Chapter 2
Clinical Agreement in Quantitative Measurements

Limits of Disagreement and the Intraclass Correlation

Abhaya Indrayan

Abstract In clinical research, comparison of one measurement technique with another is often needed to see whether they agree sufficiently for the new to replace the old. Such investigations are often analysed inappropriately, notably by using correlation coefficients, which could be misleading. This chapter describes alternatives based on graphical techniques that quantify disagreement as well as the concept of intraclass correlation.

Introduction

Assessment of agreement between two or more measurements has become important for the following reasons. Medical science is growing at a rapid rate. New instruments are being invented and new methods are being discovered that measure anatomic and physiologic parameters with better accuracy and precision, and at lower cost. Emphasis is on simple, non-invasive, safer methods that require smaller sampling volumes and can help in continuous monitoring of patients when required. Acceptance of any new method depends on a convincing demonstration that it is nearly as good as, if not better than, the established method. The problem in this case is not equality of averages but of equality of all individual values.

The term agreement is used in several different contexts. The following discussion is restricted to a setup where a pair of observations (x,y) is obtained by measuring the same characteristic on the same subject by two different methods, by two different observers, by two different laboratories, at two anatomic sites, etc. There can also be more than two. The measurement could be qualitative or quantitative. Quantitative agreement is between exact values, such as intra-ocular

A. Indrayan (✉)
Department of Biostatistics and Medical Informatics, University College of Medical Sciences, Delhi, India
e-mail: a.indrayan@gmail.com

S.A.R. Doi and G.M. Williams (eds.), *Methods of Clinical Epidemiology*,
Springer Series on Epidemiology and Public Health,
DOI 10.1007/978-3-642-37131-8_2,

pressure in two eyes, and quantitative agreement is between attributes such as the presence or absence of a minor lesion in radiographs read by two radiologists. The method of assessing agreement in these two cases is different. This chapter is on agreement in quantitative measurements. Agreement in qualitative measurements is discussed in the previous chapter.

Assessment of Quantitative Agreement

Irrespective of what is being measured, it is highly unlikely that the new method would give exactly the same reading in every case as the old method, even if they are equivalent. Some differences would necessarily arise – if nothing else, at least as many as would occur when the same method is used two times on the same subject under identical conditions. How do you decide that the new method is interchangeable with the old? The problem is described as one of quantitative agreement. This is different from evaluating which method is better. The assessment of better is done with reference to a gold standard. Assessment of agreement does not require any such standard.

Quantitative Measurements

The problem of agreement in quantitative measurement can arise in at least five different types of situations. (1) Comparison of self-reported values with instrument-measured values, for example, urine frequency and bladder capacity using a patient questionnaire and a frequency–volume chart. (2) Comparison of measurements at two or more different sites, for example, paracetamol concentration in saliva with that in serum. (3) Comparison of methods, for example, bolus and infusion methods of estimating hepatic blood flow in patients with liver disease. (4) Comparison of two observers, for example, duration of electroconvulsive fits reported by two or more psychiatrists on the same group of patients, or comparison of two or more laboratories when, for example, aliquots of the same sample are sent to two different laboratories for analysis. (5) Intraobserver consistency, for example, measurement of the anterior chamber depth of an eye segment two or more times by the same observer using the same method to evaluate the reliability of the method.

In the first four cases, the objective is to find whether a simple, safe, less expensive procedure can replace an existing procedure. In the last case, the reliability of the method is being evaluated.

Statistical Formulation of the Problem

The statistical problem in all these cases is to check whether or not a $y = x$ type of relationship exists in individual subjects. This looks like a regression setup $y = a + bx$ with $a = 0$ and $b = 1$, but that is not really. The difference is that, in regression, the relationship is between x and the average of y. In an agreement setup, the concern is with individual values and not with averages. Nor should agreement be confused with high correlation. Correlation is nearly 1 if there is a systematic bias and nearly same difference occurs in every subject. Example 1 illustrates the distinction between $y = x$ regression and agreement.

