Br Heart J 1988;**60**:177–80

Editorial

The end of the p value?

STEPHEN J W EVANS,* PETER MILLS, JANE DAWSON

From the Departments of Clinical Epidemiology and Cardiology, The London Hospital; and the British Heart Journal

The application of statistical methods to medical data has been undergoing a sea-change. This is of particular importance in cardiology because the current methods that statisticians recommend express the results of studies in terms that are directly relevant to the clinical use to which they may be put. In March 1986 the *British Medical Journal* nailed its colours firmly to the mast, telling readers that "authors . . . will be expected to calculate confidence intervals whenever the data warrant this approach"[1,2] and the *Lancet*,[3,4] *Annals of Internal Medicine*, and *American Journal of Public Health* are among other journals that have endorsed the new orthodoxy. We expect that studies reported in the *British Heart Journal* will increasingly reflect this approach. The nuts and bolts of calculating the confidence intervals of various types of data are described in a series of articles in the *British Medical Journal*,[5–9] and below we review some aspects of the approach that are particularly relevant to papers published in the *British Heart Journal*.

Towards estimation and away from hypothesis testing

The null hypothesis generally states that there is no relation between the variables under study. For example, when the change in cardiac output before and after intervention is analysed the null hypothesis proposes that the average change is zero. It follows that calculation of the p value, which is based on the null hypothesis, is frequently an inappropriate statistical method for summarising the analysis of cardiological data. Many published studies do not seriously consider the possibility that an intervention has no effect. When a test intervention has been used the question usually being asked is "how great is its effect?" rather than "does it have an effect?"

*Statistical adviser to the *British Heart Journal*.

Requests for reprints to Jane Dawson, *British Heart Journal*, BMA House, Tavistock Square, London WC1H 9JR.

This point may be illustrated by comparing cardiac output before and after administration of an inotropic drug. A paired t test with a p value starts with the hypothesis that the inotrope has no effect. It is unlikely that the drug would be under investigation if no effect on cardiac output were really expected. The questions for the clinician are "on average, how great is the change produced by the intervention" and "with what precision has the average change been estimated?" These questions are answered by the calculation of confidence intervals, whereas hypothesis testing can give only the answer "yes" or "no" to the question "Is there a change?"

Figure 1a is an example of data that arise in such a study. The paired t test gives a value of $t = 3{\cdot}3$ ($p = 0{\cdot}01$). It indicates that the rise is statistically significant but does not indicate the size of the rise. The appropriate 95% confidence interval which is shown in fig 1a is based on the mean change and two standard errors on either side of the mean and suggests the likely interval within which the true mean lies. Thus the confidence interval centred on the mean change of 0·6 l/min extends from a mean change of +0·2 l/min to one of +1·0 l/min. This implies that the true mean value could lie anywhere between 0·2 and 1·0 and that the data are unlikely to be consistent with a mean change of zero. A confidence interval that does not include zero is equivalent to a test with a statistically significant p value. When the confidence interval includes zero, as for example when the interval is from −0·2 l/min to +1·4 l/min then although the mean change remains the same, at +0·6 l/min, the possibility must be considered that the intervention causes a fall rather than a rise or that it causes no change at all. If the data arise from a smaller sample (see fig 1b in which $n = 7$ instead of $n = 9$) or if their standard deviation is larger (see fig 1c in which the standard deviation has increased by 25%) the confidence interval will be wider.

