Chapter 11

Validation of Instrument Measuring Continuous Variable in Medicine

Rafdzah Zaki

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/66151

Abstract

In medicine, accurate measurement of clinical values is vital, either at the stage of health screening, diagnosing cases or making prognosis. There are numerous instruments or machines that have been invented for the purpose of measuring various clinical variables such as blood pressure, glucose level, body temperature and oxygen level. When a new method of measurement or instrument is invented, the quality of the instrument has to be assessed. This chapter will focus on the application of statistical methods used to analyse continuous data in a method comparison study or validation study in medicine. The concept of validity and analysis in method comparison study will be discussed. This chapter also reviews the theoretical aspects of several common methods and approaches that have been used to measure agreement and reliability including the Bland-Altman limits of agreement (LoA) and intra-class correlation coefficient (ICC). Issues related to method comparison studies will be highlighted, which include the evaluation of agreement and reliability in a single study, the application of multiple statistical methods and the use of inappropriate methods in testing agreement and reliability. Finally, the importance of education in method comparison studies among medical professional will be emphasized.

Keywords: agreement, reliability, medical instrument, continuous variable

1. Introduction

In medicine, accurate measurement of clinical values is vital, either at the stage of health screening, diagnosing cases or making prognosis. For example, accurate measurement of blood pressure, heart rate and oxygen level is crucial for monitoring patients under general anaesthesia in surgery. Inaccurate measurement of these variables will result in inappropriate management of the patient, thus putting the patient's life at risk.

Most of the important variables measured in medicine are in numerical forms or continuous in nature, such as blood pressure, glucose level, oxygen level, weight, height, body temperature, creatinine level, albumin level and many other clinical values. There are numerous instruments or machines that have been invented for the purpose of measuring various variables. Some measurements are obtained by using invasive techniques and expensive procedures. Consequently, new instruments and tests are constantly being developed and fashioned to provide complementary meaningful information to the search for information, with the aim of providing cheaper, non-invasive, more convenient and safe methods. Whether a test's outcome can provide trustworthy judgements or decisions depend particularly on the measurement quality of the test [1].

When a new method of measurement or instrument is invented, the quality of the instrument has to be assessed. We want to know by how much the value of measurements obtained using new method differs from the old method, or from the gold standard. Information provided by any clinical instrument cannot be trusted and licitly used in any judgement and decision-making process if the measurement quality has not been evaluated. This is where a method comparison study or a validation study comes into medicine. Clinimetric properties indicating that the test is reliable and valid should be considered as fundamental for determining the measurement quality of any test [2]. In general, clinimetric refers to the development of methodological and statistical methods applicable in clinical medicine in order to assign numbers or scores to observable clinical events [3, 4].

1.1. Validity

An instrument is considered to be valid if it measures what it is intended to measure [3]. The term 'validity' actually has a wide range of classification and definition. In clinical research, current and accepted validity concepts include *criterion* validity, *construct* validity and *content* validity, the first two being the most relevant for performance-based tests [5].

Criterion validity is used to examine the extent to which a measurement instrument provides the same results as the gold standard [5]. This type of validity is the most powerful in terms of its usefulness, and is divided into two types: *concurrent* validity and *predictive* validity. Of these, *concurrent validity* is the most used method. This is when we are trying to compare a new measurement tool with the criterion measure, both of which are given at the same time [5, 6]. The new tool is usually simpler, cheaper or less invasive compared to the standard or currently used tools. In contrast, in *predictive validity*, the criterion will not be available until sometime in the future. When no gold standard is available, the common alternative is to use an accepted and well-grounded reference test to relate to the evaluated test [7, 8]. Generally, this form of validity is used in developing instruments that allow us to get earlier answers or to give earlier predictions than current instruments can provide [5].

Construct validity refers to the degree to which a test measures a hypothetical, non-observable construct and this validity can be established by relating the test to outcomes of other instruments [1, 5]. It is used when we dealing with more abstract variables or factors that cannot be measured directly for example level of anxiety and pain [5]. We cannot see or directly measure anxiety, but we can observe other factors related to anxiety (according to theory) such as sweaty palm and tachycardia. The proposed underlying factors are referred to as *hypothetical construct*

or simply known as *constructs* [5]. So, *construct validity* is the next best option in the absence of an acceptable gold standard. The measurement of instrument under study will be compared with other instruments that claim to measure the same construct [5].

Content validity is a closely related concept, consisting of a judgement whether the instrument samples all the relevant or important content or domains [5, 9]. Content validity can be claimed when a test logically and obviously measures what it purposes to measure [5, 6]. The relationship between the phenomenon being measured and the test score(s) is determined by a panel of experts or researchers [6].

1.2. Reproducibility

Another approach in assessing the quality of measuring instrument is to assess the *reproducibility* of the instrument. This is when we are interested to know whether the new instrument is able to produce similar values as that predicted by the old or standard instrument. In the literature, terms reproducibility is often used interchangeably with the reliability, repeatability, consistency, agreement and stability [9]. Recently, de Vet et al. [3, 10] advocated that reproducibility is the proper term to use in clinical research, making the distinction between two aspects that are important for clinical interpretation: reliability and agreement.

1.2.1. Agreement

Agreement assesses how close the results of repeated measurements are to the 'true value' or the criterion value [10]. So, agreement actually concerns accuracy or validity; more specifically, concurrent validity. An instrument with good agreement will be able to produce accurate repeated measurements in the same person [10]. Thus, agreement parameters are important in instruments that are used for evaluative purposes. In evaluative measurement instruments, the variability between individuals in a population is not important, in comparison to the variability within an individual [10]. This is because, in some clinical settings, we want to detect differences or changes within the same individual, and not how much difference is the individual's value compared to another person's, or with the population. For example, in antenatal clinics we are interested in the weight gain of a mother throughout her pregnancy, and not how much her weight differs from the others'.

Agreement parameters estimate the *measurement error* in repeated measurements. When the measurement error is large, small changes cannot be distinguished from the measurement error [10]. The smaller the measurement error, the smaller the changes that can be detected beyond the measurement error, and the more appropriate the instrument is for evaluative purposes. Thus, for an instrument to be used to evaluate changes over time, such as changes in blood pressure after receiving antihypertensive therapy, it is important for us to ensure the agreement or the accuracy of the instrument.

1.2.2. Reliability

Reliability measures the extent to which the test results can be replicated [11]. For example, if we measure body weight using a scale five times, ideally all five measurements should be the same.