Example : Very Different Values but Regression Is $\mathbf{y = x}$

The following Hb values are reported by two laboratories for the same blood samples:

Lab I (x)	11.3	12.0	13.9	12.8	11.3	12.0	13.9	12.8
Lab II (y)	11.5	12.4	14.2	13.2	11.1	11.6	13.6	12.4

$$\bar{x} = 12.5, \quad \bar{y} = 12.5, \quad r = 0.945$$

$$\hat{y} = x, \quad \text{that is, } b = 1 \quad \text{and} \quad a = 0$$

The two laboratories have the same mean for these eight samples and a very high correlation (0.945). The intercept is 0 and slope is 1.00. Yet there is no agreement in any of the subjects. The difference or error ranges from 0.2 to 0.4 g/dL. This is substantial in the context of the present-day technology for measuring Hb levels. Thus, equality of means, a high degree of correlation and regression $y = x$ are not enough to conclude agreement. Special methods are required.

The first four values of x in this example are the same as the last four values. The first four values of y are higher and the last four values are lower by the same margin. Thus, for each $x, \bar{y} = x$ giving rise to the regression $\hat{y} = x$. In this particular case, the correlation coefficient is also nearly 1.

Quantitative agreement in individual values can be measured either by limits of disagreement or by intraclass correlation. The details are as follows.

Limits of Disagreement Approach

This method is used for a pair of measurements and based on the differences $d = (x - y)$ in the values obtained by the two methods or observers under comparison. If the methods are in agreement, this difference should be zero for every subject. If these differences are randomly distributed around zero and none of the differences is large, the agreement is considered good. A graphical approach is to plot d versus $(x + y)/2$. A flat line around zero is indicative of good agreement. Depending on which is labelled x and which is y, an upward trend indicates that x is generally more than y, and a downward trend that y is more than x.

A common sense approach is to consider agreement as reasonably good if, say, 95 % of these differences fall within the prespecified clinically tolerable range and the other 5 % are also not too far from that. Statistically, when the two methods or two observers are measuring the same variable, then the difference d is mostly the measurement error. Such errors are known to follow a Gaussian distribution. Thus the distribution of d in most cases would be Gaussian. Then the limits $\bar{d} \pm 1.96 s_d$ are likely to cover differences in nearly 95 % of subjects where $\bar{d}$ is the average and s_d is the standard deviation (SD) of the differences. The literature describes them as the limits of agreement. They are actually limits of disagreement.

$$\text{Limits of disagreement: } \bar{d} - 1.96 s_d \quad \text{to} \quad \bar{d} + 1.96 s_d \tag{2.1}$$

If these limits are within clinical tolerance in the sense that a difference of that magnitude does not alter the management of the subjects, then one method can be replaced by the other. The mean difference $\bar{d}$ is the bias between the two sets of measurements and s_d measures the magnitude of random error. For further details, see Bland and Altman (1986).

The limitations of the product–moment correlation coefficient are well known. Consider the following example. Suppose a method consistently gives a level 0.5 mg/dL higher than another method. The correlation coefficient between these two methods would be a perfect 1.0. Correlation fails to detect systematic bias. This also highlights the limitations of the limits of disagreement approach. The difference between measurements by two methods is always +0.5 mg/dL, thus the SD of the difference is zero. The limits of disagreement in this case are (+0.5,+0.5). This is in fact just one value and not limits. A naive argument could be that these limits are within clinical tolerance and thus the agreement is good. To detect this kind of fallacy, plot the differences against the mean of paired values. This plot can immediately reveal this kind of systematic bias.

Table 2.1 Results on agreement between AVRG[a] and Korotkoff BP readings in 100 volunteers

	AVRG[a]		Korotkoff
Mean systolic BP (mmHg)	115.1		115.5
SD (mmHg)	13.4		13.2
Mean difference (mmHg)		−0.4	
P-value for paired *t*		>0.50	
Correlation coefficient (*r*)		0.87	
SD of difference, s_d (mmHg)		6.7	
Limits of disagreement (mmHg)		(−13.5, 12.7)	
Intraclass correlation coefficient (r_I) (formula given in next section)		0.87	

[a]Average of readings at the appearance and disappearance of the plethysmographic waveform of a pulse oximeter

