Ultrasound Obstet Gynecol 2008; **31**: 466–475
Published online 27 February 2008 in Wiley InterScience (www.interscience.wiley.com). **DOI:** 10.1002/uog.5256

Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables

J. W. BARTLETT and C. FROST

Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London, UK

KEYWORDS: agreement; measurement error; method comparsion; reliability; repeatability; reproducibility

Clinical practice involves measuring quantities for a variety of purposes, such as aiding diagnosis, predicting future patient outcomes, and serving as endpoints in studies or randomized trials. Measurements are almost always prone to various sorts of errors, which cause the measured value to differ from the true value; accordingly, studies investigating measurement error frequently appear in this and other journals.

The importance of measurement error depends upon the context in which the measurements in question are to be used. For example, a certain degree of measurement error may be acceptable if measurements are to be used as an outcome in a comparative study such as a clinical trial, but the same measurement errors may be unacceptably large to make measurements usable in individual patient management, such as screening or risk prediction.

In the past 20 years many papers have been published advocating how studies of measurement error should be analyzed, with a paper by Bland and Altman[1] being one of the most cited and well known examples. There has been much controversy concerning the choice of parameter to be estimated and reported, and consequently confusion surrounding the meaning and interpretation of results from studies investigating measurement error.

In this paper we first distinguish between the general concepts of agreement and reliability to aid researchers in considering which are relevant for their particular application. We then review the statistical methods that can be used to investigate and quantify agreement and reliability, dealing separately with the different types of measurement error study, while emphasizing the largely common techniques that should be used for data analysis. We reiterate that the judgment of whether agreement or reliability are acceptable must be related to the clinical application, and cannot be proven by a statistical test.

We highlight the fact that reliability depends on the population in which measurements are made, and not just on the measurement errors of the measurement method. We discuss the advantages of method comparison studies making at least two measurements with each measurement method on each subject. A key advantage is that the cause of a correlation between paired differences and means in the so-called Bland–Altman plot can be determined, in contrast to when only a single measurement is made with each method.

Throughout the paper, we try to emphasize that calculated values of agreement and reliability from measurement error studies are estimates of parameters, and as such we should report such estimates with CIs to indicate the uncertainty with which they have been estimated. We restrict our attention to measurements of a continuous quantity; alternative methods are required for categorical data[2].

TERMINOLOGY AND TYPES OF STUDY

One difficulty in the measurement error field is the number of different terms used to describe studies of measurement error. The terms ‘agreement’, ‘reliability’, ‘reproducibility’ and ‘repeatability’ are used with varying degrees of consistency in the medical literature. We first make clear the distinction between the statistical concepts of agreement and reliability[3].

Agreement and reliability

Agreement quantifies how close two measurements made on the same subject are, and is measured on the same scale as the measurements themselves. Two measurements of the same subject may differ for a number of reasons, depending on the conditions under which the measurements were made. In a method comparison study there will be differences because of inherent variability in each of the measurement methods, as well as potentially a bias between the measurements from the methods. If the

Correspondence to: Mr J. W. Bartlett, Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK (e-mail: Jonathan.Bartlett@lshtm.ac.uk)

Accepted: 21 December 2007

measurements are made by different observers or raters, differences may be due to biases between the observers.

Agreement between measurements is a characteristic of the measurement method(s) involved, which does not depend on the population in which measurements are made, unless bias or measurement precision varies with the true value being measured. One popular way of quantifying agreement is to estimate the 95% limits of agreement, as proposed by Bland and Altman[1]. These limits are defined such that we expect that, in the long run, 95% of future differences between measurements made on the same subject will lie within the limits.

Reliability relates the magnitude of the measurement error in observed measurements to the inherent variability in the 'error-free', 'true', or underlying level of the quantity (terms we shall use synonymously) between subjects. These measures of variability can be expressed as standard deviation (SDs), and formally reliability is defined as

$$\frac{(\text{SD of subject's true values})^2}{(\text{SD subjects' true values})^2 + (\text{SD measurement error})^2}.$$

If reliability is high, measurement errors are small in comparison to the true differences between subjects, so that subjects can be relatively well distinguished (in terms of the quantity being measured) on the basis of the error-prone measurements. Conversely, if measurement errors tend to be large compared with the true differences between subjects, reliability will be low because differences between measurements of two subjects could be due purely to error rather than to a genuine difference in their true values.

The reliability parameter is also known as an intraclass correlation (ICC), as it equals the correlation between any two measurements made on the same subject. Reliability takes values between zero and one, with a value of one corresponding to zero measurement error and a value of zero meaning that all the variability in measurements is due to measurement error. As a dimensionless quantity, it is arguably quite difficult to interpret, and deciding what value constitutes sufficiently high reliability is often made in a subjective fashion.

Repeatability and reproducibility

Repeatability of measurements refers to the variation in repeat measurements made on the same subject under identical conditions[4]. This means that measurements are made by the same instrument or method, the same observer (or rater) if human input is required, and that the measurements are made over a short period of time, over which the underlying value can be considered to be constant. Variability in measurements made on the same subject in a repeatability study can then be ascribed only to errors due to the measurement process itself.

Reproducibility refers to the variation in measurements made on a subject under changing conditions[4]. The changing conditions may be due to different measurement methods or instruments being used, measurements being made by different observers or raters, or measurements being made over a period of time, within which the 'error-free' level of the variable could undergo non-negligible change.

Study types

The first type of study we consider is a repeatability study, in which we investigate and quantify the repeatability of measurements made by a single instrument or method, and in which the conditions of measurement remain constant.

