CORRESPONDENCE

Statistical analysis of comparisons between laboratory methods

Papers reporting comparisons of alternative methods of measuring the same quantity are common in medical journals. The statistical analysis in these papers is often inappropriate so that the conclusions are unjustified.

I was first aware of this problem when a previous editor invited me to carry out an assessment of the large volume of statistics published in the Journal of Clinical Pathology. I reviewed 76 papers in the issue from January to July 1981. One of the most common errors was the use of the correlation coefficient to establish agreement between alternative methods of measurement. I found this error in eight papers (11%)—in fact, in nearly every paper reporting such a study.

Recently I was looking through an issue of the Journal and came across the paper by Lewis et al which described the SSF particle counter as a method of measuring platelet counts. These authors used regression and correlation to compare the SSF with both a visual reference method and a centrifugation method. Neither of these statistical methods is appropriate for assessing agreement between methods of measurement.

Although several issues of the Journal showed that inappropriate analysis remains a common problem in method comparison studies, the most common approach remains the correlation coefficient. Authors seem to be unaware of the large volume of statistics in a small value of r as indicating agreement, whatever the scatter diagram shows. Thus, for example, Bowen et al refer to a "close correlation" when the value of r was 0.95, even though their figure 5 shows considerable differences between the two methods they were comparing. Another common mistake is to see if there is a significant difference between the methods using a paired t test. Unless there is a notable difference between the methods, on average, this method will be less likely to give a significant result the worse the agreement is.

The paper by Calder et al is remarkable in that it gives no raw data in any form. This supposed "statistical comparison" is based solely on summaries of analyses as p values. They argued that the most reproducible method of three techniques for measuring the thickness of malignant melanomas was that with the largest p value for interobserver comparison. This conclusion is unjustified for the reason given above; the largest p value might arise from the method with the worst reproducibility. Furthermore, it is not necessarily true that the method with the best interobserver agreement is the most reproducible. In common with most research papers, the statistical comparison of studies, these authors took only one measurement per patient per observer using each technique, and so could not evaluate reproducibility. Occasionally multiple measurements on the same patients are taken and the data treated as independent observations.4 This, too, is an invalid statistical approach.

The appropriate analysis of method comparison studies is simple, being based on calculating the mean and standard deviation of the between method differences. The inappropriateness of the correlation coefficient has been pointed out many times.2,9 If these issues were even discussed in a recent paper in this Journal.10 While I suppose we may not expect authors to have read any of your previous papers, referees and editors should enforce good statistical practice as recommended in the recent statistical series by Brown and Beck11 and elsewhere.12 Ideally, statistical referees should be recruited.14

It was not all bad news in the issues of the Journal that I examined. The study by Duverlie et al was analysed by regression analysis, but the authors also calculated limits of agreement, and these were reported in the abstract.

Evaluation of new techniques is an important area of research. The Journal should adopt a more critical evaluation of such studies.

DOUGLAS G ALTMAN
Medical Statistics Laboratory,
Imperial Cancer Research Fund,
PO Box 123,
Lincoln's Inn Fields,
London WC2

The above letter pricked the editorial conscience. In mitigation it must be said that with papers where statistical analysis appears to be post hoc to the conclusions drawn (and these are fewer than might be imagined) the advice of statistical referees is usually sought.

Perhaps the key question is whether an exactly correct approach to statistics, as opposed to a popular one, is likely to alter the conclusions at the end of the day, or whether some statistical blunders have become merely the mathematical equivalents of Fowler's "study indefensible"—idioms seen, if tested by grammar, not to say what they are nevertheless well understood to mean. It could be that the best solution in borderline cases is to provide enough raw data for sceptical readers to test the analyses afresh. Below are appended some comments from the authors whose papers were singled out for criticism, though the editors are prepared to take the blame—Ed.

