14 Statistical guidelines for contributors to medical journals

DOUGLAS G ALTMAN, SHEILA M GORE, MARTIN J GARDNER, STUART J POCOCK

Introduction

Most papers published in medical journals contain analyses that have been carried out without any help from a statistician. Although nearly all medical researchers have some acquaintance with basic statistics, there is no easy way for them to acquire insight into important statistical concepts and principles. There is also little help available about how to design, analyse, and write up a whole project. Partly for these reasons much that is published in medical journals is statistically poor or even wrong.[1] A high level of statistical errors has been noted in several reviews of journal articles and has caused much concern.[2,3]

Few journals offer even rudimentary statistical advice to contributors. These guidelines (originally published in 1983) followed suggestions[1,4] that comprehensive statistical guidelines could help by making medical researchers more aware of important statistical principles, and by indicating what information ought to be supplied in a paper. Since our original article, Bailar and Mosteller published guidelines amplifying the brief section on statistics in the "Uniform requirements for manuscripts".[5,6] Other authors have since published guidelines for particular types of study.[7–12] Lang and Secic have published very comprehensive guidance.[13]

Deciding what to include in the guidelines, how much detail to give, and how to deal with topics where there is no consensus was problematic. These guidelines should thus be seen as one view of what is important, rather than as a definitive document. We did not set out to provide a set of rules but rather to give general information and advice about important aspects of statistical

design, analysis, and presentation. Those specific recommendations that we have made are mostly strong advice against certain practices.

Some familiarity with statistical methods and ideas is assumed, since some knowledge of statistics is necessary before carrying out statistical analyses. For those with only a limited acquaintance with statistics, the guidelines should show that the subject is very much wider than mere significance testing and illustrate how important correct interpretation is. The lack of precise recommendations in some places indicates that good statistical analysis requires common sense and judgement, as well as a repertoire of formal techniques, so that there is an art in statistics as well as in medicine. We hope that the guidelines present an uncontroversial view of the most frequently used and accepted statistical procedures. We have deliberately limited the scope of the guidelines to cover the more common statistical procedures. The version presented here incorporates a few additions to the original version.

Readers may find that a relevant section presents information or advice that is unfamiliar or is not understood. In such circumstances, although almost all of the topics covered may be found in the more comprehensive medical statistics textbooks,[14–19] we strongly recommend that they should seek the advice of a statistician. The absence from the guidelines of specific references is intentional: it is better to get expert personal advice if further insight is needed. Moreover, because mistakes in design cannot later be rectified, professional advice should first be obtained when planning a research project rather than when analysing the data.

These guidelines are intended to try to help authors know what is important statistically and how to present it in their papers. They emphasise that such matters of presentation are closely linked to more general consideration of statistical principles. Detailed discussion of how to choose an appropriate statistical method is not given; such information is best obtained by consulting a statistician. We do, however, draw attention to certain misuses of statistical methods.

These guidelines follow the usual structure of medical research papers: Methods, Results (analysis and presentation), and Discussion (interpretation). As a result, several topics appear in more than one place and are cross-referenced as appropriate. Statistical checklists (chapter 15) indicate the broad categories of information that should be included in a paper.

Methods section

General principles

It is most important to describe clearly what was done, including the design of the research (be it an experiment, trial, or survey) and the collection of the data. The aim should be to give enough information to allow methods to be fully understood and, if desired, repeated by others. As noted by the International Committee of Medical Journal Editors, authors should "describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results".[6]

Authors should include information on the following aspects of the design of their research:

- the objective of the research, and major hypotheses;
- the type of subjects, stating criteria for inclusion and exclusion;
- the source of the subjects and how they were selected;
- the number of subjects studied and why that number of subjects was used;
- the types of observation and the measurement techniques used (where several assessments are made for each subject, the main focus of interest should be specified).

Each type of study—for example, surveys and clinical trials—will require certain additional information.

Surveys (observational studies)

The study design should be clearly explained. For instance, the selection of a control group and any matching procedures need detailed description. It should also be clearly stated whether the study is retrospective, cross-sectional, or prospective. The procedure for selecting subjects and the achievement of a high participation rate are particularly important, as findings are usually extrapolated from the sample to some general population. It is helpful to report any steps taken to encourage participation in the survey.

