# A simple and robust alternative to Bland-Altman method of assessing clinical agreement

Abhaya Indrayan[1]
Biostatistics Consultant
Max Healthcare Institute
Saket, New Delhi 110 017
India

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. These limits only describe the dispersion of disagreements in 95% cases and do not measure the degree of agreement. The interpretation regarding the presence or absence of agreement by this method is based on whether B-A limits are within the pre-specified externally determined clinical tolerance limits. Thus, clinical tolerance limits are necessary for this method. We argue in this communication that the direct use of clinical tolerance limits for assessing agreement without the B-A limits is more effective and has tremendous merits. This nonparametric approach is simple, is robust to the distribution pattern and outliers, has more flexibility, and exactly measures the degree of clinical agreement. This is explained with the help of two examples, including setups where clinical tolerance limits can be set up to follow varying trends if required in the clinical context – a feature not available in the B-A method.

**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

SUMMARY

What is already known:

Agreement between two quantitative measurements is generally assessed by Bland-Altman (B-A) limits. This method requires pre-specification of the clinical tolerance limits.

What this study adds:
Clinical tolerance limits can be directly used for assessing agreement without calculating B-A limits. This method is nonparametric, more robust, easy, and more appealing.

What should change now:
Agreement should be assessed by the direct use of clinical tolerance limits instead of B-A limits because of its huge merits.

# A simple and robust alternative to Bland-Altman method of assessing clinical agreement

## 1 Background

Although a large body of literature exists on the methods for assessing agreement in different contexts such as reproducibility[1] and bioequivalence[2,3], many 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 of these studies generally is to find whether one method can be replaced with the other without much loss of information. If a disagreement is concluded, the source is investigated – whether it is due to lack of accuracy, due to a large variation, or for any other reason.

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[4]. 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[5].

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 differences. The value of $\bar{d}$ is an estimate of the bias of one method over the other.

Under the Gaussian assumption, which is likely to hold in this case 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"[4]. 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 limitation of the B-A limits is that they end up describing the disagreements and not measuring the degree of agreement. Even the 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. Thus is a nonparametric, flexible, robust, and simple method and is a step further than the method of coverage probability[6] and the probability of agreement[7]. This direct method may not have attracted attention because it is too simple, but it has tremendous merits, particularly in situations where clinical needs allow varying tolerance limits such as more accuracy at critical values and relaxed limits at away values. This useful feature is not available in other methods. We discuss several merits of this

direct method and illustrate them with the help of two examples, including setups where the tolerance limits can be varied for different values of the measurements.

## 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, a significant contribution of the B-A method is the plot of difference against the average of the two values, known as the B-A plot[8], 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.

Perhaps no method has universal applicability, and the B-A method also has its share of problems because of its dependence on strong assumptions that may be unrealistic in many situations. The following is a list of these problems:

1. Although the stated objective of the B-A method is to find whether one method of measurement can be replaced by another method, the B-A limits end up only describing the dispersion of disagreements in 95% of the cases, without telling whether the agreement exists or not[9], nor they measure the degree of agreement.
2. The B-A limits are valid only when the distribution of the differences is Gaussian. If the distribution is highly skewed, the limits would provide a misleading assessment of the disagreement. To remedy this, a suitable transformation such as log transformation is sometimes advocated that may work in some cases but may be tedious in some other cases such as in the case of left-skewed distribution.
3. The method is based on mean $\bar{d}$ of the differences. If some differences are large and positive, and others negative, they tend to balance out and can give $\bar{d}$ nearly equal to zero. This may provide a false sense of security about the bias and may affect the interpretation of the agreement.
4. A single genuine outlying difference, which cannot be excluded, can severely distort the mean and the SD even if most differences are small. This would undesirably inflate the B-A limits. A similar problem arises when several differences are zero or equal. Both these are distinct possibilities in an agreement setup. The limits can be wide depending on the variance of the differences, even if most differences are small, and the estimate of bias $\bar{d}$ may be distorted.
5. Although the B-A limits use 95% coverage that can be varied but 95% coverage seems to be a rule in applications. This percentage is as arbitrary as 5% level of significance, which is being severely criticized[10]. Any other coverage would be equally arbitrary.
6. The usual B-A limits require that the measurements under comparison have the same precision and it should be the same for smaller values as for larger values[11]. If one method has a higher variance than the other, or if the variance varies with the values, the B-A limits may provide misleading results.
7. The B-A method may lead to overestimation of bias in some cases[12].

