 doi:10.20944/preprints202108.0343.v1

Direct use of clinical tolerance limits for assessing agreement: A robust nonparametric approach

Abhaya Indrayan[1]
Biostatistics Consultant
Max Healthcare Institute
Saket, New Delhi 11 017
India
[1] Corresponding author: Dr. A. Indrayan, a.indrayan@gmail.com, +919810315030

Abstract

Clinical agreement between two quantitative measurements on a group of subjects is generally assessed with the help of the Bland-Altman (B-A) limits. The interpretation regarding agreement is based on whether B-A limits are within the pre-specified clinical tolerance. Thus, clinical tolerance limits are necessary for this method. We argue in this communication that such limits of clinical tolerance can be directly used for assessing agreement and plead that this nonparametric approach is simple and robust to the distribution pattern and outliers. Such direct use of clinical tolerance limits has more flexibility, and it is more effective in assessing the extent of agreement.

Keywords

Agreement analysis; Bland-Altman method; Clinical tolerance limits; Limits of agreement; Nonparametric approach; Robust method

Running title: Clinical tolerance limits for agreement

Conflict of interest: None

Funding: None

 doi:10.20944/preprints202108.0343.v1

Direct use of clinical tolerance limits for assessing agreement: A robust nonparametric approach

1 Background

Some studies consider the question "Are two measurements of a characteristic of a subject by two methods, two sites, or by two observers sufficiently agree with one another?". The objective is to find whether one can be replaced with the other without much loss of information. When the measurements are quantitative, such as hemoglobin level and creatinine level, the method of choice for assessing this agreement is the one developed by Bland and Altman (1). The method was extremely successful in making us aware that the agreement between individual values x and y cannot be inferred by equality of means, and the correlation coefficient is even worse because it is perfect 1 between x and $y = ax + b$, i.e., when all the values obtained by one method are a linear combination of the other and there is no agreement. It was also separately shown that the regression $y = x$, with intercept = 0 and regression coefficient = 1, is also not appropriate for this purpose because this too is based on means (2, p. 638)

The Bland-Altman (B-A) method requires the calculation of the limits ($\bar{d} - 2s_d$, $\bar{d} + 2s_d$), where $\bar{d}$ is the mean and s_d is the standard deviation (SD) of the individual differences $d = x - y$. These limits are popularly known as Bland-Altman limits of agreement, although they are better understood as the limits of disagreement since they are based on the differences.

Under the Gaussian assumption, which is likely to hold because x and y are measuring the same quantity and the difference is likely to be just the measurement error, nearly 95 percent of the differences are likely to be within the B-A limits. An adequate agreement is inferred when these limits are narrow in the sense that the difference within these limits "would not affect decisions on patient management" (1). Let us call such limits of indifference as clinical tolerance limits. The authors stated, "How far apart measurements can be without causing difficulties will be a question of judgment" and suggested, "Ideally, it (*the clinical tolerance limits*) should be defined in advance to help in the interpretation of the methods comparison".

The crucial aspect of the B-A limits is that their interpretation regarding agreement or no agreement entirely depends on the pre-specified limits of clinical tolerance. We argue in this communication that such limits of clinical tolerance can be directly used for assessing the extent of the quantitative agreement without calculating the B-A limits. We also discuss the merits of this direct method and illustrate the method and its merits with the help of an example.

2 Issues with the Bland-Altman Method

The B-A method has tremendous merits. Besides making us aware of the distinction between the individual agreement and the group agreement, the other significant contribution of B-A method is the plot of difference against the average of the two values, known as B-A plot (3), which gives a nice scatter. The plot of y vs. x is not so informative as most values tend to cluster along the $y = x$ line.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 August 2021 doi:10.20944/preprints202108.0343.v1

Perhaps no method has universal applicability, and the B-A method also has its share of problems. For example, see a critique of the B-A plot for age comparison (4) and a discussion on the overestimation of bias in the B-A analysis (5). There are other problems too. First is the requirement that the differences *d* follow a Gaussian distribution – if not with original values, at least with the transformed values such as after log-transformation where helpful. Whereas this is likely to happen in a large majority of cases, it may not be so in isolated cases. The second is using 95% limits, which can be criticized for being arbitrary just as the 5% level of significance is being criticized (6). Although the confidence level 95% can be varied but almost all users have stuck to this level as though this is fixed. Perhaps any other confidence level would be as arbitrary. The third is that when one measurement has a constant bias (*a*) against the other, the limits of agreement would be (*a* to *a*), which is just a single point. Fourth is that the method requires the calculation of the 95% confidence interval (CI) of the lower limit ($\bar{d} - 2s_d$) and the upper limit ($\bar{d} + 2s_d$) based on their respective standard errors – using Student *t*-distribution, which assumes Gaussian distribution of the limits. All these are minor problems and can be possibly overlooked. But the next problem is not trivial: the method depends on the mean and variance of *d*, both of which can be severely affected by even a single genuine outlier that cannot be excluded or when several values are equal. Both these are a distinct possibility in an agreement setup. The limits can be wide depending on the variance of the differences even if most differences are small. The most severe problem with this method, however, is the complete dependence of its interpretation on clinical tolerance limits as elaborated next.

