REVIEW ARTICLE ASIAN JOURNAL OF MEDICAL SCIENCES

Method comparison: Statistical measurement correlation or agreement-most appropriate tool?

Chandra Bhushan Tripathi[1], Puja Kumari Jha[2], Rachna Agarwal[3]

[1]Professor, Department of Biostatistics, [3]Professor, Department of Neurochemistry, Institute of Human Behavior and Allied Sciences, [2]Associate Professor, Department of Biochemistry, University College of Medical Sciences and GTB Hospital, New Delhi, India

Submission: 31-08-2023 Revision: 29-11-2023 Publication: 01-01-2024

ABSTRACT

In laboratory, a reliable method is very important to report precise and accurate results which are very important for clinical diagnosis and management decisions. In case the existing method needs to be replaced by new method due to poor results or financial issues, method comparison of a new measurement system/method with an established/referral method is required to see whether they agree sufficiently for the new to replace the old method or use the two interchangeably. Most commonly Pearson's product-moment correlation is used to ascertain whether two methods agree sufficiently so that the old method can be replaced by a new method. However, it is the most inappropriate method with number of limitations. In this paper, other methods are discussed with their merits and demerits with examples.

Key words: Method comparison; Pearson's product-moment correlation; Bland–Altman method; Intra class correlation coefficient; Clinical tolerance limit

Access this article online

Website: http://nepjol.info/index.php/AJMS

DOI: 10.3126/ajms.v15i1.58213

E-ISSN: 2091-0576

P-ISSN: 2467-9100

INTRODUCTION

In laboratory, agreement of methods for measurement is required when one method is replaced by another method, or a new alternative method is introduced in the system or to assess the alignment between two instruments.[1] However, a new method is evaluated by comparing it with the existing/established and relatively accurate method, not necessarily reference method, before it can be recommended as a replacement of the existing method. It is different from the calibration process, in which known quantities are measured by a new method and the result is compared with the true value or with measurements made by a highly accurate method or reference method.[2]

The logical statistical technique for comparison between existing and new method, will be to measure true value sample. If the reference method has not been used to measure the value of reference material/control with metrological traceability, there is no way to measure true value. However, it is not possible that the new method will have exactly the same reading as the existing method. On search for literature and the most common measures in the laboratories, two terms were revealed to decide that the methods are interchangeable. One, the coefficient of correlation to measure the strength of linear relationship between the values obtained from existing and new methods. Two, degree of agreement to measure the agreement between the values obtained from existing and new methods.

Address for Correspondence:
Dr. Rachna Agarwal, Professor, Department of Neurochemistry, Institute of Human Behavior and Allied Sciences, New Delhi, India.
Mobile: +91-8527350102. **E-mail:** rachna1000@gmail.com

Correlation coefficient

Product moment correlation (r) appears to be the most appropriate statistical tool, which measures the strength of the linear relationship between two values obtained by the existing and new method of testing. Many studies use it as an indicator of agreement between two measurement methods. Although this method is useful for a fixed measurement range, it is not sensitive to change in the scale of measurement.[3] In other words, altering the scale of measurement does not affect the correlation but it will certainly affect the agreement between the measurement values obtained from two methods. As the correlation coefficient measures the association between two values, not necessarily an agreement between them, it is clear that correlation and agreement are not interchangeable while comparing the existing method with new method before replacement. It is worth to mention here that a high correlation does not mean that both methods are in good agreement, whereas good agreement always shows good correlation. Actually Pearson's correlation coefficient measures only the strength of linear relationship rather than agreement and it explains that methods may have high correlation but may not have high agreement because it may occur of one method gives consistently higher value than another method.[4]

Interpretation

Product Moment Correlation, r=0.98. This indicates that Method A and Method B have a strong correlation. However, bias (difference of mean- Mean) is –15 and standard deviation (SD) is 26.50.

Hence, Mean-2SD=–68 and Mean+2SD=38

Thus new method, method B may measure 68 mg/dL below or 38 mg/dL above the old method, method A. This difference is highly significant for blood sugar estimation as it may affect the patient management. Hence, method B is not acceptable.

It shows that a high correlation need not necessarily mean that both methods agree with each other.