Clinical trials

The treatment regimens (including ancillary patient care and criteria for modifying or stopping treatment) need detailed definition. The method for allocating treatments to subjects should be

stated explicitly. In particular, the specific method of randomisation (including any stratification) and how it was implemented need to be explained. Lack of randomisation should be noted as a deficiency in design and the reasons given. Studies using deterministic allocation methods, for example based on hospital number or alteration, are not truly randomised and are unlikely to be acceptable to the *BMJ*[20] or other leading medical journals.

The use of blinding techniques and other precautions taken to ensure an unbiased evaluation of patient response should be described. The main criteria for comparing treatments, as agreed in the trial protocol, should be listed. For crossover trials the precise pattern of treatments (and any run in and wash out periods) needs explaining.

A more comprehensive list of information to include in the report of a clinical trial is given in the checklist in chapter 15. Many leading medical journals now require authors to comply with the CONSORT recommendations for reporting controlled trials.[12]

Statistical methods

All the statistical methods used in a paper should be identified. When several techniques are used it should be absolutely clear which method was used where, and this may need clarification in the results section. Common techniques, such as t tests, simple χ^2 tests, Wilcoxon and Mann–Whitney tests, correlation (r), and linear regression, do not need to be described, but methods with more than one form, such as t tests (paired or unpaired), analysis of variance, and rank correlation, should be identified unambiguously. More complex methods do need some explanation, and if the methods are unusual a precise reference should be given, preferably to a textbook (with page numbers). It may help to include brief comments on why the particular method of analysis was used, especially when a more familiar approach has been avoided. It is useful to give the name of a computer program or package used—for example, the Statistical Package for the Social Sciences (SPSS)—but the specific statistical methods must still be identified.

Results section: statistical analysis

Descriptive information

Adequate description of the data should precede and complement formal statistical analysis. In general variables which are

important for the validity and interpretation of subsequent statistical analyses should be described in most detail. This can be achieved by graphical methods, such as scatter plots or histograms, or by using summary statistics. Continuous variables (such as weight or blood pressure) can be summarised using the mean and standard deviation (SD) or the median and a percentile range—say, the interquartile range (25th to 75th percentile). The latter approach is preferable when continuous measurements have an asymmetrical distribution. The standard error (SE) is not appropriate for describing variability. For ordered qualitative data (such as stages of disease I to IV) the calculation of means and standard deviations is incorrect; instead, proportions should be reported.

Deviations from the intended study design should be described. For example, in clinical trials it is particularly important to enumerate withdrawals with reasons, if known, and treatment allocation. For surveys, where the response rate is of fundamental importance, it is valuable to give information on the characteristics of the non-responders compared with those who took part. The representativeness of the study sample will need to be investigated if it is a prime intention to extrapolate results to some appropriate population.

It is useful to compare the distribution of baseline characteristics in different groups, such as treatment groups in a randomised trial. Such differences that exist, even if not statistically significant, are real and should be properly allowed for in the analysis (see ''Complex analyses'', below). Such tests assess only the integrity of the randomisation, not whether the groups are comparable.

Underlying assumptions

Methods of analysis such as *t* tests, correlation, regression, and analysis of variance all depend to some extent on certain assumptions about the distribution of the variable(s) being analysed. Technically, these assumptions are that in some aspect the data come from a Normal distribution and if two or more groups are being compared that the variability within each is the same.

It is not possible to give absolutely the degree to which these assumptions may be violated without invalidating the analysis. But data which have a highly skewed (asymmetrical) distribution or for which the variability is considerably different across groups may require either some transformation before analysis (see ''Data transformation'', below) or the use of alternative

''distribution free'' methods, which do not depend on assumptions about the distribution (often called non-parametric methods). For example, the Mann–Whitney U test is the distribution free equivalent of the two-sample *t* test. Distribution-free methods may also be appropriate for small data sets, for which the assumptions cannot be validated adequately.

Sometimes the assumption of Normality may be especially important—for example, when the range of values calculated as two standard deviations either side of the mean is taken as a 95% ''normal range'' or reference interval. In such cases the distributional assumption must be shown to be justified.

