Chapter for the book "Methods of Clinical Epidemiology"

Measures of Clinical Agreement

A. Indrayan, PhD(OhioState),FAMS,FRSS,FASc

Former Professor and Head, Department of Biostatistics and Medical Informatics

University College of Medical Sciences, Delhi

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 invented and new methods are discovered that measure anatomical and physiological parameters with better accuracy and precision, and at a lower cost. Emphasis is on simple, noninvasive, 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, if not better, as 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 anatomical sites, etc. They can be more than two also. The measurement could be qualitative or quantitative. Quantitative agreement is between exact values such as of intra-ocular pressure in two eyes and quantitative agreement is between attributes such as presence or absence of a minor lesion in x-rays read by two radiologists. The method of assessing agreement in these two cases is different. First section is on agreement in quantitative measurements. Agreement in qualitative measurements is discussed in the next section.

I 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 each case as the old method even if they are equivalent. Some differences would necessarily arise—if nothing else, at least as much as would occur when same method is used two times on the same subject in 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.

1.1 Agreement in Quantitative Measurements

The problem of agreement in quantitative measurement can arise in at least

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

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 really is not so. 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 would be nearly one if there is a systematic bias and nearly same difference occurs in every subject. Example 1 illustrates distinction between $y = x$ regression and agreement.

Example 1: Very different values but regression is y = x

Following are Hb values 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

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

$\hat{y} = x$, i.e., $b = 1$ and $a = 0$.

The two laboratories have same mean for these eight samples and a very high correlation (0.945). The intercept is zero 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 of measuring Hb level. Thus, equality of means, a high degree of correlation and regression $y = x$ are not enough to conclude agreement. Special methods are required.

Side note: If you notice carefully, 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 same margin. Thus, for each x, $\bar{y} = x$ giving rise to the regression $\hat{y} = x$. In this particular case, the correlation coefficient also is nearly one. ■

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

1.2 Limits of Disagreement Approach

This method is 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 in 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 ds vs. $(x + y)/2$. A flat line around zero is indicative of good agreement. Depending upon which is labeled x and which 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 pre-specified clinically tolerable range and the other 5% also not too far. Statistically, note that 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 SD of the differences. The literature describes them as the limits of agreement. *They are actually limits of disagreement.*

Limits of disagreement: $\bar{d} - 1.96 s_d$ to $\bar{d} + 1.96 s_d$ (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).

Enough is known about the limitation of product-moment correlation coefficient (Indrayan 2012). If things are not clear enough, consider the following example. Suppose a method consistently gives a level 0.5 mg/dL higher than the other method. The correlation coefficient between these two methods would be prefect 1.0. Correlation fails to detect systematic bias. This also highlights the limitation of limits of disagreement approach. The difference between measurements by two methods is always +0.5 mg/dL – thus the SD of difference is zero. The limits of disagreement in this case are (+0.5, +0.5). This in fact is just one value and not limits. A naïve 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.

Example 2: Limits of disagreement between pulse oximetry and Korotkoff readings

Consider the study by Chawla et al. (1992) on systolic 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 at the (a) disappearance of the waveform on the pulse oximeter on gradual inflation of the cuff and at the (b) 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 observed to be 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 in Table 1. The scatter, the line of equality and plot of d vs. $(x+y)/2$ are shown in Figure 1. Figure 1(b) shows that the differences were large for smaller values.

TABLE 1: Results on agreement between AVRG and Korotkoff blood pressure readings in 100 volunteers

	AVRG*	**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	

* Average of readings at appearance and disappearance of plethysmographic waveform of a pulse oximeter

