Ultrasound Obstet Gynecol 2003; **22**: 85–93
Published online 9 May 2003 in Wiley InterScience (www.interscience.wiley.com). **DOI:** 10.1002/uog.122

Applying the right statistics: analyses of measurement studies

J. M. BLAND* and D. G. ALTMAN†

**St George's Hospital Medical School, London and †Cancer Research UK Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Oxford, UK*

KEYWORDS: agreement; correlation; error; measurement; regression

ABSTRACT

The study of measurement error, observer variation and agreement between different methods of measurement are frequent topics in the imaging literature. We describe the problems of some applications of correlation and regression methods to these studies, using recent examples from this literature. We use a simulated example to show how these problems and misinterpretations arise. We describe the 95% limits of agreement approach and a similar, appropriate, regression technique. We discuss the difference vs. mean plot, and the pitfalls of plotting difference against one variable only. We stress that these are questions of estimation, not significance tests, and show how confidence intervals can be found for these estimates. Copyright © 2003 ISUOG. Published by John Wiley & Sons, Ltd.

INTRODUCTION

Many research papers in imaging concern measurement. This is a topic that in the past has been much neglected in the medical research methods literature. In this paper we discuss the estimation of the agreement between two methods of measurement, and the estimation of the agreement between two measurements made by the same method, also called repeatability. In both cases we are concerned with the question of interpreting the individual clinical measurement. For agreement between two different methods of measurement, we ask whether we can use measurements by these two methods interchangeably, i.e. can the method by which the measurement was made be ignored. For two measurements made by the same method, we ask how variable measurements from a patient can be if the true value of the quantity does not change and what this measurement tells us about the patient's true or average value. In some studies repeated observations are made by the same observer or many different observers and are treated as repeated observations of the same thing. In others a small number of observers, often two, are used and systematic differences between them are explored, in which case the analysis is like that for comparing two different methods of measurement.

We avoid all mathematics, except for one formula near the end. Instead we show what happens when some simple statistical methods are applied to a set of randomly generated data, and then show how this helps the interpretation of these methods when they are used to tackle measurement problems. We illustrate these methods by examples drawn from the imaging literature. For some of these examples, rather than bother the original authors for their data, we have digitized them approximately from the published graphs, and our figures differ slightly but not in any important way from those originally published.

We shall start with a typical example of a measurement study. Borg *et al.*[1] compared single X-ray absorptiometry (SXA) with single photon absorptiometry (SPA). They produced a scatter plot for arm bone mineral density similar to that in Figure 1. This looks like good agreement, with a tight cloud of points and a high value for the correlation coefficient, $r = 0.98$. The points cluster quite closely around the line drawn through them, the regression line. But should this make us think we could use bone mineral densities measured by SXA and SPA interchangeably? In Figure 2 we have added the line of equality, the line on which points would lie if the two measurements were the same. Nearly all the points lie to the left of the line of equality. There is a clear bias: the SXA measurements tend to exceed the SPA measurements by 0.02 g/cm^2. We shall now explain why the correlation coefficient does not reflect this bias and go on to explore the interpretation of the regression line. To do this we show what happens when these methods are applied to artificially generated data, i.e. when we know what the interpretation should be. We then describe a simple

Correspondence to: Prof. D. G. Altman, Cancer Research UK Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Old Road, Headington, Oxford OX3 7LF, UK (e-mail: doug.altman@cancer.org.uk)

Accepted: 21 February 2003

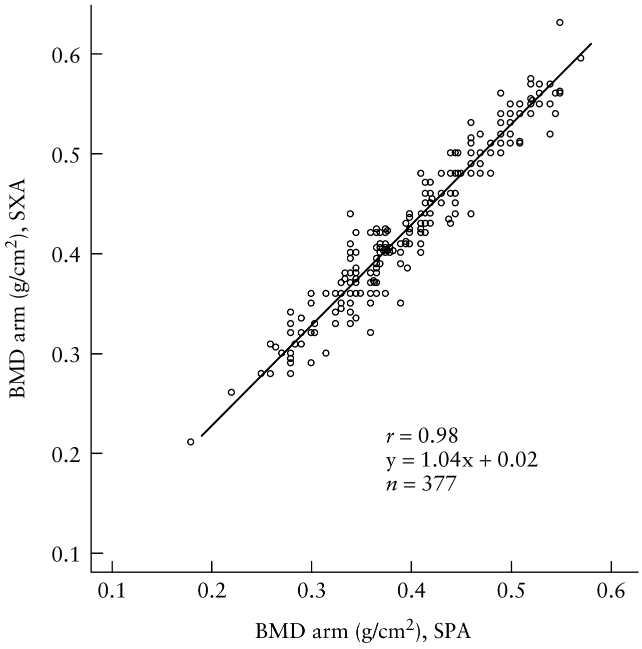

Figure 1 Arm bone mineral density (BMD) measured by single X-ray absorptiometry (SXA) and single photon absorptiometry (SPA)[1].

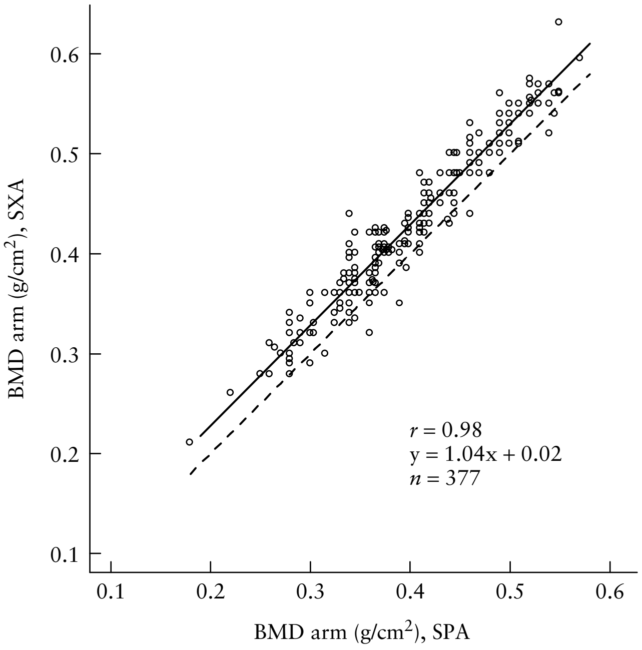

Figure 2 Data of Figure 1, with the line of equality (- - - -) added. ———, regression line.

alternative approach, limits of agreement, which avoids such problems.

