Analysis of method comparison studies

Sally Hollis
From the Statistics and Computation Unit, Hope Hospital, Salford, UK

Additional key phrases: limits of agreement; measurements; precision; accuracy

> Give a man three weapons—correlation, regression and a pen—and he will use all three.[1]

Comparisons of two methods of measurement, particularly two assays, are very common in clinical biochemistry. If one method is a 'gold standard' giving very accurate measurements, then it is sufficient to calibrate the new method against the established method using regression analysis. However, usually we cannot regard either method as giving a true measurement without any error. Traditionally, as in most areas of medicine, method comparison studies have been analysed using correlation coefficients. However, correlation measures association between two methods and is not a measure of agreement for two reasons. First, the test of significance of the correlation coefficient is irrelevant since it assesses whether the coefficient is significantly different from zero. As the two methods are measuring the same thing, the knowledge that we can confidently reject the hypothesis of no association tells us little about the agreement between the two methods. Secondly, if the two methods are to agree, the points on a scatterplot of the two methods must lie close to the line of equality, not just close to the line of best fit. For example, if method A always gives a value twice as high as method B, the correlation between the two methods will be perfect, but there is clearly not good agreement.

The problem of assessing points relative to the line of equality rather than to the line of best fit has been partially addressed by reporting the equation of the regression line in addition to the correlation. The slope and intercept of the regression line can be examined to see if the line is close to the line of equality, which has an intercept of zero and a slope of one. These requirements can even be formally tested if the standard errors of the slope and intercept are also supplied. It should be noted that statistical packages will usually give a significance test of the slope compared to zero rather than one.

So, by using a slightly more complex analysis, we can start to assess the specific concept of agreement rather than association. However there are still problems with this approach. Standard linear regression analysis assumes that the dependent (y) variable is measured with error, but that the independent (x) variable is not. This leads to two different lines of best fit, depending on the choice of the dependent variable. The magnitude of the difference between the two lines increases as the correlation between the two variables decreases. This is not just a purely statistical problem, but can give two substantially different lines (Fig. 1). This can be eliminated by the use of a method assuming errors in both variables which will give a symmetrical answer, such as Deming regression.[2]

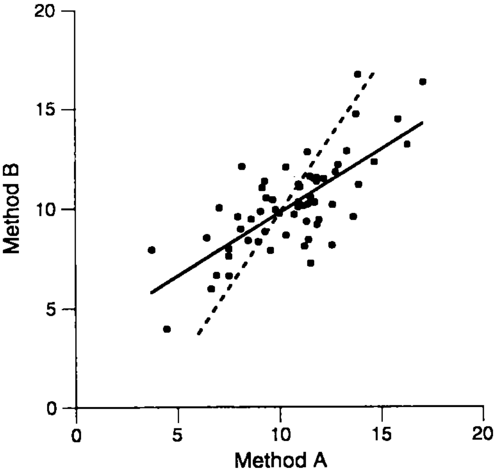

FIGURE 1. *Regression lines of y on x (solid line) and x on y (dotted line).* N = *60,* r = *0·73.*

Correspondence to: Sally Hollis, Statistics Editor, *Annals of Clinical Biochemistry*, Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK.

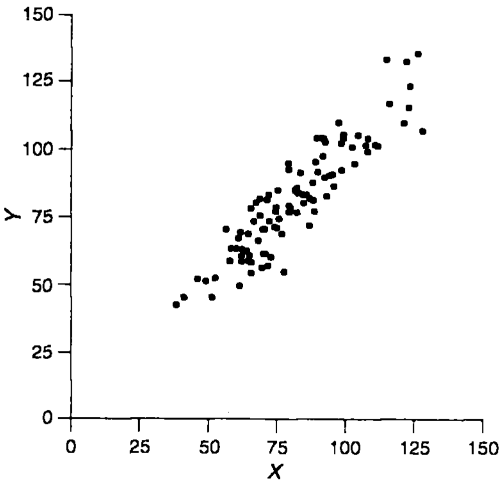

FIGURE 2. *One hundred observations randomly sampled from normal distributions with mean 80 and standard deviation 20 and a correlation of 0·9.*

A further problem with the correlation coefficient is that its value depends on the range of data sampled. Figure 2 shows 100 observations randomly sampled from normal distributions with mean 80 and standard deviation 20, and a correlation coefficient of 0·9. The correlation coefficient of the entire 100 observations is 0·91. When the data are split randomly into two equal halves, the correlation coefficients within each half, 0·88 and 0·93, are very similar to the original correlation coefficient. However, if the data is split according to whether X is above or below 80, the two coefficients are 0·72 and 0·83, considerably lower than the correlation coefficient for the entire sample. A wider range of data will always tend to give a higher correlation coefficient. In method comparison studies a positive effort is often made to include particularly extreme data for practical reasons, leading to a misleading inflated correlation coefficient.

