Original Research Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2023, Vol. 32(1) 195–206

DOI: 10.1177/09622802221137743
journals.sagepub.com/home/smm
SAGE

Use of clinical tolerance limits for assessing agreement

Patrick Taffé

Abstract

In this study, we have further extended the methodology proposed, first, by Lin et al. (2002) and, later, extended by Stevens et al. (2017, 2018), on the coverage probability/probability of agreement, by relaxing the strong parametric assumptions regarding the distribution of the latent trait and developing inference methods allowing to compute both pointwise and simultaneous confidence bands. The methodology requires repeated measurements by at least one of the two measurement methods and accommodates heteroscedastic measurement errors. It performs often very well even when one has only one measurement by one of the two measurement methods and at least five repeated measurements from the other. It circumvents some of the deficiencies of the Bland & Altman limits of agreement method and provides a more direct assessment of the agreement level.

Keywords

Agreement, limits of agreement, tolerance limits, differential bias, proportional bias, method comparison

1 Introduction

In clinical research, to assess the agreement/interchangeability between two measurement methods, when the characteristic of interest is continuous, the Bland & Altman's limits of agreement (LoA) method[1,2] is one of the most frequently used methodologies (their 1986[2] paper has been cited over 50,000 times as of May 2022). Often this is motivated by a new and perhaps less expensive or easier method of measurement against an established reference standard. To evaluate the comparability of the methods, the investigator collects repeated measurements from each method on a sample of subjects. Bland & Altman's LoA are then computed by adding and subtracting 1.96 times the estimated standard deviation from the mean differences, that is, the average bias estimate. A scatter plot of the differences versus the means of the two variables with the LoA superimposed is used to visually appraise the level of agreement. Further, a regression of the differences as a function of the means is added to the plot to indicate the direction and amplitude of the bias.[3] A final decision regarding the agreement between the two methods is based on whether the LoA is within pre-defined clinical tolerance limits.

However, it has been shown that in the presence of a proportional bias or situations where the variances of the measurement errors of each of the methods are not constant (i.e. heteroscedastic) the Bland & Altman's plot may be misleading.[4] When this is the case, the regression line may show an upward or downward trend when there is no bias or a zero slope when there is a bias. Therefore, recently, a new statistical methodology to assess the bias, precision, and agreement of a new measurement method, which circumvents the deficiencies of the Bland & Altman's LoA method, has been developed.[4,5] This methodology, however, is in the same spirit as that of Bland & Altman, and the level of agreement is not directly quantified (except for the concept of "percentage of agreement," which, however, does not depend on pre-defined tolerance

Center for Primary Care and Public Health (unisanté), Division of Biostatistics, University of Lausanne, Lausanne, Switzerland

Corresponding author:
Patrick Taffé, Center for Primary Care and Public Health (Unisanté), DFRI/Division of Biostatistics, University of Lausanne, Route de Berne 113, 1010 Lausanne, Switzerland.
Email: patrick.taffe@unisante.ch

limits). Rather, based on inspection of the Bias, Precision, and Agreement plots, the investigator has to decide whether disagreement (not agreement) is not too high for the two methods to be deemed to be interchangeable.

To more directly quantify the level of agreement, Lin et al.[6] have proposed the concept of "coverage probability," where the probability that the absolute difference between the two measurements made on the same subject is less than a predefined value is computed. Their methodology, however, allows one to assess the overall agreement and does not take into account that the level of agreement might depend on the value of the true latent trait in a continuous way. In addition, implicitly, homoscedastic measurement errors are assumed (an often too strong assumption) and the presence of a possible bias is not assessed. For these reasons, Stevens et al.[7] have extended this methodology to allow the coverage probability to depend on the value of the latent trait, as well as on the amount of bias. Later, they further extended their methodology to allow for heteroscedastic measurement errors.[8] They have called their extended agreement concept "probability of agreement."

There are, however, several important limitations to the Stevens et al.[7,8] methodology. First of all, it essentially relies on a parametric specification of the distribution of the latent trait, which is unpleasant as usually little is known about this distribution, particularly given that measurement errors make its identification difficult. In addition, the investigator has to decide over which support the index will be computed, thereby possibly extrapolating to impossible or unrealistic values of the genuine latent trait. Third, they considered constant tolerance limits, whereas often the variance of the measurement errors is heteroscedastic and increases with the value of the latent trait. Consequently, non-constant tolerance limits, which depend on the latent trait might be better suited. Fourth, they computed pointwise, and not simultaneous, confidence bands around the probability of agreement, which bears important limitations, as in practice one may be interested in assessing the agreement level at different values of the latent trait or even across an entire interval and not just at a single point of the support. Fifth, they did not show any simulation results regarding the coverage rate of their confidence band, nor used, for example, the logit transformation to avoid negative values to be included in the confidence band (as illustrated by Figure 2 in their 2018 paper[7]). Actually, their developments were based on the use of the delta method, which may not always provide reliable results (as demonstrated by the results of the present study), particularly when the function of the parameters is highly non-linear. Indeed, the same issue with the delta method has been demonstrated in a recently published study.[5] Sixth, since the probability of agreement index depends on the value of the true trait, which is by definition latent and not estimated in their approach, it is practically not possible to assess the true level of agreement. Lastly, their unconditional probability of agreement concept,[7] to assess the overall agreement level, is not a proper unconditional or marginal probability; rather it is a conditional probability evaluated at the mean value of the latent trait and therefore it does not summarize the overall agreement appropriately.

Many other methods to assess agreement/interchangeability have been proposed in the literature and we refer the interested reader to some recently published papers for a more extensive literature review.[8–10]

Therefore, in this paper, the main goal is to develop an extended methodology, based on Lin et al.[6] and Steven et al.[8] coverage probability/probability of agreement concepts, which overcomes the above-mentioned limitations. For this, we will rely on an empirical Bayes approach (to predict the value of the true latent trait), which has already been fruitfully used to overcome the limitations of Bland & Altman's LoA methodology.[4,5] The inference will be developed to build both pointwise and simultaneous confidence bands around the conditional agreement curve, such that the investigator may adopt either method according to his inference goal.