Reliability is concerned with precision. It also represents the extent to which individuals can be distinguished from each other, despite the variability of repeated measurements in one person or subject (i.e. measurement error) [10]. In contrast with agreement, reliability measures the variability between people or subjects. This measurement tells us how well the measured value in one person can be distinguished from another [10]. Thus, reliability parameters are important when measurement instruments are used for discriminative purposes; for example, to decide whether a certain value is normal or abnormal, and when the measurement from the instrument is involved in important decisions, such as whether treatment is required or not.

In clinical practice, the cut-off for normal and abnormal values is usually well established by clinical guidelines, which are produced based on extensive reviews of available evidence. Reliable instruments should be able to provide values that will allow doctors or clinicians to distinguish whether their patients are in the normal or abnormal group. For instance, if we take the blood pressure of one patient five times, all the values should be almost the same, and the values should give us an idea whether the patient's blood pressure is normal or not.

An acceptable range of reliability will vary depending on the circumstances [5]. For example, if repeated measurements of a weighing scale are found to vary around the 'true' weight by 0.5 kg, the reliability of this weighing scale would be acceptable if the measurements are only to be done on an adult population, but not reliable when used to weigh new-born babies in the hospital. This is because differences of 0.5 kg in weight in an adult represent only a very small percentage of an adult body weight, and will not affect him or her clinically. In contrast, a difference of 0.5 kg represents a large proportion of body weight for a new-born baby.

1.3. Agreement versus reliability

Although the terms 'agreement' and 'reliability' carry different meanings, they are sometime used interchangeably in medical literature. To illustrate the concept of agreement and reliability in more simple language, imagine if we have three target boards (see **Figure 1**) that show the results of five repeated measurements of body weight of the same person, using three different scales (A, B and C). The centre of each board indicates the true value. **Figure 1(A)** shows that after taking five measurements using scale A, the results of the measurements are scattered all over the target board. This suggests that the measurements are not near each other (poor reliability), and are not near their intended target or true value (poor agreement).

Figure 1(B) shows that all the five measurements from scale B appear in more or less the same location on the target board, but not in the centre of the target board. This suggests that five different measurements were almost the same (good reliability), but they did not hit the intended target (poor agreement). **Figure 1(C)** shows that all the five measurements from scale C are close to each other (good reliability), and hit the centre of the target board (good agreement).

In most clinical situations, we use the same instrument to evaluate changes over time and also to differentiate values from the normal or abnormal cut-off point (which is usually derived from population-based studies). One of the examples of this situation is in the screening of hypertension cases, and the assessment of the reduction of blood pressure post-treatment, in a clinic or health centre. Both blood pressure measurements are performed using the same blood pressure machine, or sphygmomanometer. So, agreement and reliability parameters are equally

http://dx.doi.org/10.5772/66151

important in determining the quality of instruments. In fact, it is difficult to be certain about the agreement of an instrument if the instrument is not reliable. Similarly, a precise instrument or instrument with good reliability will not necessarily measure the 'true' value. Therefore, when comparing two instruments, or methods of measurement, we should consider assessing the *repeatability* of the instrument, which covers both agreement (accuracy) and reliability (precision).

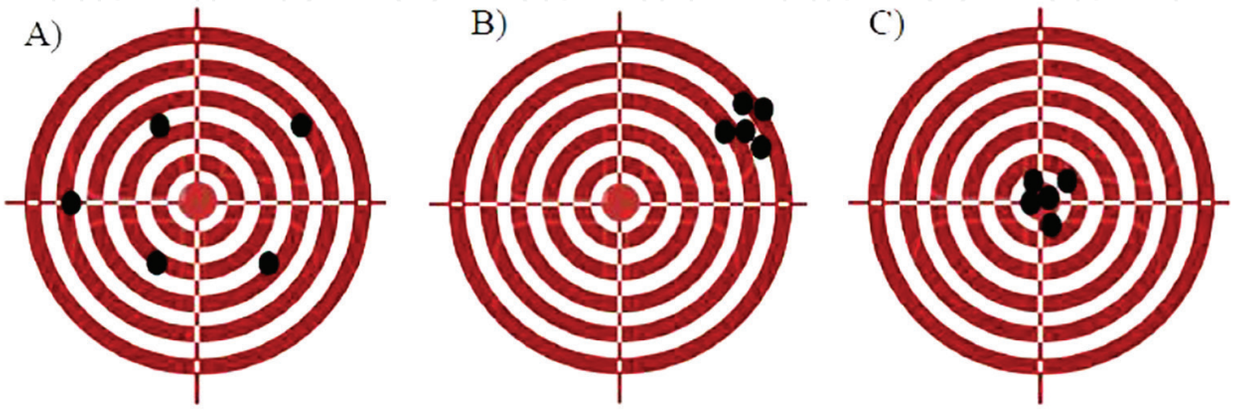

Figure 1. Results of measurements of body weight using three different scales A, B and C.

2. Issues related to method comparison studies

2.1. Evaluation of agreement and reliability in a single study

Agreement and reliability are both important in assessing the quality of instruments. An instrument with high agreement will not be useful if it is unreliable. Ideally, these parameters should be assessed together. However, recent systematic reviews showed that this is not commonly followed in practice, especially with respect to agreement studies [12, 13]. Most (71%) of the reliability studies, measured agreement at the same time [13]. However, only 30% of agreement studies found assessed reliability [12]. Researchers tend to focus on one aspect of quality when validating instruments, although there is a possibility of agreement and reliability studies being conducted separately for the same instrument. Nonetheless, it is important to ensure the reliability of the instrument first, before testing for agreement, because it is impossible to assess the agreement of an unreliable instrument.

2.2. Inappropriate application of statistical method

Thousands of validation studies have been conducted in the past. Various statistical tests have been used to test for agreement and reliability [14–16]. Some of the methods that were used were inappropriate. Correlation coefficient (r), coefficient of determination (r^2), regression coefficient and means comparison have been shown to be inappropriate for the analysis in method comparison study. This has been discussed by Altman and Bland, since the 1980s [15] and also by Daly and Bourke [17]. Reasons for why those methods are inappropriate for the analysis in method comparison study will be discussed in Sections 3 and 4.

One example of the inappropriate application of statistical methods in method comparison study is in the study to explore the suitability of existing formulas to estimate the body surface area (BSA) of new-borns [18]. The authors compared different methods of estimation of body surface area in newborn and used correlation coefficient to determine the agreement of those methods [18]. In one of their results, the authors described that the method of estimating body surface area (BSA) using the BSA-Meban was most similar to the BSA-Mean, by having a mathematically perfect correlation with $r = 1.00$ ($p = 0.000$) [18]. However, their conclusion was obviously inappropriate because the correlation coefficient only measures linear relationship, and does not suggest that the two methods give similar results.