THE INTERPRETATION OF CORRELATION COEFFICIENTS

To illustrate the interpretation of correlation, we shall start with some artificial, randomly generated data. This is not because we have no real data, but because with randomly generated data we know the answers to our questions. We generated 100 observations from a Normal distribution to represent the true value of the quantity being measured. Now we want to compare two different measurements of this true quantity. We generated two measurements, X and Y, by adding to the true value some measurement error, from a Normal distribution and independent of the true value.

This gave us artificial data representing two observations on each of a group of subjects. These observations might be measurements by two different methods, by the same method but different observers, or by the same method and the same observer. We know that they are closely related and that there is no consistent bias or tendency for X to be greater or less than Y. Figure 3 shows the artificial data, with the line of equality.

A natural approach to data like those of Figure 3 is to calculate a correlation coefficient. For these data the correlation between X and Y is $r = 0.86$, $P < 0.0001$. There are two problems with this analysis. First, correlation depends on the range and distribution of the variables and hence on the way in which the sample of subjects was chosen. Second, correlation ignores any systematic bias between the two variables. To show that correlation depends on the range of the variables, consider only subjects in a restricted range. For subjects whose true measurement would be < 10, $r = 0.60$ and for subjects whose true measurement would be ≥ 10, $r = 0.62$ (Figure 4). Both of these are less than $r = 0.86$ for all the data. If we take several pairs of measurements from the same subject, the correlation can be zero but we should not conclude from this that the two methods do not agree. It is what we should expect, because there is no variation at all in the true value. (Altman and Bland[2] discuss this and give an example.) So the correlation coefficient depends on the group of subjects selected. It should be used only if we have a representative sample of the patient population we wish to study. In

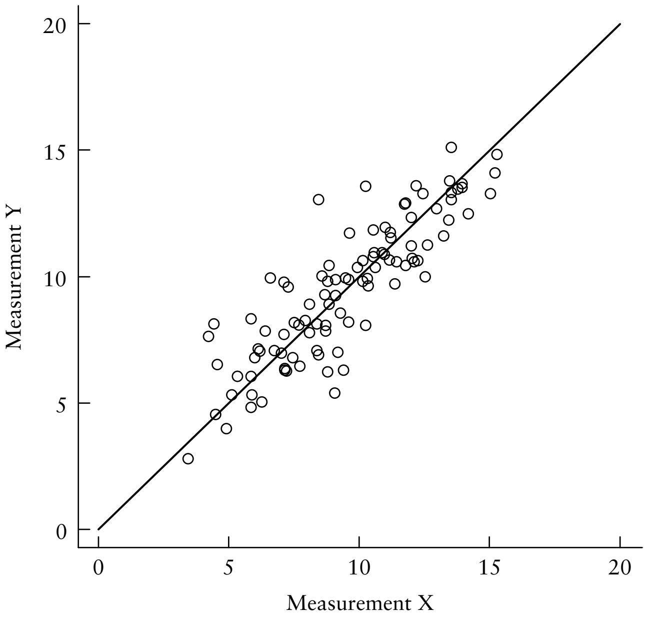

Figure 3 Artificial measurements X and Y, with the line of equality.

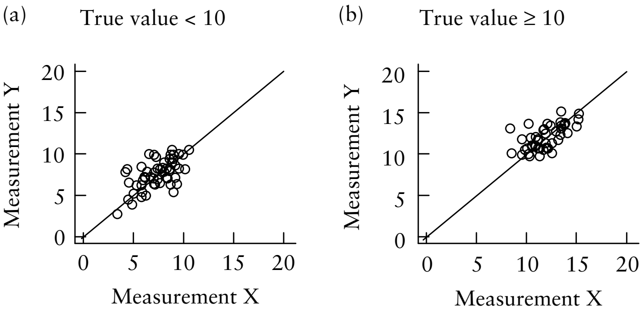

Figure 4 Artificial measurements X and Y, divided by whether the true value were < 10 (a) or ≥ 10 (b).

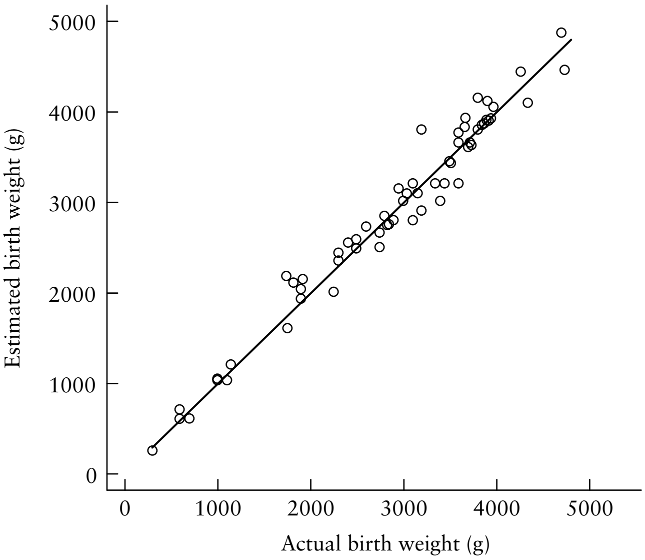

Figure 5 Final fetal weight estimated by three-dimensional ultrasound plotted against actual birth weight[3].

measurement studies, samples have often been chosen to contain far more subjects with extreme values than would a representative sample. This is often a desirable feature in such studies, as we want to explore the differences between methods over the full range of measurements, but it affects the interpretation of some statistics, including correlation coefficients. Figure 5 shows an example comparing final fetal weight, estimated by three-dimensional ultrasound, with actual birth weight[3]. The sample contains far more small birth weights than would a representative sample of births. The actual birth weight has a mean of 2930 g and the standard deviation (SD) is 1050 g, compared with the mean of 3300 (SD, 560) g observed in an unselected UK sample[4]. It is thus much more variable than a representative sample and would produce a higher correlation coefficient.