2 Methodology of the clinical tolerance limits

Consider the general measurement error model:

$$
\begin{aligned}
y_{1ij} &= \alpha_1 + \beta_1 x_{ij} + \varepsilon_{1ij}, \;\; \varepsilon_{1ij}|x_{ij} \sim N(0, \sigma^2_{\varepsilon_1}(x_{ij}; \boldsymbol{\theta}_1)) \qquad (1)\\
y_{2ij} &= \alpha_2 + \beta_2 x_{ij} + \varepsilon_{2ij}, \;\; \varepsilon_{2ij}|x_{ij} \sim N(0, \sigma^2_{\varepsilon_2}(x_{ij}; \boldsymbol{\theta}_2))\\
x_{ij} &\sim f_x(\mu_x, \sigma^2_x)
\end{aligned}
$$

where y_{1ij} be the jth replicate measurement by method 1 on individual $i, j = 1, \ldots, n_i$ and $i = 1, \ldots, N$, whereas y_{2ij} is obtained by method 2, x_{ij} is a latent variable with density f_x representing the true unknown trait, and ε_{1ij} and ε_{2ij} represent measurement errors by methods 1 and 2. It is assumed that the variances of these errors, that is, $\sigma^2_{\varepsilon_1}(x_{ij}; \boldsymbol{\theta}_1)$ and $\sigma^2_{\varepsilon_2}(x_{ij}; \boldsymbol{\theta}_2)$, are heteroscedastic and increase with the level of the true latent trait x_{ij} in a way to be precisely specified later, which depends on the vectors of unknown parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. For the reference method, for instance, method 2, $\alpha_2 = 0$ and $\beta_2 = 1$, whereas for method 1, the differential α_1 and proportional β_1 biases have to be estimated from the data.

The mean value of the latent variable x_{ij} is μ_x and its variance σ_x^2. It is assumed that the latent variable represents the true unknown but constant value of the trait for individual i and, therefore, $x_{ij} \equiv x_i$ (this assumption may be relaxed to allow, e.g. a linear trend[4]).

When method 2 is the reference standard and method 1 is the new method to be evaluated, the model reduces to:

$$\begin{gathered} y_{1ij} = \alpha_1 + \beta_1 x_i + \varepsilon_{1ij},\ \ \varepsilon_{1ij}|x_i \sim N(0, \sigma^2_{\varepsilon_1}(x_i; \boldsymbol{\theta}_1)) \\ y_{2ij} = x_i + \varepsilon_{2ij},\ \ \varepsilon_{2ij}|x_i \sim N(0, \sigma^2_{\varepsilon_2}(x_i; \boldsymbol{\theta}_2)) \\ x_i \sim f_x(\mu_x, \sigma_x^2) \end{gathered} \tag{2}$$

Note that this measurement error model is slightly different from the classical measurement error model[11] in that the heteroscedasticity depends on the latent trait and not on an observed average. In addition, we have considered a simple linear relationship between y_{1ij} and x_i to identify the differential and proportional biases. It is possible, however, to consider instead a non-linear function of x_i but in that case, the bias no longer decomposes into two components with nice interpretations.

2.1 Computation of the conditional and overall/marginal agreement

Consider the differences:

$$d_{ij} = y_{1ij} - y_{2ij} = (\alpha_1 + \beta_1 x_i + \varepsilon_{1ij}) - (x_i + \varepsilon_{2ij}) = \alpha_1 + (\beta_1 - 1)x_i + (\varepsilon_{1ij} - \varepsilon_{2ij}) \tag{3}$$

and assume that lower $C_L(x_i)$ and upper $C_U(x_i)$ clinical tolerance limits, which may depend on the true trait, have been defined a priori, that is, before seeing the data, based on clinical considerations, such that the Bernoulli variable X_{ij} will take value 1 when the difference d_{ij} lies between the two clinical tolerance limits and 0 otherwise:

$$X_{ij} = \begin{cases} 1 & \text{if } \ C_L(x_i) < d_{ij} < C_U(x_i) \\ 0 & \text{otherwise} \end{cases}$$

We have $X_{ij}|x_i \sim Bernoulli(\pi(x_i))$, where $\pi(x_i)$ is the *coverage probability*[6] or *conditional probability of agreement.*[7,8]

One may obtain the *overall/marginal agreement* by marginalizing:

$$P(C_L(x_i) < d_{ij} < C_U(x_i)) \equiv \pi = \int_{-\infty}^{\infty} \pi(x_i) f_X(x_i) dx \tag{4}$$

Given model (2) and assumptions, one may compute the *conditional probability of agreement*:

$$\begin{aligned} P(C_L(x_i) < d_{ij} < C_U(x_i)|x_i) &\equiv \pi(x_i) \\ &= P\left(\frac{C_L(x_i) - \alpha_1 - (\beta_1 - 1)x_i}{\sqrt{V(\varepsilon_{1ij} - \varepsilon_{2ij})}} < \frac{d_i - \alpha_1 - (\beta_1 - 1)x_i}{\sqrt{V(\varepsilon_{1ij} - \varepsilon_{2ij})}} < \frac{C_U(x_i) - \alpha_1 - (\beta_1 - 1)x_i}{\sqrt{V(\varepsilon_{1ij} - \varepsilon_{2ij})}} |x_i\right) \\ &= \Phi\left(\frac{C_U(x_i) - \alpha_1 - (\beta_1 - 1)x_i}{\sqrt{\sigma^2_{\varepsilon_1}(x_i; \boldsymbol{\theta}_1) + \sigma^2_{\varepsilon_2}(x_i; \boldsymbol{\theta}_2)}}\right) - \Phi\left(\frac{C_L(x_i) - \alpha_1 - (\beta_1 - 1)x_i}{\sqrt{\sigma^2_{\varepsilon_1}(x_i; \boldsymbol{\theta}_1) + \sigma^2_{\varepsilon_2}(x_i; \boldsymbol{\theta}_2)}}\right) \end{aligned} \tag{5}$$

where Φ is the standard normal cumulative distribution function.

As regards the clinical tolerance limits, $C_L(x_i)$ and $C_U(x_i)$, it is up to the investigator to define them, ideally based on clinical considerations, before seeing the data.

One simple alternative is to set constant values, which do not depend on the true latent trait:

$$\begin{gathered} C_L(x_i) = a \\ C_U(x_i) = b \end{gathered}$$

Alternatively, one may define limits that depend on the true latent trait in a specific form, for example, linear:

$$C_L(x_i) = -a - bx_i$$

$$C_U(x_i) = a + bx_i$$

where "a" represents the smallest tolerable upper limit for a zero value of the latent trait and "b" is the percentage of acceptable difference beyond the zero latent trait value.

Other choices are of course possible and the tolerance limits need not be linear.