8. When one measurement has a constant bias ($a$) against the other without any variation, the limits of agreement would be ($a$ to $a$), which is just a single point. For example, if the difference is always 5 units, the limits of agreement would be from 5 to 5.
9. The B-A limits are symmetric and governed by $\bar{d}$. The clinical tolerance limits can be asymmetric in some situations as illustrated in our example in this communication.
10. 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[13]. This assumes the Gaussian distribution of the limits which may or may not hold. In any case, the calculation of the CI makes it relatively a tedious procedure, and most method comparison studies skip this step.
11. The most severe problem with this method is the complete dependence of the interpretation of B-A limits on externally determined clinical tolerance limits. These can be directly used to assess agreement, as elaborated next. Many agreement studies end up stating the limits of agreement without specifying the clinical tolerance limit[14,15] ignoring that no agreement study with B-A methods can be complete without setting up clinical tolerance limits.
12. The B-A method does not consider variation or trends in clinical tolerance limits.

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[4] 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, in their opinion, are too wide and 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 subjectively 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[9] remarked 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[16] 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 that the prespecified clinical tolerance limits should be directly used to find the percentage of differences within these limits. Let us call this percentage agreement. This simple method of assessing agreement, in one go, can take care of almost all the problems listed earlier for the B-A limits.

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 possibly systematic differences, some differences 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 measurement of aspartate aminotransferase (AST) by two methods, 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 with $C_L$ negative and $C_U$ positive but may or may not be symmetric. We shortly give an example of such asymmetric limits. The limits $(C_L, C_U)$ can also be varied for different values of the measurements. An example of this is also given later in this communication.

Define the extent of agreement $\pi = P(C_L < d < C_U)$. The estimate of $\pi$ is the binomial proportion of the observed differences falling between $(C_L, C_U)$ – thus amenable to all kinds of statistical inference. If somebody wants to be more confident, the 95% confidence lower bound for $\pi$ 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[17]. This will give the bound below which the extent of agreement is extremely unlikely. If somebody wishes to test the hypothesis that the agreement is at least a given threshold, such as $H_0$: $\pi \geq \pi_0$, this can be easily done by the exact one-tail test on the binomial probability under the null hypothesis or, for reasonably large *n*, by referring $z = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$ to Gaussian distribution, where $p$ is the observed proportion of differences within the specified clinical tolerance limits.

This method measures the strength of agreement instead of a binary yes or no. Many researchers these days would like to measure the exact degree of agreement and interpret it in their context. This direct method is simple, 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 degree of the agreement but, in case needed, the minimal agreement would be estimated by the lower confidence bound.

Although dichotomization has its risks[18], for those who prefer binary result as agreement exists or not, 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 threshold can be chosen by the investigator depending on the clinical context. Some clinicians would want no more than 1 or 2 percent differences to go beyond the clinical tolerance for agreement, and some may be willing to tolerate deviation in10 percent cases or even higher. Such flexibility (and several other advantages as given later) is available under the direct method but not under the B-A method. 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.

The tolerance limits may be based on the expected measurement error. In this case, since positive errors are likely to be as much as negative errors, $C_L = -C_U$. 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 degree of agreement by the proposed method. The nonparametric (Hodges-Lehmann) estimate of the bias, not affected by the distribution pattern or

the outliers, is the median of the average of all pairs of the differences $(d_i + d_j)/2$ $(i < j)$[19]. There will be a total of $n(n+1)/2$ such pairs when there are $n$ subjects in the study. The confidence interval (CI) can also be obtained.