Another example of the inappropriate application of the Pearson correlation coefficient was demonstrated in one study conducted in Greece [19]. The authors aimed to assess the validity of a new motorised isometric dynamometer for measuring strength characteristics of elbow flexor muscles. They set the criteria of the Pearson correlation coefficient's (r) values >0.97 to demonstrate that high agreement occurred between measures, and with $r = 0.986$, they concluded that the new dynamometer was accurate [19].

The use of inappropriate methods for the assessment of agreement and reliability will, undoubtedly, result in an inappropriate interpretation of the results and conclusions on the quality of an instrument. Consequently, this might result in the application of invalid equipment in medical practice, and will jeopardise the quality of care given to patients. The proportion of studies with inappropriate statistical methods might reflect the proportion of medical instruments that have been validated using inappropriate methods in current clinical practice.

As found in recent systematic reviews, 19% of reliability studies [13] and 10% of agreement studies [12] used inappropriate methods, which means that there is a distinct possibility that some medical instruments or equipment used currently were validated using inappropriate methods, with consequently erroneous conclusions being drawn from these methods. This equipment, therefore, may not be as precise or accurate as believed, which could, potentially, affect the management of patients, the quality of care given to patients and, worse, it could cost lives.

Altman and Bland [15] proposed a method for agreement analysis in their original 1983 article. Later, they drew the attention of the medical professionals to this area in an article in The Lancet [20]. Their article [20] received very high citation [21]. The popularity of the Bland-Altman method was thought, owing to its simplicity, practicality and ability to detect bias, when compared to other methods [16].

The issue of which method is the best is still debatable, and almost all methods have been criticised, especially for the agreement study. Even the Bland-Altman method has been criticised. Hopkins [22] demonstrated that the Bland-Altman plot indicates, incorrectly, that there is systematic bias in the relationship between two measures. Recent study also showed that there is overestimation of bias in the Bland-Altman analysis [23].

2.3. Application of multiple methods

According to recent systematic reviews conducted, most reliability studies (86%) relied on a single statistical method to assess reliability [13], in contrast with agreement studies

http://dx.doi.org/10.5772/66151

where most of the studies (65%) used a combination of statistical methods [12]. The application of multiple or a combination of methods, particularly in the assessment of agreement, suggests that there is no consensus among researchers on which method is the best statistical method for measuring agreement. One example of the multiple application of method is in one study that testing the accuracy of peak flow meters [24]. In this study, the authors applied three statistical methods (Pearson's correlation coefficient, comparing mean (significant test) and the Bland-Altman method) to assess for agreement of peak flow meters [24].

A strong reason for using multiple methods in assessing agreement and reliability is that each statistical method has its strengths and weaknesses. The usage of multiple methods in method comparison studies has the advantage of compensating for the limitations of any one single method [14, 25]. As long as the methods chosen are appropriate for it purposes. However, another possible reason for using multiple methods is the researcher's limited understanding of the statistical methods for agreement and reliability. This is probably the reason for the application of multiple inappropriate statistical methods in a single study; for example, the use of both correlation coefficient and significance test of the difference between means, to test for agreement and reliability. Both of these methods have been clearly shown to be inappropriate statistical methods to assess agreement and reliability [15, 17].

3. Most commonly used methods to assess agreement

In 2012, Zaki et al. [12] review the statistical methods used to measure the agreement of equipment measuring continuous variables in the medical literature. The most common method to assess agreement is the Bland-Altman limits of agreement (LoA), followed by correlation coefficient (*r*), comparing means, comparing slope and intercept and intraclass correlation coefficient. However, some of these methods were inappropriate to assess agreement.

3.1. Bland-Altman limits of agreement

Bland-Altman limits of agreement were found to be the most commonly used method to assess agreement in medical literature. In 1983, Bland and Altman introduced limits of agreement (LoA) to quantify agreement [15]. They proposed a scatter plot of the differences of two measurements against the average of the two measurements, and later it becomes a graphical presentation of agreement (see **Figure 2**). Bland and Altman [20] stated that it is very unlikely for two different methods or instruments to be exactly in agreement, or give identical results for all individuals. However, what is important is how close the values obtained by the new method (predicted values) are to the gold standard method (actual values). This is because a very small difference in the predicted and the actual value will not have an effect on decisions of patient management [20]. So they started with an estimation of the difference between measurements by two methods or instruments [20]. To construct limits of agreement, first we need to calculate the mean and standard deviation of these differences. The formula for limits of agreement (LoA) is given as [20]:

$$\text{LoA} = \text{mean difference} \pm 1.96 \times (\text{standard deviation of differences}) \quad (1)$$

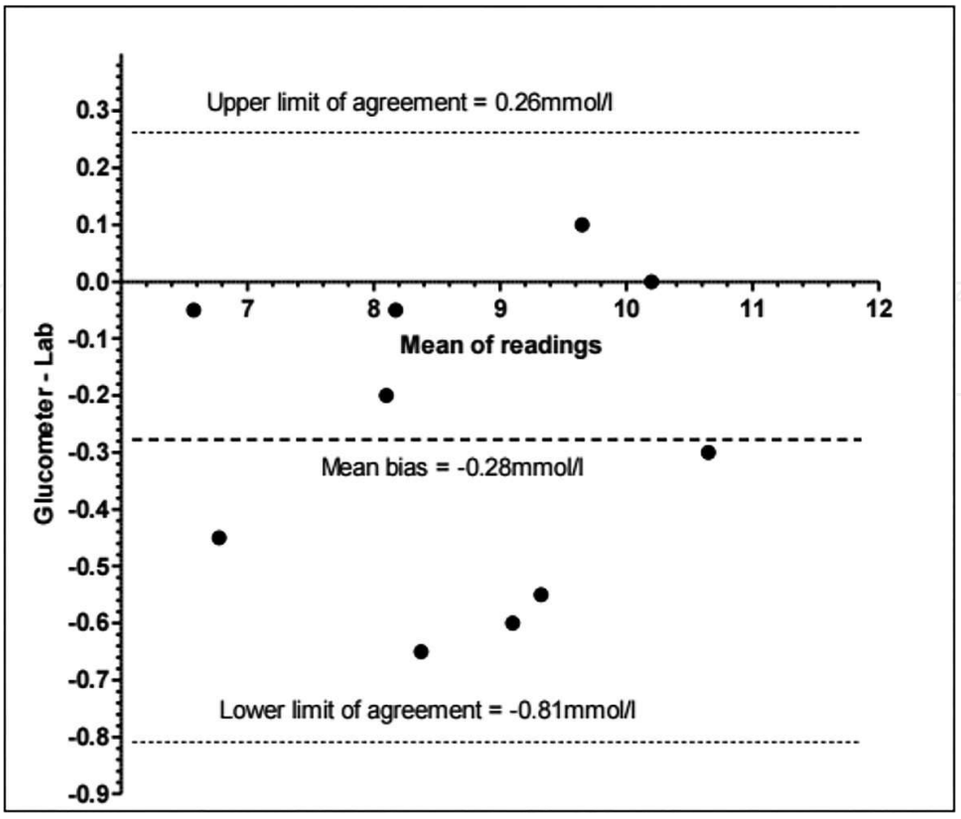

Figure 2. The Bland-Altman plot.