2.2 Estimation of the model parameters

We do not want to make strong distributional assumptions regarding the true latent trait x_i and instead of specifying a parametric distribution, and treating x_i as a nuisance parameter to be integrated out from the joint likelihood function,[12] we estimate the regression model for y_{2ij} by marginal maximum likelihood (i.e. without specifying any parametric distribution for x_i) and adopt an empirical Bayes approach[13] to predict x_i, by means of its posterior distribution (i.e. the mean of the conditional distribution of x_i given the vector $\mathbf{y}_{2i}$, which is the best linear unbiased prediction (BLUP) for x_i):

$$\hat{x}_i = E(x_i|\mathbf{y}_{2i}) \tag{6}$$

$$= \int x_i \frac{f_{y_2}(\mathbf{y}_{2i}|x_i)f_x(x_i)}{\int f_{y_2}(\mathbf{y}_{2i}|x_i)f_x(x_i)\,dx_i} dx_i \tag{7}$$

where for the sake of notational convenience we have suppressed the dependence of the density functions f_{y_2} and f_x from their parameters, which have been estimated by maximum likelihood.[4]

Notice, that with sufficient repeated observations per individual (e.g. between 5 and 10) and a sufficient number of individuals (e.g. 100) the empirical distribution of the BLUP of x_i will approximate well the true distribution, as can be verified empirically by simulations and shown theoretically.[14] Then, one proceeds to the estimation of the regression equation for y_{1ij} in (2) and of the differential α_1 and proportional β_1 biases simply by OLS after having substituted the BLUP $\hat{x}_i$ for the true unmeasured trait x_i.[4]

Estimates of the standard deviation of the measurement errors are obtained following a similar approach to that of Bland and Altman[3,4]:

$$\hat{\sigma}_{\varepsilon 1}(\hat{x}_i; \hat{\boldsymbol{\theta}}_1) = \hat{E}(|\hat{\varepsilon}^*_{1ij}|)\sqrt{\pi/2} = (\hat{\theta}_1^{(0)} + \hat{\theta}_1^{(1)}\hat{x}_i)\sqrt{\pi/2} \tag{8}$$

$$\hat{\sigma}_{\varepsilon_2}(\hat{x}_i; \hat{\boldsymbol{\theta}}_2) = \hat{E}(|\hat{\varepsilon}^*_{2ij}|)\sqrt{\pi/2} = (\hat{\theta}_2^{(0)} + \hat{\theta}_2^{(1)}\hat{x}_i)\sqrt{\pi/2} \tag{9}$$

Notice, that the form of the heterogeneity needs not to be a straight line and a fractional polynomial may be used instead if the investigator believes that the straight-line model is too restrictive. In any case, a graphical representation of $|\hat{\varepsilon}^*_{2ij}|$ versus $\hat{x}_i$ may be useful to visually check the plausibility of the straight-line model.

Once the parameters have been estimated and conditional agreement computed:

$$\hat{\pi}(\hat{x}_i) = \Phi\left(\frac{C_U(\hat{x}_i) - \hat{\alpha}_1 - (\hat{\beta}_1 - 1)\hat{x}_i}{\sqrt{\sigma^2_{\varepsilon_1}(\hat{x}_i; \hat{\boldsymbol{\theta}}_1) + \sigma^2_{\varepsilon_2}(\hat{x}_i; \hat{\boldsymbol{\theta}}_2)}}\right) - \Phi\left(\frac{C_L(\hat{x}_i) - \hat{\alpha}_1 - (\hat{\beta} - 1)\hat{x}_i}{\sqrt{\sigma^2_{\varepsilon_1}(\hat{x}_i; \hat{\boldsymbol{\theta}}_1) + \sigma^2_{\varepsilon_2}(\hat{x}_i; \hat{\boldsymbol{\theta}}_2)}}\right) \tag{10}$$

the variance needs to be computed to build pointwise and simultaneous confidence bands.

To proceed, we will rely on two different approaches, the first, the conventional multivariate delta method, and the second, based on a simulation method. In our experience, with complicated functions involving ratios of random variables, the delta method may perform poorly.[5] Also, to guarantee confidence bands comprising values between 0 and 1, the logit transform (of the conditional agreement) will be used.

When the goal is to assess the agreement level for a specific value of the latent trait, a pointwise confidence interval will be fine as it guarantees that on average 95% of the computed intervals will cover the true value. However, when interest lies in several points from the support or in the whole curve, a simultaneous confidence band is required, as it guarantees a proper coverage rate for the simultaneous inference, whatever the number of points from the support.

2.3 Computation of pointwise and simultaneous confidence bands

The variance of $\hat{\pi}(\hat{x}_i)$ may be computed by using the multivariate delta method and the uncertainty in the estimate $\hat{x}_i$ accounted for by the law of total variance.[5] However, as mentioned above and based on previous experience,[5] the performance of the delta method may not always be good and one needs to consider alternative methods to compute the variance. One way of doing this is by simulations:

- Step 1: For each individual i, 1000 values for $\hat{x}_i$ are drawn in $N(\hat{x}_i, V(\hat{x}_i - x_i))$.
- Step 2: As $\hat{\theta}_k^{(l)} \overset{asy}{\sim} N(\hat{\theta}_k^{(l)}, V(\hat{\theta}_k^{(l)}))$, $k = 1, 2$, $l = 0, 1$, 1000 values for $\hat{\theta}_1^{(0)}$ and $\hat{\theta}_1^{(1)}$, and for $\hat{\theta}_2^{(0)}$ and $\hat{\theta}_2^{(1)}$, are drawn in a bivariate Normal distribution, accounting for the $\text{cov}(\hat{\theta}_k^{(0)}; \hat{\theta}_k^{(1)})$.
- Step 3: 1000 values for $\hat{\alpha}_1$ and $\hat{\beta}_1$ are drawn in a bivariate Normal distribution, accounting for the $\text{cov}(\hat{\alpha}_1; \hat{\beta}_1)$.
- Step 4: Finally, 1000 values for $\hat{\pi}(\hat{x}_i)$ are generated for each individual i and the sample variance computed.

A pointwise 95% confidence band for $\pi(x_i)$ may be computed either directly, by referring to the 2.5 and 97.5 percentiles of the simulated distribution, or indirectly, by using the simulated variance and Normal approximation on the logit scale:

$$\begin{aligned} \text{logit}(\hat{\pi}(\hat{x}_i)) - Z_{1-\alpha/2}\sqrt{V(\text{logit}(\hat{\pi}(\hat{x}_i)))} &\leq \text{logit}(\pi(x_i)) \\ &\leq \text{logit}(\hat{\pi}(\hat{x}_i)) + Z_{1-\alpha/2}\sqrt{V(\text{logit}(\hat{\pi}(\hat{x}_i)))} \end{aligned} \tag{11}$$

where $Z_{1-\alpha/2}$ is either the quantile of the standard Normal distribution or Student distribution with $N - 6$ degrees of freedom (i.e. number of individuals minus number of parameters estimated).