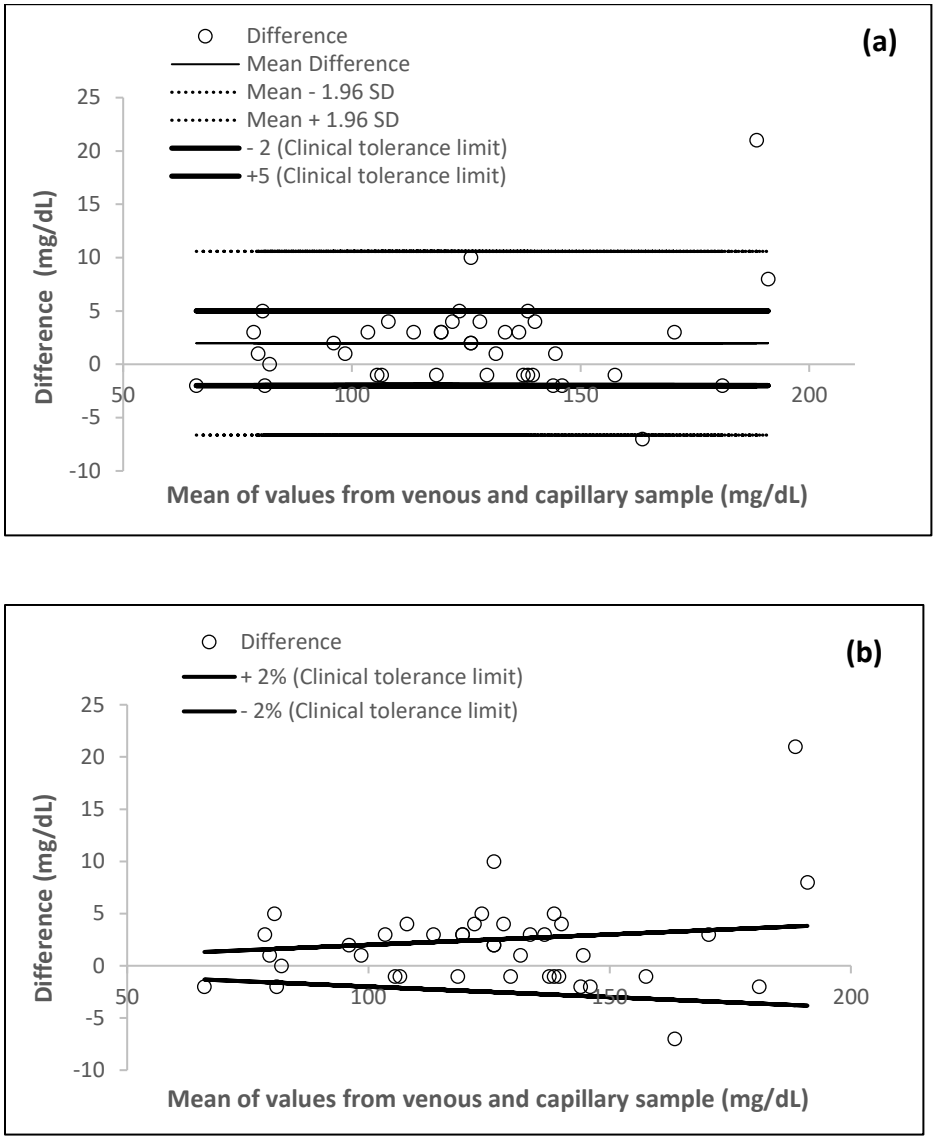



Figure 1. (a) Clinical tolerance limits (solid) and Bland-Altman limits (dotted), (b) Clinical tolerance limits for proportional difference

The B-A plot would help studying the trend and interpreting 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). There are drawn at tolerance limits instead of $\bar{d} \pm 2s_d$. If the differences are likely to be proportional to the magnitude of values, the proportional differences can be examined for agreement after setting clinical tolerance limits for the proportional differences. 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. We later show that the clinical tolerance limits can be allowed to follow a linear or nonlinear trend if desirable in a specific clinical context. Such a flexibility could give a better assessment of the agreement in some cases.