So, 95% of differences should lie within these limits. To illustrate this, we can use the data from **Table 1** (adapted from Table 12.5, interpretation and uses of medical statistics) [17], which compared the values from the glucometer and laboratory. If we apply the data from **Table 1**, the first step of the analysis is to calculate the difference and mean. The mean difference for the data is −0.28 mmol/l, and the standard deviation of difference is 0.27 mmol/l. This makes the LoA = −0.81 to 0.26 mmol/l.

Limits of agreement give us the range of how much one method is likely to differ from another. So it is all about the differences. If we are testing a new method B against the old method A, and the difference is calculated from A–B, then a positive value of limits of agreement means A > B, or new method B underestimates the new method A. If a negative value of limits of agreement means A < B, or the new method B overestimates the old method A. So, the result of Bland-Altman analysis between glucometer and laboratory values (**Table 1**) can be shown as follows:

$$\begin{aligned} &\text{Differences} &&= \text{Glucometer–Laboratory} \\ &\text{Mean difference} &&= -0.28 \text{ mmol/l} \\ &\text{Limits of Agreement} &&= -0.81\text{–}0.26 \text{ mmol/l} \end{aligned} \tag{2}$$

This means that, on average, the glucometer measures 0.28 mmol/l less than the laboratory. Also, 95% of the time the glucometer reading will be somewhere between 0.81 mmol/l below and 0.26 mmol/l above the laboratory values.

Patient	Lab value (*L*)	Glucometer (*G*)	*G–L*	Mean
1	10.20	10.20	0.00	10.20
2	8.20	8.00	–0.20	8.10
3	8.70	8.05	–0.65	8.38
4	9.60	9.70	0.10	9.65
5	9.60	9.05	–0.55	9.33
6	8.20	8.15	–0.05	8.18
7	9.40	8.80	–0.60	9.10
8	7.00	6.55	–0.45	6.78
9	6.60	6.55	–0.05	6.58
10	10.80	10.50	–0.30	10.65

Table 1. Hypothetical data of blood glucose level from a glucometer and laboratory.

3.2. Correlation coefficient

One of the favourite approaches in measuring agreement is to calculate the correlation coefficient (*r*) [11, 15, 26]. This method is the next most popular method after the Bland and Altman method [12]. The first approach in this analysis is to make a scatter plot, and then to calculate product-moment correlation coefficient [11]. To calculate the product moment correlation coefficient (*r*), variables for each pair of measurements are labelled as *X* and *Y*. The formula for the correlation coefficient *r* is given as:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}} \quad (3)$$

If we use an example from data presented in **Table 1** (to compare blood sugar levels from glucometer and laboratory values), the Pearson correlation coefficient (*r*) is 0.9798 with a 95% confidence interval of 0.9139–0.9954, and *p*-value <0.0001 (analysis using SPSS 17.0 software). The null hypothesis here is that the measurements of blood glucose level by the two methods (glucometer and laboratory) are not related linearly. With a very small *p*-value, we can reject this null hypothesis and propose an alternative hypothesis: there is a linear relationship between the measurements of glucose level by the two methods (glucometer and laboratory). Some people will interpret this as being that there is an agreement between the two instruments. This is another mistake conducted by many researchers [15].

Correlation is a measure of association, and only measures the strength of linear relationship [11]. Strong correlation does not mean strong agreement. To demonstrate the inappropriate use of correlation, let's double the value of glucometer from **Table 1** so that it is obvious that there is no agreement between the glucometer and the laboratory value (see **Table 2**). Despite this, the correlation analysis of data from **Table 2** will give exactly the same Pearson correlation coefficient (*r*) of 0.9798, with a similar 95% confidence interval (CI) of 0.9139–0.9954.

Of course, the two instruments (glucometer and laboratory measurement) do not agree, but the correlation coefficient value is still very high, suggesting a strong correlation or association.

Agreement is assessing a different aspect of relationship between two measurements as compared to the correlation coefficient. The correlation coefficient reflects the noises and the direction of a linear relationship [27, 28]. Perfect correlation occurs if all the points lie along any straight line (see **Figure 3**), and so data with poor agreement can produce a high or strong association [20, 29]. Furthermore, data covering an extensive (wide) range of values will appear to be more highly correlated than if it covers a narrow range [20]. Therefore, it is clear that correlation is not an appropriate method for testing agreement.

Lab value	Glucometer	Glucometer ×2
10.20	10.20	20.40
8.20	8.00	16.00
8.70	8.05	16.10
9.60	9.70	19.40
9.60	9.05	18.10
8.20	8.15	16.30
9.40	8.80	17.60
7.00	6.55	13.10
6.60	6.55	13.10
10.80	10.50	21.00

Table 2. Hypothetical data of blood glucose value.

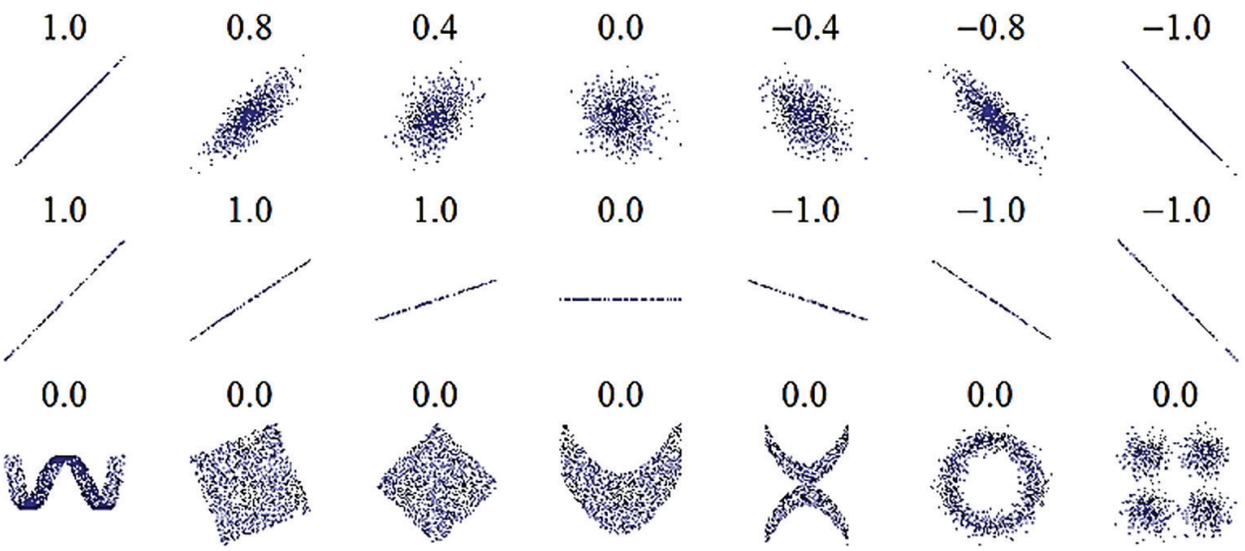

Figure 3. Correlation coefficient values, the noises and direction of a linear relationship [29].