With regard to a $100(1-\alpha)$ % *simultaneous* confidence band, the following inequality must hold[15,16]:

$$P\Big(\pi(x_i) \in \Big[\hat{\pi}(\hat{x}_i) \pm q_{1-\alpha}\sqrt{V(\hat{\pi}(\hat{x}_i) - \pi(x_i))}\Big],\ \forall x_i \in \Psi\Big) \geq 1 - \alpha \tag{12}$$

where Ψ denotes the set of x_i values of interest and $q_{1-\alpha}$ a critical value to be determined.

The probability statement (12) is equivalent to:

$$P\left(\sup_{\min \leq x_i \leq \max} \frac{|\hat{\pi}(\hat{x}_i) - \pi(x_i)|}{\sqrt{V(\hat{\pi}(\hat{x}_i) - \pi(x_i))}} \leq q_{1-\alpha}\right) \geq 1 - \alpha \tag{13}$$

where min is the minimum and max maximum value of x_i. The appropriate value of $q_{1-\alpha}$ can be found by simulating the left-hand side quantity in the probability statement (13), say a thousand times, and computing the $1-\alpha$ empirical quantile of the distribution.

As in our case, as x_i is a latent trait and not observed, $q_{1-\alpha}$ is computed as the $1-\alpha$ quantile of the sampling distribution of[5]:

$$\sup_{\min \leq \hat{x}_i \leq \max} \frac{|\hat{\pi}^{(m)}(\hat{x}_i) - \hat{\pi}(\hat{x}_i)|}{\sqrt{V(\hat{\pi}^{(m)}(\hat{x}_i))}} \tag{14}$$

where $\hat{\pi}^{(m)}(\hat{x}_i)$ represents the mth simulated value of $\hat{\pi}(\hat{x}_i)$, $m = 1, \ldots, R$, and $\hat{\pi}(\hat{x}_i)$ is the estimated value of $\pi(x_i)$ based on the sample at hand. The number R of simulations should be large enough (e.g. 1000) to get stable estimates.

Concretely, a $100(1-\alpha)$ % simultaneous confidence band for the conditional agreement is given by:

$$\hat{\pi}(\hat{x}_i) - q_{1-\alpha}\sqrt{V(\hat{\pi}(\hat{x}_i))} \leq \pi(x_i) \leq \hat{\pi}(\hat{x}_i) + q_{1-\alpha}\sqrt{V(\hat{\pi}(\hat{x}_i))} \tag{15}$$

where $q_{1-\alpha}$ is found by using the following simulation algorithm:

- Step 1: $\hat{\alpha}_1^{(m)}$ and $\hat{\beta}_1^{(m)}$ are simulated from the appropriate bivariate Normal distribution and the $\hat{x}_i^{(m)}$ are drawn in $N(\hat{x}_i, V(\hat{x}_i - x_i))$. Then, $\hat{\pi}^{(m)}(\hat{x}_i^{(m)})$ and $V(\hat{\pi}^{(m)}(\hat{x}_i^{(m)}))$ are computed for each simulation m.
- Step 2: For $m = 1, \ldots, R$ calculate:

$$q_{1-\alpha}^{(m)} = \sup_{\min \leq \hat{x}_i \leq \max} \left| \frac{\hat{\pi}^{(m)}(\hat{x}_i^{(m)}) - \hat{\pi}(\hat{x}_i)}{\sqrt{V(\hat{\pi}^{(m)}(\hat{x}_i^{(m)}))}} \right|$$

- Step 3: $q_{1-\alpha}$ is the $1-\alpha$ quantile of the empirical distribution of the $q^{(m)}_{1-\alpha}$.

To guarantee values comprised strictly between 0 and 1, the logit transform (of the conditional agreement) is used in all the calculations and the inverse transform is applied in the end to get the required band.

The performance of these alternative variance estimators and confidence bands has been studied in the simulation section below.

3 A simulation study

The goal of this simulation study is to assess the performance of the various confidence bands developed in section 2 and introduce the "Conditional agreement plot."

3.1 Performance of the confidence bands

We simulated 1000 data sets according to the following data generating process:

$$\begin{aligned} &y_{1ij}=\alpha_1+\beta_1 x_i+\varepsilon_{1ij}, \quad \varepsilon_{1ij}|x_i\sim N(0,(\theta_1^{(0)}+\theta_1^{(1)}x_i)^2\pi/2)\\ &y_{2ij}=x_i+\varepsilon_{2ij}, \qquad\qquad \varepsilon_{2ij}|x_i\sim N(0,(\theta_2^{(0)}+\theta_2^{(1)}x_i)^2\pi/2)\\ &x_i\sim U[a,\ b]\end{aligned} \tag{16}$$

where $i=,1,\ldots,N$ individuals, the number of repeated measurements $n_{2i}\sim U\{c,d\}$ by the reference method was drawn in a discrete Uniform distribution, and for the new measurement method $n_{1i}\sim U\{e,f\}$, and by varying the values of the vector of parameters $[\alpha_1\,\beta_1\,\theta_1^{(0)}\,\theta_1^{(1)}\,\theta_2^{(0)}\,\theta_2^{(1)}]'$.

For the sake of brevity, we focused on five different settings, which may be relevant for the clinical practice: (1) zero differential and no proportional biases; (2) negative differential bias with proportional bias less than 1; (3) negative differential bias with proportional bias larger than 1; (4) positive differential bias with proportional bias less than 1; (5) positive differential bias with proportional bias larger than 1. To allow comparisons with results from a previous study,[5] the following pairs of values for the differential and proportional biases were used (α_1,β_1) =(0, 1), (−6, 0.85), (−4, 1.2), (4, 0.8), (4, 1.2), and the level of the latent trait assumed to be uniformly distributed between 10 and 40. Other distributions for x_i have also been investigated (Normal and Gamma).

The number of repeated measurements from the reference standard was drawn in the three distributions: (1) $n_{2i}\sim U\{10,15\}$; (2) $n_{2i}\sim U\{5,10\}$; (3) $n_{2i}\sim U\{1,3\}$, and that from the new measurement method in the two distributions: (1) $n_{1i}\sim U\{1,3\}$; (2) $n_{1i}=1$. Regarding the number of individuals, the distribution of the latent trait, and the parameters of the variance functions, we restricted the presentation to the results obtained with $n=100$, $x_i\sim U[10,40]$, and for some selected values of $(\theta_1^{(0)},\theta_1^{(1)},\theta_2^{(0)},\theta_2^{(1)})$: (0, 0.2, 1.75, 0.08), (0.1, 0.07, 0.15, 0.09), and (1, 0.05, 2, 0.01) (these values and those for the differential and proportional biases were selected to illustrate settings where Bland & Altman's methodology provided strongly biased estimates of the differential and proportional biases).

