METHOD COMPARISON STUDIES IN MEDICINE

Rafdzah Z[1], Bulgiba A[1], Ismail NA[2]

1 Julius Centre University of Malaya, Department of Social & Preventive Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur
2 Department of Applied Statistics, Faculty of Economics & Administration, University of Malaya, Kuala Lumpur

Correspondence:

Rafdzah Zaki
Julius Centre University of Malaya,
Department of Social & Preventive Medicine,
Faculty of Medicine, University of Malaya,
50603 Kuala Lumpur, Malaysia.
Email: rafdzah@ummc.edu.my

ABSTRACT

INTRODUCTION AND OBJECTIVE:

Most of important variables measured in medicine are in numerical forms or continuous in nature. New instruments and tests are constantly being developed for the purpose of measuring various variables, with the aim of providing cheaper, non-invasive, more convenient and safe methods. When a new method of measurement or instrument is invented, the quality of the instrument has to be assessed. Agreement and reliability are both important parameters in determining the quality of an instrument. This article will discuss some issues related to methods comparison study in medicine for the benefit of medical professional and researcher.

METHOD:

This is a narrative review and this article review the most common statistical methods used to assess agreement and reliability of medical instruments that measure the same continuous outcome. The two methods discussed in detail were the Bland-Altman Limits of Agreement, and Intra-class Correlation Coefficient (ICC). This article also discussed some issues related to method comparison studies including the application of inappropriate statistical methods, multiple statistical methods, and the strengths and weaknesses of each method. The importance of appropriate statistical method in the analysis of agreement and reliability in medicine is also highlighted in this article.

CONCLUSION:

There is no single perfect method to assess agreement and reliability; however researchers should be aware of the inappropriate methods that they should avoid when analysing data in method comparison studies. Inappropriate analysis will lead to invalid conclusions and thus validated instrument might not be accurate or reliable. Consequently this will affect the quality of care given to a patient.

Keywords: *agreement, reliability, method comparison study, validation study*

Introduction

In medicine, accurate measurement of clinical values is vital. Most of important variables measured in medicine are in numerical forms or continuous in nature, such as blood pressure, body temperature, haemoglobin level, and many other clinical values. Inaccurate measurement of these variables will result in inappropriate management of the patient, thus putting the patient's life at risk.

There are numerous instruments or machines that have been invented for the purpose of measuring various variables. New instruments and tests are constantly being developed, with the aim of providing cheaper, non-invasive, more convenient and safe methods. When a new method of measurement or instrument is invented, the quality of the instrument has to be assessed. This is where a method comparison study or a validation study comes into medicine. This article will discuss some issues related to methods comparison study in medicine for the benefit of medical professional and researcher.

Agreement versus Reliability

Agreement and reliability are both important parameters in determining the quality of an instrument. To illustrate the

concept of agreement and reliability in a 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). 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 suggest 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 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 five measurements from scale C are close to each other (good reliability), and hit the centre of the target board (good agreement).

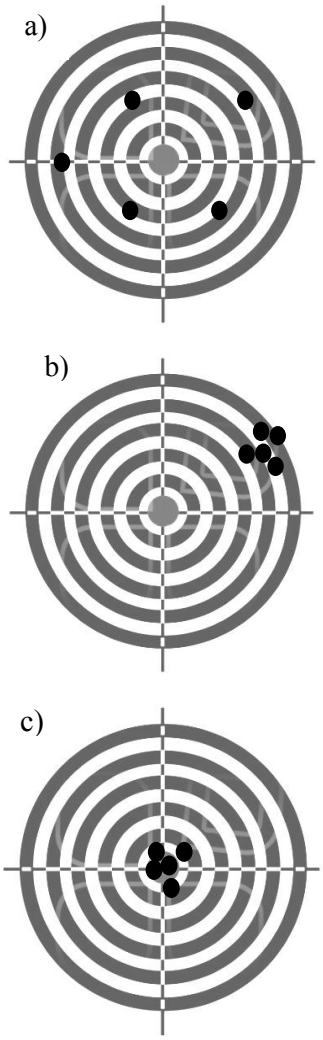

Figure 1: Results of measurements of body weight using three different scales A, B and C.

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 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 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 both agreement (accuracy) and reliability (precision).

An instrument with high agreement will not be useful if it is unreliable. Ideally, these parameters should be assessed together. However, we have conducted two systematic reviews (1, 2) and found that this is not commonly followed in practice, especially with respect to agreement studies. Most of the reliability studies (71%) also measured agreement at the same time (2). However, only 30% of agreement studies assessed reliability (1). 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.