The second type of study we consider is a method comparison study, in which measurements are made using two measurement methods on a sample of subjects. The use of two different methods means that this is a reproducibility study, but the term method comparison is used as it clearly communicates the changing conditions under which measurements have been made. In contrast to a repeatability study, systematic bias may exist between measurements made by two different methods, and their measurement errors may have different SDs.

The last type of study we consider is one in which measurements are made by different observers or raters. Again this is a type of reproducibility study. As with method comparison studies, biases may exist between observers, and their measurement SDs may differ. As we discuss below (see Measurements with observers or raters), if interest lies in quantifying the measurement error characteristics of the particular observers in one's study, exactly the same analysis methods should be used as for a method comparison study.

REPEATABILITY STUDIES

In order to investigate the repeatability of measurements, a repeatability study must, for an appropriately selected sample, make at least two measurements per subject under identical conditions. This means that the measurements must be made by the same measurement method, or the same observer or rater. The objective is to then quantify the agreement and reliability of measurements made by that particular method or observer.

Quantifying agreement

In a repeatability study we exclude the possibility of bias between measurements, so that agreement between measurements made on the same subject depends only on the within-subject SD, which measures the size of measurement errors. One way to describe the agreement is to report an estimate of the within-subject SD, which is the same as the SD of the measurement errors. An alternative is to report the SD of the differences between two measurements made on the same subject. This is equal to

$$\sqrt{2} \times \text{within\_subject SD}.$$

A further alternative is to report the estimated repeatability coefficient, which is defined by

$$1.96 \times \sqrt{2} \times \text{within\_subject SD}.$$

If the differences between two measurements made on a subject are approximately Normally distributed, in the long run we expect the absolute difference between two measurements on a subject to differ by no more than the repeatability coefficient on 95% of occasions.

To estimate the within-subject SD, we can fit a one-way analysis of variance (ANOVA) model to the data containing the repeat measurements made on subjects. ANOVAs can be fitted in all modern statistical packages. ANOVA models partition variability in data into that which can be ascribed to differences between groups, and that remaining to within groups. For a repeatability study, the groups are defined by the subjects under measurement, and this must be specified in the statistical package used. The ANOVA model estimates how much variation in measurements can be attributed to differences in the true, or 'error-free' values of subjects, with the remainder constituting measurement error.

Fitting the ANOVA model results in estimates of the between-subject and within-subject SDs (or alternatively the corresponding variances, which are the SDs squared). The estimate of the within-subject SD can be used in the above formula to give an estimate of the repeatability coefficient.

We illustrate with a study by Järvelä *et al.*, who investigated the repeatability of measurements of flow volume made by one observer using three-dimensional (3D) power Doppler ultrasonography from 22 ovaries[5]. The estimated repeatability coefficient was 3.97, meaning that the absolute difference between any two future measurements made by that particular observer on a particular subject/unit are estimated to be no greater than 3.97 on 95% of occasions. It is important to note that the repeatability of another observer may be different, because of differences in the training and ability of observers.

Because the repeatability coefficient calculated is an estimate, it is important to calculate a CI for it to indicate how precisely it has been estimated. A CI may be given automatically by statistical software, but in the Appendix we review how a 95% CI for the within-subject SD can be calculated. Such a CI can be used to find a CI for the repeatability coefficient by multiplying the CI limits by $1.96 \times \sqrt{2}$. If the CI for the within-subject variance is given by software instead, the limits must first be square rooted to give a CI for the within-subject SD.

The ANOVA model assumes that the measurement errors are statistically independent of the true 'error-free' value, and that the SD of the errors is constant throughout the range of 'error-free' values. Sometimes the SD of errors increases with the true value being measured. This should be checked by plotting paired differences between measurements against their mean, the so called Bland–Altman plot. We illustrate this later in the context of a method comparison study (see Plotting the data) and describe how such heterogeneous errors can often be dealt with by making a log transformation (see Non-uniform variability in measurement errors).

Use of the repeatability coefficient relies on the differences between measurements being approximately Normally distributed. This can be checked by a histogram or Normal plot of the paired differences in measurements on each subject. An assumption made in the construction of the CI for the within-subject SD is that the measurement errors are normally distributed. If this assumption is in doubt, the bootstrapping technique may be employed to obtain CIs[6].

Estimating reliability

The reliability of a measurement method is often of interest when measurements are to be used to differentiate between subjects or groups of subjects. For example, if we have a choice of two measurement methods that could be used to measure an outcome in a clinical trial or observational study, using the method with higher reliability will give greater statistical power to detect differences between groups, for a given sample size.

In this subsection we describe how measurements from a single method can be used to estimate the method's reliability in a given population. Later we discuss the use of reliability to compare measurement methods (see Reliability in method comparison studies) and different observers (see Interest in specific observers).

The same ANOVA model as described above (see Quantifying agreement) can be used to estimate reliability. Some statistical packages will automatically give the estimated ICC, which is the measure of reliability. Otherwise, we can substitute the estimates of the between- and within-subject SDs into the formula given above (see Agreement and reliability) to estimate the ICC.

For the flow index data of Järvelä *et al.* an estimated ICC of 0.82 was reported[5]. This means that 82% of the variability in measurements of the flow index was estimated to be due to genuine differences in flow index between ovaries, with the remaining 18% being due to errors in the measurement process and the observer involved. Because the measurements were all made by one observer, the reliability may be referred to as intraobserver reliability.

As with the repeatability coefficient, the calculated ICC is an estimate, and a 95% CI should be given. Some statistical packages give a CI, but in the Appendix we review the calculation of a 95% CI for the ICC. For the flow index data of Järvelä *et al.* the ICC estimate of 0.82 had a 95% confidence interval of 0.74 to 0.93.