When the confidence interval includes zero the

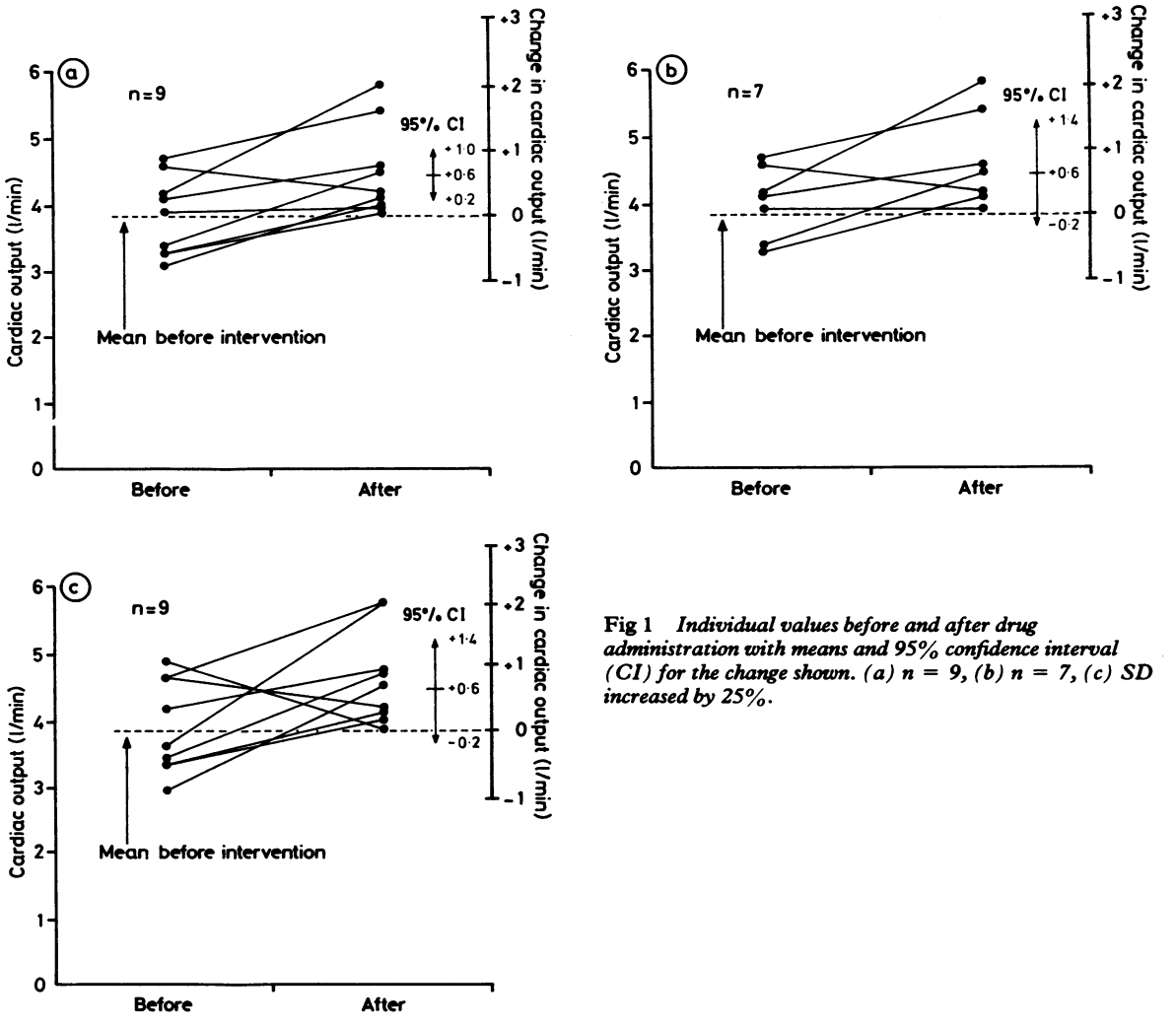

Fig 1 *Individual values before and after drug administration with means and 95% confidence interval (CI) for the change shown. (a) n = 9, (b) n = 7, (c) SD increased by 25%.*

result is equivalent to a significance test that gives a non-significant result. Relying on this feature alone is no better than the use of p values, but the upper limit of the confidence interval (+1·4) draws attention to the possibility that the average increase *might* be clinically useful.

The obvious advantage of a confidence interval is that it expresses results in the units in which the measurements were made, and so allows the reader to consider critically the clinical relevance of the results. If the sample size is small the confidence interval will be wide. The clinician must then examine the extremes of the interval. Do these extremes indicate that the clinical relevance of the results is consistent with the conclusions drawn from the analysis? If the conclusion is drawn that the drug has "no effect" because p is not statistically significant but the 95% confidence interval reaches 1·4 l/min (fig 1b) then it is clear that the drug may well have a positive effect that has not been demonstrated by this study. The confidence interval (of say +0·05 to +0·1 l/min) can also make it clear that a difference which is statistically significant (based on p values) is of no clinical relevance because the statistical significance of the result has been produced spuriously by a very large sample of say about 2000. In such a study even the upper value of the confidence interval suggests that such a change is too small to be of clinical benefit despite its statistical significance.