**Example 1: 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 levels obtained on 40 unrelated subjects by the conventional venous sample analyzed in a laboratory (Method-1) and the capillary sample analyzed by an improvised glucometer at home (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 in either direction. Higher values by this margin are already covered by +5 limit and we need to make the provision for lower values only. Thus, the clinical tolerance limits for agreement are (–2, +5) mg/dL and asymmetric in this case. These are asymmetric. In case the values 'sufficiently' agree, the glucometer, being highly convenient and quick, can replace the current method that requires venous sampling and involves a laboratory.

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

| 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 |

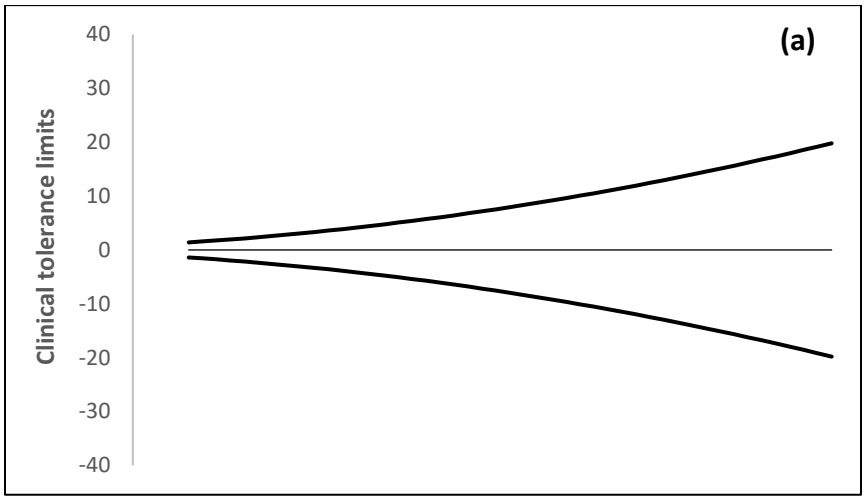
| | | | | |
|---|---|---|---|---|
| 21 | 118 | 121 | 3 | 2.54 |
| 22 | 182 | 180 | -2 | -1.10 |
| 23 | 167 | 160 | -7 | -4.19 |
| 24 | 132 | 135 | 3 | 2.27 |
| 25 | 82 | 82 | 0 | 0.00 |
| 26 | 79 | 80 | 1 | 1.27 |
| 27 | 139 | 138 | -1 | -0.72 |
| 28 | 125 | 127 | 2 | 1.60 |
| 29 | 119 | 118 | -1 | -0.84 |
| 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 |

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) (dots may not be visible as these are very close to one-another). Under the B-A method, 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 (solid in Figure 1(a)), 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 the 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 this example does not provide sufficient confidence. This conclusion is different from what was obtained earlier by the point estimate. The Hodges-Lehmann estimator of bias in this case is 1.5, with a 95% CI from 0.5 to 3.0.

Perhaps a single example based on synthetic data is not enough to demonstrate the merits of the direct method, but it illustrates how the method can be effectively used in a practical setup. The enormous merits of this method are enumerated later in this paper.

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, and the B-A limits would be calculated based on log transformation. There is no need for logarithmic transformation for using the direct method and the clinical tolerance limits can be defined in terms of percentage. For illustration, we now take equal clinical tolerance limit on both sides as –2% to +2% of the value obtained by Method-1, which we consider as the reference in this case. 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 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.