Statistical Methods of assessing Agreement and Reliability

There are several methods and approaches that have been used to measure agreement and reliability. The most common method to assess agreement found in the systematic review (1) is the Bland-Altman Limits of Agreement (LoA), followed by Correlation Coefficient (*r*), comparing means, comparing slope and intercept, and Intra-class Correlation Coefficient. According to the systematic review of reliability studies (2), various methods have also been used to estimate reliability, and among these popular methods include: Intra-class Correlation Coefficient, comparing means, Bland-Altman Limits of Agreement, and Correlation Coefficient (*r*). However, Correlation Coefficient (*r*), comparing means, and ICC have been shown to be inappropriate in assessing agreement. Whereas, in the analysis of reliability, Correlation Coefficient (*r*), Bland-Altman Limits of Agreement and comparing means were thought to be inappropriate.

Agreement Analysis

In 1983, Bland and Altman introduced Limits of Agreement (LoA) to quantify agreement (3). Bland and Altman (4), 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 (4). So they started with an estimation of the difference between measurements by two methods or instruments (4). The formula for Limits of Agreement (LoA) is given as (4):

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

The 95% Limits of Agreement is dependent on the assumptions that the mean and standard deviation of the differences are constant throughout the range of measurement, and the distribution of these differences follow approximately a normal distribution (3). It is important to check for these assumptions (3). Altman and Bland (1983) proposed a scatter plot of the differences of two measurements against the average of the two measurements, and a histogram of the differences, to check for these assumptions (3). Initially, the scatter plot is only to check the assumption and not the analysis of agreement, but then it becomes a graphical presentation of agreement (see Figure 2).

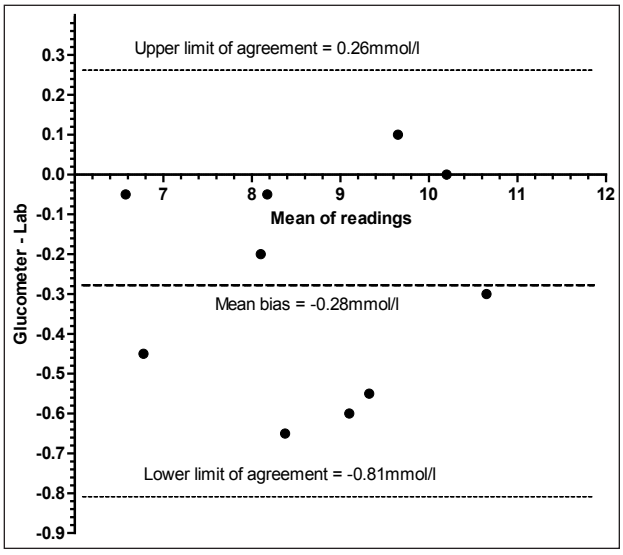

Figure 2: The Bland-Altman Plot

Reliability Analysis

The Intra-class Correlation Coefficient was originally proposed by Sir Ronald Aylmer Fisher (5, 6). He was a statistician from England, and Fisher's exact test was one of his well-known contributions to statistics (5, 7). The earliest ICCs were modifications of the Pearson Correlation Coefficient (8). 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 (9, 10).

The general formula for ICC is given as (8):

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

Values obtained from ANOVA table:

Measurement error, (δ_E^2) Mean square of Error, MS_E

$$\text{Subject variability, } \delta_S^2 = \frac{\text{Mean square of Subject, } MS_S - \text{ Mean square of Error, } MS_E}{\text{Number of observer}}$$

ICC is a ratio of variances derived from ANOVA, so it is unit-less. The closer this ratio is to 1.0, the higher the reliability (8). Chinn (1991) recommended that any measure should have an Intra-class Correlation Coefficient of at least 0.6 to be useful Chinn, 1991). Rosner (11) suggested the interpretation of ICC as shown in Table 1:

Table 1: Interpretation of ICC

ICC value	Interpretation
< 0.4	poor reliability
0.4 ≤ ICC < 0.75	fair to good reliability
≥ 0.75	excellent reliability

Is the most popular method the best?

Agreement Analysis

Although the Bland-Altman Limits of Agreement is the most popular method used to assess agreement, there are a few issues and limitation related to it of which medical researchers should be aware of.