Hypothesis tests

The main purpose of hypothesis tests (often less accurately referred to as ''significance tests'') is to evaluate a limited number of preformulated hypotheses. Other tests, which are carried out because they have been suggested by preliminary inspection of the data, will give a false impression because in such circumstances the calculated P value is too small. For example, it is not valid to test the difference between the smallest and largest of a set of several means or proportions without making due allowance for the reason for testing that particular difference; special ''multiple comparison'' techniques are available for making pairwise comparisons among several groups. However, where three or more groups are compared which have a natural ordering, such as age groups or stages of cancer, the data should be analysed by a method that specifically evaluates a trend across groups.

It is customary to carry out two-sided hypothesis tests. If a one-sided test is used this should be indicated and justified for the problem in hand.

The presentation and interpretation of results of hypothesis tests are discussed in later sections. The use of confidence intervals in addition to hypothesis tests is strongly recommended—see next section and chapters 1 and 3.

Confidence intervals

Most studies are concerned with estimating some quantity, such as a mean difference or a relative risk. It is desirable to calculate the confidence interval around such an estimate. This is a range of values about which we are, say, 95% confident that it includes

the true value. There is a close relation between the results of a test of a hypothesis and the associated confidence interval: if the difference between groups is significant at the 5% level then the associated 95% confidence interval excludes the zero difference. The confidence interval conveys more information because it indicates a range of values for the true effect which is compatible with the sample observations (see also "Interpretation of hypothesis tests", below, and chapter 3).

Confidence intervals reveal the precision of an estimate. A wide confidence interval points to lack of information, whether the difference is statistically significant or not, and is a warning against overinterpreting results from small studies.

In a comparative study, confidence intervals should be reported for the differences between groups, not for the results of each group separately.

Paired observations

It is essential to distinguish the case of unpaired observations, where the comparison is between measurements for two different groups—for example, subjects receiving alternative treatments—from that of paired observations, where the comparison is between two measurements made on the same individuals in different circumstances (such as before and after treatment). For example, where with unpaired data the two sample t test would be used, with paired data the paired t test should be used instead. Similarly, the Mann–Whitney U test for unpaired data is replaced by the paired Wilcoxon test, and the usual χ^2 test for 2×2 tables is replaced by McNemar's test. It should always be made clear which form of test was used. Likewise the method for calculating a confidence interval differs from that for unpaired observations (see chapters 4, 5, 6, and 7).

The same distinction must be made when there are three or more sets of observations. All of the statistical methods mentioned in this section may be generalised to more than two groups; in particular, paired and two-sample t tests generalise to different forms of analysis of variance.

Units of analysis

Often several measurements are made on the same patient, but the focus of interest usually remains the patient. The simplest case is when researchers study a part of the human anatomy

which is in duplicate, such as eyes, but sometimes very many measurements can be taken on a single patient. Multiple counting of individual patients can lead to seriously distorted results. In particular, it inflates the sample size and may lead to spurious statistical significance. The patient is the unit of investigation and thus should be the *unit of analysis*. (A related issue is discussed in the following section.)

By contrast, groups are sometimes the focus of interest. For example, in a ''cluster'' randomised trial groups such as hospital wards or general practices may be randomised to different interventions. In such studies it is wrong to analyse data for individual patients as if they were independent observations. Here the cluster is the correct unit of analysis.

Repeated measurements

A common study design entails recording serial measurements of the same variable(s) on the same individual at several points in time. Such data are often analysed by calculating means and standard deviations at each time and presented graphically by a line joining these means. The shape of this mean curve may not give a good idea of the shapes of the individual curves. Unless the individual responses are very similar it may be more valuable to analyse some characteristic of the individual profiles, such as the time taken to reach a peak or the length of time above a given level. This would also help to avoid the problems associated with multiple hypothesis testing (see ''Many hypothesis tests'', below).

