on a particular occasion for a patient who remains in the trial. The second is because a patient is no longer in the trial. A withdrawal occurs when a patient has had an unwanted effect, or shown deterioration in health which has led the clinician in charge to withdraw the patient from the trial. A second type of withdrawal, more properly called a loss, is when a patient ceases to be in the trial for a reason which is unassociated with the treatment, or the effects of the treatment on their general well-being. Such a loss might be because the patient has moved away, or because they have had some other event in their lives unassociated with the condition or the treatment which has meant that they cannot continue in the trial. Many analyses ignore patients whose data are incomplete, but this is unsatisfactory, and the following sections attempt solutions to the problem (see also Chaput de Saintonge & Vere, 1982).

Missing data

There are two reasons why the data is missing for a patient, the first being that the patient did not attend on this particular occasion to be able to have the measurements made, but subsequently the patient continued to have measurements made and was not withdrawn from the trial or experiment. Secondly the patient may have been available, but the individual result was lost or a particular blood sample was not taken so that the data is missing even though the patient was seen at that particular time. This missing data may be distributed randomly, in the sense that the actual value which is missing may not have any association with the reason for it being missing, or it may be associated with the reason for it being missing. If a patient tends to be at one extreme or another then it may be that a particular instrument was unable to register that value so that the data is missing. Ideally the fact that the data was at one extreme or another is recorded but in practice not always so! It may be the case that patients who do not turn up at a clinic for intermediate measurements are non-complaint and do not wish to be asked about compliance, or feeling very well and therefore see no need to visit their doctors, or feeling so ill that travel to a clinic is impossible.

What can be done about the missing data? Assuming for the moment that the data is missing in a random fashion it is reasonable to attempt to estimate

the missing values. Any use of missing value estimation must be clearly stated with exact methods and the consequences for the analysis. One particularly useful computer program for examining whether data is indeed missing in a random fashion is BMD-PAR (Dixon *et al.*, 1981), which, having confirmed that it is random as far as can be seen from the data, is able to go on and estimate the missing values. The methods available for estimating missing values are many and may be regarded as being essentially of two types. The first is to use a data value which is representative of all the data so that the value may be replaced quite simply by the mean (or median) value for that variable across all patients and all treatment groups. The second approach is to use the knowledge about the other data from that patient to help obtain a better estimate than the overall mean for that data value. For example the mean or median value for that individual for this variable may be used. If all other patients have a different mean on this occasion compared with others then this difference can be used to adjust the estimated value using both that individual's data on other occasions and other individuals' data on this occasion. An even more sophisticated approach is to try and use other variables which correlate well with the variable which is missing provided that they are available for the case in question. A regression equation is obtained using the variable which is missing as the dependent variable and taking as few other independent variables as possible to obtain the equation. The equation is calculated from those individuals whose data is present and then the independent variables for the missing case are used to predict the missing value. In the event of the value being biased or in the event of the data being missing in a non random way then there are dangers in this method and the whole analysis may be very complex if not impossible. A possible approach is to use the missing value estimation technique as if the value were unbiased and then see the effect on the analysis of allowing the estimated value to increase or decrease in the direction in which the biases are thought to have occurred. It will be difficult to justify presentation of an analysis in which the use of values estimated by such techniques were made, but if the result of the analysis is not changed by using such extreme values then the investigator is very much more confident that the analysis is robust to such exaggerated assumptions about the nature of the missing data. On the other hand, if the result of the analysis is changed noticeably by using missing value estimation techniques and, where appropriate, allowing for possible bias in one direction or another, the investigator must in the reports warn that the analysis may be biased because of the missing data. It takes some courage to state that the trial is not able to be used to draw firm conclusions because of the missing data.

The usual approach of investigators is not nearly as

rigorous as this and it could be said that there is no gain in doing such 'fancy' analyses since if they don't change the conclusions it wasn't worth it and if they do no-one is prepared to say so. It seems to me that unless scientists are prepared to state the limitations of their studies there will be a perpetuation of pseudoscience and the value of clinical trial reports will continue to decline. It is perhaps an explanation for some of the contradictory reports which appear in the literature.