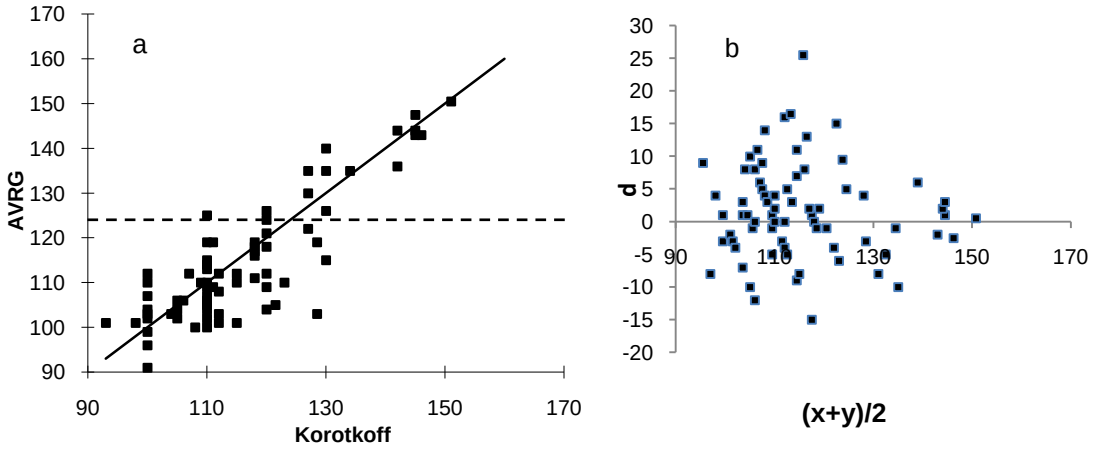

FIGURE 1: (a) Scatter of the pulse oximeter and Korotkoff readings (AVRG = average of readings at disappearance and reappearance of waveform) (b) Plot of d vs. $(x+y)/2$

Despite the means being nearly equal and r very high, the limits of disagreement (Table 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 (Bland and Altman 1986), 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. Thus, the final decision is clinical rather than statistical when this procedure is used.■

1.3 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, values of two different variables are obtained on a series of subjects. For example, you can have weight and height of 20 girls of age 5–7 years. You can also have weight of 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 relationship in this case. Now consider 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 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

$$\text{Intraclass correlation: } \rho_I = \frac{\sigma_M^2}{\sigma_M^2 + \sigma_e^2}, \qquad (2)$$

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 to only two methods. These could be 3 or more. In my weight example, you can compare agreement among 5 machines by taking weight of each of 15 persons on these 5 machines.

The estimate of ρ_I is easily obtained by setting up usual analysis of variance (ANOVA) table. If there are M methods under comparison, the ANOVA table would like as Table 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: 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

Little algebra yields

$$\text{Estimate of intraclass correlation } r_I = \frac{MSA - MSE}{MSA + (n-1)MSE}. \qquad (3)$$

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

In terms of the available values, the computation of the intraclass correlation coefficient 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.

Intraclass correlation coefficient (a pair of readings):

$$r_I = \frac{2\Sigma_i (x_{i1} - \bar{x})(x_{i2} - \bar{x})}{\Sigma_i (x_{i1} - \bar{x})^2 + \Sigma_i (x_{i2} - \bar{x})^2}, \quad (4)$$

where x_{i1} is the measurement on the ith subject (i = 1, 2, …, 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 of 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 more than 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 in 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 (4) is for comparing two methods or two raters. You may use this correlation for several measurements. For example, you may have wave amplitude of electrical waves at M = 6 different sites in brain of each of n = 40 persons. For multiple raters or multiple methods,

Intraclass correlation coefficient (several readings):

$$r_I = \frac{\Sigma_i \Sigma_{j \neq k} (x_{ij} - \bar{x})(x_{ik} - \bar{x})}{(M-1)\Sigma_i \Sigma_j (x_{ij} - \bar{x})^2}; \quad i = 1, 2, \ldots, n; \quad j, k = 1, 2, \ldots, M; \quad (5)$$

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.

For grading of the strength of agreement, following cutoffs can be used.

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

1.3 An Alternative Simple Approach to Agreement Assessment

None of the two methods described in the preceding sections is perfect. Let us first look at their relative merits and demerits and then I propose an alternative method, which may also be not 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 intraclass correlation coefficient 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 a subjective element in this approach also.
3. Intraclass correlation is unit free, easy to communicate, and interpretable on a scale of zero to one as "no agreement" to "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. 5).
6. Intraclass correlation can also be used for comparing one group of raters with another group. Suppose you have 4 male assessors and 3 female assessors. Each subject is measured by all the seven. You can compare intraclass correlation obtained for male assessors with the one obtained for female assessors. In fact, you can have separate set of subjects for assessment by males and another set of subjects to be assessed by females.

A review of the literature suggests that researchers prefer the limits of disagreement approach to the intraclass correlation coefficient 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 help in better appreciation of the procedure to assess quantitative agreement:

1. As mentioned earlier, the limits of disagreement $\bar{d} \pm 1.96 s_d$ 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, whereas I prefer to call them limits of disagreement.
2. The intraclass correlation coefficient too is subject to sampling fluctuation. For assessing agreement, the relevant quantity is the lower bound of r_l. This can be obtained by the method described by Indrayan and Chawla (1994). Their method for computing the intraclass correlation coefficient is based on ANOVA, but that gives the same result as obtained by the formula given in Eq. (4).
3. Though 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 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 to compare with a gold which gives a fixed target value for each subject. For agreement with gold, see Lin et al. (2002).