Some people use the *coefficient of determination* (r^2) parameter as a measure of agreement. One example of the application of this method is in a recent study [30] on the accuracy evaluation of point-of-care glucose analysers in the Saudi Arabian market. The authors compare the blood glucose readings from five different types of glucose analysers with the results from laboratory analysis. Their aim was to test the accuracy of the devices. In one of their results, the authors described that the Nova StatStrip device showed an excellent performance that almost agreed and correlated perfectly with the lab results, because the $r^2 = 0.99$ [30]. However, the use of the *coefficient of determination* (r^2) is inappropriate because r^2 is obtained from correlation coefficient r, which is a wrong method to measure agreement. *Coefficient of determination* (r^2) is used to state the proportion of variance in the dependent variables that is explained by the regression equation or model [17]. The more closely the points are dispersed around the regression line in the scatter plot, the higher the proportion of variation explained by the regression line, thus the greater the value of r^2 [11]. So it applies a similar concept to the correlation coefficient.

3.3. Comparing means

The third most popular method used is comparing means of readings from two instruments [12]. In this method, the means of readings from two instruments are compared. The test of significance is then carried out to test the null hypothesis that there is no difference between the means of readings from the two instruments.

In assessing agreement, the same measurement of similar subjects will be taken using different instruments. Therefore, a paired *t*-test is usually used to test the hypothesis. Here, we want to know whether the difference observed is the true difference or has only occurred by chance when there was really no difference in the population. If the difference is truly occurring, and the null hypothesis is not true, then the alternative hypothesis must be true. So, in this case, the alternative hypothesis is that there is a significant difference between the mean of reading from the two instruments.

People have interpreted non-significance results to mean that there is not enough evidence to show that the two means differ (i.e. no differences), thus there is an agreement between the two groups, and vice versa. An example of this inappropriate approach is in a study conducted in Sweden on the assessment of left ventricular volumes, using simplified 3-D echocardiography and computed tomography [31]. However, the paired t-test with non-significant results does not indicate agreement. The reason for this is that the value of mean is affected by the value of each piece of data, especially when there is an outlier. Distribution of differences between the instruments can lead to a difference in means being non-significant. It is possible that poor agreement between the two instruments can be hidden in the distribution of differences, and thus the two methods can appear to agree [17]. To illustrate this example, we have a hypothetical dataset comparing the measurements from standard instrument A, with the new instruments B and C (**Table 3**).

From the dataset (**Table 3**), it is obvious that the two new instruments (B and C) do not agree with the standard instrument A. The mean and standard deviation for the three groups are all

Patient	A	B	C
1	1	1	5
2	2	3	4
3	3	2	3
4	4	5	2
5	5	4	1

Table 3. Hypothetical dataset for instruments A, B and C.

the same: the mean is equal to 3.0 and standard deviation is equal to 1.58. If we compare the readings from instruments A and B, using a paired *t*-test, the results will be:

$$\begin{aligned} &\text{Mean difference (confidence interval)} = 0\ (-1.24\text{–}1.24\ ) \\ &\text{Standard deviation of differences} = 1.0 \\ &p\text{-value} = 1.0 \end{aligned} \tag{4}$$

So, from this analysis, we can conclude that there is no difference between the mean reading of instruments A and B. If we are saying that non-significant results indicate an agreement, this suggests that there is an agreement between instruments A and B. However, we know that this is not true. Similarly, the result will be not significant when we compare the mean reading of instruments A and C, where the results will be:

$$\begin{aligned} &\text{Mean difference (confidence interval)} = 0\ (-3.93\text{–}3.93\ ) \\ &\text{Standard deviation of differences} = 3.16 \\ &p\text{-value} = 1.0 \end{aligned} \tag{5}$$

Again, this does not suggest that there is an agreement between instruments A and C. The inappropriate application of the test of significance, as a test for agreement, has also been discussed earlier in the article by Altman and Bland [15]. What matters in agreement is that each reading from the standard instrument should be repeated by the second instrument. We are not interested in the mean of readings by each instrument, but are interested in each individual reading. Therefore, comparing means using a significance test is not an appropriate method for assessing agreement.

3.4. Intra-class correlation coefficient

The intra-class correlation coefficient (ICC)is another popular method found used to test for agreement [12]. ICC was devised initially to assess the relationship between variables within classes, or reliability. However, it was then used to assess agreement, to avoid the problem of linear relationship being mistaken for agreement in the product moment correlation coefficient (r) [26, 32]. Different assignments of measurements of X and Y, in the calculation of the correlation coefficient (r), would produce different values of r. To overcome some of the limitations of the correlation coefficient (r), the ICC averages the correlations among all the possible ordering of the pairs [33]. The ICC also extends to more than two observations, in contrast with the correlation coefficient (r) [11]. In general, the ICC is a ratio of two variances:

http://dx.doi.org/10.5772/66151

$$\text{ICC} = \frac{\text{Variance owing to rated subjects}}{\text{Variance owing to subjects} + \text{Eerror}} \quad (6)$$

The value of the ICC can theoretically vary from 0 to 1, where 0 indicates no reliability or disagreement in the agreement study. The ICC of one indicates perfect reliability, or perfect agreement. There are different types of ICC that have been described by Shrout and Fleiss [33]. McGraw and Wong [35] expanded the Shrout and Fleiss system to include two more general forms of ICC. Weir [34] summarised different types of ICC, based on models introduced by Shrout and Fleiss [33] and McGraw and Wong [35] (see **Table 4**).