Withdrawals

It is vital that those who are seen as being withdrawn for treatment-related reasons play a full part in the analysis, otherwise the analysis may be severely biased, and an example of such a bias is discussed by Pocock (1982).

Withdrawals may be treated in the same way as missing values in order to be able to be confident that the result of the analysis of the trial is not dependent on those values having been left out. A better method is when treatment failure or success can be used as the outcome variable so that a withdrawal can be included in the analysis and regarded as a treatment failure. In long-term studies this may be particularly relevant if the time between entry to the trial and occurrence of a failure is itself of interest. In this case the methods of analysis using life tables, which are conventionally used for analysing survival, can be used for data which consists of treatment success or treatment failure. Losses which are neither treatment successes nor treatment failures are handled naturally in this type of analysis. They are included in the analysis until the time they are 'lost' from the trial (they are at risk of failing until then) and their time in the trial is regarded as 'censored'. The papers by Peto *et al.* (1976, 1977) give an excellent description of the methods of design and analysis involving long-term follow-up of patients. If the outcome is neither categorical nor 'time in the trial', then withdrawals are not handled as easily. It may then be important to attempt to make estimates of the value that someone would have had at the end of the trial had they remained in the trial. For example a trial may show that one treatment shows a slight and steady improvement and another treatment shows a fairly dramatic deterioration in a subset of the patients but not much effect in the others. Those patients in whom the dramatic deterioration is seen will be withdrawn from the trial. If the comparison is then made using only those patients completing the trial it is very severely biased. A first alternative is to carry forward the last measurement made on that patient through to subsequent occasions so that if comparisons are made between the beginning and end of the trial then the value that is used for the person who has been withdrawn during the trial is the last available measure-

ment for them. For an example of this in a trial report see Chaput de Saintonge *et al.* (1982). If there has been a trend which can be seen to be linear it may even be reasonable to use the regression approach for an individual patient in order to obtain a predictive value at the end of the trial. As mentioned above it may not be totally justified to present the exact results of the analysis using somewhat arbitrarily obtained estimated values at the end of the trial. However if the result of the analysis is changed by this method, this fact should be presented and the conclusions from the analysis of the trial have to be seriously weakened. The particular techniques that I would recommend for analysing such trends within a patient are those of non-parametric regression. An example of such methods may be found in Conover (1980) or Hollander & Wolfe (1978). A median regression line is obtained by looking at the slopes of all lines between all possible pairs of points and picking the median slope of these lines. This has the effect of being insensitive to single outlying values and for clinical data will provide a more robust regression line than the conventional least squares approach. There are other more sophisticated methods which may be found in Huber (1981) as discussed above.

Lost measurements

It is these data which are most frequently thrown away in a clinical trial. For example six measurements made at monthly intervals during a single treatment period in a clinical trial are made, but the only measurements that are reported on and analysed are the beginning and end measurements. What can be done about using the data at the intermediate time periods in order to be able to improve our understanding of what is happening in the trial? There are several possible approaches to this problem. The first is to look for a time trend for each individual patient and may be called the 'regression' approach. The second is to look for consistent time differences rather than a particular trend over time, which may be called the 'repeated measures analysis of variance' approach. The third is to try the methods of time series analysis. For all of these approaches the first step is to examine the residuals when individual variation has been taken out as mentioned above.

What is not permitted is to treat these observations as independent. With some computer programs it is particularly easy to do an incorrect analysis testing each of the six visits at which measurements are made for each patient as a 'case' or observation unit. If there are twenty patients in each group it is not possible to do a *t*-test using 120 observations versus 120 observations, which is what may be done by the computer program if the data is incorrectly structured for the analysis. Even if six *t*-tests are done with 20 values

v 20 values at each, adjustment of the P values needs to be made.

The regression approach

Taking the differences between the first and last measurements in a trial is essentially obtaining a regression line through two points! A technique sometimes applied erroneously in a trial (and frequently done in other kinds of experiments) is to obtain a regression with time using all the observations made. This approach is wrong in the same way as the t -test between two groups of 120 observations. This is not only statistically invalid but can lead to very wrong conclusions. It is statistically invalid because the variation within an individual is likely to be less than that between individuals. The successive measurements on an individual are generally correlated with one another rather than being independent which is the assumption required for regression. Consequently a separate regression line must be obtained for each individual. The slope obtained for each individual could then be used in a simple statistical test to compare two or more treatment groups. The book by Oldham (1968) gives a fully illustrated discussion of this point in Chapter 8.