Repeated measurements of the same variable on one individual under the same experimental conditions, known as *replicate readings*, should not be treated as independent observations when comparing groups of individuals. Where the number of replicates is the same for all subjects analysis is not difficult; in particular, analysis of variance is used where t tests would have been applied to unreplicated data. If the number of replicates varies among individuals, a full analysis can be very complex. The use of the largest or smallest of a series of measurements (such as maximum blood pressure during pregnancy) may be misleading if the number of observations varies widely among individuals.

Data transformation

Many biomedical variables have distributions which are positively skewed, with some very high values, and they may require

mathematical transformation to make the data appropriate for analysis. In such circumstances the logarithmic (log) transformation is often applicable, although occasionally other transformations (such as square root or reciprocal) may be more suitable.

After analysis it is desirable to convert the results back into the original scale for reporting. In the common case of log transformation, the antilog of the mean of the log data (known as the ''geometric mean'') should be used. The standard deviation or standard error must not be antilogged, however; instead, confidence limits on the log scale can be antilogged to give a confidence interval on the original scale. A similar procedure is adopted with other transformations when there is a single sample, but back transformation of the confidence limits for a difference between sample means makes sense only for the log transformation (see chapter 4).

If a transformation is used it is important to check that the desired effect (such as an approximately Normal distribution) is achieved. It should not be assumed that the log transformation, for instance, is necessarily suitable for all positively skewed variables.

Outliers

Observations that are highly inconsistent with the main body of the data should not be excluded from the analysis unless there are additional reasons to doubt their credibility. Any omission of such outliers should be reported. Because outliers can have a pronounced effect on a statistical analysis, it can be useful to analyse the data both with and without such observations to assess how much any conclusions depend on these values.

Correlation

It is preferable to include a scatter plot of the data for each correlation coefficient presented, although this may not be possible if there are several variables. When many variables are being investigated it is useful to show the correlations between all pairs of variables in a table (correlation matrix), rather than quoting just the largest or significant values.

For data which are irregularly distributed the rank correlation can be calculated instead of the usual Pearson ''product moment'' correlation (r). Rank correlation can also be used for variables that are constrained to be above or below certain

values—for example, birth weights below 2500 g—or for ordered categorical variables. Rank correlation is also preferable when the relation between the variables is not linear, or when the values of one variable have been chosen by the experimenter rather than being unconstrained.

The correlation coefficient is a useful summary of the degree of linear association between two quantitative variables, but it is one of the most misused statistical methods. There are several circumstances in which correlation ought not to be used. It is incorrect to calculate a simple correlation coefficient for data which include more than one observation on some or all of the subjects, because such observations are not independent. Correlation is inappropriate for comparing alternative methods of measurement of the same variable because it assesses association, not agreement. The use of correlation to relate change over time to the initial value can give grossly misleading results.

It may be misleading to calculate the correlation coefficient for data comprising subgroups known to differ in their mean levels of one or both variables—for example, combining data for men and women when one of the variables is height.

Regression and correlation are separate techniques serving different purposes and need not automatically accompany each other. The interpretation of correlation coefficients is discussed below (''Association and causality'').

Regression

It is highly desirable to present a fitted regression line together with a scatter diagram of the raw data. A plot of the fitted line without the data gives little further information than the regression equation itself. It is useful to give the values of the slope (with its standard error) and intercept and a measure of the scatter of the points around the fitted line (the residual standard deviation). A confidence interval may be constructed for a regression line and prediction intervals for individuals based on the fitted relationship. The lines joining these values are not parallel to the regression line but curved, showing the greater uncertainty of the prediction corresponding to values on the horizontal (x) axis away from the bulk of the observations (see chapter 8).

Regression on data including distinct subgroups can give misleading results, particularly if the groups differ in their mean level of the dependent (y) variable. More reliable results may be obtained by using analysis of covariance (see chapter 8).

Regression and correlation are separate techniques serving different purposes and need not automatically accompany each other. The interpretation of regression analysis is discussed below (''Prediction and diagnostic tests'').

Survival data

The reporting of survival data should include graphical or tabular presentation of life tables, with details of how many patients were at risk (of dying, say) at different follow up times (see chapter 9). The life table or actuarial survival curve deals efficiently with the ''censored'' survival times which arise when patients are lost to follow up or are still alive; their survival time is known to be only at least so many days. To avoid misinterpretation of the unreliable later part of the curve, it may help to truncate the survival curve when there are only a few (say five) subjects still at risk. The calculation of mean survival time is inadvisable in the presence of censoring and because the distribution of survival times is usually positively skewed.