Because we use the same ANOVA model to estimate reliability as we describe for estimating agreement, the same assumptions apply. Additionally, calculation of 95% CIs for the ICC estimate relies on Normality of the true 'error-free' values.

Reliability depends on population heterogeneity

The reliability of a measurement method depends upon the heterogeneity of the population in which the measurements are made. From the definition of reliability given above (see Agreement and reliability), we see that the heterogeneity of subjects' true 'error-free' values in

the population, measured by the between-subject SD, affects the value of reliability. Thus, whereas agreement between repeat measurements is a characteristic of the method or instrument (assuming the distribution of measurement errors is uniform across the range of true values), reliability depends on both the magnitude of measurement errors and the true heterogeneity in the population in which measurements are made.

The point is best illustrated with a hypothetical example. Suppose a new method for measuring volumes has a within-subject SD of 10 cm^3, which does not vary with the underlying value being measured. Suppose we perform a study to estimate the technique's reliability, and we sample from a heterogeneous population in which the between-subject SD in true volumes is 20 cm^3. This would give a reliability or ICC of $\frac{20^2}{20^2 + 10^2} = 0.8$. Now suppose instead we sample our subjects from a more homogeneous population, in which the SD of subjects' true values is 10 cm^3, the same as the variability of the within-subject error. In this case, the reliability (ICC) is equal to $\frac{10^2}{10^2 + 10^2} = 0.5$.

If studies report only an estimate of the reliability (ICC), readers can only make use of the estimate if the population in which the reader intends to use such measurements has equal heterogeneity. We suggest that investigators report estimates of between- and within-subject SD, in addition to the ICC estimate. In this way, readers can judge whether the measurement method will be sufficiently reliable for their application, in which the heterogeneity between subjects may be different.

METHOD COMPARISON

Before we use a new measurement method or technique in clinical practice, we must ensure that the measurements it gives are 'sufficiently similar' to those generated by the measurement method currently used, i.e. that measurements made with the old method are reproducible using the new method. If the measurements from the two methods are sufficiently close and the management of a patient on their basis would be the same, the new method could replace the established method in clinical practice, perhaps because the new method is cheaper or less invasive to conduct. The decision as to what constitutes 'sufficiently similar' therefore depends on how the measurements are to be used.

Agreement between methods

If the two methods' measurements are made on the same metric or scale, we can quantify their agreement. To investigate and quantify the agreement between measurements made by two methods, we must at a minimum measure a sample of subjects using both the established measurement method and the newly proposed method. The data from such a study therefore consist of pairs of measurements from each subject, with the pair containing the subjects' measurements from the two measurement methods. Later we advocate making two measurements on each subject using each method but, because the one measurement per method is the most common design, we begin by describing its analysis.

Plotting the data

The first step to analyzing such a dataset is to plot the data. The simplest plot is of subjects' measurements from the new method against those from the established method (or vice versa). If both measurements were completely free from error, we would expect the points to lie on the diagonal line of equality. Overlaying the line of equality can be used to examine if there is bias; if more data points are above the line than below, this suggests that the method on the vertical axis gives larger measurements on average.

Investigators sometimes show the line of best fit for this plot and report a statistical significance test for whether the slope of the plot differs from the line of equality. As noted by Bland and Altman in a recent paper in this journal[7], we would expect the best-fit line to have a slope shallower than one if the method plotted on the horizontal axis contains any measurement error, and so a significance test of the hypothesis that the true slope is equal to one is not a good idea.

Although it contains the same information as a scatter plot with the line of equality, visual assessment of the disagreements between the measurements from two methods is often more easily done by plotting the difference in a subject's measurements from the two methods against the mean of their measurements, as first suggested by Altman and Bland[8]. Indeed, this is now commonly referred to as the Bland–Altman plot, and is frequently shown in publications of measurement error studies.

We illustrate the plot using an example from the imaging literature. Ruano *et al.* examined fetal lung volume measurement in eight cases with congenital diaphragmatic hernia and 25 controls without pulmonary malformation, immediately before termination[9]. Fetal lung volume was measured using 3D ultrasound imaging and also post mortem by water displacement. We have approximately digitized their plot of differences against means and it is reproduced here as Figure 1, using only data from the 25 controls.

The plot shows the difference between lung volumes calculated using ultrasound examination and water displacement against the mean of these two values. Unfortunately, Ruano *et al.* did not label their plot or table of results to indicate which method's measurements were subtracted from which. It is obviously important to be consistent and always subtract the same method's measurement from the other's. The solid line indicates the mean of the paired differences – its distance from zero provides an estimate of the bias between the two methods. Labeling which method's measurements have been subtracted from which enables readers to see which method, on average, gave larger measurements. The

dashed lines indicate the estimated limits of agreement (and their CI limits), which we describe shortly.

The plot can be used to visually inspect the differences between measurements made by the two methods. The variability of the differences between the two methods indicates how well the methods agree. If the variability of the paired differences is uniform along the range of measurements, and there is no relationship between the difference and mean, we can quantify the agreement between the methods by estimating the limits of agreement. Having described the calculation of the limits of agreement, we illustrate how the assumptions for limits of agreement may be violated and how they can be dealt with.