An Alternative Simple Approach

The limits of disagreement approach just described is based on average difference and has the limitation applicable to all averages. For example, this approach does not work if the bias or error is proportional. Fasting blood glucose level varies from 60 to 300 mg/dL or more. Five percent of 60 is 3 and of 300 is 15. Limits of disagreement approach considers them different and ignores that both are 5% and proportionately same. Also, if one difference is 10 and the other is 2, not necessarily proportional, limits of disagreement consider only the average. Individual differences tend to be overlooked. A few unusually large differences distort average and are not properly accounted except by disproportional inflation of 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 limits of disagreement approach also, albeit for average. These clinical limits of indifference can be absolute or in terms of percentage. If not more than a prespecified, say, 5% of individual differences are beyond these limits in a large sample, you can be safe in assuming adequate agreement. This does not require any calculation of 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 vs. x can track that the differences are systematic or random.

Example 3: Agreement between two methods of measuring fasting blood glucose level

Consider the following data on fasting blood sugar level 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

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

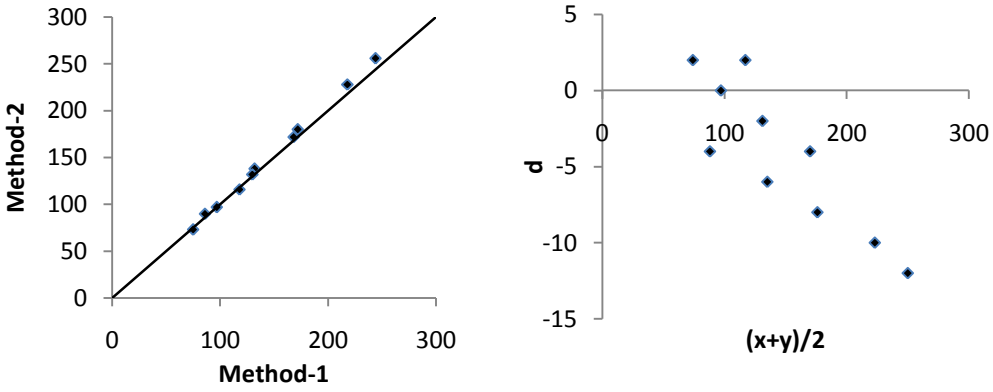

FIGURE 2: (a) y vs x plot for data in Example 3, and (b) d vs. $(x+y)/2$ plot for the same data

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

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 larger error for larger values that proportionate considerations would allow. Also, these are based on average and do not adequately consider individual differences. If one 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 tolerance limit and you can impose an additional condition that none should exceed, say, by 10% of the base value. Since based on individual differences and not average, this alternative approach may be more appealing too.

II Agreement in Qualitative Measurements

Assessing optical disc characteristics by two or more observers, results of Lyme disease serological testing by two or more laboratories, and comparison of x-ray images with Doppler images are examples of the problem of qualitative agreement. The objective is to find the extent of agreement between two or more methods, observers, laboratories, etc. In some cases, for example in comparison of two laboratories, agreement has the same interpretation as **reproducibility**. In the case of comparison of observers, it is termed **interrater reliability**. In all these cases only one group of subjects is assessed twice. Thus this is a matched pair setup. Quantitative agreement was discussed in the previous section. Now consider qualitative agreement.

2.1 The Meaning of Qualitative Agreement

For simplicity, restrict for the time being to the presence or absence of a

characteristic assessed by two observers on the same group of subject. An example is presence or absence of a lesion in x-rays read by two observers. Suppose the observations are as given in Table 3.

TABLE 3: Presence or absence of a lesion assessed by two radiologists in x-rays of 60 suspected cases

	Observer-1		
Observer-2	**Present**	**Absent**	**Total**
Present	29	7	36
Absent	13	11	24
Total	42	18	60

The two observers agree on a total of 40 cases in this example. This is the sum of the frequencies in the leading diagonal cells. In the other 20 cases, the observers do not agree. Apparently, the agreement = 40/60 = 66.7%. But part of this agreement is due to chance, which might happen if both are dumb observers and randomly allocate subjects to present and absent categories. This chance agreement is measured by the cell frequencies expected in the diagonal when the observer's ratings are independent of one another. These expected frequencies are obtained by multiplying the respective marginal totals and dividing by the grand total, as obtained for calculating the chi-square.

For the data in Table 3, the chance-expected frequencies are 36×42/60 = 25.2 and 24×18/60 = 7.2 in the two diagonal cells. The total of these two is 32.4. Agreement on so many cases is expected by chance alone. Thus, agreement in excess of chance is in only 40 – 32.4 = 7.6 cases. The maximum possible excess is 60 – 32.4 = 27.6. A popular measure of agreement is the ratio of the observed excess to the maximum possible excess, in this case 7.6/27.6 = 0.275 or 27.5%. Thus, the two observers in this case do not really agree much on rating of x-rays for the presence or absence of lesion. Most of their agreement is due to chance.