There can be a number of limitations in calculating correlation for method comparison:[4]

1. *r* measures the strength of a relation between the values, not the agreement between them (Table 1)
2. A change in scale of measurement does not affect the correlation, but it affects the agreement. For example, if the temperature is measured in centigrade and the Fahrenheit scale is plotted, a correlation of r=1 does not mean that both measurements agree with each other (R1)
3. Correlation depends on the range of the true value quantity in the sample. If this is wide, the correlation will be higher than if it is narrow
4. Data showing poor agreement can produce a high correlation (Table 1).

Example: The example has been demonstrated in Table 2, Figures 1 and 2.

Bland and Altman method

There are multiple limitations in using a correlation coefficient. In view of this, it is grossly incorrect method to use for evaluation of agreement between two methods;

Table 1: Comparison of methods A and B for blood glucose estimation by Bland–Altman analysis

S. No.	Method A (mg/dL)	Method B (mg/dL)	Difference (A-B)	Mean (A+B)/2
1.	100	109	–9	104.5
2.	105	110	–5	107.5
3.	110	113	–3	111.5
4.	120	124	–4	122
5.	150	159	–9	154.5
6.	140	154	–14	147
7.	140	150	–10	145
8.	160	168	–8	164
9.	70	72	–2	71
10.	80	62	18	71
11.	90	122	–32	106
12.	100	80	20	90
13.	150	181	–31	165.5
14.	200	259	–59	229.5
15.	250	275	–25	262.5
16.	130	148	–18	139
17.	350	320	30	335
18.	400	434	–34	417
19.	450	479	–29	464.5
20.	500	587	–87	543.5
Mean	189.75	205.30	–15.55	197.53
SD	129.98	145.34	26.50	137.24

SD: Standard deviation

Table 2: Comparison of methods A and B for blood glucose estimation by Pearson's correlation

S. No.	Range: 70–120 mg/dL		Range: 70–500 mg/dL	
	Method A	Method B	Method A	Method B
1.	100	109	70	72
2.	105	110	80	62
3.	110	113	90	122
4.	120	124	100	80
5.	80	62	150	181
6.	90	122	200	259
7.	100	80	250	275
8.	130	148	300	380
9.	70	72	350	320
10.	80	62	400	479
11.	90	122	500	587

there was a need of an alternative way of assessing the degree of agreement between any two methods of measurement. Bland and Altman have proposed a method of comparison, based on the discrepancy between the measurements of two methods/equipment, when applied to the same person/sample. However, before comparing any two methods of measurement, Precision i.e. the repeatability of new method for measuring the analyte must be checked. If the new method has poor precision, the possibility of agreement between old and new methods will be poor.[2]

Bland–Altman analysis involves the following steps:[4]

1. Define the acceptable difference (bias) between the values of analyte obtained from the two methods under comparison by which the management of patient will not be changed. For example, in the measurement of Serum Glucose difference of 10 mg/dL is not going to make any difference as far as clinical management is concerned. However, even a difference of 0.5 mg/dL in serum creatinine estimation is not acceptable. Hence mean acceptable bias may be different for different analytes
2. Examine the data by simply plotting the results of one method against those of other and calculating the correlation coefficient (r). r>0.95 means there is strong association between two measurements. However, this cannot explain the agreement between two methods under comparison.
3. Calculate the bias (difference) between the values of analyte obtained from the two methods performed on the same set of samples, which is relative bias. The mean and SD of the bias (difference) are calculated from the data obtained, which is the estimate of error
4. Calculate the average of two measurements obtained from the same samples as neither of the two methods is a reference method. Since true value is not known, it may give the estimate of true value. If the first method is a reference method, can use values measured by that method as true value instead of the mean of the two measurements[1]
5. A scatter diagram is plotted with the difference between the methods on Y axis and the average of the two measurements on X axis. On graph, the mean bias (difference) is represented by a solid line and mean+2SD and mean–2SD by interrupted lines (Figure 3).

Thus, New method, Method B may measure 86 mg/dL below or 38 mg/dL above the old method Method A. This difference is highly significant for Blood Sugar estimation as it may affect the patient management. Hence, Method B is not acceptable.