The second problem with correlation is that it looks at the degree of association, not agreement. If we have a third measurement Z, obtained by adding 2.0 to X, this will consistently overestimate the true value by 2 units. However, the correlation of Y with Z is the same as its correlation with X, 0.86 (Figure 6). While the correlation between Y and Z is the same as that between Y and X, the agreement is not. Thus high correlation does not imply close agreement; it is blind to the possibility of bias. The clear bias in Figure 2 is a good example. (Of course, in a real example we would not know which method were biased compared with the true value.)

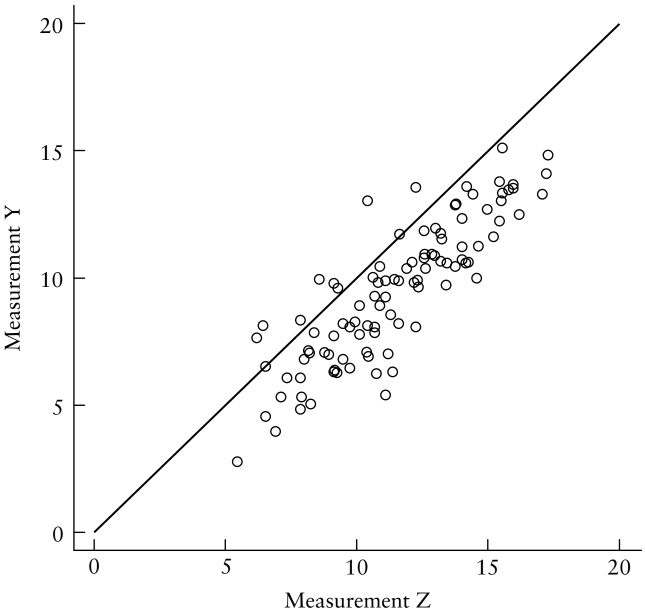

Figure 6 Artificial measurements Y and Z, with the line of equality.

Bias can be very large indeed. For example, Bakker *et al.*[5] investigated the agreement between renal volume measurements by ultrasound and by magnetic resonance imaging. Their data for 40 kidneys are shown in Figure 7, in the form that they used. There is a clear and significant mean difference of about 25% between the two methods. A correlation coefficient would completely miss this difference and would thus be highly misleading. We think that the consistent difference is even clearer if the data are presented as a scatter diagram (Figure 8).

When dealing with intraobserver variation using the same method of measurement, where the repeated observations are made by the same observer on the same subject, there should not be any consistent bias. We can then use correlation, again provided the sample is representative. But if we are comparing two different methods of measurement, there may well be a consistent bias, and the correlation coefficient could be quite misleading. Correlation will tell us something about whether the two methods are measuring the same underlying quantity, i.e. about the validity of the two methods, but not about their agreement and whether they can be used interchangeably.

Correlation is thus inappropriate for the study of agreement between different methods of measurement. Despite this, people do it.

REGRESSION LINES

Some applications of regression are also inappropriate. It is often thought that, as the data should cluster around the line of equality for good agreement, the regression line should be similar to the line of equality. This is

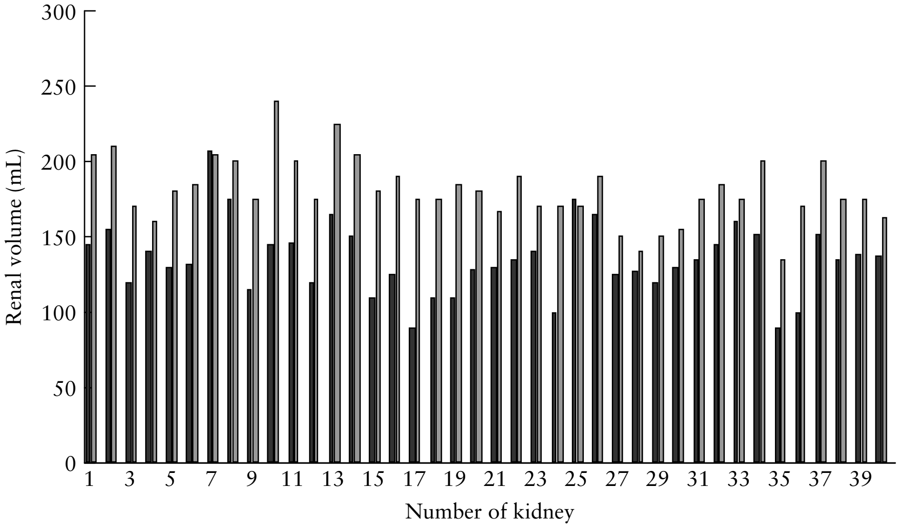

Figure 7 Renal volume measurements by ultrasound (■) and magnetic resonance (▨) imaging[5].