When the effects of two different drugs on cardiac output are being compared the appropriate test is an independent samples *t* test and the equivalent 95% confidence interval may also be calculated. In fig 2, drug A gives the same results as shown in fig 1a, while

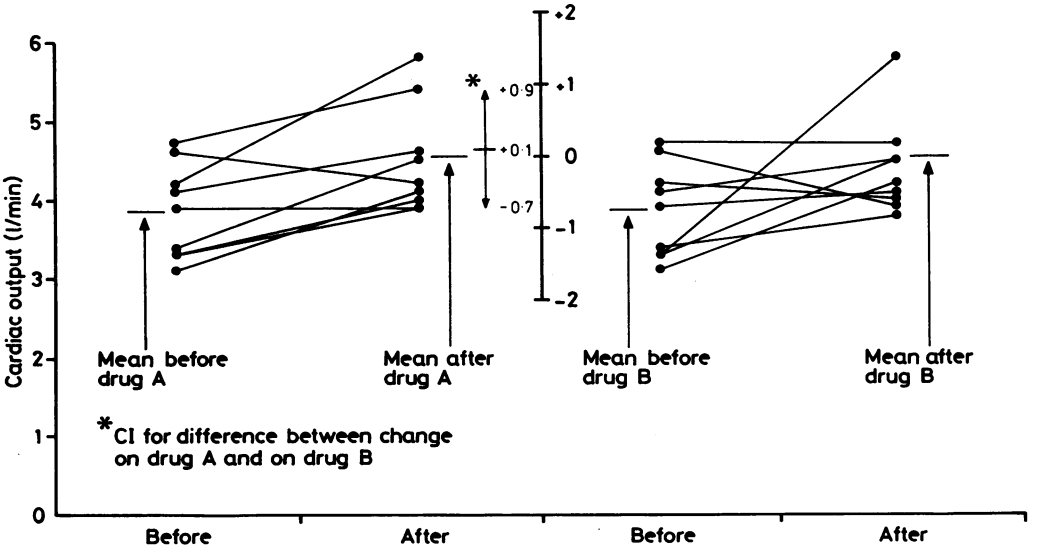

Fig 2 *Individual and mean values before and after administration of two drugs. The 95% confidence interval for difference between the changes with drug A and the changes with drug B is −0·7 to +0·9 l/min.*

the effect of drug B on cardiac output is "not statistically significant" when assessed by a paired *t* test. The relevant question is no longer simply whether each drug alters cardiac output but whether the change with drug A (δCO_A) is importantly different from that with drug B (δCO_B). The 95% confidence interval for *the difference between the changes with A and with B* (δCO_{AB}) is −0·7 to +0·9 l/min. This shows that although the change with A is statistically significant ($p < 0·05$) and that with B is not statistically significant, there is insufficient evidence that the change with A is different from the change with B (because the 95% confidence interval for the difference in changes between A and B includes zero). At the same time confidence intervals show that potentially clinically important differences (for example of 0·9 l/min) between the drugs may not have been detected because the sample sizes were too small to produce significant p values.

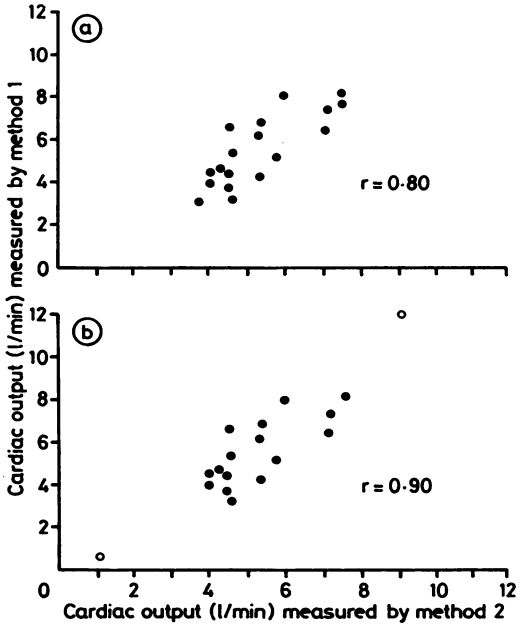

Fig 3 *(a) Traditional scatter diagram for comparison of two methods of measuring cardiac output. (b) Two outlier values "improve" the correlation.*

Thus confidence intervals provide all the information that significance tests give us and also indicate the clinical relevance of the information. Confidence intervals require little more calculation than the appropriate significance test.