Limits of agreement

The limits of agreement give a range within which we expect 95% of future differences in measurements between the two methods to lie. To estimate them, we first calculate the mean and SD of the paired differences, which for the data in Figure 1 are 0.08 cm^3 and 2.80 cm^3 respectively. If the paired differences are Normally distributed, we can calculate limits within which we expect 95% of paired differences to fall as

$$\text{mean difference} - 1.96 \times \text{SD(differences)}$$
$$\text{mean difference} + 1.96 \times \text{SD(differences)}.$$

For the data in Figure 1 this gives

$$0.08 - 1.96 \times 2.80 = -5.408$$
$$0.08 + 1.96 \times 2.80 = 5.568.$$

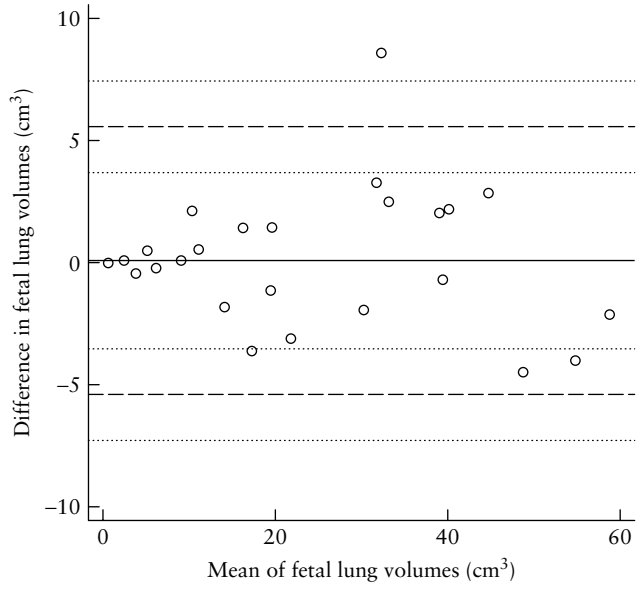

Figure 1 Differences in lung volume measured by three-dimensional ultrasound examination and by water displacement against their means, from study by Ruano *et al.*[9]. Solid line represents mean; upper dashed line shows the mean +1.96 SD and lower dashed line the mean −1.96 SD, each with 95% CI (dotted lines).

Based on these data, we therefore expect the difference in lung volume as measured by the two methods to lie between $-5.41\ cm^3$ and $5.57\ cm^3$ for 95% of future measurements.

It is important to realize that the limits of agreement calculated in this fashion are just estimates, and as with any type of estimate it is essential to quantify and report how precisely the limits are estimated through calculation of 95% CIs.

If the paired differences are Normally distributed, the standard error of the limits of agreement is approximately equal to $\frac{1.71\text{SD}}{\sqrt{n}}$, where n denotes the number of subjects in the study, and SD is the estimated SD of the paired differences[10]. For the data in Figure 1, the standard error of the limits of agreement is

$$\frac{1.71 \times 2.80}{\sqrt{25}} = 0.958.$$

We can then calculate approximate 95% CIs for the limits of agreement by taking each limit plus or minus 1.96 standard errors

$$-5.408 \pm 1.96 \times 0.958 = (-7.286, -3.530)$$
$$5.568 \pm 1.96 \times 0.958 = (3.690, 7.446).$$

The estimated limits of agreement indicate how large the disagreements between the methods' measurements will be on 95% of occasions. The decision about whether the methods agree sufficiently well must then be made depending on the context in which the measurements will be used.

Bias between methods

In contrast to the repeatability coefficient, which assumes no bias exists between measurements, the limits of agreement method relaxes this assumption. The mean of the paired differences tells us whether on average one method tended to underestimate or overestimate measurements relative to the measurements of the second method, which we refer to as a bias between the methods.

In the data in Figure 1 from Ruano *et al.*, there was little suggestion of a bias between the two methods, because the mean difference between the methods was close to zero, at 0.08 cm^3. We can perform a statistical significance test to assess whether there is evidence of bias by performing a one-sample t-test of the mean differences (a paired t-test of the measurements from each method), whereby we test the hypothesis that the true mean of the differences is zero, corresponding to no bias between the methods. For the data in Figure 1, this gives $P = 0.89$, confirming that there was no evidence of a bias between 3D ultrasound lung volumes and those made using water displacement.

We can also calculate an approximate 95% CI for the bias, by adding and subtracting 1.96 standard errors from

the mean of the paired differences. The standard error for the mean difference is given by

$$\frac{\text{SD(differences)}}{\sqrt{n}}$$

where, as before, n is the number of paired differences.

Non-uniform variability in measurement errors

The limits of agreement method assumes that the SD of the methods' differences is uniform throughout the range of measurements. Sometimes this will not be the case. In particular, frequently the SD of the differences will increase with the mean. In this case, for small values the limits of agreement will be wider than necessary, whereas for larger values the limits will be too narrow.

Examination of the plot of differences against mean from the study by Ruano *et al.* (Figure 1) suggests that the variability of the differences may be larger when the value being measured is larger. The SD of the paired differences below the median of the means was 1.50 cm^3 whereas it was 3.82 cm^3 for those above the median.

This problem can often be overcome by taking logarithms of the measurements of the two methods[1]. If taking the difference in the logarithm of the methods' measurements results in constant variability, we can calculate the limits of agreement and CIs for the limits in the usual way.

Because the difference of two logarithms is equal to the logarithm of the ratio, we can back-transform all the estimates made on the log scale by exponentiating. Because of this, after back-transformation the limits of agreement correspond to the ratio of one method's measurements to the other's. For the data in Figure 1, the mean of the difference in logged measurements was 0.004, with SD 0.109. The limits of agreement on the log scale are therefore −0.210 to 0.218. Exponentiating these gives the geometric mean ratio of 1.004 (0.4% higher), with 95% limits of agreement for the ratio of 0.81 (19% lower) to 1.24 (24% higher). The 95% CIs can similarly be calculated on the log scale before back-transforming to the ratio scale. Without knowing which measurements were made by which method, we cannot determine whether the ratio is for measurements by ultrasound imaging to those by water displacement, or vice versa.