The interpretation of $\bar{d} \pm 2s_d$ for assessing the agreement crucially depends on whether these limits are within the range of clinical tolerance. If the limits are within clinical tolerance, the agreement is considered to exist, otherwise not. Thus, this gives a binary result. Bland and Altman (1) give an example of PEFR (peak expiratory flow rate) measured by two methods and obtained the 'limits of agreement' from –79.7 l/min to +75.5 l/min which, they say, would be unacceptable for clinical purposes. Similarly, in their second example on oxygen saturation measured by two methods, they obtained (– 2.0 to 2.8) as the limits of agreement and called them 'small enough' in the sense of clinically unimportant and concluded that the agreement exists. Although they advised setting up the clinical tolerance limits in advance to help in the interpretation of the methods comparison, a conclusion regarding agreement or the lack of it was reached in both of their examples without pre-specifying the clinical tolerance limits. Giavarina (7) remarked that the B-A method defines the interval of agreement but "does not say whether those limits are acceptable or not", and that "Acceptable limits must be defined a priori based on clinical necessity, biological considerations, or other goals". An Editorial in the British Journal of Anaesthesia (8) also mentioned in the context of B-A limits that "The question of how small is small depends on the clinical context". Thus, the B-A limits of agreement are relevant for assessing agreement only when the clinical tolerance limits are predefined.

3 Direct Use of the Clinical Tolerance Limits: A Simple, Nonparametric, Robust, and More Appealing Alternative for Assessing Agreement

We propose direct use of prespecified clinical tolerance limits to find the percentage of differences within these limits and call this percentage agreement. Consider a pair of medical measurements (*x*, *y*) on a random sample of *n* subjects. The natural parameter of interest is the extent of agreement between the two measurements. Because of random fluctuations and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 August 2021 doi:10.20944/preprints202108.0343.v1

possibly systematic differences, some difference between the observed values of *x* and *y* will almost invariably occur. Suppose the clinicians decide that this difference should not be less than C_L or more than C_U for it to be acceptable as of no clinical consequence. For example, in the case of aspartate aminotransferase (AST), if these limits are set at ±2 U/L, a difference within these limits will be considered as having no clinical significance. (C_L, C_U) are the clinical tolerance limits and they would be around zero but may or may not be symmetric.

Define the extent of agreement $\pi = P(C_L < d < C_U)$. The estimate of π is the binomial proportion of the observed differences falling between (C_L, C_U). If somebody wants to be more confident, the 95% lower confidence bound for π can be obtained by one of the several methods but the Wilson score method can be recommended, which is implementable and generally considered to perform better (9). This will give the limit below which the proportion agreement is extremely unlikely.

This method measures the extent of agreement instead of a binary yes or no. Although dichotomization has its risks (10), for those who prefer binary result as agreement exists or not, we suggest a cut-off a little later. However, many researchers these days would like to measure the exact extent of agreement instead of binary yes or no and interpret it in their context. This direct method is simpler, nonparametric, and immediately tells the percentage agreement. The information regarding the percentage of the differences within and beyond tolerance is more useful in deciding whether the agreement is adequate, and this would assess clinical agreement in the true sense since it is based on clinical tolerance limits. This method uses all the individual differences and not their mean and SD. Perhaps many clinicians would prefer to use the percentage agreement to estimate the extent of agreement but, in case needed, the minimal agreement would be estimated by the lower confidence bound.

For those who prefer binary results, we recommend that at least 90% of differences should be within the clinical tolerance limits to conclude an adequate agreement. In place of 90%, any other desired percentage can be chosen by the investigator depending on the clinical context. Some clinicians would want no more than 1 or 2 percent values go beyond the clinical tolerance for agreement, and some may be willing to tolerate 10 percent or even higher deviation. Such flexibility is available under the direct method but not under the B-A method.