Confidence Interval for Limits of Agreement

Limits of agreement is actually just an estimate of the values which apply to the whole population (4). So, whatever value of limits of agreement are obtained from a study, they only apply to that study population. If a similar study was repeated in a different study population, this second sample would give different limits of agreement. Therefore, to infer the limits of agreement to the whole population, a 95% confidence interval (CI) of the upper and lower limit of agreement should be calculated, as suggested by Bland and Altman (4). The 95% confidence intervals can be calculated by finding the appropriate point of the *t* distribution with *n* - 1 degrees of freedom and the standard deviation of the difference, SD (4):

$$\text{CI for upper limit of agreement} = \text{Mean Bias} + (1.96(\text{SD}) \pm t\sqrt{\frac{3\text{SD}^2}{\text{n}}});$$

$$\text{CI for lower limit of agreement} = \text{Mean Bias} - (1.96(\text{SD}) \pm t\sqrt{\frac{3\text{SD}^2}{\text{n}}});$$

However, this is rarely practised by researchers. Out of 178 papers reviewed earlier (1) that used the Bland-Altman method to assess agreement, only one paper considered the 95% confidence interval of limits of agreement. Bland and Altman are also aware of this problem and regret that these confidence intervals are seldom quoted (12). Theoretically, without reporting the confidence interval, their conclusion about the agreement of methods measured can only be applied to the measurement during the research, and cannot be inferred to clinical practice.

This issue has also been discussed in detail by Hamilton and Stamey (2007), who suggested that Limits of Agreement only provide a reference interval, and can be misleading if the Confidence Interval (CI) is not considered (13). They concluded that Limits of Agreement should never be used as the decisive factor in concluding agreement between two instruments (13).

Interpretation of Bland-Altman Limits of Agreement

One of the reasons why the Bland-Altman Method is so popular is its simplicity (14). Although the interpretation of limits of agreement seems to be simple and easy, medical researcher should be aware of the appropriate way of interpreting the Bland-Altman analysis. Mistakes or inappropriate interpretation of limits of agreement can occur as found in the following published article.

In 2005, a study tested the agreement of three peak flow meters (A, B and C) using three statistical methods (Pearson's Correlation Coefficient, t-test, and the Bland-Altman method) (15). For peak flow meters A and B, the limits of agreement were found to be 40 l/min to 60 l/min. The authors interpreted this as the differences between peak flow meter A and B to range from 40–60 l/min (15). They did not comment whether peak flow B would overestimate the value of peak flow A, which is the most important clinical finding desired. Furthermore, the overall conclusions on the agreement of the peak flow meters were made based on a paired t-test.

In fact Bland and Altman themselves made a mistake in the interpretation of the limits of agreement in one of their earlier publications (4), where they compared the readings between a large peak flow meter (PEFR) and mini peak flow meter. By plotting the difference (Large PEFR – mini PEFR) against the mean, the upper limit of agreement was 75.5 l/min and the lower limit of agreement was -79.7 l/min (4). Their interpretation was that the mini peak flow meter may be 80 l/min below or 76 l/min above the large peak flow meter. However, because the difference was calculated from Large PEFR – mini PEFR, the positive difference means that the mini PEFR underestimates the large PEFR, and the negative difference means that the mini PEFR overestimates the large PEFR. So, the appropriate interpretation should be that the mini PEFR may be 80 l/min above or 76 l/min below the large PEFR.

Thus, a mix of negative and positive values of limits of agreement might confuse some researchers. In addition, imagine if we apply the 95% confidence interval for the limits of agreement. This would create further confusion and make the Bland-Altman method appear to not be as straightforward as originally thought. Therefore, medical researcher should put an effort to really understand this method and interpret the result appropriately.

Proportional Bias

Hopkins (2004) demonstrated that the Bland-Altman plot indicates incorrectly that there is a systematic bias in the relationship between two measures (16). Using a fixedly generated data, Hopkins clearly showed the proportional bias produced in the Bland-Altman plot, but not in the regression (ordinary least squares method) analysis. If a slope of regression line fitted to the Bland-Altman plot differs significantly from zero, it is argued that proportional bias exists (17). Using randomly generated data, Hopkins showed that proportional bias was produced in the Bland-Altman plot, but not in the regression (ordinary least squares method) analysis, and concluded that the Bland-Altman plot should not be used to make conclusions about bias for any instrument (16). He added that bias in the Bland-Altman plots was not restricted to calibrated instruments, but could arise as an artefact of random error between measures that have not been calibrated (16). Commenting on Hopkins' article, Batterham (2004) favoured the ordinary least squares regression technique, rather than the Bland-Altman limits of agreement (18).