An alternative which is slightly more complex is to use a weighted analysis. This gives more weight to those slopes which have all data values close to the regression line than to those with greater scatter. (The weight is in inverse proportion to the variance about the regression line). Residuals from the regression must be examined for each patient in the trial and for all patients taken together. A 'General Linear Model' described by Nelder & Wedderburn (1972), may be used to do this analysis using all the data with appropriate tests for commonality of the slopes within and between treatment groups. Such a model is linear in the parameters but can take account of non-linearity in the variables. For example a polynomial or a logarithmic transformation can be used and the residuals may have a named distribution which is not necessarily Gaussian. The computer program GLIM by Baker & Nelder (1978) is the best source of such a general approach.

A practical technique which is, however, not well-loved by many statisticians is to use non-parametric regression as mentioned above to obtain the slope for each patient and to use the slope so obtained in a simple comparison of treatments. Again the residuals should be examined as in any of the methods of fitting the data. This method deals with missing data from whatever cause but does not work well when very few (less than 4) points are available. Although it may be better than just using two points at the beginning and end of the trial.

Repeated measures analysis of variance and covariance

This method is an adjustment of the standard methods of Analysis of Variance (ANOVA) or Analysis of Covariance (ANCOVA) to allow for the lack of independence of serial measurements made on the same individual. It can be used to analyse crossover trials as well as parallel group trials. Where there are no missing data and the design is balanced over time the computer program BMD-P2V is very good, while for unbalanced and missing data BMD-P3V is able to carry out the analysis.

More details of this method are given in Winer (1971). It is worth remarking that analysis of unbalanced data (whether the imbalance is due to missing data or to imbalance in the design) is fraught with danger as has been said in relation to missing data. However in dealing with clinical data it is rare to find perfect balance and the statistician who merely says 'dangerous—do not analyse' is of little help to the clinician. It is better to proceed in co-operation or the Clinician may go his or her own way blissfully unaware of the possibilities of misleading conclusions.

Conclusions

The use of the word 'data' in a clinical trial is etymologically suspect; it is never 'given', except to a statistician, it is more usually won with effort, tears, sweat and if not the blood of the investigator generally that of the patient!

There remains considerable scope for complete investigation of it as well as the call to minimise its unnecessary collection. When a trial is seen in the context of aiding clinical decision making (Spicer, 1982) the data which do not yield significant results on their own might be combined with data from other sources in an informal, if not a formal way using Bayesian methods. This approach might make exaggerated claims for treatment success less likely when the success is dependent on biased selection of a sample, but may also allow treatment success to be seen when it has not reached statistical significance in several trials which are too small. For an example of pooling data across trials see Baber & Lewis (1982). It has even been suggested (Vere, personal communication) that everyone should show the data they plan to throw away to an independent party who might suggest uses for it which had been missed by the owner!