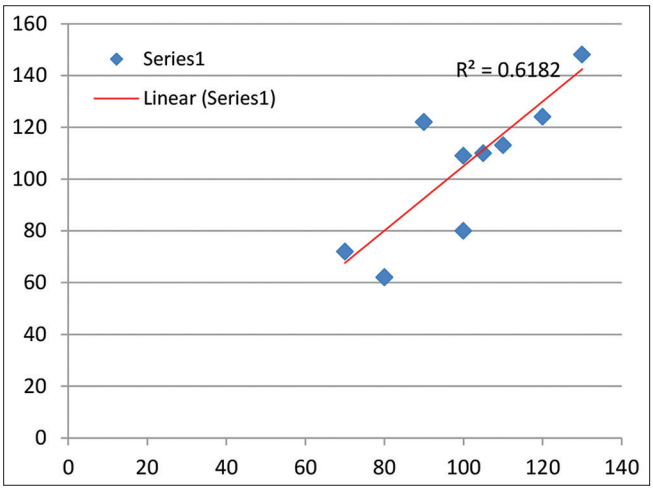

Figure 1: Product moment Correlation, r=0.618 was between Method A and Method B, in samples whose value ranged between 70 mg/dL and 120 mg/dL

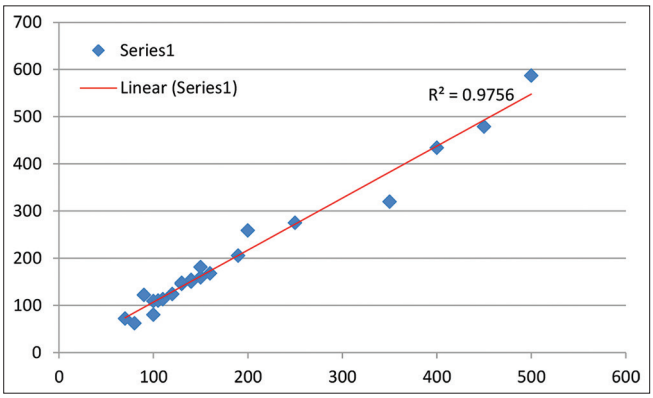

Figure 2: Product Moment Correlation, r=0.973 between Method A and Method B, in samples whose value ranged from 70 to 500 mg/dL. This indicates that Method A and B are highly correlated in wider range as compared to lower range

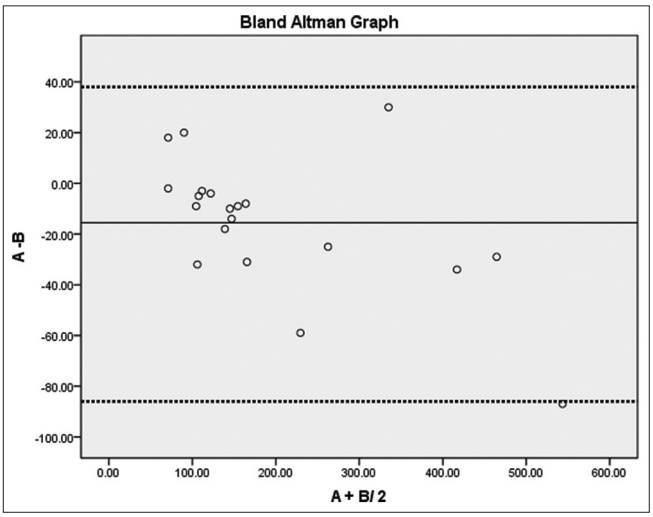

Figure 3: Scatter plot for bland and altman analysis of data of Table 3: Bias (difference-mean) is –15 and standard deviation is 26.50, Mean-2SD= –86 and Mean+2SD=38

Acceptance criteria

- If the limits of agreement (LOA) are within the acceptable value defined at the beginning of the analysis, then it can be accepted that both methods agree and therefore interchangeable
- The values (mean±2SD) are known as LOA or agreement interval,[1] within which 95% of the differences of the second method, compared to the first method must lie. When the differences in the value obtained from two methods performed on the same samples, are clustered around zero, with little variability, then this indicates that new method can replace the old method.

Examples are shown in Tables 3 and 4 with Figures 3 and 4.

Thus new method, Method C may measure 13 mg/dL below or 12 mg/dL above the old method A. This difference is acceptable as this may not affect patient management. Hence, method C is acceptable.