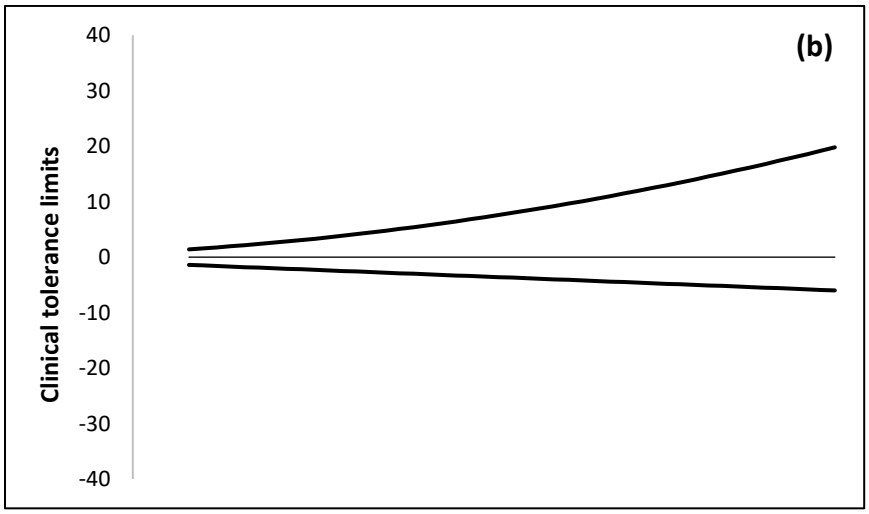



Figure 2. Varying clinical tolerance limits: (a) increasing percentage with the values of the measurement, (b) smaller limit for negative differences and larger for positive differences in terms of percentage

Since the primary objective in the agreement setup is to assess the agreement with respect to the clinical tolerance and not to describe the extent of disagreement, in place of changing the B-A limits for different setups, we focus on the rationality in setting up the clinical tolerance limits. An extremely nice feature of the direct method is its flexibility for setting up the clinical tolerance limits. These can be based on the clinical implications of the difference between the two methods under comparison. The limits do not have to be constant or fixed proportion through the range of values as considered in Example 1 but can be allowed to vary and can follow a linear or nonlinear trend. Such variation in tolerance limits is easy to implement with our method but not with the B-A method. For example, in the case of blood glucose levels, one may allow 2% difference for low values level, gradually increasing to 3% for middling values, and 4% for high values (Figure 2(a)). We have shown symmetric limits in this figure, but they do not have to follow a symmetric trend for negative and positive differences. If the clinical context allows, the lower (or the higher) limit can be set at a constant rate such as 2%, and the upper (or lower) limit following an increasing or decreasing trend such as in Figure 2(b). In the case of blood pressure (BP), a clinician may want fairly accurate readings when the level is around the threshold such as 90 mmHg for diastolic BP because of implications regarding prescribing a treatment, but may be willing to tolerate higher differences for lower and higher levels. A similar situation is illustrated in the following example.

**Example 2: Agreement between systolic blood pressure (BP) readings by sphygmomanometer and automatic blood pressure monitor**

This example illustrates the method of direct use of clinical tolerance limits when these vary with the value of the measurement. The data are from Table 1 in the Bland-Altman[13] paper for comparing the systolic BP reading by an observer (J1) and by an automatic monitor (S1) on 85 subjects. It is for these two readings that the authors assessed agreement by B-A limits. These limits for this data are enormously wide from –54.7 mmHg to +22.2 mmHg (Figure 3(a)). These limits are inconsequential for our purpose and stated for the sake of completeness.

Suppose the clinicians are not willing to accept more than a 3-mmHg difference between the readings by two methods under comparison when the systolic BP is around the threshold of 135 mmHg because that has implications for instituting a treatment, but a larger difference can be allowed when the reading is away from this value because a larger difference at low or high values does not much affect the clinical assessment. This can give clinical tolerance limits of the shape given in Figure 3(b). In this figure, the clinical tolerance limits are drawn at (3 + 10% of the difference from 135) mmHg on either side. Only 15 out of 85 differences are within these limits of clinical tolerance – the strength of agreement is only 17.6%, and the 95% confidence lower bound is 12.0%.