2.2 Cohen Kappa

In terms of notations, the foregoing procedure to measure agreement between qualitative characteristics on the same scale is given by

$$\text{Cohen kappa: } \kappa = \frac{\Sigma O_{ii} - \Sigma(O_{i\bullet}O_{\bullet i}/n)}{n - \Sigma(O_{i\bullet}O_{\bullet i}/n)}, \qquad (6)$$

where O_{ii} is the cell frequency in the ith row and ith column (diagonal element),

$O_{i\bullet}$ is the marginal total in the ith row,

$O_{\bullet i}$ is the marginal total in the ith column, and

n is the grand total.

The first term in the numerator is the observed agreement and the second term is the chance agreement. Thus, the numerator is the agreement in excess of chance. The denominator is its maximum possible value. Kappa in the formula given in Eq. (6) is for a general $I \times I$ table, i.e., the rating is not necessarily restricted to present-absent type of dichotomy but could also be into three or four or a larger number of categories.

Kappa is used mostly to assess how close the agreement is to one rather than how far is it from zero. The following scale is suggested.

Kappa	Strength of agreement
<0.3	Poor
0.3–0.5	Fair
0.5−0.7	Moderate
0.7−0.9	Good
>0.9	Excellent

Example 4: Cohen kappa for agreement between the results of two laboratories

Detection of intrathecal immunoglobulin G (IgG) synthesis is important in patients with suspected multiple sclerosis. Isoelectric focusing is a method used for detection of intrathecal IgG synthesis. Let this be assessed as positive, doubtful, and negative by two laboratories on 129 patients. The results are in Table 4.

TABLE 4: Assessment of intrathecal synthesis by two laboratories

	Laboratory-1			
Laboratory-2	**Positive**	**Doubtful**	**Negative**	**Total**
Positive	36	5	3	44
Doubtful	7	12	6	25
Negative	1	4	55	60
Total	44	21	64	129

In this case, by Eq. (6),

$$\kappa = \frac{(36+12+55) - (44\times 44/129 + 25\times 21/129 + 60\times 64/129)}{129 - (44\times 44/129 + 25\times 21/129 + 60\times 64/129)}$$

$$= \frac{54.155}{80.155} = 0.68.$$

Side note: Generally speaking, a kappa value equal to 0.68 is adequate to conclude fair agreement but, in this example, the investigation is on reproducibility between laboratories. If both the laboratories are using standardized tools and methods, the agreement should be close to 1. Thus, the reproducibility of the method of isoelectric focusing between the laboratories for assessing intrathecal synthesis cannot be considered good in this case despite a not so disappointing value of kappa. ■

The following comments regarding Cohen kappa may be helpful:

1. There are some other measures of agreement for qualitative variables. These are discussed by Agresti (1992). He also describes other agreement assessment models such as log–linear, Rasch, and latent class models.
2. Cohen kappa is valid for nominal categories only. Ordinal or metric categories are considered nominal by this measure and the order is ignored. The other assumptions are that (a) the subjects are independent; (b) the observers, laboratories, or methods under comparison operate independently of one another; and (c) the rating categories are mutually exclusive and exhaustive. These conditions are easily fulfilled in most practical situations.
3. Rare though, you may sometimes find reference to **weighted kappa**. In this, the cells are assigned a weight according to the degree of disagreement they exhibit. Thus, cells in the diagonal, since they are in full agreement, get zero weight. Off–diagonal cells get varying weight depending upon either your perceived importance of the involved cells or on some objective criterion such as quadratic weight. All this makes kappa too complex and possibly not as useful.
4. The kappa value is +1 for complete agreement and 0 if the agreement is the same as expected by chance. However, the value does not become –1 for complete disagreement. Thus, *it is a measure of the extent of agreement but not of disagreement*.
5. For the formula of variance of kappa for large samples, see Fleiss et al. (1969). This variance can be used to construct a CI and to test a hypothesis on the value of kappa. Standard statistical software have provision to do these calculations.
6. Kappa values from different studies may not be comparable as the value also varies with the number of subjects in categories relative to the total. If one study has 30% subjects in category A and another study has 50%, the value of kappa will differ even if the extent of agreement is the same.

7. Kappa does not distinguish between +/− discordance and −/+ discordance. Both get the same weight. Thus kappa can be 0.9 in 100 subjects when all discordances are +/− type and none of −/+ type.

References

Agresti A (1992). Modelling patterns of agreement and disagreement. *Stat Methods Med Res* 1: 201–218.

Bland JM, Altman DG (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* i: 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.

Fleiss JL, Cohen J, Everitt BS (1969). Large sample standard errors of kappa and weighted kappa. *Psychol Bull* 72: 323–327.

Indrayan A (2012). Medical Biostatistics, 3rd edn. Chapman&Hall/ CRCPress.

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.