However, Ludbrook (2002) claimed that the presence of bias in the analysis was a result of some kind of statistical assumption (17). Ludbrook (2010) recommended that a linear regression line be fitted to the Bland–Altman plot to check for this bias (19). It was argued that, if the slope of the regression line fitted to the Bland–Altman plot is not significantly different from zero then the proportional bias is absent(19). Thus we should not be worried about any artifactual bias. However, recent study (20) showed that testing the slope of regression line of the Bland–Altman plot does not remove the artifactual bias in the prediction.

The main concern about the proportional bias is that this will result in artefactual bias in the prediction. The predicted bias will consist of artefact and real bias, which cannot be differentiated by the researcher (16). Therefore the Bland–Altman method should be used with caution and should be complemented by other methods.

Reliability Analysis

Intra-class Correlation Coefficient or ICC is the most popular method used to assess the reliability of medical instruments. There are a few concerns regarding the application of ICC in evaluating reliability:

Choosing appropriate type of ICC

There are different types of ICC, and confusion exists regarding which ICC to use (8). Muller and Buttner (2004) demonstrated that different types of ICC may result in quite different values for the same dataset, under the same sampling theory (21). So it is important to determine which type of ICC is suitable, depending on the purpose of the analysis. Weir (2005) suggested some issues that should be considered when choosing an ICC test:

(a) One- or two-way model:
- For the one-way model each subject is assumed to be assessed by different raters, and the raters are also assumed to be selected from the population. This model allows for situations where all subjects are not rated by all raters. In this model, all sources of error are lumped together. A one-way model should be considered when information on which raters rated the subject is not known (8).
- The two-way model assumes that each subject was assessed by the same raters, and requires raters to be crossed with subjects (i.e. each rater rates all subjects). The two-way model allows the error to be devised into random and fixed errors (8, 22).

(b) Random- or fixed-effect model
- In a fixed-effects model, the levels of variable are fixed or specified in advance (11). The fixed factor is considered when all levels of the factor of interest are included in the analysis. Raters are considered as fixed effects, but items/subjects are treated as random effects (no generalization beyond the sample). So, there is no attempt to generalise the result on reliability (8).
- Under a random-effects model, both factors (raters and items/subjects) are viewed as random effects (11). Random factor is considered when the analysis is to be generalised to other levels (8).

(c) Single or mean score (8):
- Single Measures ICC should be reported if only a single measure on a subject was taken.
- If two or more trials were measured on a subject, then Average Measures ICC should be reported. The Averaged Measures ICC will always be higher than the Single Measures ICC

Between-subjects variability

The ICC is influenced greatly by between-subjects variability. If the ICC is applied to data from a group of individuals with a wide range of the measured characteristics, the value of the ICC will indicate higher reliability, compared to the same analysis when applied to a group of data with a narrow range of the same characteristic (8). However, according to Weir (2005) this is an unfair criticism, because the ICC is not meant to provide an index of absolute measurement error (8). In general, the ICC is a ratio and does not quantify precision.

Single or Multiple methods?

According to both our systematic reviews published recently (1, 2), most reliability studies (86%) relied on a single statistical method to assess reliability, in contrast with agreement studies where most of the studies (65%) used a combination of statistical methods (see Table 2). A strong case for using multiple methods in assessing agreement and reliability is because each statistical method has its own strengths and weaknesses. The usage of multiple methods has the advantage of compensating for the limitations of any one single method. As long as the methods chosen are appropriate for it purposes. Luiz and Szklo (2005) suggested that more than one statistical method to assess agreement may be reported usefully, since no strategy seems to be fool proof (23). Similarly, in reliability studies, it was suggested that no single reliability estimate should be used for reliability studies, and a combination of methods was more likely to provide more information on the reliability of an instrument (9).

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 (3, 24).

Table 2: Single versus multiple methods

	AGREEMENT (N=210)	RELIABILITY (N=42)
Overall:		
Multiple methods	137 (65%)	6 (14%)
Single method	73 (35%)	36 (86%)
p<0.0001		
According to year:		
2007		
Multiple methods	n=70	n=26
Single method	43 (61%)	6 (23%)
p=0.0002	27 (39%)	20(77%)
2008		
Multiple methods	n=70	n=7
Single method	46 (66%)	0
p=0.0009*	24 (34%)	7 (100%)
2009		
Multiple methods	n=70	n=9
Single method	48 (69%)	0
p<0.0001*	22 (31%)	9 (100%)
(*Fisher's exact)		

Application of Inappropriate Statistical Methods

The proportion of studies with inappropriate statistical methods, found in both earlier systematic reviews, will reflect the proportion of medical instruments that have been validated using inappropriate methods in current clinical practice. As found in the earlier systematic reviews, eight (19%) of reliability studies (2) and twenty (10%) of agreement studies (1) 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. Inappropriate application of statistical methods in method comparison studies also reflects the lack of knowledge in this area among medical researchers. This is alarming and it is important for clinicians or medical researchers to be aware of this.