Method comparison and the null hypothesis

Many investigations in cardiology compare two methods of measuring the same variable—for example cardiac output determined by Doppler echocardiography and by the Fick principle at cardiac catheterisation. In the past, such data have been summarised by correlation or regression coefficients and calculation of a p value (fig 3a). In both these methods the null hypothesis is tested. But the null hypothesis, which states there is no association between two variables, is not relevant to the

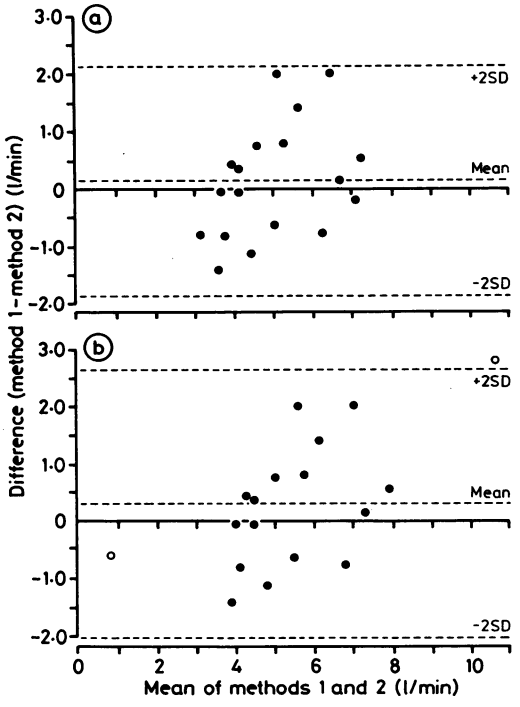

Fig 4 *(a) Diagram showing differences versus mean for data in fig 3a. (b) The two outliers (circles) in fact produce a worse agreement between the methods.*

measurement of the same variable by two different methods.[10] The magnitude of the correlation coefficient is strongly influenced by the range of values under study. In addition, its "significance" is increased simply by increasing the number of subjects studied. The apparently stronger correlation between the variables for the data shown in fig 3b ($r = 0{\cdot}90$ *v* $r = 0{\cdot}80$) is purely the result of the inclusion of two outliers. The correlation coefficient gives neither the magnitude of any possible discrepancy between the two methods nor whether such discrepancy is consistent over the range of values.

Like confidence intervals the method of analysis advocated by Bland and Altman[10] emphasises clinical relevance, which is determined by understanding the extent to which the two methods give different results—not by confirming that they show a little better than chance agreement when used to measure cardiac output. So fig 4a shows that method 1 gives slightly higher values than method 2 (the mean of the difference is higher than zero) and fig 4b shows that inclusion of the outlying values reduces the agreement rather than improves it (the standard deviation has increased from 1·00 to 1·17).

The key questions are what is the variability of a single observation (including its measurement error) and how much disagreement is there between the two methods of measurement? It is also important to be aware of systematic variation in the answers over the range of interest; for example when the two sets of measurements are examined does one method yield high values at the upper end of the range and low values at the lower end of the range? If it does, are these discrepant values genuine or are they spurious? Investigation of the methods together with some understanding of the possible clinical applications will be necessary to decide which is the better method of measurement. Lastly, while the two methods may agree over a wide range of values including those of normal individuals, does this degree of agreement between the techniques extend into the range of values commonly encountered in disease?

For some time now the *British Heart Journal* has been informally persuading authors who inappropriately use correlation and regression coefficients to use the method of Bland and Altman to examine agreement between methods. The fact that, over many years, correlation and regression have been misused is no reason to perpetuate a bad practice. Comparison of the r values obtained in different studies is meaningless. In addition, the *British Heart Journal* also recommends that confidence intervals should be given where relevant for studies that assess the effects of interventions, compare the effects of different drugs, or evaluate non-invasive techniques.

References

1 Langman MJS. Towards estimation and confidence intervals. *Br Med J* 1986;**292**:716.
2 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J* 1986;**292**:746–50.
3 Anonymous. Report with confidence. *Lancet* 1987; **i**:488.
4 Bulpitt CJ. Confidence intervals. *Lancet* 1987;**i**:494–7.
5 Gardner MJ, Altman DG. Statistics in medicine: confidence intervals: estimating with confidence. *Br Med J* 1988;**296**:1210–1.
6 Altman DG, Gardner MJ. Statistics in medicine: calculating confidence intervals for regression and correlation. *Br Med J* 1988;**296**:1238–42.
7 Morris JA, Gardner MJ. Statistics in medicine: calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *Br Med J* 1988;**296**:1313–6.
8 Machine D, Gardner MJ. Statistics in medicine: calculating confidence intervals for survival time analyses. *Br Med J* 1988;**296**:1369–71.
9 Campbell MJ, Gardner MJ. Statistics in medicine: calculating confidence intervals for some non-parametric analyses. *Br Med J* 1988;**296**:1454–6.
10 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**i**:307–10.