Comparison between treatment groups of the proportion surviving at arbitrary fixed times can be misleading and is generally less efficient than the comparison of life tables by a method such as the logrank test. Methods for calculating estimates of survival and confidence intervals are given in chapter 9.

When there are sufficient deaths one can show how the risk of dying varies with time by plotting, for suitable equal time intervals, the proportion of those alive at the beginning of each time interval who died during that interval. Adjusting for patient factors which might influence prognosis is possible using regression models appropriate to survival data (see next section).

Comparison of survival between the group of individuals who respond to treatment and the group who do not is misleading and should never be performed.

Complex analyses

In many studies the observations of prime interest may be influenced by several other variables. These might be anything that varies among subjects and which might have affected the outcome being observed. For example, in clinical trials they might include patient characteristics or signs and symptoms. Some or all of the covariates can be combined by appropriate multiple regression techniques to explain or predict an outcome variable, be it a

continuous variable (blood pressure), a qualitative variable (post-operative thrombosis), or the length of survival (using, respectively, multiple linear regression, multiple logistic regression, or proportional hazards (Cox) regression analysis). Even in randomised clinical trials investigators may need assurance that the treatment effect is still present after simultaneous adjustment for several risk factors. When models are used to obtain estimates adjusted for other variables, it should be made clear which variables were adjusted for, on what basis they were selected, and, if relevant, how continuous variables were treated in the analysis.

Multivariate techniques, for dealing with more than one outcome variable simultaneously, really require expert help and are beyond the scope of these guidelines. Any complex statistical methods should be communicated in a manner that is comprehensible to the reader. It may help to place technical material in an appendix.

Results section: presentation of results

Presentation of summary statistics

Mean values should not be quoted without some measure of variability or precision. The standard deviation (SD) should be used to show the variability among individuals and the standard error of the mean (SE) to show the precision of the sample mean (see chapter 3: appendix 1). It must be made clear which is presented.

The use of the symbol $\pm$ to attach the standard error or standard deviation to the mean (as in $14{\cdot}2 \pm 1{\cdot}9$) causes confusion and should be avoided. Several medical journals do not now allow its use. The presentation of means as, for example, 14·2 (SE 1·9) or 14·2 (SD 7·4) is preferable. Confidence intervals are a good way of providing a reasonable indication of uncertainty of sample means, proportions, and other statistics. For example, a 95% confidence interval for the true mean is from about two standard errors below the observed mean to two standard errors above it (see chapter 4). Confidence intervals are more clearly presented as 10·4 to 18·0 (see chapter 3) than by use of the $\pm$ symbol.

When paired comparisons are made, such as when using paired *t* tests, it is important to give the mean and standard deviation of the differences between the observations or the standard error of the mean difference as appropriate (see chapter 3: appendix 1).

For data that have been analysed with distribution free methods it is more appropriate to give the median and a central range,

covering, for example, 95% of the observations, than to use the mean and standard deviation (see ''Descriptive information'', above). Non-parametric confidence intervals can be calculated (see chapter 5). Likewise, if analysis has been carried out on transformed data, the mean and standard deviation of the raw data will probably not be good measures of the centre and spread of the data and should not be presented.

When percentages are given, the denominator should always be made clear. For small samples, the use of percentages is unhelpful. When percentages are contrasted it is important to distinguish an absolute difference from a relative difference. For example, a reduction from 25% to 20% may be expressed as either 5% or 20%.

Results for individuals

The overall range is not a good indicator of the variability of a set of observations as it can be strongly affected by a single extreme value and it increases with sample size. If the data have a reasonably Normal distribution the interval two standard deviations either side of the mean will cover about 95% of the observations, but a percentile range is more widely applicable to other distributions (see ''Descriptive information'', above).

Although statistical analysis is concerned with average effects, in many circumstances it is important also to consider how individual subjects responded. Thus, for example, it is very often clinically relevant to know how many patients did not improve with a treatment as well as the average benefit. An average effect should not be interpreted as applying to all individuals (see also ''Repeated measurements'', above).