Notice that $n_{2i}\sim U\{1,3\}$ means on average two repeated measurements by the reference standard, which is an extremely unfavorable scenario given that the methodology requires several measurements per individual by at least one of the two measurement methods. Simulation results and conclusions for various sample sizes and levels of heteroscedasticity were similar.

As regards the clinical tolerance limits, for the sake of simplicity, they have been set to constant values and do not depend on the true latent trait:

$$C_L(x_i)=-5$$

$$C_U(x_i)=5$$

Results of the simulations regarding coverage rates (nominal set at 95%) of the pointwise and simultaneous confidence bands are presented in Tables 1 and 2. In Table 1, the cells contain coverage rates obtained by the Normal approximation, quantiles of the simulation distribution, and delta method (separated by slashes).

Coverage rates are generally very good when the variance has been computed by the simulation method and the Normal approximation is used, and at least five to 10 repeated measurements per individual are available by the reference method. With few repeated measurements per individual, the coverage rate of the method based on the quantiles of the simulation

Table 1. Coverage rates (nominal 95%) of the pointwise confidence band (Normal approximation/quantiles of the simulation distribution/delta method).

Differential bias	Proportional bias	$n_{2i} \sim U\{10, 15\}$	$n_{2i} \sim U\{5, 10\}$	$n_{2i} \sim U\{1, 3\}$	$\theta_1^{(0)}, \theta_1^{(1)}, \theta_2^{(0)}, \theta_2^{(1)}$
$n_{1i} \sim U\{1, 3\}$					
0	1	94.6/94.7/88.0	94.7/94.9/88.2	92.6/93.2/84.9	0, 0.2, 1.75, 0.08
−6	0.85	94.8/93.8/90.5	93.7/92.2/89.0	87.9/83.3/81.7	0.1, 0.07, 0.15, 0.09
−4	1.2	94.1/94.0/90.2	93.5/93.5/89.7	90.1/91.0/86.4	1, 0.05, 2, 0.01
4	0.8	94.9/94.7/89.7	94.6/94.4/89.1	93.6/92.7/87.6	0, 0.2, 1.75, 0.08
4	1.2	94.6/93.7/94.0	95.8/94.9/95.3	96.1/95.1/95.1	0, 0.2, 1.75, 0.08
$n_{1i} = 1$					
0	1	94.6/94.7/88.2	94.1/94.3/87.4	91.7/92.6/83.4	0, 0.2, 1.75, 0.08
−6	0.85	94.9/93.6/90.1	95.3/93.8/91.6	92.0/86.2/85.7	0.1, 0.07, 0.15, 0.09
−4	1.2	94.4/94.0/90.5	94.8/94.8/91.2	91.4/91.8/86.9	1, 0.05, 2, 0.01
4	0.8	95.3/95.0/90.2	94.8/94.6/89.7	93.8/93.1/87.8	0, 0.2, 1.75, 0.08
4	1.2	95.2/94.1/94.2	96.7/95.3/95.8	97.3/96.3/96.2	0, 0.2, 1.75, 0.08

Table 2. Coverage rates (nominal 95%) of the simultaneous confidence band.

Differential bias	Proportional bias	$n_{2i} \sim U\{10, 15\}$	$n_{2i} \sim U\{5, 10\}$	$n_{2i} \sim U\{1, 3\}$	$\theta_1^{(0)}, \theta_1^{(1)}, \theta_2^{(0)}, \theta_2^{(1)}$
$n_{1i} \sim U\{1, 3\}$					
0	1	98.3	97.3	91.8	0, 0.2, 1.75, 0.08
−6	0.85	99.4	99.1	98.4	0.1, 0.07, 0.15, 0.09
−4	1.2	98.9	98.5	94.4	1, 0.05, 2, 0.01
4	0.8	98.3	97.6	94.4	0, 0.2, 1.75, 0.08
4	1.2	99.5	99.5	97.0	0, 0.2, 1.75, 0.08
$n_{1i} = 1$					
0	1	99.5	97.5	91.2	0, 0.2, 1.75, 0.08
−6	0.85	99.7	99.2	99.6	0.1, 0.07, 0.15, 0.09
−4	1.2	99.6	99.1	97.4	1, 0.05, 2, 0.01
4	0.8	99.7	97.7	96.0	0, 0.2, 1.75, 0.08
4	1.2	99.6	99.5	98.4	0, 0.2, 1.75, 0.08

distribution has not performed as well as that based on the simulated variance and Normal approximation. Globally, the coverage rate of the delta method is not good.

For example, when $n_{1i} = 1$, $n_{2i} \sim U\{10, 15\}$, differential bias is 0 and proportional bias 1, and the clinical tolerance limits have been prespecified as $C_L(x_i) = -5$ and $C_U(x_i) = 5$, the coverage rate of the pointwise 95% confidence band using the Normal approximation and quantiles of the simulation distribution is 94.6% and 94.7%, whereas it is only 88.2% for the delta method.

Coverage rates are generally on the conservative side with at least five to 10 repeated measurements per individual by the reference method.

3.2 The conditional agreement plot

To introduce the "Conditional agreement plot," the following data generating process has been considered:

$$y_{1i} = 4 + 0.8x_i + \varepsilon_{1i}, \quad \varepsilon_{1i}|x_i \sim N(0, (0.2x_i)^2) \tag{17}$$

$$y_{2ij} = x_i + \varepsilon_{2ij}, \;\; \varepsilon_{2ij}|x_i \sim N(0, (1.75 + 0.08\,x_i)^2)$$

$$x_i \sim U[10, 100]$$

where $i = 1, \ldots, 100$ and the number of repeated measurements per individual i has been set to $n_{1i} = n_{2i} = 5$. The new method (method 1) has a differential bias of 4 and a proportional bias of 0.8. In addition, the variance of the measurement errors from method 1 is larger than that of reference method 2.

Consider, first, the setting where the clinical tolerance limits have been set to constant values and do not depend on the true latent trait:

$$C_L(x_i) = -5$$

$$C_U(x_i) = 5$$

In Figure 1, a scatter plot of the simulated data versus the true latent trait is provided.

On the scatter plot of the simulated data the proportional bias of the new method is clearly apparent, as well as the heteroscedasticity of the measurement errors.