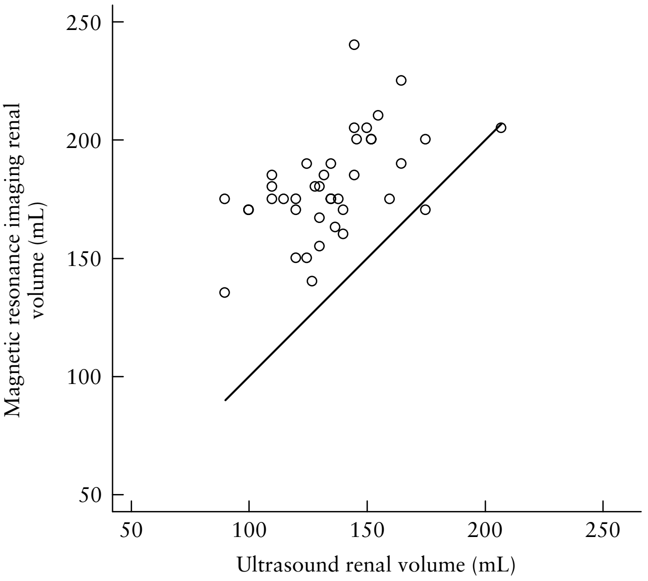

Figure 8 Renal volume measurements by ultrasound and magnetic resonance imaging[5] (data of Figure 7) presented as a scatter diagram, with the line of equality.

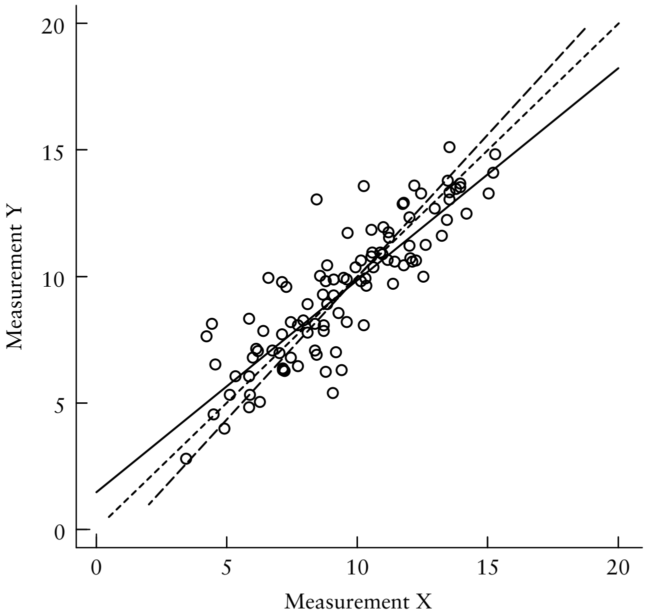

Figure 9 Artificial measurements X and Y with the line of equality (- - - -) and regression lines of both Y on X (——) and X on Y (– – –).

not so. For our randomly generated data, which do lie about the line of equality, the regression line of Y on X is shown in Figure 9; the equation is Y = 1.49 + 0.84 X. The regression line does not coincide with the line of equality, which has equation Y = 0.0 + 1.0 X. It does not go through the origin and its slope is less than one. The 95% confidence interval (CI) for the slope is 0.74–0.94. The slope is therefore significantly different from 1.0. Similarly, the intercept has a 95% CI of 0.51–2.47 and is significantly different from 0.0. The cause of the discrepancy is that regression attempts to predict the observed Y from the observed X, not the true Y from the true X. Measurement errors in X reduce the slope of the line and so raise the lower end of the line and lower the upper end, so that the intercept is increased above zero. Figure 9 also shows the line for the regression with X as the dependent variable and Y as independent, X = 1.12 + 0.89 Y. Again, the slope is significantly less than one (95% CI, 0.78–0.99) and the intercept significantly greater than zero (95% CI, 0.09–2.15). On the scatter diagram the line is plotted with Y as the dependent variable, and so its slope is shown as 1/0.89 and appears > 1.0, but for a method comparison a regression analysis is expected to give a slope < 1.0.

So when we regress measurements by one method on measurements by another, we expect that the slope will be less than one and the intercept greater than zero, whichever way round we do the regression. A slope < 1.0 and an intercept > 0.0 thus do not tell us anything. This is not always understood by researchers.

For example, Bankier *et al.*[6] compared subjective visual grading vs. objective quantification with macroscopic morphometry and thin-section computed tomographic (CT) densitometry in pulmonary emphysema (Table 1). These measurements were not on the same scale, but all scales had a common point at zero. Bankier *et al.* interpret this table thus: 'All but one of the CIs did not contain zero, which is suggestive of systematic overestimation of emphysema when compared with objective measurements.'[6]. We disagree. This is what we would expect to see if there were no such bias.

Others have tested the null hypothesis that the slope is equal to 1.0. For example, Tothill *et al.*[7] studied absorptiometers used for measuring total body bone and soft tissue. Table 2 shows some of their results looking at the relationship between the known density in a phantom (artificial model) with the measured density. The tests of significance are irrelevant and unnecessary, because the null hypotheses of zero intercept and unit slope are not expected to be true. Tothill *et al.*[7] also cite Prior *et al.*[8] who compared body composition measurements by a Hologic QDR1000W with a four-compartment hydrodensitometry model. The regression equation of percent fat for the four-compartment model (y) on percent fat measured by the Hologic 1000 (x) was $y = 3.30 + 0.85x$. This was interpreted as 'the [Hologic] 1000 underestimated fat in the leanest women and overestimated it in the fattest'[7]. As we have seen, we expect the intercept to exceed zero and the slope to be less than one when there is no relationship between the error and the magnitude, as with our artificial data. This interpretation is therefore incorrect. It is a well-known instance of regression towards the mean[9,10]. We shall consider an appropriate application of regression later.

Table 1 Results of a study of subjective visual grading vs. objective quantification with macroscopic morphometry and thin-section computed tomographic (CT) densitometry in pulmonary emphysema[6]: linear regression results

Reader	*Subjective score and densitometric measurement*	*Subjective score and morphometric measurement*
1	0.350, 1.059	0.629, 1.365
2	– 0.008, 0.598	0.443, 1.147
3	0.002, 0.658	0.854, 1.038

Data are 95% CIs for the intercepts of regression lines.

Table 2 Regression equations for correlations between measured and nominal bone mineral density in hardboard plus aluminium whole body phantom[7]

	Intercept, a	*Slope,* b	*Correlation,* r
Hologic QDR 4500A			
Legs	0.22	0.77	0.990
Arms	0.11	0.88	0.998
Spine	0.08	0.82	0.986
Lunar expert			
Legs	0.09	0.96	0.999
Arms	0.05	0.99	0.999
Spine	0.10	0.85	0.990

All intercepts (a) are significantly higher than zero and all slopes (b) are significantly lower than 1.0 ($P < 0.05$).