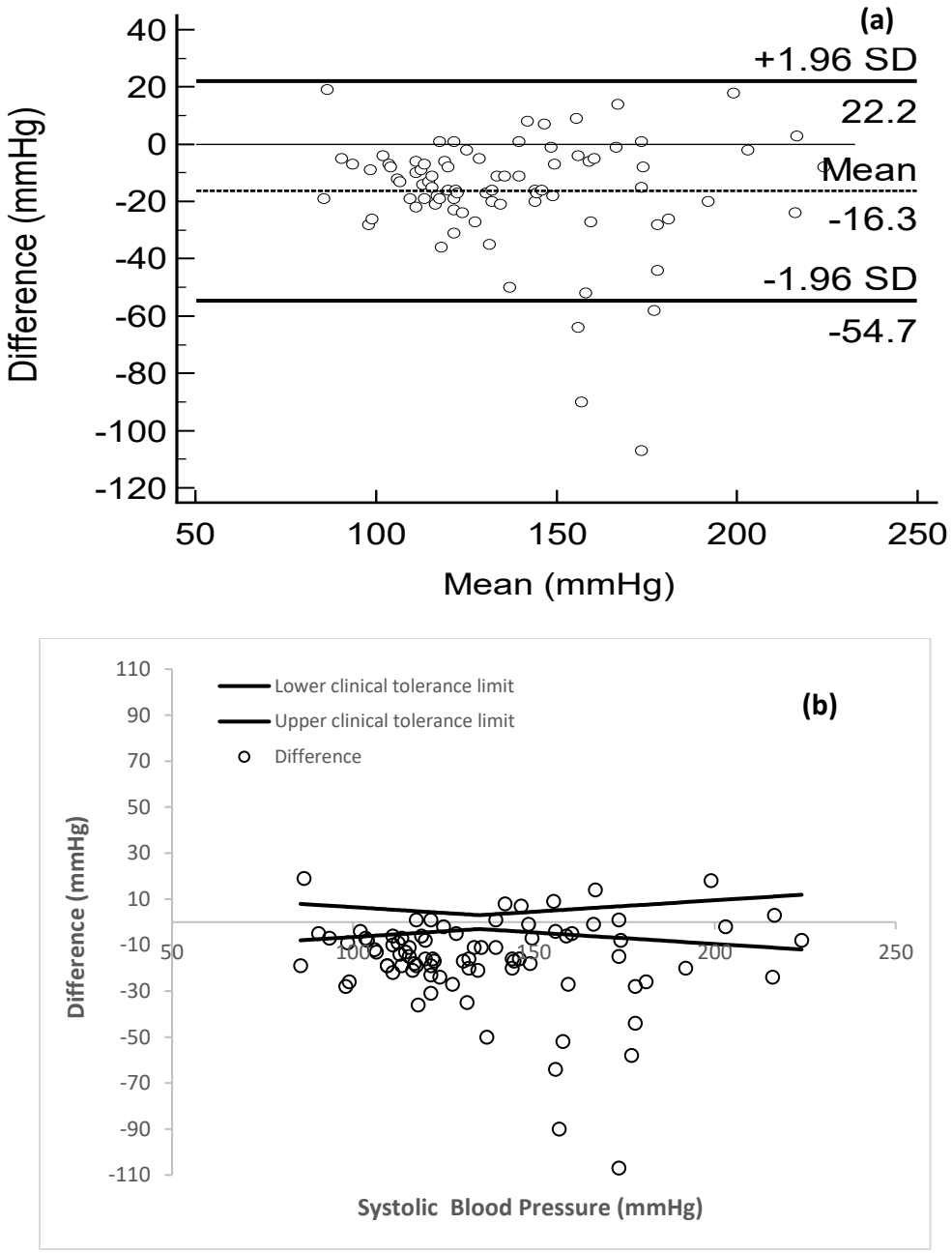



Figure 3. (a) Bland-Altman limits for the agreement between readings by sphygmomanometer and automatic monitor (data from Table 1 of Bland-Altman 1998), (b) Agreement with clinical tolerance limits where the limits vary with the value of the measurements

## 4 Discussion

Statistical methods for assessing agreement have been extensively reviewed[6,20.21]. Dunn and Roberts[22] and Alanen[23] have provided a critique of the B-A method. But the B-A method continues to be extremely popular with more than 100,000 results in Google search and more

than 1100 documents in PubMed database (2022). Many workers have spent their time and energy in explaining the method and in working out its extensions for different setups[5,11,13,16,24,25] so much so that a set of guidelines for reporting agreement studies was also proposed[26]. The literature is so huge that it is not feasible to review all of that here.