The Importance of Appropriate Statistical Method in Medicine

Patient Care

In clinical situations, the duty of a doctor is to provide the best care or treatment for their patients. Most of the time, doctors have to decide what is the best available option for their patients. In some cases, this may involve life and death decisions; for example, deciding to thrombolyse patient with myocardial infarction in an Accident and Emergency department. Doctors have to assess a patient thoroughly and, assisted by information from some medical equipment such as electrocardiogram (ECG) and blood pressure machines, before the decision to thrombolyse the patient can be made.

In 2009, a study to assess the accuracy and precision of five currently available blood glucose meters in South Africa was conducted (25). The study compared five glucometers that utilise different analytical techniques (reflectometry or amperometry), and all the glucometers were calibrated (25). The authors found that although all the devices showed satisfactory precision, there was substantial discordance when their results were compared to a laboratory reference (25). Only three out of the five glucometers fulfilled the criteria suggested by the International Standardisation Organisation. All meters demonstrated significant deviation from the American Diabetes Association guidelines, as more than 60% of the measurements exceeded the recommended percentage of deviation (25).

It is well-known that both type 1 and type 2 diabetes show a direct relationship between the degree of glucose control and the risk of systemic complications (26). Many clinical organisations such as the American Diabetes Association promote the self-monitoring of blood glucose, because it allows diabetic patients to achieve and maintain specific glycaemic goals (26). The variability observed with the accuracy of glucometers can impact patient care in different settings, some of which include the diabetic patient on insulin in a home care or a clinical setting. Most of the time, glucose determinations and insulin adjustments are made according to glucometer readings. Inaccuracies can lead to misclassification of hypoglycaemic or hyperglycaemic episodes. It is, therefore, imperative that glucometer values are accurate and precise. Otherwise, a failure in this regard may lead to critical medical errors.

The variation amongst these glucometers found in the study (25) were probably a result of the improper evaluation of the glucometer in the validation study. This suggests that there is a necessity for proper evaluation, and it is important to be sure that appropriate statistical methods for the validation of the instrument has been used in any research or clinical situation.

Evidence-Based Medicine

The practice of Evidence-Based Medicine (EBM) has been promoted to ensure the best quality of care is given to the patient. One example is in the treatment of hypertension. According to the most recent National Institute of Clinical Excellence (NICE) *Clinical Guidelines on Hypertension* (27), antihypertensive drug treatment should be offered to people of any age with stage 2 hypertension. Stage 2 hypertension is defined as a patient with blood pressure of 160/100 mmHg or higher, and whose subsequent ambulatory blood pressure monitoring (ABPM), daytime average or home blood pressure monitoring (HBPM) average blood pressure, is 150/95 mmHg or higher (27).

The recommendation from the guidelines was derived from the views of experts, patients, carers and industry, and includes the best available evidence (from research) (27). Without doubt, researchers must have used some instrument to measure blood pressure in the process of producing evidence. However, which instrument was used in their studies: the automatic blood pressure machine or manual sphygmomanometer? Were these machines validated, and if the machines were validated, which statistical method was used? If the instruments used were not validated, or were validated using inappropriate statistical methods, we can actually question the quality of the evidence from such studies. A lack of precision and validity of an instrument in research may result in invalid evidence. The main goal of research, especially in epidemiological studies, is about applying the evidence to the population for practice. Appropriate statistical analysis is actually the “root” of Evidence-Based Medicine.

Conclusion

Although there is no single perfect method, researchers should be aware of the inappropriate methods that they

should avoid when analysing data in method comparison studies (i.e. to assess agreement and reliability). This is important because inappropriate analysis will lead to invalid conclusions and thus validated instrument might not be accurate or reliable. This will result in inaccuracy of prediction or diagnosis, and inappropriate management or treatment. Consequently this will affect the quality of care given to a patient and, most importantly, inappropriate treatment might put the patient's life at risk. Poor quality of care will also jeopardise the doctor-patient relationship. Inaccurate measurements cannot be used as an excuse for making any mistake in the management of patients. Therefore it is vital to ensure the validity of an instrument, and appropriate statistical methods should be applied in a validation study. In other words, appropriate statistical methods should be used when testing agreement and reliability of an instrument.