Presentation of results of hypothesis tests

Hypothesis tests yield observed values of test statistics which are compared with tabulated values for the appropriate distribution (Normal, t, χ^2 etc.) to derive associated P values. It is desirable to report the observed values of the test statistics and not just the P values. The quantitative results being tested, such as mean values, proportions, correlation coefficients, should be given whether the test was significant or not. It should be made clear precisely which data have been analysed. If symbols, such as asterisks, are used to denote levels of probability, these must be defined and it is helpful if they are the same throughout the paper.

P values are conventionally given as $<0{\cdot}05$, $<0{\cdot}01$, or $<0{\cdot}001$, but there is no reason other than familiarity for using these particular values. Exact P values (to no more than two significant figures), such as $P = 0{\cdot}18$ or $0{\cdot}03$, are more helpful. It is unlikely to be necessary to specify levels of P lower than 0·0001. Calling any value with $P > 0{\cdot}05$ ''not significant'' is not recommended, as it may obscure results that are not quite statistically significant but do suggest a real effect (see ''Interpretation of hypothesis tests'', below). When quoting P values it is important to distinguish $<$ (less than) from $>$ (greater than). P values between two limits should be expressed in logical order—for example, $0{\cdot}01 < P < 0{\cdot}05$ where P lies between 0·01 and 0·05. P values given in tables need not be repeated in the text.

The interpretation of hypothesis tests and P values is discussed below (''Interpretation of hypothesis tests'').

Figures (graphical presentation)

Graphical display of results is helpful to readers, and figures that show individual observations are to be encouraged. Points on a graph relating to the same individual on different occasions should preferably be joined, or symbols used to indicate the related points. A helpful alternative is to plot the difference between occasions for each individual.

The customary ''error bars'' of one standard error above and below the mean depict only a 67% confidence interval, and are thus liable to misinterpretation; 95% confidence intervals are preferable. The presentation of such information in figures is subject to the same considerations as discussed above (''Presentation of summary statistics''). Figures are most valuable when they display data that are too complex to put into a table. At the other extreme, a figure that displays, say, only two or three means with their standard errors or confidence intervals is often a waste of space; either more information should be added, such as the raw data, or the summary values should be put in the text or a table instead. Tables are also preferable if the data values are likely to be used by others in subsequent analyses (including meta-analysis).

Scatter diagrams relating two variables should show all the observations, even if this means slight adjustment to accommodate duplicate points. These may also be indicated by replacing the plotting symbol by the actual number of coincident points.

Tables

It is much easier to scan numerical results down columns rather than across rows, and so it is better to have different types of information (such as means and standard errors) in separate columns. The number of observations should be stated for each result in a table. Tables giving information about individual patients, geographical areas, and so on are easier to read if the rows are ordered according to the level of one of the variables presented.

Numerical precision

Spurious precision adds no value to a paper and even detracts from its readability and credibility. Results obtained from a calculator or computer usually need to be rounded. When presenting means, standard deviations, and other statistics the author should bear in mind the precision of the original data. Means should not normally be given to more than one decimal place more than the raw data, but standard deviations or standard errors may need to be quoted to one extra decimal place. It is rarely necessary to quote percentages to more than one decimal place, and even one decimal place is often not needed. With samples of less than 100 the use of decimal places implies unwarranted precision and should be avoided. Note that these remarks apply only to presentation of results—rounding should not be used before or during analysis. It is sufficient to quote values of t, χ^2 and r to two decimal places.

Miscellaneous technical terms

It is impossible to define here all statistical terms. The following comments relate to some terms which are frequently used in an incorrect or confusing manner.

Correlation should preferably not be used as a general term to describe any relationship. It has a specific technical meaning as a measure of association, for which it should be reserved in statistical work.

Incidence should be used to describe the rate of occurrence of new cases of a given characteristic in a study sample or population, such as the number of new notifications of cancer in one year. The proportion of a sample already having a characteristic is the prevalence.

Non-parametric refers to certain statistical analyses, such as the Mann–Whitney U test; it is not a characteristic of the observations themselves.