Figure 2, left, shows the true Tolerance limit plot, which serves as a benchmark for the comparison with the Tolerance limit plot (Figure 3 left) in which the unknown true trait has been predicted by the empirical Bayes method (i.e. BLUP of x_i), and right the true conditional probability of agreement computed using formula (5).

In this example, due to the presence of a proportional bias and heteroscedasticity of measurement errors, the true conditional probability of agreement is not constant and decreases with the level of the true latent trait value (x_i). The overall/marginal agreement has simply been estimated by the proportion of observations contained within the tolerance limits and the 95% confidence interval computed by the Agresti and Coull method.[17]

Based on the estimates of the parameters of model (2), the conditional probability of agreement and BLUP of the true latent trait x_i have been computed. In Figure 3, we have drawn, left, the scatter plots of the differences versus the BLUP of x_i, called the "Tolerance limit plot," and, right, the "Conditional probability of agreement plot."

The Tolerance limits plot from Figure 3 is very similar to that of Figure 1, which illustrates that already with five repeated measurements per individual the BLUP of x_i generally provides a very good approximation of the true latent trait.[4] This plot may be useful to get a first glimpse of the agreement based on the predefined tolerance limits. On the Conditional agreement plot, one can see that the estimated (in green) and true values (in red) of the conditional agreement are pretty similar, except maybe for low values of the latent trait. We have added the 95% pointwise confidence bands (CB) computed by the simulation and delta methods. Simulation results in Table 1 have shown that the coverage rate of the simulation method was much better than that of the delta and the reason is well illustrated in this example where the pointwise CB calculated by the simulation method is wider than that computed by the delta method. As expected, the simultaneous CB is much wider than the pointwise and allows one to assess the probability of agreement for as many points from the support as desired.

For the sake of comparison, in Figure 4 we have also computed the Bland & Altman's LoA and Taffé's agreement plots.

In comparison with the Conditional probability of agreement plot, which directly provides the level of agreement, the investigator must carefully inspect the LoA on the Bland & Altman's and Taffé's agreement plots to assess the degree of agreement. In this respect, the percentage of agreement (defined as $\%\hat{A} = 1 - (Z_{1-\alpha/2}\hat{\sigma}_d + |bias_i|)\,/\,\hat{x}_i$, which is roughly proportional to the half-width of the LoA plus the absolute amount of bias, see Taffé[5]), in the Agreement plot, may be helpful. Comparing the percentage of the agreement curve with the Conditional agreement plot, it is interesting to notice that for low values of the latent trait the two figures disagree: on the Agreement plot the level of agreement drops down for values of the latent trait below 25, whereas it is the highest on the Conditional agreement plot. Actually, this is not surprising as the percentage of agreement penalizes the width of the LoA by the absolute value of

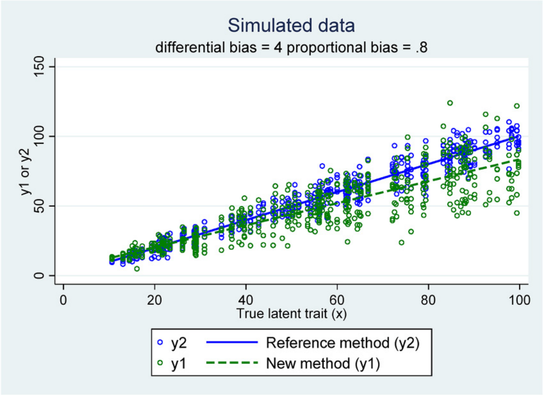

Figure 1. Scatter plot of the simulated data.

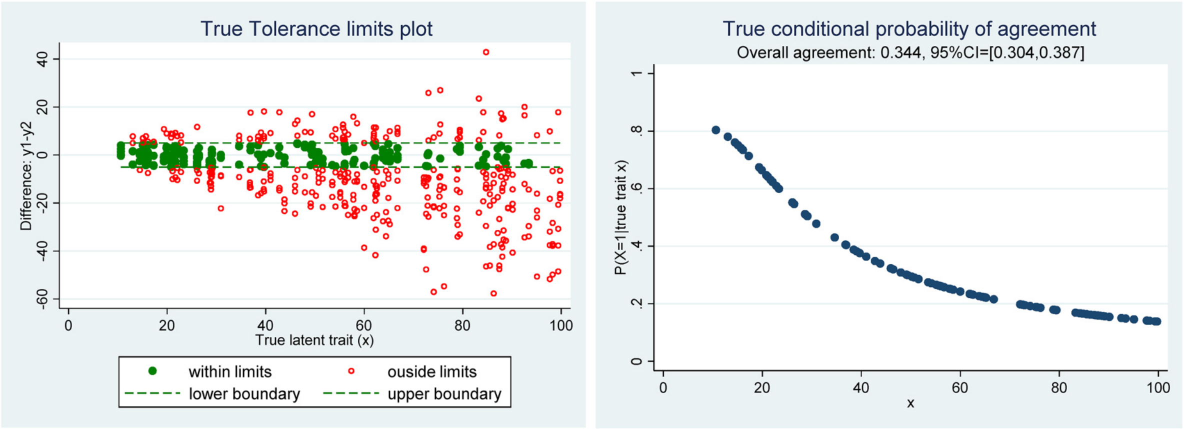

Figure 2. (Left) Scatter plot of the differences y_1–y_2 versus the true latent trait with tolerance limits, (right) scatter plot of the true conditional probability of agreement.

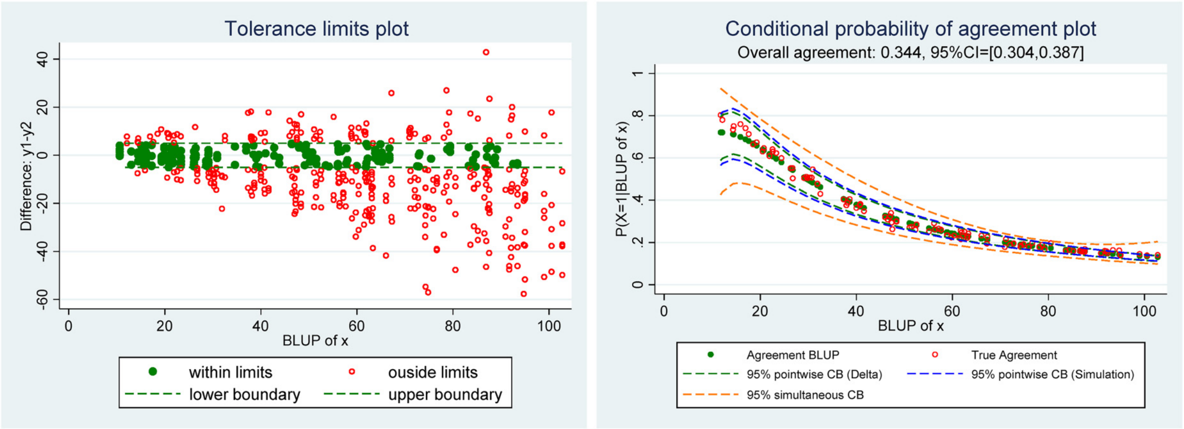

Figure 3. (Left) Scatter plot of the differences y_1–y_2 versus the BLUP of *x* with tolerance limits, (right) conditional probability of agreement plot.

the bias, which is proportionally high for low values of the latent trait, whereas the tolerance limits have been set to constant values, irrespective of the latter.