These problems have been noted in the statistical literature for some time, and in 1983 Altman and Bland published a paper highlighting these problems and suggesting an alternative method of analysis.[3] This paper was published in *Statistician*, a journal aimed at professional statisticians and not widely read in the medical community. In 1986 a more accessible version of this paper was published in the *Lancet*,[4] ensuring that it reached a wide clinical audience, and had a considerable impact on the analysis of method comparison studies in the medical literature. In 1992 the *Lancet* paper had been cited over 600 times.[5] It has been generally accepted that the approach Bland and Altman suggested is the appropriate technique for analysing method comparison studies and articles advocating this approach or variations of it have since appeared in many journals including the *Annals of Clinical Biochemistry*.[6] One clinical biochemistry journal has even reproduced the entire article verbatim.[7]

The cornerstone of Bland and Altman's approach is examination of the differences between the two methods. They suggest a plot of the difference between the two methods against the average of the two methods. This allows the assessment of bias (do the differences differ systematically from zero?) and error (how much do the differences vary?). Figure 3 is a difference plot of the data used in Fig. 2. It is apparent from this display that there is no systematic bias (the differences are symmetrical about zero), that the majority of differences lie between − 15 or + 15 and that there is no obvious relationship between the difference and the average.

Statistically, if there is no relationship between the difference and the average, the agreement between the two methods can be summarized using the mean and standard deviation of the

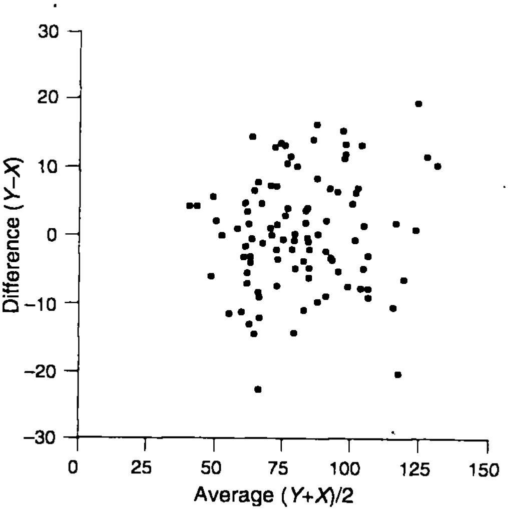

FIGURE 3. *Difference plot of 100 observations randomly sampled from normal distributions with mean 80 and standard deviation 20 and a correlation of 0·9.*

differences. The accuracy can be assessed by a test of whether the mean difference is zero (easily determined using a 95% confidence interval for the mean difference between the two methods) and the precision described by the 'limits of agreement', the confidence interval for *individual* differences between the two methods. The limits of agreement are given by the mean difference ± twice the standard deviation of the differences. For small samples it is advisable to use the two-tailed 95% value from the t-distribution on $(n-1)$ degrees of freedom as the multiplier of the standard deviation in the calculation of the limits of agreement.

For the data shown in Fig. 3, the differences have a mean of 0·08 with a standard deviation of 8·42. The standard error of the mean difference (SEM) is calculated by dividing the standard deviation of the differences by the square root of the number of observations, in this case $\text{SEM} = 8{\cdot}42/\sqrt{100} = 0{\cdot}84$. The 95% confidence interval for the mean difference is given by the mean ± 2 × SEM = 0·08 ± 2 × 0·84 = − 1·76 to 1·61 (again the appropriate value from the t-distribution may be used instead of 2 for a more accurate answer). This confidence interval includes zero so there is no evidence of systematic bias. The limits of agreement are given by mean ± 2 × standard deviation = 0·08 ± 2 × 8·42 = − 16·9 to 16·8. The difference between X and Y will be between − 17 and 17 for about 95% of cases. Does this represent acceptable agreement? There is no statistical answer to this question—it is a matter of clinical judgement. Ideally, the acceptable limits of agreement should be chosen when the study is planned, before any data have been collected.

If there does appear to be a relationship between the difference and the average, then the standard limits of agreement will not be appropriate since the observed range of differences depends on the average value. The first approach in this situation is to try using a log transformation of the data, since this will often remove the relationship so that the limits of agreement can be anti-logged to give a range of percentages of the average rather than absolute values. No other form of transformation is recommended since only logs can be sensibly back-transformed in this manner. Figure 4 is a difference plot for two immunoassays. The mean difference and the limits of agreement are shown on the plot. There appears to be a tendency for the spread of the differences to increase as the average increases. If this is the case then the limits of agreement will be under-estimated for low values and over-estimated for high values. The log-difference plot for these data (Fig. 5) shows that the differences are now more evenly distributed. The limits of agreement for the logged data are − 0·37 and 0·21, indicating that for about 95% of samples, the value of assay 2 will be between 0·43 and 1·62 times the value of assay 1. The values of assay 2 may differ from assay 1 by about 60% above or below.