One of the reasons that the B-A method became so popular is its simplicity. Calculate the limits of (dis)agreement based on the mean and SD of the differences, compare them with the clinical tolerance limits, and you are done. The method of directly using the clinical tolerance limits advocated in this paper is even more simple. Just calculate the percentage of differences falling within the clinical tolerance limits and you are done. The emphasis is on agreement with respect to the clinical tolerance and not on the dispersion of the differences. Nearest to this method is the probability of agreement proposed by Steven et al.[7]. Another close method is the total deviation index estimated by tolerance intervals[27]. The latter is for concordance and not exactly for agreement. Both these methods are based on distributional assumptions, particularly the Gaussian, and require mean and SD. The direct method is nonparametric and simpler.

The direct method has several merits. Besides not requiring worrying about the distribution of the differences, it obviates the need to calculate ($\bar{d} \pm 2s_d$) limits. Also, there is no need to calculate the CIs of these lower and upper limits, which are messy, particularly for repeated measures[13]. Our method is not affected by heteroscedasticity either. In case needed, the heteroscedasticity can be used to determine differential clinical tolerance limits for different values. 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% coverage in the case of the B-A limits. There is a flexibility to call ≥90% agreement as good, 80%-90% as tolerable and 70%-80% unsatisfactory, and <70% as dismal. Such categorization is not possible with 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 unless required for extraneous reasons. Most importantly, this method is more flexible as asymmetric clinical tolerance limits or following a specified trend can be easily 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. The sample size requirement will be the same as for estimating a population proportion and there is no need to follow the intricate procedure suggested by Lu et al.[25] for the B-A method. When such overriding merits of assessing the agreement by directly using the clinical tolerance limits as proposed are realized, extensions to various setups, such as for repeated measures and meta-analysis, can be developed over time. These could be relatively much easier to use in applications yet achieve the desired objective.

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 our 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 method based on the B-A limits.

**Funding;** None

**Conflict of interest:** None

**References**

1. Lin L, Torbeck LD. Coefficient of accuracy and concordance correlation coefficient: New statistics for methods comparison. PDA J Pharm Sci Technol. 1998;52:55-59. https://journal.pda.org/content/52/2/55.long
2. Schall R, Williams RL. Towards a practical strategy for assessing individual bioequivalence. Food and Drug Administration Individual Bioequivalence Working Group. J Pharmacokinet Biopharm. 1996 Feb;24(1):133-49. doi: 10.1007/BF02353513. https://pubmed.ncbi.nlm.nih.gov/8827586/
3. Lin LI. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. Stat Med. 2000;19:255-270. doi: 10.1002/(sici)1097-0258(20000130)19:2<255::aid-sim293>3.0.co;2-8.https://pubmed.ncbi.nlm.nih.gov/10641028/
4. 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
5. Indrayan A, Malhotra RK. Medical Biostatistics (Fourth Edition). Boca Raton, FL: CRC Press. 201: p 638.
6. Lin L, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: Models, issues, and tools. J Am Stat Assoc. (Theory and Methods). 2002;97: 257-270. https://s2.smu.edu/ngh/stat6385/Linetal2002.pdf
7. Stevens NT, Steiner SH, MacKay RJ. Assessing agreement between two measurement systems: An alternative to the limits of agreement approach. Stat Methods Med Res. 2017;26:2487-2504. doi: 10.1177/0962280215601133. Epub 2015 Sep 2. PMID: 26335274. https://pubmed.ncbi.nlm.nih.gov/26335274/
8. 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. Stat Med 2016; 35:2328–2358. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6872
9. Giavarina D. Understanding Bland Altman analysis. Biochemia Medica (Zagreb) 2015; **25**:141–151. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4470095/

10. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p < 0.05". Am Stat 2019; **73:**Suppl 1, 1–19. https://www.tandfonline.com/doi/full/10.1080/00031305.2019.1583913?scroll=top&needAccess=true

11. Taffe P. When can the Bland-Altman limits of agreement method be used and when it should not be used. J Clin Epidemiol. 2021; **37**:176-181. https://doi.org/10.1016/j.jclinepi.2021.04.004

12. Zaki R, Bulgiba A, Ismail NA. Testing the agreement of medical instruments: Overestimation of bias in the Bland-Altman analysis. Prev Med. 2013; 57:Suppl, S80-S82. https://www.sciencedirect.com/science/article/abs/pii/S0091743513000078

13. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999; **8**:135–160. https://pubmed.ncbi.nlm.nih.gov/10501650/

14. Mantha S, Roizen MF, Fleisher LA, Thisted R, Foss J. Comparing methods of clinical measurement: Reporting standards for Bland and Altman analysis. Anesth Analg. 2000;90:593-602. doi: 10.1097/00000539-200003000-00018. https://pubmed.ncbi.nlm.nih.gov/10702443/

15. Dewitte K, Fierens C, Stöckl D, Thienpont LM. Application of the Bland-Altman plot for interpretation of method-comparison studies: A critical investigation of its practice. Clin Chem. 2002;48:799-801; author reply 801-2. https://pubmed.ncbi.nlm.nih.gov/11978620/

16. Myles PS, Cui J. Using the Bland-Altman method to measure agreement with repeated measures. Br J Anaesth. 2007; **99:**309–311. https://academic.oup.com/bja/article/99/3/309/355972

17. Newcombe RG. Two-sided confidence intervals for the single proportion: Comparison of seven methods. Stat Med. 1998; 17:857-872. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.408.7107&rep=rep1&type=pdf

18. Fedorov V, Mannino F, Zhang R. Consequences of dichotomization. Pharma Stat. 2009; **8**:50–61. https://onlinelibrary.wiley.com/doi/10.1002/pst.331

19. Hollander M, Wolfe DA. Nonparametric Statistical Methods. John Wiley & Sons, 1973; p 33.

20. Watson PF, Petrie A. Method agreement analysis: A review of correct methodology. Theriogenology. 2010;73:1167-1179. doi: 10.1016/j.theriogenology.2010.01.003. https://www.sciencedirect.com/science/article/pii/S0093691X10000233?via%3Dihub

21. Barnhart HX, Yow E, Crowley AL, Daubert MA, Rabineau D, Bigelow R, Pencina M, Douglas PS. Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting. Stat Methods Med Res. 2016;25:2939-2958. doi: 10.1177/0962280214534651. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1017.6500&rep=rep1&type=pdf

22. Dunn G, Roberts C. Modelling method comparison data. Stat Methods Med Res. 1999;8:161-179. doi: 10.1177/096228029900800205.https://pubmed.ncbi.nlm.nih.gov/10501651/

23. Alanen E. Everything all right in method comparison studies? Stat Methods Medical Res. 2012; 21:297-309. doi: 10.1177/0962280210379365. https://pubmed.ncbi.nlm.nih.gov/20716590/

24. Hofman CS, Melis RJ, Donders AR. Adapted Bland-Altman method was used to compare measurement methods with unequal observations per case. J Clin Epidemiol. 2015; **68**:939–943. https://www.jclinepi.com/article/S0895-4356(15)00112-2/pdf

25. 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. Int J Biostat. 2016;**12**. Published online. doi: https://doi.org/10.1515/ijb-2015-0039

26. Kottner J, Audigé L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol. 2011; **64**:96-106. doi: 10.1016/j.jclinepi.2010.03.002. https://pubmed.ncbi.nlm.nih.gov/21130355/

27. Escaramís G, Ascaso C, Carrasco JL. The total deviation index estimated by tolerance intervals to evaluate the concordance of measurement devices. BMC Med Res Methodol. 2010; **10:**31. https://doi.org/10.1186/1471-2288-10-31. https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-10-31