In an agreement setup, the tolerance limits should ideally be based on expected measurement error but can also be based on the clinical implication for managing a patient. If a researcher wants to add a condition, such as no difference should be more than two times the upper or lower tolerance limit, that can also be done in this method. Any big difference, howsoever isolated, raises the alarm regarding the agreement, and this method can be used to raise such an alarm.

Unlike the B-A limits, the clinical tolerance limits to be used in our method do not have to be symmetric with respect to any value – they can be ($-a$ to $+b$) where $a \neq b$ and a or b can be zero depending upon the clinical context. We shortly give an example of such asymmetric limits. The bias $\bar{d}$ and the variation s_d can be obtained in case those are of interest for a particular problem although these are not needed for assessing the extent of agreement by the proposed method. The B-A plot would be helpful for studying the trend and for interpretation of the results with the direct method also. In the case of asymmetric tolerance limits, the plot will be as shown by solid lines in Figure 1(a). The only difference is that these lines are drawn at tolerance limits instead of $\bar{d} \pm 2s_d$. If the difference is likely to be proportional to the magnitude of values, the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 August 2021 doi:10.20944/preprints202108.0343.v1

proportional difference can be examined for the agreement after setting clinical tolerance limits for the proportional difference. In that case, the plot would be as in Figure 1(b). These plots are based on the following example we have made up to illustrate the direct method.

Figure 1(a) Clinical tolerance limits (solid) and Bland-Altman limits (dotted) (b) Clinical tolerance limits for proportional difference

Example: Agreement in fasting blood glucose level measured by the conventional venous sampling and a new glucometer reading of capillary level

Consider the fictitious values in Table 1 of fasting blood glucose level obtained on 40 unrelated subjects by the conventional venous sample analyzed on an autoanalyzer (Method-1) and the capillary sample analyzed by an improvised glucometer (Method-2) that claims to provide adjusted values to match with the venous values. Since the capillary level is known to be higher, the company claims that the values given by their glucometer can be higher despite adjustment but will not exceed venous values by more than 5 mg/dL in at least 90% cases. The clinicians may be willing to accept this kind of error in view of the distinct advantage of capillary sampling. Suppose the anticipated random variation is not more than 2 mg/dL. Thus, the clinical tolerance limits for agreement are (–2, +5) mg/dL in this case. In case the values 'sufficiently' agree, the glucometer, being highly convenient and quick, can replace the current method that requires venous sampling.

In this made-up example, we have intentionally chosen asymmetric clinical tolerance limits to illustrate the direct method for this situation too, but symmetric limits can also be chosen.

Table 1. Values of fasting blood glucose level by two methods

The mean of the differences in Table 1 is 1.98 mg/dL and SD = 4.39 mg/dL. Thus, the B-A limits of agreement are (–6.81, +10.76). These are plotted as dotted lines in Figure 1(a). It is up to the researcher to interpret these as sufficiently trivial or not, and conclude the agreement or its lack, based on subjective assessment. Perhaps most would say that these are too wide, and the values given by the new glucometer do not agree with the values given by the venous sample.

When the predefined clinical tolerance limits of (–2, +5) are applied, 36 (90%) of 40 values are within these limits in our example. Thus, the agreement exists by this criterion, which is ostensibly more stringent in this case relative to the B-A limits of agreement. The conclusion now reached is different than by the B-A method despite stricter limits. The B-A method also does not provide the strength of agreement, which is assessed as 90% by the direct method in this example. If one wishes to add another condition such as no difference should be more than 10 mg/dL then one value with a difference of 21 mg/dL puts a question mark. A value as high as this raises suspicion that something wrong has happened with this reading. This could be the culprit for the B-A method also as it severely affects the $\bar{d}$ and s_d. If we exclude this value, the B-A limits of agreement become (–4.85, +7.83), which still seem unacceptably wide for agreement setup in this case but he agreement by the direct method remains good at 36/39 = 92.3%. When all the values are considered, the 95% Wilson lower bound tells that the agreement is extremely unlikely to be less than 81% in the concerned population. If the criterion is at least 90% agreement between the venous and capillary values of fasting blood glucose, the agreement in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 August 2021 doi:10.20944/preprints202108.0343.v1

this example does not provide sufficient confidence. This conclusion is different from what was obtained earlier by the point estimate.