Although agreeing with Dr Altman that mean and standard deviation of between methods differences is the best procedure for analysis in comparability studies, the regression analysis used by us in this one aspect of our evaluation of a new platelet counter also included measurements of slope and intercept as well as paired t-test, as has been recommended in authoritative guidelines for the evaluation of haematological instruments.4 We considered this to be adequate for its purpose as the data were evenly distributed throughout the range of measurements analysed. Furthermore, we acknowledged that the comparability study was insubstantial because of the well recognised variability of the platelet counts and centrifugation methods which were used for comparison.

Consequently, as stated in the discussion section of the paper and in the abstractive, the conclusions were restricted to the findings on linearity, coincidence correction, sample volume control, stability and limits of carry over. No significance was placed on the comparability.

SM LEWIS,
RAMANAN,
F KUBOTA
UK National External Quality Assessment Scheme for Haematology,
Royal Postgraduate Medical School, Hammersmith Hospital, Du Cane Road, London W12 ONN

Appropriate statistical analysis is an important part of any study and this point has been well made by Dr Altman. The appropriateness, however, must be judged in the context of the study and the substance of the data. Thus in comparing an established method with a new technique the calculation of mean method difference may be entirely appropriate. When two separate estimates of a single notional entity are methods, however, and where the aim is simply to see whether both these independent methods reflect a common underlying parameter, then the mean method difference may be a significant clinical measure of difference, and that the divergence is more pronounced at higher levels. Some of the reasons for this variation between the two estimates were alluded to in the text. We calculated that the divergence is "limits of agreement" for the two methods to be 1-50 to 1-77. The clinical importance of these figures is uncertain and they were not included in our report. We aimed to present the reader with what we believed to be a pertinent analysis of the data, albeit in fairly uncharted territory.

It is important to recognise that statistical analysis is only one of the tools which authors can use to reduce a mass of data to an understandable level. This is not an end in itself. When method comparisons are involved and especially when there is no absolute certainty as to what is right, these analyses may not even be the most important part of the discussion. While authors may describe and exhibit, the ultimate decision will be made in the laboratory where theory will be tempered by practical reality. We trust that the Journal will continue to reflect this reality.

D BOWEN,
I CAVILLE
University of Wales College of Medicine, Department of Haematology, Heath Park, Cardiff

I would certainly agree with Doug Altman's comments that statistical analysis in many papers is inappropriate. I share his dismay at the lack of raw data presented in many papers, which severely limits the ability of readers to assess for themselves the validity of any conclusions reached. I would also agree with Doug Altman that a high correlation does not mean that two methods agree and give the same results. A high correlation with a high "r" value simply means that there is a relation between the two variables.

Our paper published last year in the Journal of Clinical Pathology used least squares linear regression to assess correlation (relation) between two methods, but the regression analysis was not used to assess agreement between them. Indeed, to quote from the
discussion of our paper: “This study has shown that although there is a significant correlation between the turbidimetric clotting technique and radial immunodiffusion for plasma fibrinogen assay, discrimination between hyperlipidaemic patients and a group of healthy control subjects is better by the immunochromatographic assay.”

Our conclusion was that the significance of plasma fibrinogen concentration as a risk factor for vascular disease may be method dependent and that one method could certainly not be used in place of the other to assess cardiovascular risk. As our paper did not use regression analysis or the correlation coefficient to assess agreement by alternative methods of measurement I am unsure as to why our paper was quoted by Doug Altman in the first place.

I would like to state as one who has submitted papers for publication to a number of scientific journals that I have had a more critical eye passed over my work with respect to its scientific and statistical content from referees of the Journal of Clinical Pathology than from many other journals to which I have submitted.

MARTYN L KNAPP
Department of Chemical Pathology,
St. Peter's Hospital,
Chertsey,
Surrey

The letter from Douglas Altman highlights a common problem in medical and biological research, namely, a lack of communication between researcher and statistician. Few departments have a resident statistician and most researchers have to rely on books with generic titles like Statistics for Medical Research. While these give good details on the calculations of each test, we have yet to find any which give clear information as to when a particular test should be used, its limitations, etc. We would be grateful to know if such a book exists.