Shrout and Fleiss [33]	Computational formula	McGraw and Wong [35]	Model
1, 1	$\frac{MS_B - MS_W}{MS_B + (k-1)MS_W}$	1	One-way random
1, *k*	$\frac{MS_B - MS_W}{MS_B}$	*K*	One-way random
	Use 3, 1	*C*, 1	Two-way random
	Use 3, *k*	*C*, *k*	Two-way random
2, 1	$\frac{MS_S - MS_E}{MS_S + (k-1)MS_E + \left(k\left(MS_T - MS_E\right)/n\right)}$	*A*, 1	Two-way random
2, *k*	$\frac{MS_S - MS_E}{MS_S + \left(k\left(MS_T - MS_E\right)/n\right)}$	*A*, *k*	Two-way random
3, 1	$\frac{MS_S - MS_E}{MS_S + (k-1)MS_E}$	*C*, 1	Two-way fixed
3, *k*	$\frac{MS_S - MS_E}{MS_S}$	*C*, *k*	Two-way fixed
	Use 2, 1	*A*, 1	Two-way fixed
	Use 2, *k*	*A*, *k*	Two-way fixed

MS_B, Between-subjects mean square; MS_E, error mean square; MS_S, subjects mean square; MS_T, trials mean square; MS_W, within-subjects mean square.

Table 4. Different types of ICC.

Shrout and Fleiss [33] suggested three main models: Model 1 is a one-way fixed model; Model 2 is a two-way random model; and Model 3 is a two-way fixed model. The model is represented in the format of ICC (a, b). The value of '*a*' can be 1, 2 or 3 (this depends on the three main models). For value '*b*', when $b = 1$, this suggests single measures ICC and $b = k$ suggests averaged measures ICC [34].

In the ICC model suggested by McGraw and Wong [35], the designation '*C*' refers to consistency and '*A*' refers to absolute agreement. The '*A*' model considers both fixed and systematic error, whereas the '*C*' model only considers fixed error [34, 35].

Although a total of 10 ICC models were summarised by Weir [34], there are similarities in some of the ICC formula for different types of ICC. This is because the difference between the random model and the fixed model is not in the calculation but in the interpretation of the ICC [35]. According to Shrout and Fleiss [33], there is only one ICC that measures the extent of absolute agreement, and that is ICC (2, 1), which is based on the two-way random-effects ANOVA (analysis of variances) [14, 33]. This model is similar to ICC (*A*, 1), as suggested by McGraw and Wong [35].

The ICC (*C*, 1) for consistency simply compares the consistency between trials. For example, for the hypothetical data from **Table 5**, will produce ICC (*C*, 1) = 1.0, which is interpreted as a perfect agreement. However, the absolute agreement ICC, or ICC (*A*, 1), compares both the consistency between trails and the agreement between ratings. So, the same pairs of data from **Table 5** will produce ICC (*A*, 1) = 0.67, which suggests some degree of disagreement (or moderate agreement).

Patient	First reading	Second reading
1	2	4
2	4	6
3	6	8

Table 5. Hypothetical dataset of repeated measurements from instrument A.

However, the use of ICC in assessing agreement has been criticised by Bland and Altman [32]. In testing the agreement of instruments, the new method will usually be compared to the standard instrument [32]. The aim of testing is to ensure that the new method will produce the same measurements as the standard instrument (i.e. good agreement). This can also mean that the new method is designed to provide similar predictions of measurement as the standard instrument. So, there is clear ordering of the two variables, where the measurements from the standard instrument are usually denoted as *X* and measurements of the new method are denoted as *Y*.

The ICC also ignores the ordering and treats both methods as a random sample from a population of methods [32]. In an agreement study, there are two specific methods that will be compared, not two instruments chosen at random from some population. Another assumption in the ICC model, which is quite unjustified in methods comparison study, is that the measurement error of both methods has to be the same [32]. The main purpose in testing agreement is to identify the measurement error of the new instrument in comparison to the standard instrument. Another issue with ICC is that it is influenced by the range of data. If the variance between subjects is high, the reliability will certainly appear to be high [14].

3.5. Comparing slopes and *y*-intercepts

Often, in testing for agreement, the slope is tested against one. The argument is that if the two methods or instrument are equivalent (i.e. if it measures the same variable of the same subject, both instruments will give the same reading), thus the slope of the straight line will be one [15].

http://dx.doi.org/10.5772/66151

Straight line equation will show the relationship between two variables, and can be expressed as: $y = \alpha + \beta x$, where 'y' is the predicted or expected value for any given value of 'x', while 'α' is the intercept of the straight line with the y-axis and 'β' is the slope. The values of both 'α' and 'β' are constant. The slope 'β' is also called the *regression coefficient*, and measures the amount of change in the 'y' variable for a unit change in 'x' [11]. So, if instrument A measures 'y', and instrument B measures 'x' and if $y = x$, the slope of the straight line equation is equal to one. It is true that the straight line of $y = x$ will always have slope of 1. However, this is not always true in reverse, because for a line with a slope of 1, the straight line could be $y = x$, or could be $y = \alpha + x$. Therefore, testing the slope is equal to 1, is also an inappropriate method of testing agreement. When the test of slope is equal to 1 is significant, some people proceed to test the y-intercept. Theoretically, if slope is 1 and y-intercept is 0, then y will be equal to x ($y = x$). However, testing both slope and intercept to assess agreement is not so popular compared to other methods.

Bland and Altman [36] also suggested that the old measurement (y) can be regressed on the new measurement (x), and then one can calculate the standard error of a prediction of the old value from the new. This can be used to estimate predicted value from old measurements for any observed value of new measurement, with a confidence interval, which is also known as a *prediction interval* [36]. However, the problem is that the prediction interval is not constant; it is smaller in the middle, and wider towards the extremes [36].

4. Most commonly used methods to assess reliability

Various methods have also been used to estimate reliability. Recent review of reliability study in medicine found that among popular methods used include: intra-class correlation coefficient, comparing means, Bland-Altman limits of agreement, and correlation coefficient (r).

4.1. Intra-class correlation coefficient

The intra-class correlation coefficient (ICC) is the most popular method used to assess reliability of medical instruments [13]. The ICC was originally proposed by Fisher [37, 38]. He was a statistician from England, and Fisher's exact test was one of his well-known contributions to statistics [37, 39]. The earliest ICCs were modifications of the Pearson correlation coefficient [34]. However, the modern version of ICC is now calculated using variance estimates, obtained from the analysis of variance or ANOVA, through partitioning of the total variance between and within subject variance [14].