Fasting blood glucose level has a vast range of values, say, from 60 to 400 mg/dL, depending on the condition of the person at the time of the test. It is likely in this case that the difference between venous and capillary readings will increase as the values increase. Thus, the proportional difference may be more appropriate but there is no need for logarithmic transformation for using the direct method. For illustration, we now take equal clinical tolerance limit on both sides as –2% to +2% of the value obtained by Method-1. For agreement, these limits should be narrow since only the random variation is expected in this setup, and we have chosen ±2% for illustration. For these limits, the plot of the tolerance range is as shown in Figure 1(b). Now only 19 (47.5%) differences are within these limits. Generally, this low agreement would not be acceptable, and we can conclude with this criterion that the agreement is poor for the proportional changes.

4 Discussion

There must be enormous merit in the B-A method that gives more than 100,000 results in Google search and more than 1100 documents in PubMed database (11). Many workers have spent their time and energy in explaining the method and in working out its extensions for different setups (2, 8, 12, 13, 14). The literature is so huge that it is not feasible to review all of that here.

The method of directly using the clinical tolerance limits is simple, does not require worrying about the distribution of the differences, and obviates the need to calculate ($\bar{d} \pm 2s_d$) limits. Also, there is no need to calculate the CI of the lower and upper limits, which are messy, particularly for repeated measures (12). The percentage of agreement is a natural parameter, and its estimate is immediately available that can be used to interpret the adequacy of the agreement. For a binary result, a minimum of 90% agreement can be used to infer that the agreement is sufficient. This percentage is as arbitrary as 95% for the CI in the case of the B-A method. The clinical tolerance limits do not depend on the variance of the differences whereas the limits of the agreement do. Also, this method of directly using the clinical tolerance limits is more robust as there is no need to worry about how outliers or constant values of the differences are affecting the $\bar{d}$ and the s_d, and there is no need to estimate the bias or the slope with respect to the average value unless they are needed for extraneous reasons. This method is more flexible also as asymmetric clinical tolerance limits can be used if required in the clinical context. Since the proposed approach is based on individual differences, and not the average and SD of the differences, this may be more appealing too. When such overriding merits of assessing the agreement by directly using the clinical tolerance limits as proposed are realized, extensions to various setups can be developed in the course of time.

We may further recommend for agreement analysis that the individual differences should be thoroughly examined irrespective of the method used to assess agreement. The possibility of a good agreement for low or middling values and poor agreement for high values, or the vice-versa, cannot be ruled out, and this will not be detected either by the B-A limits of agreement or by the direct method. This is a limitation of both the methods. Also, when we conclude a ‘good’ agreement, the range of values under study should be specified. Extrapolation much beyond the range actually studied is always fraught with unknown uncertainties.

5 Conclusion

Direct use of clinical tolerance limits is a hugely preferable method for assessing agreement between two quantitative measurements on the same subjects because this method is natural, robust, nonparametric, and more flexible compared to the B-A limits.

Funding; None

Conflict of interest: None

References

1. Bland JM, Altman DG. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *Lancet* 1986; **i**: 307–310. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(86)90837-8/fulltext
2. Indrayan A, Malhotra RK. *Medical Biostatistics* (Fourth Edition) 2018. Boca Raton, FL: CRC Press.
3. Francq BG, Govaerts B. How to Regress and Predict in Bland-Altman Plot? Review and Contribution Based on Tolerance Intervals and Correlated-Errors-in-Variables Models. *Statistics in Medicine* 2016; **35**: 2328–2358. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6872
4. Fish R. *Bland-Altman Plot for Age Comparisons*. http://derekogle.com/fishR/2017-04-20-Modified_BlandAltmanPlot - Accessed 6 August 2021
5. Zaki R, Bulgiba A, Ismail NA. Testing the Agreement of Medical Instruments: Overestimation of Bias in the Bland-Altman Analysis. *Preventive Medicine* 2013; **57** Suppl: S80-S82. https://www.sciencedirect.com/science/article/abs/pii/S0091743513000078
6. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond *"p* < 0.05". *The American Statistician* 2019; **73** Suppl 1: 1–19. https://www.tandfonline.com/doi/full/10.1080/00031305.2019.1583913?scroll=top&needAccess=true
7. Giavarina D. Understanding Bland Altman Analysis. *Biochemia Medica (Zagreb)* 2015; **25**: 141–151. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4470095/
8. Myles PS, Cui J. Using the Bland-Altman Method to Measure Agreement with Repeated Measures. *British Journal of Anaesthesia* 2007; **99**: 309–311. https://academic.oup.com/bja/article/99/3/309/355972
9. Newcombe RG. Two-sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. *Statistics in Medicine* 1998; **17**: 857-872. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.408.7107&rep=rep1&type=pdf