Figure 2 shows a plot of the ratio of the measurements by the two methods to their mean, using a logarithmic vertical scale, with the estimated 95% limits of agreement and their 95% CIs. This shows that the assumption of a uniform SD for the paired differences (now ratios) is now more reasonable.

Alternatives to log transformation can also be used to try to create uniform variability. For example, Kusanovic *et al.*[11] estimate limits of agreement using paired differences in measurements as a percentage of the average of the two measurements. The key point is that a transformation should be used which creates transformed 'paired differences' that are approximately Normally distributed and have approximately uniform SD across the range of means.

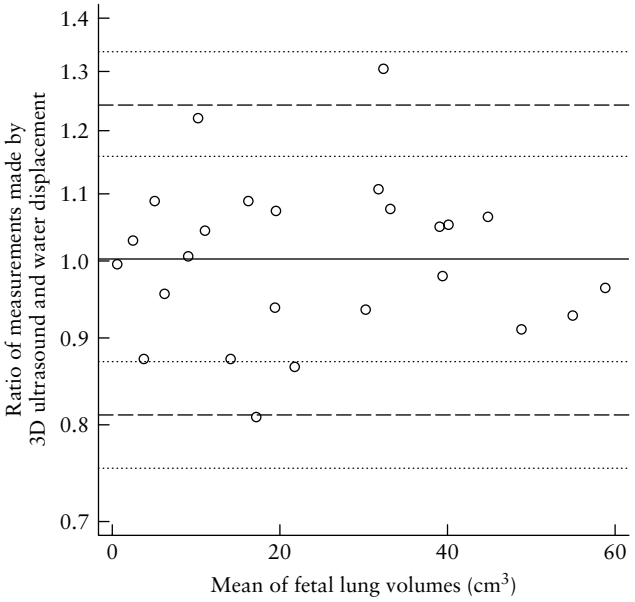

Figure 2 Ratio of lung volumes measured by three-dimensional (3D) ultrasound examination and by water displacement (log scale) against mean of measurements, based on data published by Ruano *et al.*[9]. Solid line represents mean; upper dashed line shows the mean +1.96 SD and lower dashed line the mean −1.96 SD, each with 95% CI (dotted lines).

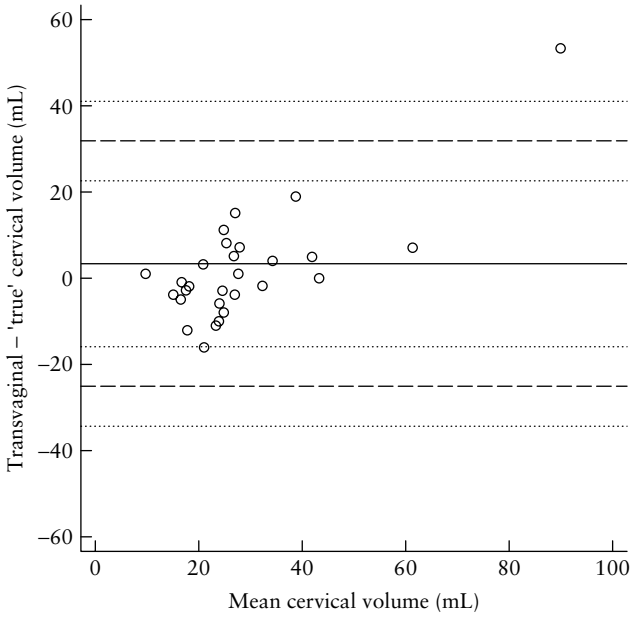

Figure 3 Paired differences between cervical volume measured by transvaginal three-dimensional ultrasound examination and by water displacement, from study by Farrell *et al.*[12]. Estimated correlation 0.76, $P < 0.0001$. Solid line represents mean; upper dashed line shows the mean +1.96 SD and lower dashed line the mean −1.96 SD, each with 95% CI (dotted lines).

Association between difference and mean

It is possible for there to be an association between the paired differences and means. An example of this

is found in a study by Farrell *et al.*, who examined the agreement between cervical volume, as measured using 3D ultrasound imaging, and the 'true' cervical volume, calculated using water displacement, in 28 women before hysterectomy[12]. We have digitized the data from their Bland–Altman plot of the difference between the transvaginal ultrasound volume and 'true' volume against the mean, and reproduce it here in Figure 3. There appears to be a positive association between the paired differences and means.

We can perform a statistical test to assess the evidence for a linear association, either testing whether the correlation coefficient between the paired differences and means differs significantly from zero or by linear regression of the differences against the means. The former is known as Pitman's test in the context of comparing the SD of measurements from two methods, which for the data in Figure 3 is highly statistically significant ($P < 0.001$).

Having found evidence for an association between paired differences and means, how should we proceed? Rather than fitting regression models for the difference as a function of the mean[10], we believe consideration should be given to the cause of the association.

A linear association (correlation) between the paired differences and means occurs when the SD of measurements from one method differs from the SD of measurements from the other method. There are therefore at least two possible causes of an observed association between paired differences and means.

The first is that there is real association between the difference in measurements from the two methods and true value being measured, i.e. the bias between methods changes over the range of true values. The second is that the within-subject SDs of the two methods differ. This will happen in the absence of changing bias if a new method has smaller or larger measurement errors than the standard method. This situation is thus entirely feasible, and would commonly occur when a substantially superior or inferior measurement method is compared with a standard technique.

If the cause of the correlation is unequal within-subject SDs and we were able to plot the paired differences against the true value being measured, there would be no correlation. The inconsistency between the Bland–Altman plot and a plot against true values occurs because the paired means contain measurement errors, in contrast to the true values.