The general formula for ICC is given as [34]:

$$\text{ICC} = \frac{\text{Subject variability } (\delta_S^2)}{\text{Subject variability } (\delta_S^2) + \text{Mmeasurement error } (\delta_E^2)} \tag{7}$$

Values obtained from ANOVA table:

Measurement error, $\delta_E^2 = \text{Mean square of Error, } MS_E$

$$\text{Subject variability, } \delta_S^2 = \frac{\text{Mean square of Ssubject, } MS_S - \text{Mmean square of Eerror, } MS_E}{\text{Number of observer}} \tag{8}$$

There is no ordering of the repeated measures and can be applied to more than two repeated measurements [5]. ICC is a ratio of variances derived from ANOVA, so it is unit less. The ratio is closer to 1.0, the higher the reliability [34]. Suppose, for example, that we measure carbon monoxide level for 10 patients using a same instrument three times. The hypothetical data are shown in **Table 6**. From the data, an ANOVA table can then be developed as in **Table 7**.

Patient	First reading	Second reading	Third reading	Mean
1	6	7	8	7
2	4	5	6	5
3	2	2	2	2
4	3	4	5	4
5	5	4	6	5
6	8	9	10	9
7	5	7	9	7
8	6	7	8	7
9	4	6	8	6
10	7	9	8	8

Table 6. Hypothetical data of repeated measurements of carbon monoxide level.

Source of variation	Sum of squares	Degree of freedom	Mean square
Patients	114	9	12.67
Raters/instrument	20	2	10
Error	10	18	0.56
Total	144	29	

Table 7. Analysis of variance summary table.

From **Table 7**, the value of ICC can be calculated:

$$\begin{aligned} \text{Measurement error, } \delta_E^2 &= MS_E = 0.56 \\ \text{Subject variability, } \delta_S^2 &= \frac{MS_S - MS_E}{\text{Number of observer}} \\ &= \frac{12.67 - 0.56}{3} = 4.04 \\ \text{ICC} &= \frac{4.04}{4.04 + 0.56} = 0.88 \end{aligned} \tag{9}$$

The interpretation is that 88% of the variance in the measurements results from the 'true' variance among patients. However, note that this is according to the 'classical' definition of reliability. There are different forms of ICC depending on various assumptions or criteria. Chinn [40] recommended that any measure should have an intra-class correlation coefficient

of at least 0.6 to be useful [40], whereas Rosner [41] suggested the interpretation of ICC as shown in **Table 8**.

ICC value	Interpretation
<0.4	Poor reliability
$0.4 \leq ICC < 0.75$	Fair to good reliability
≥0.75	Excellent reliability

Table 8. Interpretation of ICC.

4.2. Comparing means/mean difference

Second most popular method that has been used to assess reliability is to compare means of two sets of measurements (either using t-test or looking at the mean difference) [13]. Since reliability involves repeated measurement of the same subject, a paired t-test is usually applied. However, the paired t-test only gives information about differences between the means of two sets of data, and not about individual differences [14]. As in the explanation in Section 3.3, on assessing agreement, comparing means is also not a suitable method of assessing reliability.

4.3. Bland-Altman method

The Bland-Altman limits of agreement (LoA) also has been used as a method to assess reliability. Bland and Altman [20] suggested that LoA are suitable for the analysis of repeatability of a single measurement method. However, the use of LoA to evaluate reliability has been criticised, as it only estimates reliability when there are two observations for each subject [20]. This breaches the concept of reliability, that allows repeated (more than two) numbers of observations per subject [11]. Although Bland and Altman [42] suggested methods to deal with multiple measurements in calculating the LoA, this method is more suitable for the analysis of agreement rather than reliability. They proposed calculating the mean of the replicated measurements by each instrument, for each subject [42]. Then, these pairs of means could be used to compare the two instruments using the limits of agreement [42]. The use of LoA in the analysis of reliability also has been criticised by Hopkins [43], who gave reasons why LoA is not the best method to use for reliability analysis. According to Hopkins [43], the values of the LoA can result in up to a 21% bias, and this depends on the degrees of freedom of the reliability study (i.e. number of participants and trials).

4.4. Correlation coefficient

As discussed in Section 3.2, the correlation coefficient provides information about the association and the strength of linear relationship. Correlation will not detect any systematic or fixed errors, and it is possible to have two sets of scores that are highly correlated, but not repeatable [14]. Therefore, it is recommended that the correlation coefficient should not be used in isolation for measuring reliability [14, 44]. Furthermore, the correlation coefficient also breaches the concept of reliability, as it only estimates reliability when there are only two observations for each subject [11].

5. Summary

Agreement signifies the accuracy of certain instruments, whereas reliability indicates precision. It is imperative that all medical instruments are accurate and precise. Otherwise, a failure may lead to critical medical errors. Therefore, there is a necessity for the proper evaluation of all medical instruments, and it is important to be sure that the appropriate statistical method has been used. Preferably, agreement and reliability should be assessed together in a validation study.

Simplicity, practicality (or interpretability) and ability of a certain method to detect systematic bias are among the important factors when choosing a method to evaluate agreement. Nonetheless, the ability of the method in detecting bias is still the priority. The use of multiple methods has the advantage of compensating for the limitations of any single method. However, the application of multiple inappropriate statistical methods should be avoided.

Finally, the inappropriate analysis in the method comparison study is a cause for concern in the medical field and cannot be ignored. It is important for medical researchers and clinicians from all specialties to be aware of this issue because inappropriate statistical analyses will lead to inappropriate conclusions. Thus, jeopardising the quality of the evidence, which may in turn, influences the quality of care given to the patients. Educating medical researchers on methods in validation study and clear recommendations and guidelines on how to perform the analysis will improve their knowledge in this area, and help reducing the problem of inappropriate statistical analysis. It is also important to involve statisticians who are able to understand in depth of various statistical methods in the medical education program. Consulting statisticians or inviting them as part of the medical research team also could help in reducing mistake in statistical analysis.

Acknowledgements

I would to express my utmost gratitude to Prof. Dr. Awang Bulgiba and Prof. Dr. Noor Azina Ismail from the University of Malaya for their support, guidance and assistance when I was conducting research in this area. Also for their motivation and support in my career development.

Author details

Rafdzah Zaki

Address all correspondence to: rafdzah@hotmail.com

Department of Social & Preventive Medicine, Faculty of Medicine, Julius Centre University of Malaya, Kuala Lumpur, Malaysia

References

[1] Portney, L.G, Watkins, M.P. Foundations of clinical research: Applications to practice. New Jersey: Prentice-Hall; 2000.

[2] Feinstein, A.R. Clinimetrics. New Haven: Yale University Press; 1987.

[3] de Vet, H.C.W, Terwee, C.B, Bouter, L.M. Current challenges in clinimetrics. Journal of Clinical Epidemiology. 2003;**56**(12):1137–1141.

[4] de Vet, H.C.W, Terwee, C.B, Bouter, L.M. Clinimetrics versus psychometrics: two sides of the same coin. Journal of Clinical Epidemiology. 2003;**56:**1146–1147.