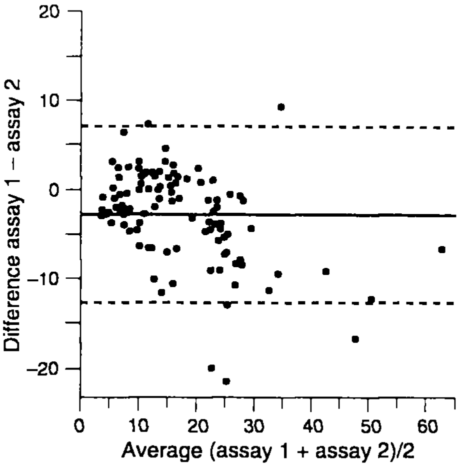

FIGURE 4. *Difference plot of 116 samples measured by two immunoassays with the mean difference (solid line) and limits of agreement (dashed lines).*

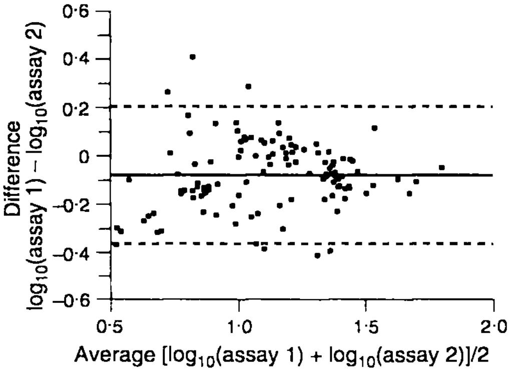

FIGURE 5. *Log-difference plot of 116 samples measured by two immunoassays with the mean difference (solid line) and limits of agreement (dashed lines).*

There is no definitive rule which describes when the logarithm of the differences should be used. Remember that the aim is to determine limits of agreement that are valid across the whole range of values. It will sometimes be impossible to give simple limits of agreement, if there is a relationship that is not removed by log transformation or if there are outliers which substantially affect the standard deviation. In these cases a difference plot should be shown, but the limits of agreement may have to be described in general terms rather than statistically determined.

Plotting the differences expressed as a percentage of the average, as recommended by Pollock *et al.*,[6] can be very helpful in illustrating the magnitude of differences but often it is not possible to summarize the percentage differences in a simple manner across the range of the data, and an alternative method of statistical analysis is necessary.

If either or both of the methods have poor precision (values are not reproducible) then the agreement between the methods will be reduced due to this additional error. The relative precision of the two methods can be assessed by making duplicate measurements by each method and comparing the variances of these duplicates.[8] This should be repeated at various different levels if the precision varies over the range of observed values. Once it has been established that two methods show reasonable agreement, it is important to examine the relative precision of the methods to determine whether one method is superior.

The *Journal*'s requirement is that difference plots and appropriate limits of agreement are used to analyse method comparison studies. Although correlation and linear regression have become the common means of presenting and analysing such studies, the method is flawed and the *Annals of Clinical Biochemistry* will no longer regard it as acceptable.

REFERENCES

1 Anon. The anomaly that won't go away. *Lancet* 1978; ii: 978

2 Deming WE. *Statistical Adjustment Of Data*. New York: Wiley, 1943: 184

3 Altman DG, Bland JM. Measurement in medicine: the analysis of comparison studies. *Statistician* 1983; **32**: 307–17

4 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of agreement. *Lancet* 1986; i: 307–10

5 Bland JM, Altman DG. Comparing methods of clinical measurement—a citation classic commentary on statistical methods for assessing agreement between 2 methods of clinical measurement. *Curr Cont/Clin Med* 1992; **40**: 8

6 Pollock MA, Jefferson SG, Kane JW, Lomax K, MacKinnon G, Winnard CB. Method comparison—a different approach. *Ann Clin Biochem* 1992; **29**: 556–60

7 Bland JM, Altman DG. Statistical methods for assessing agreement between measurements. *Biochim Clin* 1987; **11**: 399–404

8 British Standards Institution. *Accuracy (Trueness and Precision) of Measurement Methods and Results: Use in Practice of Accuracy Values*. BS ISO 5725 Part 6. London: BSI, 1994

Accepted for publication 20 June 1995