10. Fedorov V, Mannino F, Zhang, R. Consequences of Dichotomization. *Pharmaceutical Statistics* 2009; **8**: 50–61. https://onlinelibrary.wiley.com/doi/10.1002/pst.331

11. PubMed. https://pubmed.ncbi.nlm.nih.gov/?term=%22Bland-Altman+method%22 .- Accessed 11 July 2021.

12. Bland JM, Altman DG. Measuring Agreement in Method Comparison Studies. *Statistical Methods in Medical Research* 1999; **8**: 135–160. https://pubmed.ncbi.nlm.nih.gov/10501650/

13. Lu MJ, Zhong WH, Liu YX, Miao HZ, Li YC, Ji MH. Sample Size for Assessing Agreement Between Two Methods of Measurement by Bland-Altman Method. *International Journal of Biostatistics* 2016; **12**. Published online. DOI: https://doi.org/10.1515/ijb-2015-0039

14. Hofman CS, Melis RJ, Donders AR. Adapted Bland-Altman Method was Used to Compare Measurement Methods with Unequal Observations per Case. *Journal of Clinical Epidemiology* 2015; **68**: 939–943. https://www.jclinepi.com/article/S0895-4356(15)00112-2/pdf

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 August 2021 doi:10.20944/preprints202108.0343.v1

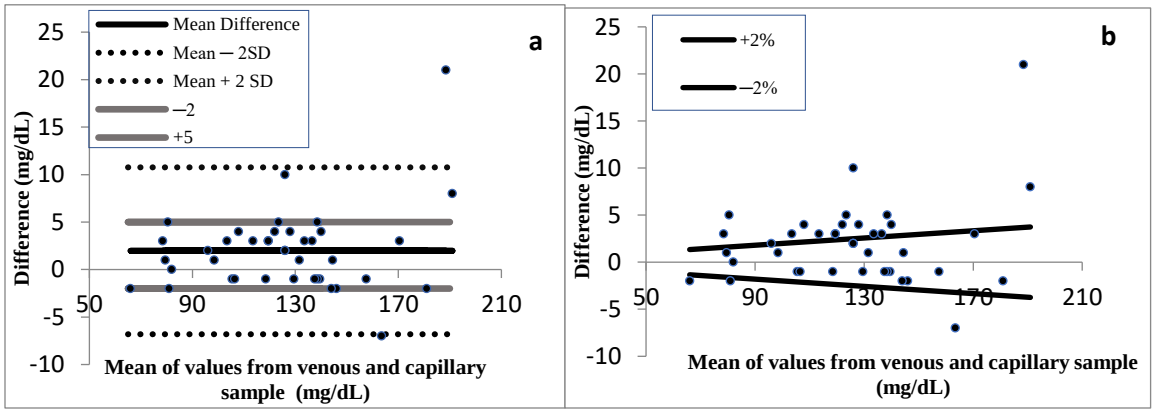

Figure 1(a) Clinical tolerance limits (solid) and Bland-Altman limits (dotted) (b) Clinical tolerance limits for proportional difference

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 August 2021 doi:10.20944/preprints202108.0343.v1

Table 1. Values of fasting blood glucose level by two methods

Subject No.	Fasting blood glucose level (mg/dL)		Difference (mg/dL)	Percentage difference (%)
	Method-1	Method-2		
1	106	110	4	3.77
2	82	80	-2	-2.44
3	121	126	5	4.13
4	95	97	2	2.11
5	178	199	21	11.80
6	147	145	-2	-1.36
7	135	138	3	2.22
8	140	139	-1	-0.71
9	112	115	3	2.68
10	126	130	4	3.17
11	130	129	-1	-0.77
12	106	105	-1	-0.94
13	187	195	8	4.28
14	77	80	3	3.90
15	120	124	4	3.33
16	118	121	3	2.54
17	67	65	-2	-2.99
18	136	141	5	3.68
19	98	99	1	1.02
20	102	105	3	2.94
21	118	121	3	2.54
22	182	180	-2	-1.10
23	167	160	-7	-4.19
24	132	135	3	2.27
25	82	82	0	0.00
26	79	80	1	1.27
27	139	138	-1	-0.72
28	125	127	2	1.60
29	119	118	-1	-0.84
30	78	83	5	6.41
31	131	132	1	0.76
32	145	143	-2	-1.38
33	169	172	3	1.78
34	158	157	-1	-0.63
35	144	145	1	0.69
36	138	137	-1	-0.72
37	121	131	10	8.26
38	107	106	-1	-0.93
39	125	127	2	1.60
40	138	142	4	2.90