Advantage of Bland Altman analysis

- Graphical approach will check whether there is any trend. An increase in difference for higher values leads to poor agreement in two methods for higher value samples.
- The Bland–Altman plot can also be useful for studying the trend and interpretation.

Limitation of Bland Altman analysis

- It quantifies the bias and LOA, within which 95% of values of the difference between two methods should lie. However, it cannot say, if agreement is sufficient to use a method. Bias will be considered significant if line of equality is not within the confidence interval of the mean difference. Hence, LOA expected to be defined a priori, based on biologically and analytically relevant criteria.[1]
- The graphical representation of data points obtained from both the methods tested on the same samples will show proportional bias due to variability of the difference in values obtained by two methods.[1]

Table 3: Comparison of methods A and B for blood glucose estimation by Bland–Altman analysis

S. No.	Method A	Method B	Difference (A-B)	Mean (A+B)/2
1.	100	109	–9	104.5
2.	105	110	–5	107.5
3.	110	113	–3	111.5
4.	120	124	–4	122
5.	150	159	–9	154.5
6.	140	154	–14	147
7.	140	150	–10	145
8.	160	168	–8	164
9.	70	72	–2	71
10.	80	62	18	71
11.	90	122	–32	106
12.	100	80	20	90
13.	150	181	–31	165.5
14.	200	259	–59	229.5
15.	250	275	–25	262.5
16.	130	148	–18	139
17.	350	320	30	335
18.	400	434	–34	417
19.	450	479	–29	464.5
20.	500	587	–87	543.5
Mean	189.75	205.30	–15.55	197.53
SD	129.98	145.34	26.50	137.24

SD: Standard deviation

Detection of proportional bias

Scatter gram is constructed of differences of values obtained from two methods under comparison on averages of these values, followed by calculating best the line of best fit for linear regression on this scattergram. In case the slope of regression is different from zero, then these two methods have proportional bias.[5] This may lead to overestimation of bias.

Modified Bland–Altman method

In classical Bland–Altman method, the difference between values calculated by two methods is plotted on Y axis. In the modified Bland–Altman method, differences are expressed as percentage, i.e. A-B/mean ×100 on Y axis. Such modification is useful in case there is an increase in variability of the differences as the concentration of the measurand increases due to the constant coefficient of variation across a range of concentration, whereas for constant differences across concentrations of measurand, the unit difference between the values provides a better representation of the difference between the two measurements obtained from old and new methods. Hence, in former case, modified BA plot for evaluation is preferable, whereas, in latter case classical Bland Altman method is better. Hence, both the plots may be considered for better evaluation of two methods.[1]

Intra-class correlation (ICC) coefficient

Method comparison is done before replacing the old method by new method, there is a need to have a reliable measure that should reflect both correlation and agreement between measurements done by old and new methods.[6] Pearson's correlation coefficient only measures the strength of linear relationship, whereas Bland–Altman plot is a graphical method of measuring the agreement between two methods with few limitations. Hence both the methods are not ideal measures of reliability.

Any new method replacing an old method should be assessed for reliability before using it in the system. Reliability is defined as the extent to which measurements can be replicated, i.e., extent of correspondence between

Table 4: Comparison of method A and C for blood glucose estimation by Bland Altman analysis

S. No.	Method A	Method C	Difference (A-C)	Mean (A+C)/2	% A-C/Mean x 100
1.	109	100	9	104.5	8.6
2.	110	105	5	107.5	4.6
3.	113	110	3	111.5	2.7
4.	120	124	4	122	3.3
5.	159	150	9	154.5	5.8
6.	144	140	4	142	2.8
7.	140	150	–10	145	–6.9
8.	160	162	–2	161	–1.2
9.	70	72	–2	71	–2.8
10.	82	80	–2	81	–2.5
11.	90	92	–2	91	–2.2
12.	100	108	–8	104	–7.7
13.	150	151	–1	150.5	–0.7
14.	200	209	–9	204.5	–4.4
15.	250	255	–5	252.5	–2.0
16.	138	130	8	134	6.1
17.	354	350	4	352	1.1
18.	404	400	4	402	1.0
19.	450	459	–9	454.5	–2.0
20.	500	507	–7	503.5	–1.3
Mean	192.15	192.70	–0.55	192.43	3.49
SD	129.52	131.95	6.21	130.23	4.29