References

- ANDERSON, S., AUQUIER, A., HAUCK, W.W., OAKES, D., VANDAELE, W. & WEISBERG, H.I. (1980). *Statistical Methods of Comparative Studies*. New York: Wiley.
- BABER, R.J. & LEWIS, J.A. (1982). Confidence in results of beta-blocker postinfarction trials. *Br. med. J.*, **284**, 1749–1750.
- BAKER, R.J. & NELDER, J.A. (1978). *The GLIM System Release 3*. Oxford: Numerical Algorithms Group.
- BARNETT, V. & LEWIS, T. (1978). *Outliers in Statistical Data*. New York: Wiley.
- BOX, G.E.P. & COX, D.R. (1964). An Analysis of Transformations. *J. Roy. Statist. Soc. B.*, **26**, 211–252.
- CHAPUT DE SAINTONGE, D.M., LEVIND, D.F., TEMPLE SAVAGE, I., BURGESS, G.W.S., SHARP, J., MAYHEW, S.R., SADLER, M.G., MOODY, R., GRIFFITHS, R., GRIFFITHS, S. & MEADOWS, G. (1982). Trial of three-day and ten-day courses of amoxycillin in otitis media. *Br. med. J.*, **284**, 1078–1081.
- CHAPUT DE SAINTONGE, D.M. & VERE, D.W. (1982). Measurement in clinical trials. *Br. J. clin. Pharmacol.*, **13**, 775–783.
- CONOVER, W.F. (1980). *Practical Nonparametric Statistics*. New York: Wiley.
- COOK, R.D. (1977). Detection of influential observations in linear regression. *Technometrics.*, **19**, 15–18.
- van DOBBEN DE BRUYN, E. (1968). *Cumulative Sum Tests*. London: Griffin.
- DANIEL, C. & WOOD, F.S. (1980). *Fitting Equations to Data*, 2nd Ed. New York: Wiley.
- DIXON, W.J. (1981). *BMDP Statistical Software*. Berkeley: University of California Press.
- DRAPER, N.R. & SMITH, H. (1981). *Applied Regression Analysis* 2nd ed. New York: Wiley.
- EDWARDS, A.W.F. (1972). *Likelihood*. Cambridge: Cambridge University Press.
- FREIMAN, J.A., CHALMERS, T.C., SMITH, H. & KUEBLER, R.R. (1978). The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *New Engl. J. Med.*, **298**, 690–694.
- GORE, S.M. & ALTMAN, D.G. (1982). *Statistics in practice*. London: British Medical Association.
- HOLLANDER, M. & WOLFE, D.A. (1973). *Nonparametric statistical methods*. New York: Wiley.
- HUBER, P.J. (1981). *Robust statistics*. New York: Wiley.
- LEWIS, J.A. & ELLIS, S.H. (1982). A statistical appraisal of post-infarction beta-blocker trial. *Primary Cardiology*, Suppl. No 1.
- MANTEL, N. & HAENSZEL, W. (1959). Statistical aspects of analysis of data from retrospective studies. *J. Nat. Cancer Inst.*, **22**, 719–748.
- MARAZZI, A. (1980). ROBETH: a subroutine library for robust statistical procedure. Compstat. Proceedings of Computational Statistics, Eds Barritt, M.M. & Wishart, D. Vienna: Physica-Verlag.
- MARDIA, K.V., KENT, J.T. & BIBBY, J.M. (1979). *Multivariate analysis*. New York: Academic Press.
- NELDER, J.A. & WEDDERBURN, R.W.M. (1972). Generalized linear models. *J. Roy. Statist. Soc. A*, **135**, 370–384.
- OLDHAM, P.D. (1968). *Measurement in medicine*, pp 190–202. London: English Universities Press.
- PETO, R., PIKE, M.C., ARMITAGE, P., BRESLOW, N.E., COX, D.R., HOWARD, S.V., MANTEL, N., McPHERSON, K., PETO, J. & SMITH, P.G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of the patient (I). *Br. J. Cancer*, **34**, 585–611.
- PETO, R., PIKE, M.C., ARMITAGE, P., BRESLOW, N.E., COX, D.R., HOWARD, S.V., MANTEL, N., McPHERSON, K., PETO, J. & SMITH, P.G. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of the patient (II). *Br. J. Cancer*, **35**, 1–39.
- PETTIT, B.R., KING, G.S. & EVANS, S. (1980). The potential of multivariate discriminant analysis in the antenatal detection of neural tube defects. *Clin. Chim. Acta*, **102**, 191–198.
- POCOCK, S.J. (1982). Statistical aspects of clinical trial design. *Statistician*, **31**, 1–18.
- SCHWARTZ, D., FLAMANT, R. & LELLOUCH, J. (1980). *Clinical trials*. New York: Academic Press.
- SPICER, C. (1982). Statistical decision theory and clinical trials. *Br. J. clin. Pharmacol.*, **14**, (in press).
- TUKEY, J.W. (1977). *Exploratory data analysis*. Reading, Mass: Addison-Wesley.
- VERE, D.W. (1982). Logic and inference. *Br. J. clin. Pharmacol.*, **14**, 171–176.
- WINER, B.J. (1971). *Statistical principles in experimental design* 2nd ed. Tokyo: McGraw-Hill Kogakusha.