THE 95% LIMITS OF AGREEMENT

When we wrote our first paper on comparing methods of measurement[2], we were mainly concerned with exposing these potentially misleading approaches. However, we thought that we should also suggest a valid method of analysis. Our proposal was the 95% limits of agreement method. We started with the differences between measurements by two methods, which we thought we should summarize. We calculated the mean and SD of these differences. Then we calculated the mean difference ± 2 (or, more precisely, 1.96) SDs. We would then expect 95% of differences between measurements by two methods to lie between these limits.

For the artificial X and Y data of Figure 3, the differences Y – X have mean – 0.06 and SD 1.46. Hence the 95% limits are $-0.06-(1.96\times1.46)=-2.92$ and $-0.06+(1.96\times1.46)=2.80$. Hence, we can say that for 95% of individuals, a measurement by Method Y would be between 2.92 units less and 2.80 units greater than a measurement by Method X. We thought that this approach was so obvious and so clearly answered the question that it needed no justification and we therefore did not go into detail. In a later paper[11] we elaborated the idea and gave a worked example.

For a recent practical example, Cicero *et al.*[12] compared cervical length at 22–24 weeks of gestation measured by transvaginal and transperineal-translabial ultrasonography. Their data are shown in Figure 10. The limits of

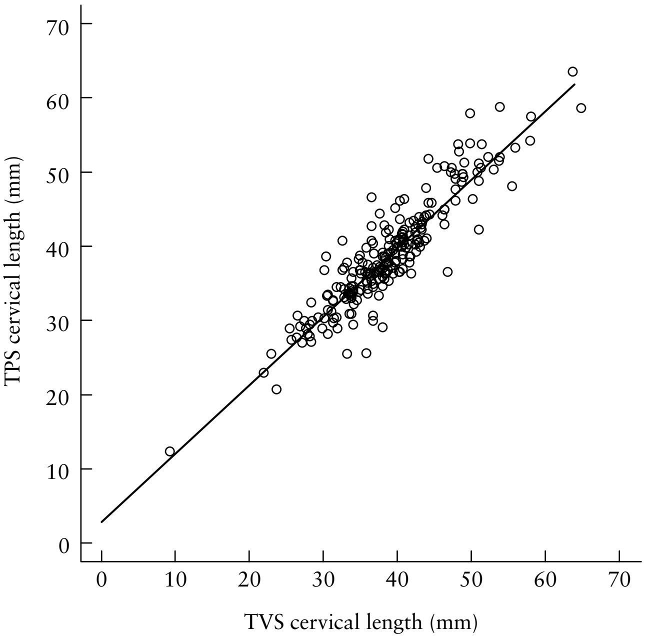

Figure 10 Cervical length at 22–24 weeks of gestation measured by transvaginal (TVS) and transperineal-translabial (TPS) ultrasonography[12].

agreement were quoted as − 5.8 mm to 6.1 mm[12]. Interestingly, these authors also quoted $r = 0.934$, $P < 0.0001$, and Figure 10 shows the regression line, not the line of equality[12]; old habits die hard.

How small the limits of agreement should be for us to conclude that the methods agree sufficiently is a clinical, not a statistical, decision. This decision should be made in advance of the analysis.

The 95% limits of agreement depend on some assumptions about the data: that the mean and SD of the differences are constant throughout the range of measurements, and that these differences are from an approximately Normal distribution. To check these assumptions we proposed two plots: a scatter diagram of the difference against the average of the two measurements and a histogram of the differences[2]. For the X and Y data, these are shown in Figure 11. The mean and SD appear uniform through the range of measurements and the differences appear to follow a Normal distribution, as they were artificially generated to do.

We also suggested adding the 95% limits of agreement and the mean difference to the scatter plot, as shown in Figure 12. About 95% of points should lie within the limits. In Figure 12, 93% are within the 95% limits and 7% are outside. Cicero *et al.*[12] showed such a plot (Figure 13). In this graph there are many overlapping points, and there are in fact 15/231 = 93.5% of the points within the 95% limits.

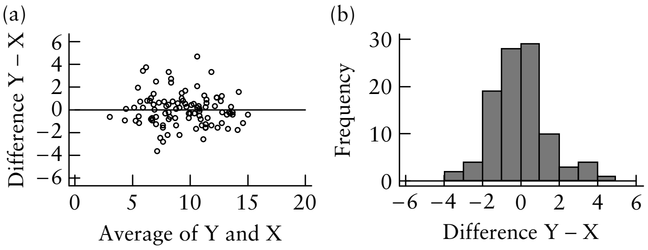

Figure 11 Plots of difference against mean (a) and histogram of differences (b) for the artificial data X and Y.

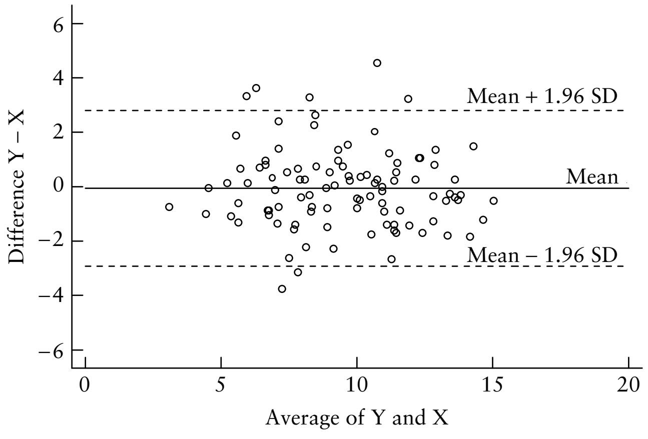

Figure 12 Plot of difference against mean for the artificial data X and Y (as in Figure 11a), with mean difference and 95% limits of agreement indicated.