Parameter should not be used in place of "variable" to refer to a measurement or attribute on which observations are made. Parameters are characteristics of distributions or relationships in the population which are estimated by statistical analysis of a sample of observations.

Percentiles—When the range of values of a variable is divided into equal groups, the cut-off points are the median, tertiles, quartiles, quintiles, and so on; the groups themselves should be referred to as halves, thirds, quarters, fifths, etc.

Sensitivity is the ability of a test to identify a disease when it really is present—that is, the proportion positive of those who have the disease. *Specificity* is the ability of a test to identify the absence of a disease when the disease really is not present—that is, the proportion negative of those who do not have the disease. See also "Prediction and diagnostic tests", below.

Further guidance on terminology is given by Lang and Secic.[13]

Discussion section: interpretation

Interpretation of hypothesis tests

A hypothesis test assesses, by means of the probability P, the plausibility of the observed data when some "null hypothesis" (such as there being no difference between groups) is true. The P value is the probability that the observed data, or a more extreme outcome, would have occurred by chance—that is, just due to sampling variation—when the null hypothesis is true. If P is small one doubts the null hypothesis. If P is large the data are plausibly consistent with the null hypothesis, which thus cannot be rejected. P is not, therefore, the probability of there being no real effect.

Even if there is a large real effect a non-significant result is quite likely if the number of observations is small. Conversely, if the sample size is very large, a statistically significant result may occur when there is only a small real effect. Thus statistical significance should not be taken as synonymous with clinical importance.

The interpretation of the results of hypothesis tests largely follows from the above. A significant result does not necessarily indicate a real effect. There is always some risk of a false positive finding; this risk diminishes for smaller P values. Furthermore, a

non-significant result (conventionally $P > 0{\cdot}05$) does not mean that there is no effect but only that the data are compatible with there being no effect. Some flexibility is desirable in interpreting P values. The 0·05 level is a convenient cut-off point, but P values of 0·04 and 0·06, which are not greatly different, ought to lead to similar interpretations, rather than radically different ones. The designation of any result with $P > 0{\cdot}05$ as not significant may thus mislead the reader (and the authors); hence the suggestion above (''Presentation of results of hypothesis tests'') to quote actual P values.

Confidence intervals are extremely helpful in interpretation, particularly for small studies, as they show the degree of uncertainty related to a result—such as the difference between two means—whether or not it was statistically significant. Their use in conjunction with non-significant results may be especially enlightening.

Many hypothesis tests

In many research projects some tests of hypotheses relate to important comparisons that were envisaged when the research was initiated. Tests of hypotheses which were not decided in advance are subsidiary, especially if suggested by the results. It is important to distinguish these two cases and give much greater weight to the tests of those hypotheses that were formulated initially. Other tests should be considered as being only exploratory—for forming new hypotheses to be investigated in further studies. One reason for this is that when very many hypothesis tests are performed in the analysis of one study, perhaps comparing many subgroups or looking at many variables, a number of spurious positive results can be expected to arise by chance alone, which may pose considerable problems of interpretation. Clearly, the more tests that are carried out the greater is the likelihood of finding some significant results, but the expected number of false-positive findings will increase too. One way of allowing for the risk of false-positive results is to set a smaller level of P as a criterion of statistical significance.

A more complex problem arises when tests of significance are carried out on dependent (correlated) data. One example of this is in the analysis of serial data (discussed above—''Repeated measurements''), when the same test is performed on data for the same variable collected from the same subjects at different

times. Another is where separate analyses of two or more correlated variables are carried out as if they were independent; any corroboration may not greatly increase the weight of evidence because the tests relate to similar data. For example, diastolic and systolic blood pressures behave very similarly, as may alternative ways of assessing patient response generally. Very careful interpretation of results is required in such cases.

Association and causality

A statistically significant association (obtained from correlation or χ^2 analysis) does not in itself provide direct evidence of a causal relationship between the variables concerned. In observational studies causality can be established only on non-statistical grounds; it is easier to infer causality in randomised trials. Great care should be taken in comparing variables which both vary with time, because it is easy to obtain apparent associations which are spurious.