Unfortunately, while statisticians use sentence like, “unless there is a negligible real difference between the methods, on average, this method will be less likely to give a significant result than the other agreement,” their science will remain a deep mystery to most medical researchers.

SR PLASTOW,
CJ CALDER,
AP CAMPBELL
Department of Pathology,
Medical School,
University of Birmingham,
Birmingham

Although our paper was not individually mentioned, it was listed in the group of papers where the analysis was stated to be inappropriate. The primary method of analysis we used was regression analysis between the prothrombin time methods. This was chosen by the World Health Organisation’s recommended procedure for prothrombin time standardisation based on historical regression analysis. The use of an alternative system would have been confusing to the reader.

JM THOMSON,
DA TABERNER,
I POLLER
UK Reference Laboratory,
Anticoagulant Reagents and Control,
Wirushima Hospital,
Manchester

Dr Altman is concerned about the statistical methods used to evaluate new techniques which compare quantity. We applied the method suggested to the data shown in fig 2 of our paper to which he referred. We calculated the mean differences and the standard deviation of the relative intensities (1%) values obtained for the individual PGA isozymogens by both methods.

The following results were obtained: PGA1: mean 0-0, SD 2-0; PGA2: mean -0.2, SD 4-0; PGA3: mean -1.4, SD 2-4; PGA2: mean 2-4, SD 6-7. This provides further evidence that PGA patterns obtained by both methods are similar.

PGA isozymogens are determined visually. These limits of agreement are acceptable for this purpose. We feel, therefore, that while Dr Altman’s criticism is justified, reanalysis of our data does not affect our paper.

RW TEN KATE
Elizabeth Gaithus Haarlem,
Boorhavestraat 32, 2235 BC Haarlem,
Postbus 417, 2000 AK
The Netherlands

Rectal gelatin coating: a marking tool for pathologists

We describe an economical and user-friendly technique for assessing the often forgotten but essential radial excursion margin for use in rectal carcinoma. For years pathologists have placed great weight on resection margins distal from colorectal cancer which has only very recently been shown to be of no value. The radial excursion margin is sometimes confusingly called the lateral resection margin,4 has been studied in academic colorectal units using accurate but elaborate and time consuming drawing and labelling techniques. This remains an inappropriate technique for the busy district general hospital pathologist. Despite this, with advances in surgical technique in the rectum, their surgeons will want to know about this margin.

Marking the resection margin is not new, but marking with India ink is messy, allows only one colour, and has a habit of spreading all over the specimen, failing to confine itself to the margin lines, and taking a long time to dry. Our gelatin solution, originally devised for breast marking,5 dries instantly on contact with the tissue, does not run down tissue planes, is easily recognised microscopically, and permits differential marking of the anterior, posterior, and two lateral margins.

We receive all rectal specimens fresh from theatre in polyethylene bags. Bottles of plain gelatin solution and bottles of gelatin mixed with tumeric, India ink, and Bismark brown are melted under a running tap, and gelatin is applied to four margins with an artist’s brush. Plain gelatin is cosinophilic. This is used to mark the anterior margin. Gelatin mixed with India ink is used for the posterior margin, gelatin containing Bismark brown is brushed on the lateral margin, and turmeric in gelatin on the left lateral margin. The opened rectum is then pinned to a cork board and fixed in formalin for 24 hours. It is then unpinned and floated in formalin for a further 24 hours. The tumour is then sliced at 2 mm intervals and the points of maximum penetration are sampled to include the radial margin. Complex drawings are not necessary and we are able to tell the surgeon how far the tumour is from which point on the radial margin it approaches the anterior, or posterior or particular nature of the gelatin pigments under the microscope.

Our method costs £0.07 per case, is simple and rapid, and adds less than five minutes to the cut up time. It is therefore eminently suitable for both the academic unit and the hardpressed district general hospital pathologist, who wish to provide their increasingly sophisticated surgical colleagues with useful rectal cancer reports.