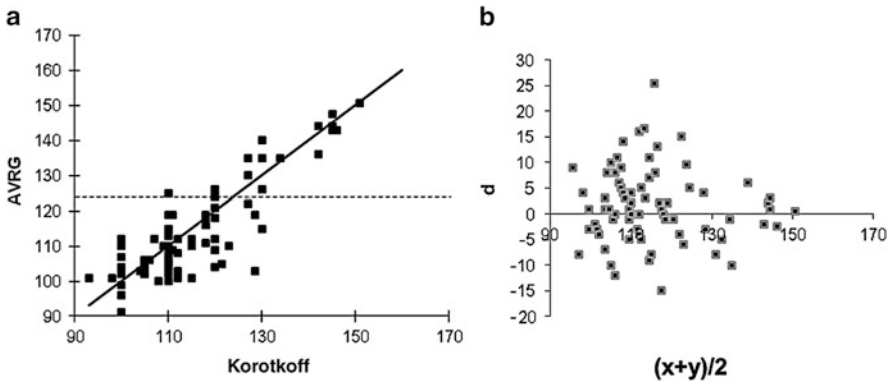

Fig. 2.1 (**a**) Scatter of the pulse oximeter based (*y*) and Korotkoff based (*x*) readings of systolic blood pressure. For pulse oximeter based readings the average of readings at the disappearance and reappearance of the waveform respectively were used (labelled AVRG in the *left panel*). (**b**) Plot of *d* versus $(x + y)/2$ (d = difference between y and x which are defined as above)

Example : Limits of Disagreement Between Pulse Oximetry and Korotkoff Readings

Consider the study by Chawla et al. (1992) on systolic blood pressure (BP) readings derived from the plethysmographic waveform of a pulse oximeter. This method could be useful in a pulseless disease such as Takayasu syndrome. The readings were obtained (a) at the disappearance of the waveform on the pulse oximeter on gradual inflation of the cuff and (b) at the reappearance on gradual deflation. In addition, BP was measured in a conventional manner by monitoring the Korotkoff sounds. The study was done on 100 healthy volunteers. The readings at disappearance of the waveform were generally higher and at reappearance generally lower. Thus, the average (AVRG) of the two is considered a suitable value for investigating the

agreement with the Korotkoff readings. The results are shown in Table 2.1. The scatter, the line of equality and the plot of d versus $(x + y)/2$ are shown in Fig. 2.1. Figure 2.1b shows that the differences were large for smaller values.

Despite the means being nearly equal and r very high, the limits of disagreement (Table 2.1) show that a difference of nearly 13 mmHg can arise between the two readings on either side (average of pulse oximetry readings can give either less or more than the Korotkoff readings). These limits are further subject to sampling fluctuation, and the actual difference in individual cases can be higher. Now it is for the clinician to decide whether a difference of such magnitude is tolerable. If it is, then the agreement can be considered good and pulse oximetry readings can be used as a substitute for Korotkoff readings, otherwise they should not be used. Thus, the final decision is clinical rather than statistical when this procedure is used.

Intraclass Correlation as a Measure of Agreement

Intraclass correlation is the strength of a linear relationship between subjects belonging to the same class or the same subgroup or the same family. In the agreement setup, the two measurements obtained on the same subject by two observers or two methods is a subgroup. If they agree, the intraclass correlation will be high. This method of assessing an agreement was advocated by Lee et al. (1989).

In the usual correlation setup, the values of two different variables are obtained on a series of subjects. For example, you can have the weight and height of 20 girls aged 5–7 years. You can also have the weight of the father and mother of 30 low birthweight newborns. Both are weights and the product–moment correlation coefficient is a perfectly valid measure of the strength of the relationship in this case. Now consider the weight of 15 persons obtained on two machines. Any person, say number 7, may be measured by machine 2 first and then by machine 1. Others may be measured by machine 1 then by machine 2. The order does not matter in this setup as the interest is in finding whether the values are in agreement or not.

Statistically, intraclass correlation is that part of the total variance that is accounted for by the differences in the paired measurements obtained by two methods. That is,

$$\text{Intraclass correlation:} \quad \rho_I = \frac{\sigma_M^2}{\sigma_M^2 + \sigma_e^2} \tag{2.2}$$

where σ_M^2 is the variance between methods if methods are to be compared for agreement and σ_e^2 is the error variance. This formulation does not restrict us to only two methods. These could be three or more. In the weight example, you can compare agreement among five machines by taking the weight of each of the 15 persons on these five machines.