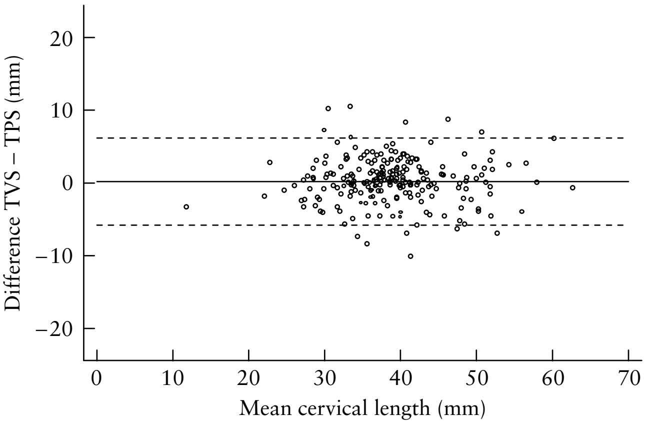

Figure 13 Difference against mean plot for cervical length measured by transvaginal (TVS) and transperineal-translabial (TPS) ultrasonography[12] (data of Figure 10).

To our chagrin, the histogram does not seem to have been adopted with the same enthusiasm, but the scatter plot alone is a reasonable check. Also to our chagrin, many researchers seem to think that the plot is the analysis. It is not, of course, but only a check on the assumptions of the limits of agreement.

The assumptions are not always met and checking is essential. In their study of fetal lung volume measurement using three-dimensional ultrasound, Bahmaie *et al.*[13] produced a difference against mean plot for measurements by two different observers (Figure 14). This shows a divergence as the magnitude increases, making the limits of agreement suspect. They would be too wide for small measurements and too narrow for large ones. Often the differences increase in size proportionally to the size of the measurement. We can resolve this difficulty by analyzing the logarithm of the measurement rather than the measurement itself. This leads to limits of agreement in the form of proportions of the measurement rather than in the original units[11]. Another, similar, solution is to find the 95% limits for the difference as a percentage of the average of the two methods[14].

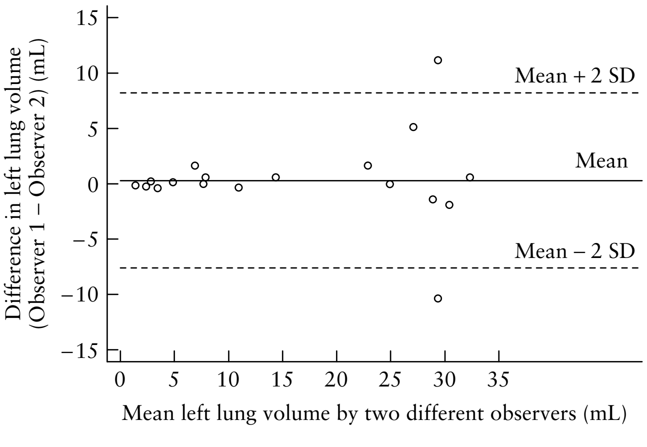

Figure 14 Difference against mean plot for measurements of fetal lung volume by two different observers using three-dimensional ultrasound[13].

The reason for plotting the difference against the average, rather than either of the measurements singly, is that when there is no real relationship between difference and magnitude, Y − X and X will be negatively correlated. Subjects for whom the X measurement is larger than their true value will tend to have negative Y − X differences, while subjects for whom the X measurement is smaller than their true value will tend to have positive Y − X differences. The subjects with the largest X measurements are likely to include those whose X measurement is above the true value; subjects with small X measurements are likely to include those with X below the true value. Hence Y − X will go down as X goes up. Similarly, Y − X and Y will be positively correlated. However, when X and Y have the same SD, as they should if they are measurements of the same thing, Y − X and Y + X should not be correlated at all in the absence of a true relationship[15]. For the X and Y data, where there is no relationship between difference and magnitude, these correlations are shown in Table 3. In this example, Y − X and X are negatively correlated and Y − X and Y are positively correlated, and these correlations are both statistically significant. In contrast, Y − X and the average of Y and X have a very small correlation, which is not significant. We can see this in the plots of difference against X and against Y (Figure 15).

Plots of difference against one measurement can be seriously misleading. Figure 5 shows final fetal weight estimated by three-dimensional ultrasound against actual birth weight[3]. There is no evidence of any tendency for the points to divert from the line of equality in the scatter diagram. The authors also plotted the percentage error (100 × (ultrasound – actual)/actual, essentially the difference) against the actual birth weight (Figure 16). The downward trend in the graph is readily apparent. If there were really a relationship such that ultrasound overestimated for low actual weights and underestimated for high ones, as Figure 16 suggests, this would be apparent in Figure 5 also. It is not. Consider the following from a comparison of magnetic resonance (MR) imaging with conventional arthrography:

Table 3 Correlations, using artificial data X and Y, of difference Y − X with the single measurements and with their average when there is no relationship between difference and magnitude

	Correlation with difference Y − X	
	r	P
Average of X & Y	− 0.06	0.6
X	− 0.31	0.002
Y	+ 0.21	0.04

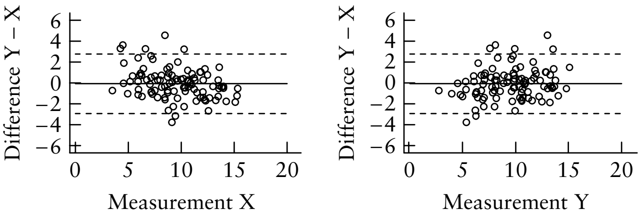

Figure 15 Plots of difference against each measurement for the artificial data X and Y, with mean difference and 95% limits of agreement indicated.