SD: Standard deviation

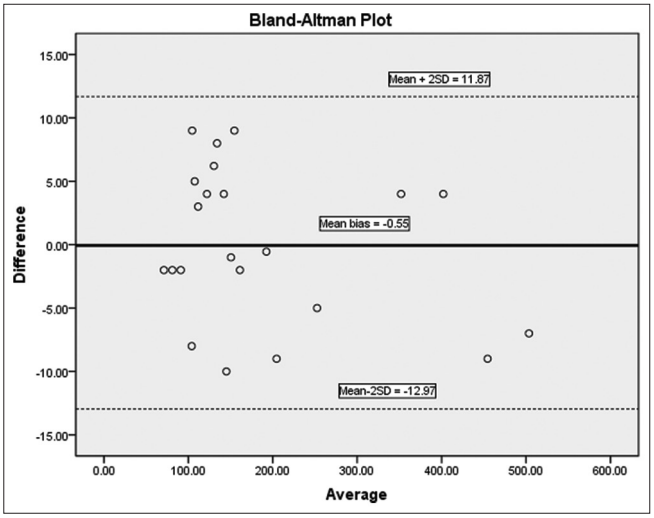

Figure 4: Scatter plot for bland and altman analysis of data of Table 4: Bias (difference- Mean is –0.55 and standard deviation is 6.21, Mean-2SD=–12.97 and mean+2SD=11.87

two methods for measuring the same variables. The simple descriptive method for measuring the agreement is to calculate the percent of subjects showing exact agreement, i.e., both methods give identical measurement or percent of subjects showing agreement within selected number of units, for example, ±4. The percentage of subjects showing agreement within selected level is informative but it ignores that out of total agreement, certain amount of agreement can be expected by chance alone Therefore, a statistical index, i.e., the ICC is needed to eliminate the expected chance agreement. It includes both degrees of correlation and agreement between the measurements.[7]

Mathematically reliability is calculated by dividing true variance by true variance plus error variance where variance is square root of standard deviation.

Reliability index=true variance/true variance+error variance; variance=SD^2

ICC coefficient is a reliability index.

Interpretation

Reliability index (r_I)

- If the r_I is '0', it means that the extent of agreement between two methods are no better than the expected chance alone.
- If the r_I is positive, it means that the extent of agreement is more than the expected chance alone.
- If the r_I is negative, it means that true agreement is less than the expected chance agreement.
- Ideally, if r_I is 1, the new method can replace the old method/reference method. However, practically it is difficult to have this value. Hence Burdock et al., advised that if r_I is 0.75 then only agreement can be considered.[7] Lee et al., also suggested that good agreement can only be achieved if lower limit of 95% CI of ICC (r_I) is more than or equal to 0.75.[8] Hence statistically significant ICC does not always indicate the high agreement if ICC is small as a small ICC may be statistically significant due to large sample size.
- As a rule of thumb at least 30 samples should be taken

Examples are shown in Tables 5 and 6.

Table 5: Interpretation of reliability index

Reliability index	Interpretation
<0.5	Poor reliability
0.5–0.75	Moderate reliability
0.75–0.9	Good reliability
>0.9	Excellent reliability

Table 6: Comparison of method of HPLC (Method A) and turbidometery (Method B) for estimation of hemoglobin estimation by intra-class correlation

S. No.	Method A (true value)	Method B	Error (A-B)
1.	5.7	4.2	1.5
2.	6	5.5	0.5
3.	6.9	6.4	0.5
4.	6.1	5.9	0.2
5.	4.9	3.8	1.1
6.	6.7	6.8	–0.1
7.	5.2	3.6	0.6
8.	6.3	5.9	0.4
9.	9.6	10.2	–0.6
10.	5.7	4.7	1.0
11.	8.4	7.8	1.0
12.	6.6	5.7	0.9
13.	4.8	5.8	–1.0
14.	5	4	1.0
15.	5	3.6	1.4
16.	4.5	3.3	1.2
17.	5.3	4.3	1.0
18.	8	8.7	–0.7
19.	5.1	3.8	1.3
20.	4.4	4.2	0.2
Mean	6.01	5.41	0.57
SD	1.38	1.86	0.72

SD: Standard deviation

ICC can be used for evaluating interrater, test–retest, and intr-arater reliability. An example is shown in Tables 5 and 6.

$$\text{Reliability Index} = \frac{1.88}{1.88 + 0.52} = 0.78$$

Method A and B have good reliability.