The estimate of ρ_I is easily obtained by setting up the usual analysis of variance (ANOVA) table. If there are M methods under comparison, the ANOVA table would look like Table 2.2. The number of subjects is n in this table and other notations are self-explanatory. E(MS) is the expected value of the corresponding mean square.

Table 2.2 Structure of ANOVA table in agreement setup

Source	df	Mean square (MS)	E(MS)
Methods (A)	$M - 1$	MSA	$\sigma_e^2 + n\sigma_M^2$
Subjects (B)	$n - 1$	MSB	$\sigma_e^2 + M\sigma_S^2$
Error	$(M - 1)(n - 1)$	MSE	σ_e^2

A little algebra yields the estimate of the intraclass correlation r_I:

$$r_I = \frac{\text{MSA} - \text{MSE}}{\text{MSA} + (n-1)\text{MSE}} \tag{2.3}$$

This can be easily calculated once you have the ANOVA table. Statistical software will give you the value of the intraclass correlation directly.

In terms of the available values, the computation of the intraclass correlation coefficient (ICC) is slightly different from that of the product–moment correlation coefficient. In the agreement setup, the interest is in the correlation between two measurements obtained on the same subject and is obtained as follows.

ICC (a pair of readings):

$$r_I = \frac{2\Sigma_i(x_{i1} - \bar{x})(x_{i2} - \bar{x})}{\Sigma_i(x_{i1} - \bar{x})^2 + \Sigma_i(x_{i2} - \bar{x})^2} \tag{2.4}$$

where x_{i1} is the measurement on the ith subject ($i = 1,2,\ldots,n$) when obtained by the first method or the first observer, x_{i2} is the measurement on the same subject by the second method or the second observer, and $\bar{x}$ is the overall mean of all $2n$ observations. Note the difference in the denominator compared with the formula for the product–moment correlation.

This was calculated for the systolic BP data described in Example 2 and was found to be $r_I = 0.87$. A correlation >0.75 is generally considered enough to conclude good agreement. Thus, in this case, the conclusion on the basis of the intraclass correlation is that the average of readings at disappearance and appearance of the waveform in pulse oximetry in each person agrees fairly well with the Korotkoff readings for that person. This may not look consistent with the limits of disagreement that showed a difference up to 13 mmHg between the two methods. The two approaches of assessing agreement can sometimes lead to different conclusions.

Equation (2.4) is used for comparing two methods or two raters. This correlation can be used for several measurements. For example, you may have the wave amplitude of electrical waves at $M = 6$ different sites in the brain of each of $n = 40$ persons. For multiple raters or multiple methods, ICC (several readings):

$$r_I = \frac{\Sigma_i\Sigma_{j\neq k}(x_{ij} - \bar{x})(x_{ik} - \bar{x})}{(M-1)\Sigma_i\Sigma_j(x_{ij} - \bar{x})^2}, \quad i = 1,2,\ldots,n, \quad j,k = 1,2,\ldots,M \tag{2.5}$$

where n is the number of subjects and M is the number of observers or the number of methods to be compared. The mean $\bar{x}$ is calculated on the basis of all Mn observations.

Table 2.3 Cutoffs for grading the strength of agreement

Intraclass correlation	Strength of agreement
<0.25	Poor
0.25–0.50	Fair
0.50–0.75	Moderate
0.75–0.90	Good
>0.90	Excellent

For grading of the strength of agreement, the cutoffs shown in Table 2.3 can be used.

An Alternative Simple Approach to Agreement Assessment

Neither of the two methods described in the preceding sections is perfect. Let us first look at their relative merits and demerits and then propose an alternative method, which may also not be perfect but is relatively simple.

Relative Merits of the Two Methods

Indrayan and Chawla (1994) studied the merits and demerits of the two approaches in detail. The following are their conclusions on the comparative features of the two methods:

1. The ICC does not depend on the subjective assessment of any clinician. Thus, it is better to base the conclusion on this correlation when the clinicians disagree on the tolerable magnitude of differences between two methods (or two observers). And clinicians seldom agree on such issues.
2. The 0.75 threshold to label an intraclass correlation high or low is arbitrary, although generally acceptable. Thus, there is also a subjective element in this approach.
3. Intraclass correlation is unit free, easy to communicate, and interpretable on a scale of zero (no agreement) to one (perfect agreement). This facility is not available in the limits of disagreement approach.
4. A distinct advantage of the limits of disagreement approach is its ability to delineate the magnitude of individual differences. It also provides separate estimates of bias $(\bar{d})$ and random error (s_d). This bias measures the constant differences between the two measurements and random error is the variation around this bias. Also, this approach is simple and does not need much calculation.
5. The limits of disagreement can be evaluated only when the comparison is between two measurements. The intraclass correlation, on the other hand, is fairly general and can be used for comparing more than two methods or more than two observers (Eq. 2.5).