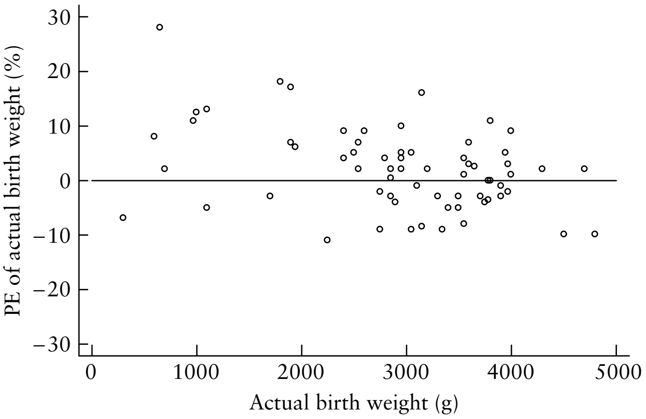

Figure 16 Fetal weight measured by three-dimensional ultrasound and actual birth weight: percentage error (PE) against actual birth weight[3] (data of Figure 5).

> 'For all parameters, the difference between the measurements of the two modalities depended on the magnitude of the measurements. By using arthrography as the standard, a slope test indicated overestimation with MR imaging at small measurements and an underestimation at large measurements (all $P < 0.001$)'[16].

If the difference MR minus arthrography were regressed on arthrography, we would expect them to be negatively related even in the absence of a true relationship between difference and magnitude[15].

The 95% limits of agreement method has been widely cited and quite widely used[17], though many who cite it do not appear to have read the paper. For example, in the MR vs. arthrography study cited above, the authors state in the methods section:

> 'For each parameter, agreement between MR imaging and arthrography was investigated using the method of Bland and Altman [1986]. Arthrography was considered to be the standard and differences between methods were calculated and plotted. A slope test was used to assess whether these differences varied systematically over the range of measurements'[16].

The results section of the paper contains no limits of agreement, but rather correlation and rank correlation coefficients with *P*-values! As for plotting difference against a standard measurement, we actually wrote:

> 'It would be a mistake to plot the difference against either value separately because the difference will be related to each, a well-known statistical artefact'[11].

APPROPRIATE USE OF REGRESSION

We mentioned earlier that there is an appropriate use of regression in the evaluation of agreement. This is particularly useful when the two methods of measurement have different units, as in the study of subjective visual grading vs. objective quantification with thin-section CT densitometry[6] described above. In this case we could not simply replace a measurement by one method with a measurement by the other, as they are not measuring the same quantity. However, we could predict what the measurement by the old method would be, given the new method. If this method agrees well with the old-method measurement, then the two methods give similar information and we could replace the new by the old. We start by regressing the measurement by the old method on the measurement by the new method. We can use this regression equation to estimate a predicted old-method measurement for any observed value by the new method. Of course, this will give the mean old-method value for subjects with this particular new-method value; it does not take the variation between subjects into account. We take this into account by calculating a range of possible values for the old-method value on this subject, called a 95% prediction interval. This gives us something akin to the limits of agreement. The problem is that it is not constant, being smallest near the middle of the range and wider as we get further towards the extremes. This effect is quite marked for small samples, but not for large. For the simulated X and Y data, regarding Y as the old or standard method and X as the new produces Figure 17. Here the spreading out is very small and hard to see. The average width of this prediction interval is 5.7, 2.85 on either side of the prediction. This is very similar to the width of the 95% limits of agreement, −2.92 to 2.80. If the 95% prediction interval has the width that we would find acceptable in the 95% limits of agreement, we could switch to the new method of measurement.

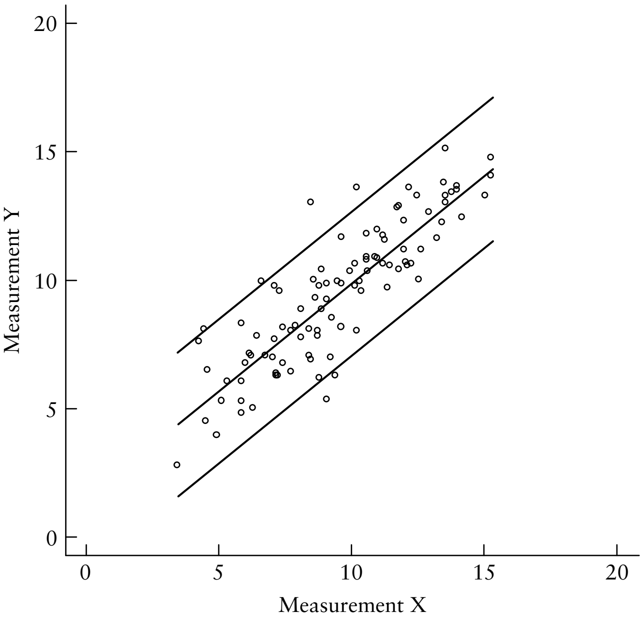

Figure 17 Regression of artificial data Y on X, with prediction limits.

MEASUREMENTS MADE USING THE SAME METHOD

Repeated measurements may be made by a single observer using the same method to investigate the measurement error, intraobserver variation, or by different observers to investigate the variation between them, interobserver variation. Interobserver variation is a complex subject and we shall not go into details here. We may have two observers and analyze them as if they were two different methods of measurement, with the potential for a consistent bias between them, as in Figure 14. We may have many observers and estimate the variance between them.

Repeated measurements by the same method can be analyzed in a manner analogous to the limits of agreement. The main difference is that the average difference should be zero, which simplifies things. If there is a consistent difference, between the first and second measurement for example, these are not true replicates of the same measurement and we cannot use them to investigate measurement error. Because there is no consistent bias, correlation can be used in the analysis of such data, provided there is a population from which the sample can be regarded as a representative sample. This is often not the case in the study of clinical measurements, where samples are often chosen to include more subjects with extremely high or low values than would a representative sample, but it is often the case in the study of measurements derived from questionnaire scales as found in psychology. In this case, the correlation coefficient is a measure of the information content of the measurement[18]. However, even when it is appropriate, the correlation coefficient does not help us to interpret a clinical measurement on a given patient. To do this we need to consider the variability between repeated measurements on the same subject. If we calculate the SD of the differences between pairs of repeated measurements, we can calculate 1.96 × SD. This gives the repeatability coefficient, which is the difference that will be exceeded by only 5% of pairs of measurements on the same subject[19]. It is thus directly comparable to the 95% limits of agreement. Thus we can use this to compare the agreement which a new method of measurement would have with a standard method, with the agreement which the new method would have with itself.