To fix this issue, one may define tolerance limits that depend on the true latent trait in a specific form, for example, linearly:

$$C_L(x_i) = -bx_i$$

$$C_U(x_i) = bx_i$$

To illustrate, in Figure 5 we have set $C_L(x_i) = -0.15x_i$ and $C_U(x_i) = 0.15x_i$, such that the tolerance limits are narrower for small values and wider for large values of the latent trait.

Clearly, now, the percentage of the agreement curve and the Conditional agreement plot give the same message, the level of agreement is best for values of the latent trait around 25.

A worked example based on real data is presented in the Supplemental Appendix.

6 Discussion

In this study, we have further extended the methodology proposed, first, by Lin et al.[6] and, then, extended by Stevens et al.,[8] on the coverage probability/probability of agreement concepts, by relaxing the strong parametric assumptions regarding the distribution of the latent trait. In addition, we have extended the tolerance limits concept, which is

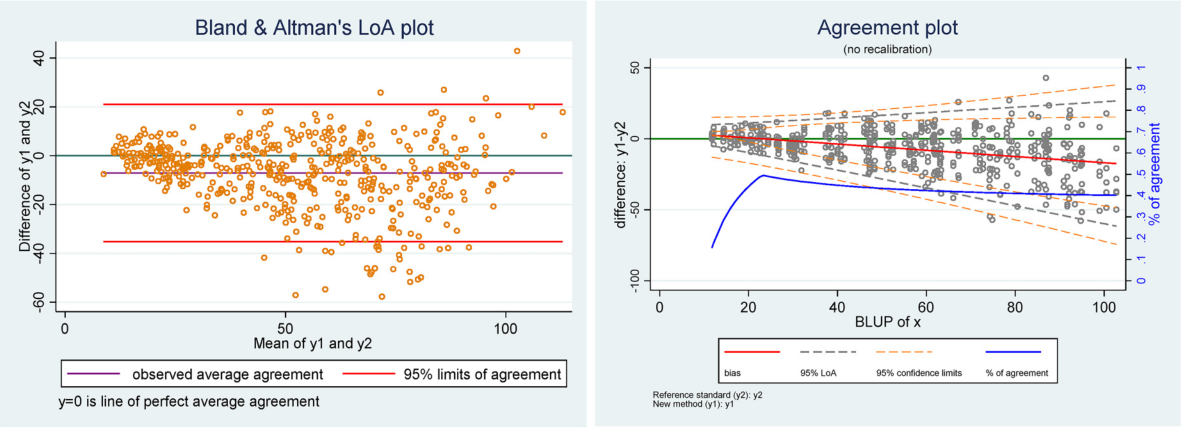

Figure 4. (Left) Bland & Altman's LoA plot, (right) agreement plot.

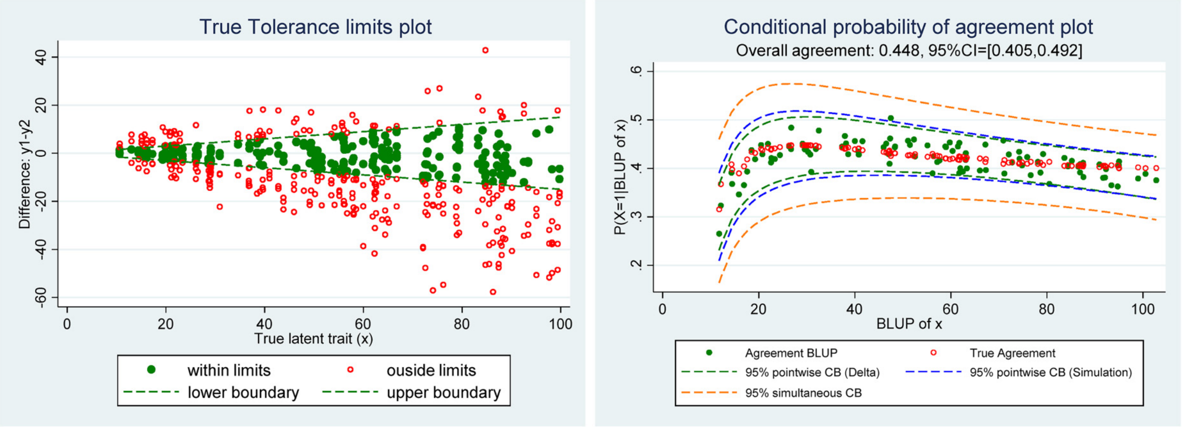

Figure 5. (Left) Scatter plot of the differences y_1–y_2 versus the BLUP of *x* with non-constant tolerance limits, (right) conditional probability of agreement plot.

allowed to depend on the true latent trait, and developed an inference theory based on different variance estimation methods. The focus has been on developing simultaneous, and not simply pointwise, confidence bands to allow for simultaneous inferences, that is, to allow the inference for the whole curve and not only for a specific value of the latent trait. This is particularly relevant from a clinical perspective, as it may turn out that the probability of agreement is high enough only for a limited range of the values of the latent trait. In addition, by superimposing on the same plot the conditional agreement curves obtained from several competing new measurement methods, along with their simultaneous confidence bands, one may assess which of these competing methods performs best and for which values of the latent trait.

We have investigated two different methods to compute the variance of the conditional agreement, the standard delta method, and a simulation method. Simulation results (in Table 1) have shown that the delta method has not performed well, whereas the simulation method performed very well when there were at least five to 10 repeated measurements per individual by the reference method and only one measurement by the new method. This poor result of the delta method has already been observed in Taffé[5] and it is recommended to abandon it for the simulation method. It is worth mentioning that to avoid impossible negative values in the CB, the logit transform should be used when using the Normal approximation or delta method. Also, in some settings, the coverage rate of the pointwise confidence band based on the quantiles of the simulation distribution has not performed as well as that based on the simulated variance and Normal approximation, particularly with few repeated measurements per individual. The reason is not clear and should be investigated in further research.