Prediction and diagnostic tests

Even when regression analysis has indicated a statistically significant relationship between two variables, there may be considerable imprecision when using the regression equation to predict the numerical level of one variable (y) from the other (x) for individual cases. The accuracy of such predictions cannot be assessed from the correlation or regression coefficient but requires the calculation of the prediction interval for the estimated y value corresponding to a specific x value (see chapter 8). The regression line should be used only to predict the y variable from the x variable, and not the reverse.

A diagnostic test with a high sensitivity and specificity may not necessarily be a useful test for diagnostic purposes, especially when applied in a population where the prevalence of the disease is very low. It is useful here to calculate the proportion of subjects with positive test results who actually had the disease (known as the *positive predictive value*). Note that there is no consensus on the definition of ''false-positive rate'' or ''false-negative rate''; it should always be made clear exactly what is being calculated, and this can best be illustrated by a 2×2 table relating the test results to the patients' true disease status.

A similar diagnostic problem arises with continuous variables. The classification as ''abnormal'' of values outside the ''normal

range'' for a variable is common, but if the prevalence of true abnormality is low most values outside the normal range will be normal. The definition of abnormality should be based on both clinical and statistical criteria.

Weaknesses

It is better to address weaknesses in research design and execution, if one is aware of them, and to consider their possible effects on the results and their interpretation than to ignore them in the hope that they will not be noticed.

Concluding remarks

The purpose of statistical methods is to provide a straightforward factual account of the scientific evidence derived from a piece of research. The skills and experience needed to design suitable studies, carry out sensible statistical analyses, and communicate the findings in a clear and objective manner are not easy to acquire. While we hope that these guidelines help authors to avoid statistical pitfalls, we reiterate our earlier advice to seek the advice of a statistician when possible.

1 Altman DG. Statistics in medical journals. *Stat Med* 1982; **1**:59–71.
2 Andersen B. *Methodological errors in medical research*. Oxford, Blackwell Science, 1990.
3 Altman DG. The scandal of poor medical research. *BMJ* 1994;**308**:283–4.
4 O'Fallon JR, Dubey SB, Salsburg DS, *et al*. Should there be statistical guidelines for medical research papers? *Biometrics* 1978;**34**:687–95.
5 Bailar JC, Mosteller F. Guidelines for statistical reporting in articles for medical journals: amplifications and explanations. *Ann Intern Med* 1988;**108**:266–73.
6 International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. *JAMA* 1997;**277**:927–34.
7 Simon R, Wittes RE. Methodologic guidelines for reports of clinical trials. *Cancer Treat Rep* 1985;**69**:1–3.
8 Epidemiology Work Group of the Interagency Regulatory Liaison Group. Guidelines for the documentation of epidemiologic studies. *Am J Epidemiol* 1981;**114**:609–13.
9 Lichtenstein MJ, Mulrow CD, Elwood PC. Guidelines for reading case-control studies. *J Chron Dis* 1987;**40**:893–903.
10 Murray GD. Statistical guidelines for the *British Journal of Surgery*. *Br J Surgery* 1991;**78**:782–4.
11 Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, *et al*. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* 1996;**276**:637–9.
12 Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF for the QUOROM group. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Lancet* 1999;**354**:1896–900.

13 Lang TA and Secic M. *How to report statistics in medicine. Annotated guidelines for authors, editors, and reviewers*. Philadelphia: American College of Physicians, 1997.
14 Altman DG. *Practical statistics for medical research*. London: Chapman & Hall, 1991.
15 Armitage P, Berry G. *Statistical methods in medical research*. 3rd edn. Oxford: Blackwell Science, 1994.
16 Bland M. *An introduction to medical statistics*. 3rd edn. Oxford: Oxford University Press, 2000.
17 Campbell MJ, Machin D. *Medical statistics. A commonsense approach*. 3rd edn. Chichester: John Wiley, 1999.
18 Colton T. *Statistics in medicine*. Boston: Little, Brown, 1974.
19 Pocock SJ. *Clinical trials. A practical approach*. Chichester: John Wiley, 1983.
20 Altman DG. Randomisation. *BMJ* 1991;**302**:1481–2.