[5] Streiner, D.L, Norman, G.R. Health measurement scales. Apractical guide to their development and use. 2nd ed. Oxford: Oxford University Press; 1995.

[6] Haynes, S.N, Richard, D.C.S, Kubany, E.S. Content validity in psychological assessment: a functional approach to concepts and methods. Psychological Assessment. 1995;7:238–247.

[7] Baxter, P. Invalid measurement validity. Developmental Medicine & Child Neurology. 2005;**47**:291.

[8] Lambert, H.C, Gisel, E.G, Wood-Dauphinee, S. The functional assessment of dysphagia: psychometric standards. Physical & Occupational Therapy in Geriatrics. 2002;**19**:1–14.

[9] Innes, E, Straker, L. Validity of work-related assessments. Work. 1999;**13**:125–152.

[10] de Vet, H.C.W, Terwee, C.B, Knol, D.L, Bouter, L.M. When to use agreement versus reliability measures. Journal of Clinical Epidemiology. 2006;**56**:1033–1039.

[11] Fay, M.P. Random marginal agreement coefficients: rethinking the adjustment for chance when measuring agreement. Biostatistics. 2005;**6**(1):171–180.

[12] Zaki, R, Bulgiba A, Ismail R, Ismail N.A. Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review. PLoS One. 2012;**7**(5):e37908. DOI: 10.1371/journal.pone.0037908

[13] Zaki, R, Bulgiba, A, Nordin, N, Ismail, N.A. A systematic review of statistical methods used to test for reliability of medical instruments measuring continuous variables. Iranian Journal of Basic Medical Science. 2013;**16**(6):803–807.

[14] Bruton, A., Conway, J.H, Holgate, S.T. Reliability: what is it, and how is it measured? Physiotherapy. 2000;**86**(2):94–99.

[15] Altman, D.G, Bland, J.M. Measurement in medicine: the analysis of method comparison studies. The Statistician. 1983;**32**:307–317.

[16] Bahareh, A, Saeed, H, Ramin, H. Comparison of Harris benedict and Mifflin-St Jeor equations with indirect calorimetry in evaluating resting energy expenditure. Indian Journal of Medical Science. 2008;62(7): 283–290.

[17] Daly, L.E, Bourke, G.J. Interpretation and use of medical statistics. 5th ed. Oxford: Blackwell Science Ltd.; 2000.

[18] Ahn, Y, Garruto, R.M. Estimations of body surface area in newborns. Acta Paediatrica (Oslo, Norway: 1992). 2008;**97**(3):366–370.

[19] Milias, G.A, Antonopoulou, S, Anthanasopoulos, S. Development, reliability and validity of a new motorized isometric dynamometer for measuring strength characteristics of elbow flexor muscles. Journal of Medical Engineering & Technology. 2008;**32**(1):66–72.

[20] Bland, J.M, Altman, D.G. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307–310.

[21] Bland, J.M, Altman, D.G. Agreed statistics: measurement method comparison. Anesthesiology. 2012;**116**(1):182–185.

[22] Hopkins, W.G. Bias in Bland-Altman but not regression validity analyses. Sportscience. 2004;**8:**42–46.

[23] Zaki, R, Bulgiba, A, Ismail, N.A. Testing the agreement of medical instruments: overestimation of bias in the Bland-Altman analysis. Preventive Medicine. 2013;**57**(Suppl):S80–S82.

[24] Nazir, Z, et al. Revisiting the accuracy of peak flow meters: a double-blind study using formal methods of agreement. Respiratory Medicine. 2005;**99**:592–595.

[25] Luiz, R.R, Szklo, M. More than one statistical strategy to assess agreement of quantitative measurements may usefully be reported. Journal of Clinical Epidemiology. 2005;**58**(4):215–216.

[26] Lee, J, Koh, D, Ong, C.N.Statistical evaluation of agreement between two methods for measuring a quantitative variable. Computers in Biology and Medicine. 1989;**19**(1):61–70.

[27] Lin, L.I. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. Statistics in Medicine. 2000;**19**(2):255–270.

[28] Bland, J.M. An Introduction to medical statistics. 2nd ed. Oxford: Oxford University Press; 1995.

[29] Wikipedia. The Free Encyclopedia. Correlation [Internet]. Available from: http://en.wikipedia.org/wiki/Correlation. [Accessed: 30 January 2009]

[30] Hanbazaza, S.M, Mansoor, I. Accuracy evaluation of point-of-care glucose analyzers in the Saudi market. Saudi Medical Journal. 2012;**33**(1):91–92.

[31] Mårtensson, M, Winter, R, Cederlund, K, Ripsweden, J, Mir-Akbari, H, Nowak, J, Brodin, L. Assessment of left ventricular volumes using simplified 3-D echocardiography and computed tomography – a phantom and clinical study. Cardiovascular Ultrasound. 2008;**6**(26)DOI: 10.1186/1476-7120-6-26.

[32] Bland, J.M, Altman, D.G. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Computers in Biology and Medicine. 1990;**20**(5):337–340.

[33] Shrout, P, Fleiss, J. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin. 1979;**86**(2):420–428.

[34] Weir, J.P. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. Journal of Strength and Conditioning Research. 2005;**19**(1):231–240.

[35] McGraw, K, Wong, S. Forming inferences about some intraclass correlation coefficients. Psychological Methods. 1996;**1:**30–46.

[36] Bland, J.M, Altman, D.G. Applying the right statistics: analyses of measurement studies. Ultrasound in Obstetrics and Gynecology. 2003;**22**(1):85–93.

[37] Wikipedia. The Free Encyclopedia. Ronald A Fisher [Internet]. Available from: https://en.wikipedia.org/wiki/Ronald_A._Fisher. [Accessed: 12 June 2011]

[38] Fisher, R.A. Statistical methods for research workers. Edinburgh: Oliver and Boyd; 1925.

[39] Fisher, J. R. A. Fisher: The life of a scientist. New York: Wiley; 1978.

[40] Chinn, S. Repeatability and method comparison. Thorax. 1991;**46**:454–456.

[41] Rosner, B. Fundamentals of biostatistics. 6th ed. Duxbury: Thomson Brooks/Cole; 2006.

[42] Bland, J.M, Altman, D.G. Measuring agreement in method comparison studies. Statistical Methods in Medical Research. 1999;**8**(2):135–160.

[43] Hopkins, W.G. Measures of reliability in sports medicine and science. Sports Medicine. 2000;**30**(1):1–15.

[44] Neveu, D, Aubas, P, Seguret, F, Kramar, A, Dujols, P. Measuring agreement for ordered ratings in 3 × 3 tables. Methods of Information in Medicine. 2006;**45**(5):541–547.

IntechOpen