Unfortunately, as explained by Dunn and Roberts[13], with only one measurement per method per subject, we cannot determine which of these two possibilities is the cause. To proceed we must make an additional assumption, either that the within-subject SDs of the two methods are equal, so that the correlation is attributed to changing bias, or that the two within-subject SDs are unequal, and that there is, in truth, no relationship between bias and true value (of course, we may have a combination of both).

For the purposes of illustration, we proceed with the example from Farrell *et al.*[12] by assuming that the observed correlation is due to a difference in the within-subject SDs of the two methods. A positive correlation between differences and means will occur when the method with less measurement error is subtracted from measurements of the method with larger error. For the data in Figure 3, the correlation is positive, suggesting the 3D ultrasound volumes contain larger errors than the volumes measured by water displacement, which seems entirely feasible. Indeed, Farrell and colleagues refer to the volume measured by water displacement as the true cervical volume, implying that such measurements were assumed to contain no errors.

Under an assumption that there is no changing bias, we can still calculate valid 95% limits of agreement, as done by Farrell *et al.*, because the limits of agreement method does not assume that the two methods' within-subject SDs are equal.

In order to assess from the data whether changing bias is occurring, we must make at least two measurements with each method on each subject. This study design has often been advocated[1,10,13], but is unfortunately rarely adopted[7]. Below we describe how such studies can be analyzed to assess whether a correlation between paired differences and means is due to changing bias between the methods (see Analysis with two measurements from each method).

Reliability in method comparison studies

As discussed previously, reliability may be a useful parameter with which to compare two different measurement methods. To estimate each method's reliability, we must make at least two measurements of each subject with each of the two methods. The repeat measurements from each method can then be analyzed as two separate repeatability studies (see Estimating reliability), giving estimates of each method's reliability, which can be compared.

An advantage of using reliability to compare measurement methods is that it can be used to compare methods when their measurements are given on different scales or metrics, as the reliability ICC is a dimensionless ratio. Because reliability depends on the heterogeneity of the true error-free values in the sampled population (see Reliability depends on population heterogeneity), it is essential that reliability ICCs are compared only if they have been estimated from the same population.

Investigators sometimes report a single reliability ICC estimate from method comparison studies in which it appears that only a single measurement per subject per method is available[9,12,14]. It would appear that the reported ICC estimates are obtained using a one-way ANOVA model, as described above (see Estimating reliability). The one-way ANOVA model assumes that there are no systematic biases between measurements within subjects, and that the within-subject SDs are equal for all measurements (exchangeability). When measurements are from two different methods, both of these assumptions may be false: bias may exist between

the methods and their within-subject SDs (repeatability) may differ. If either assumption is untrue, the model is misspecified and the resulting estimates of within and between subject SD are estimates of different parameters[15]. The resulting reliability ICC estimate then has a different interpretation from the usual one. We therefore advise against reporting the ICC estimate from fitting the simple one-way ANOVA to data from two different measurement methods.

Analysis with two measurements from each method

As already stated, making two measurements with each method permits an investigation of whether bias between the methods is constant or whether the method's measurement error SDs differ. First, the measurements from each method can be analyzed separately as two repeatability studies, using the methods described above, giving estimates of the repeatability coefficient and reliability ICC for each method.

The Bland–Altman plot of paired differences against means can be constructed by using the mean of a subject's two measurements from each method in place of the usual single measurement. As with the Bland–Altman plot based on a single measurement from each method, an observed correlation between differences and means could be due to either changing bias or a difference in measurement error variances between the methods.

We now describe a simple approach to assess whether any bias between measurements made by a new method changes with the 'true value' as measured by the existing method. We denote by $\overline{new_i}$ the mean of the i^{th} subject's two measurements made using the new method and we denote by $existing_{ij}$ the j^{th} measurement made on subject i by the existing measurement method. A plot of $\overline{new_i} - existing_{i1}$ against $existing_{i2}$ can then be used to investigate whether the bias between the new and existing methods changes with the 'true value' as measured by the existing method. It is crucial that $\overline{new_i} - existing_{i1}$ is plotted against $existing_{i2}$, rather than against $existing_{i1}$, because the latter would produce a spurious negative association owing to the common measurement error in $\overline{new_i} - existing_{i1}$ and $existing_{i1}$. It is for the same reason that plotting paired differences between measurements from two methods against the measurements from one method is inappropriate in method comparison studies[16].

An association between values of $\overline{new_i} - existing_{i1}$ and $existing_{i2}$ indicates that the bias between the methods changes with the true value as measured by the existing method. A statistical test of this hypothesis can be performed by fitting a linear regression for $\overline{new_i} - existing_{i1}$ with $existing_{i2}$ as the explanatory variable. The P-value for the test that the slope in this regression differs from zero indicates the strength of evidence against the null hypothesis of constant bias.

If evidence of bias changing with true value is found, we may be interested in estimating the rate at which the bias between the methods changes. We can estimate this rate by dividing the slope estimate from the above regression model by our estimate of the existing method's reliability. This quantity is an estimate of the increase in bias (new method minus existing method) for a one-unit increase in the true value as measured by the existing method.

The choice to plot/regress $\overline{new_i} - existing_{i1}$ against $existing_{i2}$ is arbitrary. We could equally have plotted $\overline{new_i} - existing_{i2}$ against $existing_{i1}$, and would have obtained a different P-value and estimate of the rate at which the bias changes. This arbitrary choice is undesirable both because we will get two different answers, and because it means the procedure is statistically inefficient.