Clinical tolerance limit

In this approach, clinical tolerance limits are predefined in such a manner that a difference within these limits will not have any clinical significance and are called as percentage of agreement. There will be low concentration (C_L) or high concentration (C_U), which are called clinical tolerance limit. These agreement limits are based on the expected measurement error or can be based on clinical implications for managing a patient. It signifies that any difference below C_L or more than C_U will have clinical consequences. These values would be around zero but may or may not be symmetric.[9]

Advantage

- Direct use of clinical tolerance limits ensures the exact extent of agreement and this would assess clinical agreement in the true sense as it is based on clinical tolerance limits
- This method uses all the individual differences and not their mean and SD
- In this method, percentage can be used depending on the clinical context as some clinicians will not accept a difference of more than 5% beyond the clinical tolerance for a particular analyte but accept difference of around 10% for another analyte.

CONCLUSION

Pearson's correlation coefficient measures the association between measurements, but not necessarily an agreement between them, which can be obtained from Bland–Altman analysis. However, Bland–Altman analysis cannot say if agreement is sufficient to be acceptable and change in method will relate with the clinical tolerance limits acceptable in, whereas ICC reflects both degree of correlation and agreement between measurements by calculating the reliability index. However, there are different types of ICC and may be more tedious as it requires mathematical calculations. Hence, the use of clinical tolerance limits is a preferable method for assessing agreement between measurements on the same subjects as it is a nonparametric robust method which is more flexible than others.

ACKNOWLEDGMENT

Not applicable.

REFERENCES

1. Mantha S. Statistical analysis for comparison of two methods of measurement. In: Visweswara Rao K, editors. Biostatistics: A Manual of Statistical Methods for Use in Health, Nutrition and Anthropology. 2nd ed. New Delhi: Jaypee Brothers, 2007. p. 537-545.
   https://doi.org/10.2307/2533701
2. Bland JM and Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307-310.
   https://doi.org/10.1016/s0140-6736(86)90837-8
3. Jeyaseelan L and Rao PS. Statistical measures of clinical agreement. Natl Med J India. 1992;5(6):286-290.
4. Indrayan A and Chawla R. Clinical agreement in quantitative measurements. Natl Med J India. 1994;7(5):229-234.
   https://doi.org/10.1007/978-3-642-37131-8_2
5. Bunce C. Correlation, agreement, and bland-altman analysis: Statistical analysis of method comparison studies. Am J Ophthalmol. 2009;148(1):4-6.

https://doi.org/10.1016/j.ajo.2008.09.032

6. Giavarina D. Understanding Bland Altman analysis. Biochem Med (Zagreb). 2015;25(2):141-151.
   https://doi.org/10.11613/BM.2015.015
7. Ludbrook J. Confidence in Altman-Bland plots: A critical review of the method of differences. Clin Exp Pharmacol Physiol. 2010;37(2):143-149.
   https://doi.org/10.1111/j.1440-1681.2009.05288
8. Koo TK and Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155-163.
   https://doi.org/10.1016/j.jcm.2016.02.012
9. Indrayan A. A simple and robust alternative to Bland-Altman method of assessing clinical agreement. Pharmacol Toxicol eJ. 2022;Available at SSRN 4189799.
   https://doi.org/10.2139/ssrn.4189799

Authors' Contributions:
CBT- Concept of review; **PKJ-** Drafting of manuscript, submission of manuscript; **RA-** Concept and design of study, implementation, manuscript drafting.

Work attributed to:
Institute of Human Behavior and Allied Sciences, Department of Neurochemistry, Delhi.

Orcid ID:
Chandra Bhushan Tripathi - https://orcid.org/0000-0003-1092-0620
Puja Kumari Jha - https://orcid.org/0000-0002-4662-9897
Rachna Agarwal - https://orcid.org/0000-0003-2604-9809

Source of Support: Nil, **Conflicts of Interest:** None declared.