The coverage rate of the simultaneous CB (in Table 2) has been found to be quite conservative in almost all the settings investigated. This result has already been observed in Taffé[5] and by others in the setting of longitudinal data and mixed models.[18,19] However, in the latter studies, the authors did not have to deal with the tricky issues posed by the use of a predicted latent variable as one of the regressors. Further research should strive to improve this.

Notice that, as shown in Taffé,[5] the repeated measurements need not be from the reference standard, and the estimation method may be easily adapted to the setting where the repeated measurements come from the new method. This is a great asset of the proposed methodology, as sometimes it may turn out to be easier to perform many measurements by the new measurement method. Nevertheless, requiring repeated measurements by one of the two methods might discourage the applied researcher to use our methodology. However, this is necessary for statistical identification. Indeed, when the variance of the measurement errors of each measurement method is not constant or their ratio is unknown, which is usually the case in the biomedical field (the variance of measurement errors often increases as the latent trait increases), having only one measurement by each of the two measurement methods does not allow one to identify all the parameters of the model (2).[11]

The marginal probability of agreement has simply been estimated by the proportion of observations between the two tolerance limits and its confidence interval computed by the method of Agresti and Coull,[17] although other methods such as that of Wilson may also have been used. We, also, have investigated the performance of our methodology for other distributions of the latent trait (Normal and Gamma, with skewness 1 and kurtosis 4.5, results not shown), and different parameter values for the heteroscedasticity. It still performed very well. This is the great asset of the empirical Bayes approach, which can accommodate virtually any distribution of the latent trait, given there are enough repeated measurements per individual.

We have seen that depending on how the tolerance limits have been defined (based or not on the value of the latent trait), the conditional agreement may be similar in shape to the percentage of agreement proposed by Taffé.[5] As mentioned above, the former is based on pre-defined tolerance limits, whereas the latter depends on the width of the LoA and the amount of bias, and does not require the investigator to set tolerance limits. Which one should be preferred depends on the information available a priori to the investigator for setting the limits. We have illustrated that the definition of the tolerance limits may have an important leverage effect regarding the level of the conditional agreement calculated, whereas the percentage of the agreement depends solely on the variability and bias found in the data. It is recommended to compute both measures of agreement and thoroughly inspect the plots before deciding on the agreement.

Finally, it is important to emphasize that our modeling strategy rests on the assumption that the individual latent trait is constant within individuals, that is, $x_{ij} \equiv x_i$. This means that the repeated measurements should ideally be taken in sequence within a time interval where this assumption is sensible. It is theoretically possible to extend the methodology to other settings where the latent trait has a time trend.[4] However, in that case, the simple and convenient decomposition of the bias into differential and proportional components is probably not sensible, and more sophisticated models should be developed.

In summary, we have extended the methodology proposed by Lin et al.[6] and Stevens et al.,[7,8] on the coverage probability/probability of agreement, by relaxing the strong parametric assumptions regarding the distribution of the latent trait and developing an inference method allowing us to compute pointwise and simultaneous CBs. The methodology requires repeated measurements by at least one of the two methods and can accommodate heteroscedastic measurement errors. It performs often very well even when one has only one measurement by one of the two methods and five to 10 repeated measurements from the other. This methodology will be made available in a future Stata package.

Acknowledgements

I would like to thank Dr Abhaya Indrayan for some stimulating remarks.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental material

Supplemental material for this article is available online.

ORCID iD

Patrick Taffé https://orcid.org/0000-0002-2049-6651

References

1. Altman DG and Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician* 1983; **32**: 307–317.

2. Bland JM and Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1**: 307–310.
3. Bland JM and Altman DG. Measuring agreement in method comparison studies. *Stat Meth Med Res* 1999; **8**: 135–160.
4. Taffé P. Effective plots to assess bias and precision in method comparison studies. *Stat Meth Med Res* 2018; **27**: 1650–1660.
5. Taffé P. Assessing bias, precision, and agreement in method comparison studies. *Stat Meth Med Res* 2020; **29**: 778–796.
6. Lin L, Hedayat AS, Sinha B, et al. Statistical methods in assessing agreement: models, issues, and tools. *JASA* 2002; **97**: 257–270.
7. Stevens NT, Steiner SH and MacKay RJ. Assessing agreement between two measurement systems: an alternative to the limits of agreement approach. *Stat Meth Med Res* 2017; **26**: 2487–2504.
8. Stevens NT, Steiner SH and MacKay RJ. Comparing heteroscedastic measurement systems with the probability of agreement. *Stat Meth Med Res* 2018; **27**: 3420–3435.
9. Lin L. Overview of agreement statistics. *J Biopharm Stat* 2008; **18**: 126–144.
10. Barnard HX, Yow E, Crowley AL, et al. Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting. *Stat Meth Med Res* 2016; **25**: 2939–2958.
11. Dunn G. *Statistical evaluation of measurement errors: design and analysis of reliability studies*. 2nd ed. London: Arnold, 2004.
12. Nawarathna LS and Choudhary PK. A heteroscedastic measurement error model for method comparison data with replicate measurements. *Stat Med* 2015; **34**: 1242–1258.
13. Verbeke G and Molenberghs G. *Linear mixed models in practice*. Lecture Notes in Statistics 126, Springer, 1997.
14. Jiang J. Asymptotic properties of the empirical BLUP and BLUE in mixed linear models. *Stat Sinica* 1998; **8**: 861–885.
15. Ruppert D, Wand MP and Carroll RJ. *Semiparametric regression*. Cambridge University Press, 2003.
16. Gsteiger S, Bretz F and Liu W. Simultaneous confidence bands for nonlinear regression models with application to population pharmacokinetic analyses. *J Biopharmaceut Stat* 2011; **21**: 708–725.
17. Agresti A and Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. *Am. Stat* 1998; **2**: 119–126.
18. Sun J, Raz J and Faraway J. Confidence bands for growth and response curves. *Stat Sinica* 1999; **9**: 679–698.
19. Mun J and Chun H. Effective simultaneous confidence bands for repeated measurements in linear mixed-effect models. *J Stat Comput Simulat* 2014; **84**: 1748–1760.
20. Taffé P, Halfon P and Halfon M. A new statistical methodology overcame the defects of the Bland & Altman method. *J Clin epi* 2020; **124**: 1–7.