A non-arbitrary and more efficient alternative is to plot/regress $\overline{new_i} - existing_i^*$ against $existing_i^*$, where

$$existing_i^* = \overline{existing} + \lambda\left(\overline{existing_i} - \overline{existing}\right),$$

and

$$\lambda = \frac{existing\_between\_subjectSD^2}{existing\_between\_subjectSD^2 + \frac{existing\_within\_subjectSD^2}{2}}.$$

$\overline{existing}$ denotes the sample mean of all measurements made using the existing method, and the between- and within-subject SDs are the values estimated using an ANOVA model for the repeat measurements made with the existing method; $existing_i^*$ is an unbiased estimate of a subject's true existing method level, which allows for the measurement error in the mean of the existing method's measurements. As before, if the bias between the methods does not change with true level according to the existing method, we expect the slope in the regression of $\overline{new_i} - existing_i^*$ against $existing_i^*$ to be zero. Again, this hypothesis can be tested by testing whether the slope in the regression is different from zero. (This hypothesis test ignores the estimation of the between- and within-subject SDs for the existing method, and so the Type I error rate for this test will be increased. That is, changing bias will be wrongly detected more often than the nominal Type I error rate when, in truth, the bias is constant.)

The model underlying these analyses is a type of structural equation model, and as such can be fitted efficiently using software for structural equation models. The interested reader is referred to Dunn and Roberts[13] for further details.

If the analysis suggests that the bias (if any) between the methods is constant across the range of true values as measured by the existing method, we can calculate an estimate of the bias and the estimated 95% limits of agreement for the difference between two measurements from the two methods. An estimate of the bias is given by the mean of all the measurements made by one method minus the mean of the measurements by the second method. If the within-subject SDs of the two methods are different, the SD of the difference between two measurements made by the two methods is

$$\sqrt{existing\_within\_subjectSD^2 + new\_within\_subjectSD^2}$$

which we estimate using the estimated within-subject SDs from each method, obtained from fitting separate

ANOVA models to the measurements from each method. The estimated 95% limits of agreement are then given by the estimated bias plus and minus 1.96 times the estimated SD of paired differences.

MEASUREMENTS WITH OBSERVERS OR RATERS

Often measurements are made by observers or raters. Usually, measurements made by two different observers are less similar than are two measurements made by the same observer. Just as with two methods, measurements from two observers may differ systematically due to bias between the observers (an observer 'effect'), and their measurement errors may also have different SDs. For example, measurements from an observer who can make more precise measurements will have a smaller SD than those made by a less precise observer.

If measurements in the future are to be made by different observers, we need to describe and quantify the differences between such measurements in order to judge whether differences are genuine or may be due to measurement error. As with a method comparison study, the ideal way to study this is for each observer to make at least two measurements of a sample of subjects. The design of such a study and the type of statistical analysis that are appropriate should be guided by whether interest lies in a particular set of observers, or whether we are interested in drawing inferences about a wider population of potential observers or raters.

Interest in specific observers

If future measurements are to be made by a specific set of observers, the measurement error study should involve each of these observers making measurements of a sample of subjects, ideally with at least two measurements per observer per subject. We can then use exactly the same methods as described for method comparison, treating different observers as different measurement methods.

If each observer makes two or more measurements on each subject, we can examine whether there is bias or changing bias between observers. We can eliminate the bias of future measurements (in expectation) by adjusting measurements using the corresponding estimated biases.

We can also estimate repeatability coefficients and intraobserver reliability ICCs for each observer, using the one-way ANOVA model described previously (see Estimating reliability). It may be of interest to know which observers are more reliable, and if differences in reliability can be related to observer characteristics, such as levels of experience or training[17].

If we are willing to assume that biases between observers are constant, we can fit a so-called two-way mixed-effects model to such a dataset, allowing for a random subject effect and fixed observer effects. The observer-effects estimates indicate the direction, magnitude and evidence for bias between observers. Furthermore, such models can be extended to allow the measurement error SD to differ between observers.

Interest in a population of observers

The observers in a measurement error study can often be considered as a random sample of observers from a larger population of potential observers who may be used in future studies or clinical practice. In this case, we are not interested in drawing conclusions about the particular observers in the measurement error study, but only in the information that they provide about the population of potential observers.

In this situation, it is important that a reasonable number of observers is used in the measurement error study. For example, if the measurement error study uses only two observers, very little can be concluded about the larger population of observers, because we effectively have a sample size of just two.

Because the observers in the measurement error study are considered a random sample from the population of potential observers, we analyze such a study using a model that treats the observer 'effect' as a random effect. We thus use a two-way random-effects model, with random subject effects and random observer effects.

Fitting a two-way random-effects model gives estimates of the between-subject SD, between-observer SD and the measurement error SD (assumed to be the same for different observers). The between-observer SD represents variability in measurements due to observers, or biases between the observers. From such estimates one can calculate the SD of the difference in two measurements, either made by the same observer or by two different observers.

We can estimate the reliability of measurements made by the same observer (intraobserver reliability) by:

$$\frac{between\_subjectSD^2 + between\_observerSD^2}{between\_subjectSD^2 + between\_observerSD^2 + measurement\_errorSD^2}.$$

The reliability of two measurements made by different observers, or the interobserver reliability, can be estimated by

$$\frac{between\_subjectSD^2}{between\_subjectSD^2 + between\_observerSD^2 + measurement\_errorSD^2}.$$

Unless the between-observer SD is zero, the intraobserver ICC will be higher than the interobserver ICC. This is because biases between observers act to make measurements from different observers less similar; it is more difficult to distinguish between subjects on the basis of measurements made by two different observers than if the subjects had been measured by the same observer.