References

1. Zaki R, Bulgiba A, Ismail R*, et al.* 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.
2. Zaki R, Bulgiba A, Nordin N, Ismail NA. A Systematic Review of Statistical Methods Used to Test for Reliability of Medical Instruments Measuring Continuous Variables. *Iranian Journal of Basic Medical Sciences* 2013; 16:803-807.
3. Altman DG, Bland JM. Measurement in Medicine: the analysis of method comparison studies. *The Statistician* 1983; 32:307-317.
4. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Biochimica Clinica* 1987; 11:399-404.
5. Wikipedia. The Free Encyclopedia. 2011 [cited 2011 12 June]. Available from: en.wikipedia.org/wiki/Ronald_A._Fisher.
6. Fisher RA. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd; 1925.
7. Fisher J. *R. A. Fisher: The Life of a Scientist*. New York: Wiley; 1978.
8. Weir JP. Quantifying test-retest reliability using the Intraclass Correlation Coefficient and the SEM. *Journal of Strength and Conditioning Research* 2005; 19(1):231-240.
9. Bruton A, Conway JH, Holgate ST. Reliability: What is it, and how is it measured? *Physiotherapy* 2000; 86(2):94-99.
10. Wikipedia. The Free Encyclopedia. 2011 [cited 2011 12 June]. Available from: http://en.wikipedia.org/wiki/Intraclass_correlation.
11. Rosner B. *Fundementals of Biostatistics*. 6th ed. Duxbury: Thomson Brooks/Cole; 2006.
12. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. *Ultrasound in Obstetrics & Gynecology* 2003; 22 (1):85-93.
13. Hamilton C, Stamey J. Using Bland Altman to assess agreement between two medical devices – don't forget the confidence intervals! *Journal of Clinical Monitoring and Computing* 2007; 21:331-333.
14. Cohen MD, Jennings SG. Agreement and reproducibility of subjective methods of measuring faculty time distribution. *Academic Radiology* 2002; 9(10):1201-1208.
15. Nazir Z, Razaq S, Mir S, Anwar M, Al Mawlawi G, Sajad M, *et al.* Revisiting the accuracy of peak flow meters: a double-blind study using formal methods of agreement. *Respiratory Medicine* 2005; 99:592-595.
16. Hopkins WG. Bias in Bland-Altman but not regression validity analyses. *Sportscience* 2004; 8:42-46.
17. Ludbrook J. Statistical techniques for comparing measurers and methods of measurement: A critical review. *Clinical and Experimental Pharmacology and Physiology* 2002; 29(7):527-536.
18. Batterham AM. Commentary on bias in Bland-Altman but not regression validity analyses. *Sportscience* 2004; 8:47-49.
19. Ludbrook J. Confidence in Altman-Bland plots: a critical review of the method of differences. *Clinical and Experimental Pharmacology Physiology* 2010; 37:143-149.
20. Zaki R, Bulgiba A, Ismail NA. Testing the agreement of medical instruments: Overestimation of bias in the Bland–Altman analysis. *Preventive Medicine* 2013; 7:S80-S82.
21. Muller R, Buttner P. A critical discussion of intraclass correlation coefficients. *Stat Med* 1994 Dec 15-30; 13(23-24):2465-2476.
22. Shoukri MM, Pause CA. *Statistical Methods for Health Sciences*. 2nd ed. Boca Raton, Florida: CRC Press; 1999.
23. Luiz RR, 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.
24. Daly LE, Bourke GJ. *Interpretation and Use of Medical Statistics*. 5th ed. Oxford: Blackwell Science Ltd.; 2000.
25. Essack Y, Hoffman M, Rensburg M, Van Wyk J, Meyer CS, Erasmus R. A comparison of five glucometers in South Africa. *Journal of Endocrinology, Metabolism and Diabetes of South Africa* 2009; 14(2):102-105.
26. Cohen M, Boyle E, Delaney C, Shaw J. A comparison of blood glucose meters in Australia Diabetes Research and Clinical Practice. *Diabetes research and clinical practice* 2006; 71:113-118.
27. NICE. National Institute for Health and Clinical Excellence UK clinical guideline 127 London: National Institute for Health and Clinical Excellence UK; 2011 [cited 2012 26 April]. Available from: www.nice.org.uk/guidance/CG127.