6. Intraclass correlation can also be used for comparing one group of raters with another group. Suppose you have four male assessors and three female assessors. Each subject is measured by all seven assessors. You can compare intraclass correlation obtained for male assessors with that obtained for female assessors. You can have one set of subjects for assessment by males and another set of subjects for assessment by females.

A review of the literature suggests that researchers prefer the limits of disagreement approach to the ICC approach for comparing two methods. A cautious approach is to use both and come to a firm conclusion if both give the same result. If they are in conflict, defer a decision and carry out further studies.

The following comments might provide better appreciation of the procedure to assess quantitative agreement:

1. As mentioned earlier, the limits of disagreement $\bar{d} \pm 1.96 s_{\mathrm{d}}$ themselves are subject to sampling fluctuation. A second sample of subjects may give different limits. Methods are available to find an upper bound to these limits. For details, see Bland and Altman (1986). They call them limits of agreement, but perhaps they should be called limits of disagreement.
2. The ICC too is subject to sampling fluctuation. For assessing agreement, the relevant quantity is the lower bound of r_{I}. This can be obtained by the method described by Indrayan and Chawla (1994). Their method for computing the ICC is based on ANOVA, but that gives the same result as obtained by Eq. (2.4).
3. Although not specifically mentioned, the intraclass correlation approach assumes that the methods or observers under comparison are randomly chosen from a population of methods or observers. This is not true when comparing methods because they cannot be considered randomly chosen. Thus, the intraclass correlation approach lacks justification in this case. However, when comparing observers or laboratories, the assumption of a random selection may have some validity. If observers or laboratories agree, a generalized conclusion about consistency or reliability across them can be drawn.
4. Intraclass correlation is also used to measure the reliability of a method of measurement as discussed briefly by Indrayan (2012).
5. Both these approaches are applicable when both the methods could be in error. As mentioned earlier, these methods are not appropriate for comparing with a gold standard that gives a fixed target value for each subject. For agreement with a gold standard, see Lin et al. (2002).

An Alternative Simple Approach

The limits of disagreement approach just described is based on the average difference and has the limitations applicable to all averages. For example, this approach does not work if the bias or error is proportional. Fasting blood glucose levels vary from 60 to 300 mg/dL or more. Five percent of 60 is 3 and of 300 is 15. The limits of

Table 2.4 Data on fasting blood sugar levels in 10 blood samples

Method 1 (x)	86	172	75	244	97	218	132	168	118	130
Method 2 (y)	90	180	73	256	97	228	138	172	116	132
$d = x - y$	−4	−8	+2	−12	0	−10	−6	−4	+2	−2
5 % of x	4.30	8.60	3.75	12.20	4.85	10.90	6.60	8.40	5.90	6.50

disagreement approach considers them to be different and ignores that both are 5 % and proportionately the same. Also, if one difference is 10 and the other is 2, and they are not necessarily proportional, the limits of disagreement consider only the average. Individual differences tend to be overlooked. A few unusually large differences distort the average and are not properly accounted except by disproportional inflation of the SD.

To account for small and big individual differences as well as proportional bias, it may be prudent to set up a clinical limit that can be tolerated for individual differences without affecting the management of the condition. Such limits are required anyway for the limits of disagreement approach, albeit for the average. These clinical limits of indifference can be absolute or in terms of a percentage. If not more than a prespecified percentage (say 5 %) of individual differences are beyond these limits in a large sample, you can safely assume adequate agreement. This does not require any calculation of the mean and SD. You may like to add a condition such as none of the differences should be more than two times the limit of indifference. Any big difference, howsoever isolated, raises alarm. A plot of y versus x can track that the differences are systematic or random.