CONCLUSIONS

In this paper we have distinguished between the concepts of agreement and reliability. Neither parameter is superior to the other, as they describe different characteristics of the measurement process. The choice of what to report in a particular study should be guided by how measurements are to be used in the future, and also by the fact that readers may want to use a measurement method in a

different population. We have highlighted the fact that the reliability of a method depends not only on the size of the measurement errors but on the heterogeneity of true values in the population in which measurements are made. As measurement techniques potentially may be used in a variety of settings (e.g. clinical trials, screening) and different populations, it is advisable to report estimates of between- and within-subject SDs.

We have outlined which methods we believe are appropriate for the analysis of repeatability studies, method comparison studies, and studies with measurements made by different observers or raters. In particular, we do not believe a single reliability coefficient should be used for method comparison studies. If the reliability of two methods are to be compared, each method's reliability should be estimated separately, by making at least two measurements on each subject with each measurement method.

We have discussed how an association between paired differences and means may not necessarily be caused by changing bias between two methods. Such an association may also be caused by a difference in the methods' measurement error SDs, but with only one measurement per subject per method it is not possible to determine which of these is the cause. Method comparison studies should therefore make at least two measurements per subject per method. This permits an investigation of the source of any association between paired differences and means for measurements made by the two methods, and also allows the repeatability and reliability of each method to be estimated.

Finally, where measurements involve an observer or rater, measurement error studies must use an adequate number of observers if interest lies in making inferences about a wider population of observers.

REFERENCES

1. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1**: 307–310.
2. Fitzmaurice G. Statistical methods for assessing agreement. *Nutrition* 2002; **18**: 694–696.
3. De Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol* 2006; **59**: 1033–1039.
4. National Institute for Standards and Technology. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results. http://physics.nist.gov/Pubs/guidelines/contents.html [18 December 2007].
5. Järvelä IY, Sladkevicius P, Tekay AH, Campbell S, Nargund G. Intraobserver and interobserver variability of ovarian volume, gray-scale and color flow indices obtained using transvaginal three-dimensional power Doppler ultrasonography. *Ultrasound Obstet Gynecol* 2003; **21**: 277–282.
6. Efron B, Tibshirani RJ. *Introduction to the Bootstrap*. Chapman & Hall/CRC: New York, 1993.
7. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. *Ultrasound Obstet Gynecol* 2003; **22**: 85–93.
8. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician* 1983; **32**: 307–317.
9. Ruano R, Martinovic J, Dommergues M, Aubry MC, Dumez Y, Benachi A. Accuracy of fetal lung volume assessed by three-dimensional sonography. *Ultrasound Obstet Gynecol* 2005; **26**: 725–730.
10. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999; **8**: 135–160.
11. Kusanovic JP, Nien JK, Goncalves LF, Espinoza J, Lee W, Balasubramaniam M, Soto E, Erez O, Romero R. The use of inversion mode and 3D manual segmentation in volume measurement of fetal fluid-filled structures: comparison with Virtual Organ Computer-aided AnaLysis (VOCAL^{TM}). *Ultrasound Obstet Gynecol* 2008; **31**: 177–186.
12. Farrell T, Cairns M, Leslie J. Reliability and validity of two methods of three-dimensional cervical volume measurement. *Ultrasound Obstet Gynecol* 2003; **22**: 49–52.
13. Dunn G, Roberts C. Modelling method comparison data. *Stat Methods Med Res* 1999; **8**: 161–179.
14. Gerards FA, Twisk JWR, Bakker M, Barkhof F, Van Vugt JMG. Fetal lung volume: three-dimensional ulrasonography compared with magnetic resonance imaging. *Ultrasound Obstet Gynecol* 2007; **29**: 533–536.
15. Lyles RH, Chambless LE. Effects of model misspecification in the estimation of variance components and intraclass correlation for paired data. *Stat Med* 1995; **14**: 1693–1706.
16. Bland JM, Altman DG. Comparing methods of measurement: why plotting differences against standard method is misleading. *Lancet* 1995; **346**: 1085–1087.
17. Pajkrt E, Mol BWJ, Boer K, Drogtrop AP, Bossuyt PMM. Intra- and interoperator repeatability of the nuchal translucency measurement. *Ultrasound Obstet Gynecol* 2000; **15**: 297–301.
18. Searle SR, Casella G, McCulloch CE. *Variance Components*. John Wiley & Sons: Chichester, 1992.

APPENDIX

Confidence intervals for one-way ANOVA

Here we reproduce standard results for the calculation of CIs for two parameters in the one-way ANOVA model, in which two measurements are available from each subject[18].

We denote by n the number of subjects in the study, and by SSB and SSW the between- and within-subject sums of squares from the one-way ANOVA table. For a study with two measurements per subject, the mean between- and within-sums of squares are then given by

$$MSB = \frac{SSB}{n-1},\ MSW = \frac{SSW}{n}.$$

Denote by $\chi^2_{n,p}$ the $100 \times p^{\text{th}}$ centile of the Chi-square distribution with n degrees of freedom. Then a 95% CI for the within-subject SD is given by

$$\left(\sqrt{\frac{SSW}{\chi^2_{n,0.975}}}, \sqrt{\frac{SSW}{\chi^2_{n,0.025}}}\right).$$

Defining $F = \frac{MSB}{MSW}$, and denoting by $F_{n-1,n,p}$ the $100 \times p^{\text{th}}$ centile of the F-distribution with $n-1$ and n degrees of freedom, a 95% CI for the reliability/ICC is given by

$$\left(\frac{\left(\frac{F}{F_{n-1,n,0.975}}\right)-1}{\left(\frac{F}{F_{n-1,n,0.975}}\right)+1}, \frac{\left(\frac{F}{F_{n-1,n,0.025}}\right)-1}{\left(\frac{F}{F_{n-1,n,0.025}}\right)+1}\right).$$