In our 1986 paper[11] we advocated a design where each method would be used twice on each subject, so that limits of agreement between the two methods and coefficients of repeatability for each method separately could be compared. We regret that this has not been more widely adopted by researchers.

CONFIDENCE INTERVALS FOR THE 95% LIMITS OF AGREEMENT

Another feature which we stressed in our 1986 paper was that agreement is a question of estimation, not hypothesis testing. Estimates are usually made with some sampling error, and limits of agreement are no exception. We showed how to estimate CIs for the limits of agreement. Another regret is that these CIs are seldom quoted. For the data of Cicero *et al.*[12] the mean difference was 0.2 mm with SD 3.0 mm, giving 95% limits of agreements -5.8 to $+6.1$ mm. There were 231 cases. The standard error of the limits is approximately $\sqrt{3s^2/n}$. This gives $\sqrt{3 \times 3.0^2/231} = 0.34$. The 95% CI for the limits of agreement is given by ± 1.96 standard errors $= 0.67$, so for the lower limit the CI is -6.5 to -5.1 and for the upper limit the 95% CI is $+5.4$ to $+6.8$. Not so hard, really!

Many studies are done using far fewer subjects than were included in the study of Cicero *et al.*[12] and the CIs would therefore be much wider.

CONCLUSIONS

The limits of agreement approach is fundamentally very simple and direct. Provided its assumptions of uniform mean and SD are met, it can be carried out by anyone with basic statistical knowledge. It provides statistics that are easy to interpret in a meaningful way. It can be extended to many more complex situations[20], when distributions are not Normal, when difference is related to magnitude, when there are repeated measurements on the same subject, either paired or not, and when there are varying numbers of observations on subjects. There is also a non-parametric version.

ACKNOWLEDGMENTS

We are very grateful to Dr S. Cicero and Prof. K. H. Nicolaides for the cervical length data.

REFERENCES

1. Borg J, Møllgaard A, Riis BJ. Single x-ray absorptiometry: performance characteristics and comparison with single photon absorptiometry. *Osteoporos Int* 1995; **5**: 377–381.
2. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician* 1983; **32**: 307–317.
3. Schild RL, Fimmers R, Hansmann M. Fetal weight estimation by three-dimensional ultrasound. *Ultrasound Obstet Gynecol* 2000; **16**: 445–452.
4. Brooke OG, Anderson HR, Bland JM, Peacock JL, Stewart CM. Effects on birth weight of smoking, alcohol, caffeine, socioeconomic factors, and psychosocial stress. *BMJ* 1989; **298**: 795–801.
5. Bakker J, Olree M, Kaatee R, de Lange EE, Moons KGM, Beutler JJ, Beek FJA. Renal volume measurements: accuracy and repeatability of US compared with that of MR imaging. *Radiology* 1999; **211**: 623–628.
6. Bankier AA, De Maertelaer V, Keyzer C, Gevenois PA. Pulmonary emphysema: Subjective visual grading versus objective quantification with macroscopic morphometry and thin-section CT densitometry. *Radiology* 1999; **211**: 851–858.
7. Tothill P, Hannan WJ, Wilkinson S. Comparisons between a pencil beam and two fan beam dual energy X-ray absorptiometers used for measuring total body bone and soft tissue. *Br J Radiol* 2001; **74**: 166–176.
8. Prior BM, Cureton KJ, Modlesky CM, Evans EM, Sloniger MA, Saunders M, Lewis RD. In vivo validation of whole body composition estimates from dual-energy X-ray absorptiometry. *J Appl Physiol* 1997; **83**: 623–630.
9. Bland JM, Altman DG. Statistics Notes. Regression towards the mean. *BMJ* 1994; **308**: 1499.
10. Bland JM, Altman DG. Statistics Notes. Some examples of regression towards the mean. *BMJ* 1994; **309**: 780.
11. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **i**: 307–310.
12. Cicero S, Skentou C, Souka A, To MS, Nicolaides KH. Cervical length at 22–24 weeks of gestation: comparison of transvaginal and transperineal-translabial ultrasonography. *Ultrasound Obstet Gynecol* 2001; **17**: 335–340.
13. Bahmaie A, Hughes SW, Clark T, Milner A, Saunders J, Tilling K, Maxwell DJ. Serial fetal lung volume measurement using three-dimensional ultrasound. *Ultrasound Obstet Gynecol* 2000; **16**: 154–158.
14. Linnet K, Bruunshuus I. HPLC with enzymatic detection as a candidate reference method for serum creatinine. *Clin Chem* 1991; **37**: 1669–1675.
15. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* 1995; **346**: 1085–1087.
16. Jaramillo D, Galen T, Winalski CS, DiCanzio J, Zurakowski D, Mulkern RV, McDougall PA, Villegas-Medina OL, Jolesz FA, Kasser JR. Legg-Calvé-Perthes disease: MR imaging evaluation during manual positioning of the hip – comparison with conventional arthrography. *Radiology* 1999; **212**: 519–525.
17. Bland JM, Altman DG. This week's citation classic: Comparing methods of clinical measurement. *Current Contents* 1992; **CM20**(40): 8.
18. Bland JM, Altman DG. Statistics Notes. Measurement error and correlation coefficients. *BMJ* 1996; **313**: 41–42.
19. British Standards Institution. *Precision of Test Methods 1: Guide for the Determination and Reproducibility for a Standard Test Method (BS 597, Part 1)*. BSI: London, 1975.
20. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999; **8**: 135–160.