Example : Agreement Between Two Methods of Measuring Fasting Blood Glucose Levels

Consider the data in Table 2.4. Suppose method 1 is the current standard although this can also be in error. Method 2 is extremely cheap and gives instant results. Suppose also that clinicians are willing to accept 5 % error in view of the distinct advantages of method 2. Note that this indifference is a percentage and not an absolute value.

In these data, the y versus x plot is on a fairly straight line (Fig. 2.2a) but the plot of d versus $(x + y)/2$ (Fig. 2.2b) shows an aberration with a large number of points on the negative side and following an increasing trend. This shows lack of agreement according to the limits of disagreement approach. This really is not the case as explained next.

None of the differences exceed the clinical limit of indifference of 5 % in this sample. Thus, method 2 can be considered in agreement with method 1 although a larger sample is required to be confident. However, most differences are negative, indicating that method 2 generally provides lower values. The average difference is 4.2 mg/dL in absolute terms and nearly 3 % of y in relative terms. This suggests the correction factor for bias. If you decide to subtract 3 % of the level obtained by method 2, you can reach very close to the value obtained by method 1 in most cases. Do this as an exercise and verify it for yourself.

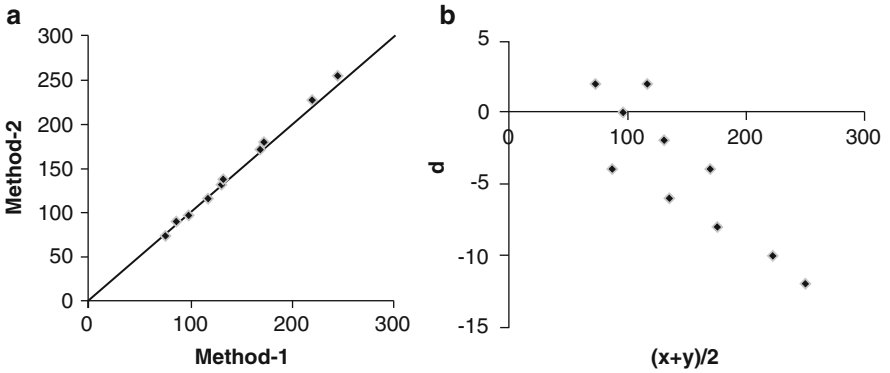

Fig. 2.2 (**a**) y versus x plot for data in Example 3, and (**b**) d versus $(x + y)/2$ plot for the same data. The variables x and y are the results of glucose measurements on the same sample by two different methods (mg/dl) and the difference between the results of two methods is given by d

Now forget about 5 % tolerance, and note that some differences are small and some are quite large in Example 3. The value of $s_d = 4.85$ in this case. Thus, the limits of disagreement are $-4.2 \pm 2 \times 4.85$, or -13.9 to +5.5. These limits may look too wide and beyond clinical tolerance, particularly on the negative side. These limits do not allow a larger error for larger values that proportionate considerations would allow. Also, these are based on an average and do not adequately consider individual differences. If 1 out of 20 values shows a big difference, this can distort the mean and inflate the SD, and provide unrealistic limits of disagreement. The alternative approach suggested above can be geared to allow not more than 5 % individual differences beyond the tolerance limit and you can impose an additional condition that none should exceed, say, by 10 % of the base value. Since it is based on individual differences and not on an average, this alternative approach may be more appealing.

Bibliography

Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307–310

Chawla R, Kumarvel V, Girdhar KK, Sethi AK, lndrayan A, Bhattacharya A (1992) Can pulse oximetry be used to measure systolic blood pressure? Anesth Analg 74:196–200

Indrayan A (2012) Medical biostatistics, 3rd edn. Chapman & Hall/CRC Press, Boca Raton

Indrayan A, Chawla R (1994) Clinical agreement in quantitative measurements. Natl Med J India 7:229–234

Lee J, Koh D, Ong CN (1989) Statistical evaluation of agreement between two methods for measuring a quantitative variable. Comput Biol Med 19:61–70

Lin L, Hedayat AS, Sinha B, Yang M (2002) Statistical methods in assessing agreement: models, issues and tools. J Am Stat Assoc 7:257–270

http://www.springer.com/978-3-642-37130-1

Methods of Clinical Epidemiology
Doi, S.A.R.; Williams, G.M. (Eds.)
2013, XVI, 282 p., Hardcover
ISBN: 978-3-642-37130-1