Title: Effective plots to assess bias and precision in method comparison studies.

Authors: Taffé P

Journal: Statistical methods in medical research

Year: 2018 Jun

Issue: 27

Volume: 6

Pages: 1650-1660

DOI: 10.1177/0962280216666667

Effective plots to assess bias and precision in method comparison studies

Patrick Taffé

Institute for Social and Preventive Medicine, University of Lausanne, Switzerland
Patrick.Taffe@chuv.ch

Abstract

Bland and Altman's limits of agreement (LoA) have traditionally been used in clinical research to assess the agreement between different methods of measurement for quantitative variables. However, when the variances of the measurement errors of the two methods are different, Bland and Altman's plot may be misleading; there are settings where the regression line shows an upward or a downward trend but there is no bias or a zero slope and there is a bias.

Therefore, the goal of this paper is to clearly illustrate why and when does a bias arise, particularly when heteroscedastic measurement errors are expected, and propose two new plots, the "bias plot" and the "precision plot", to help the investigator visually and clinically appraise the performance of the new method. These plots do not have the above-mentioned defect and still are easy to interpret, in the spirit of Bland and Altman's LoA.

To achieve this goal we rely on the modeling framework recently developed by Nawarathna and Choudhary, which allows the measurement errors to be heteroscedastic and depend on the underlying latent trait. Their estimation procedure, however, is complex and rather daunting to implement. We have, therefore, developed a new estimation procedure, which is much simpler to implement and, yet, performs very well, as illustrated by our simulations.

The methodology requires several measurements with the reference standard and possibly only one with the new method for each individual.

Keywords

Limits of agreement, differential bias, proportional bias, Bland-Altman's plot, method comparison, measurement, empirical Bayes, BLUP.

Corresponding author

Patrick Taffé, Institute for Social and Preventive Medicine (IUMSP), Biopôle 2, Route de la Corniche 10, 1010 Lausanne, Switzerland

tel.: +41 (0) 21 314 72 56

fax: +41 (0) 21 314 49 54

E-mail: Patrick.Taffe@chuv.ch

1 Introduction

Bland and Altman's limits of agreement (LoA) have traditionally been used in clinical research to assess the agreement between different methods of measurement for quantitative variables.[1] Typically, the investigator wishes to assess a new cheaper or simpler method of measurement against the established reference standard. For that purpose he disposes of one or several measurements by each method on every subject in the study. Then, Bland and Altman's LoA are computed by ± 1.96 times the estimated standard deviation of the differences and a scatter plot of the differences versus the mean of the two variables with the LoA superimposed is used to visually appraise the degree of agreement and quantify the magnitude. Often, a regression of the differences versus the mean is added to the plot to enhance its reading and assess the direction of the bias.[2]

When the variances of the measurement errors of each method are different, which is probably often the case, Bland and Altman's plot, however, may be misleading. Indeed, there are settings where the regression line shows an upward or a downward trend and there is no bias, whereas in others despite a zero slope there is a bias. This problem has been previously described in published literature but, to our best knowledge, no other simple to use and effective plots to visually appraise bias and precision have been proposed.[3-8]

The literature on measurement errors is abundant and our goal is not to survey this literature (see Nawarathna and Choudhary[9], and references therein for a recent survey). Rather, we will reconsider the problem of the estimation of the bivariate mixed effects model recently proposed by Nawarathna and Choudhary[9], which extends previously published methods to the setting of heteroscedastic measurement errors, particularly when heteroscedasticity is a function of the latent trait.

We have developed a new two-step estimation procedure, based on an empirical Bayes approach, which is much simpler to implement than the one adopted by Nawarathna and Choudhary[9], and yet performs very well as illustrated by our simulation results.

Therefore, the goals of this paper are to thoroughly investigate under what circumstances Bland and Altman LoA are reliable, and when this is not the case, present and illustrate a new two-step estimation procedure to identify and quantify the amount of differential and proportional bias, develop a method of recalibration in order to use the new recalibrated measurement method and compare its accuracy with that of the reference standard, and finally propose two new plots, the "bias plot" and the "precision plot", to help the investigator visually and clinically appraise the performance of the new method. These plots do not suffer the issues related to the Bland and Altman LoA when variances are unequal, and still are easy to interpret, in the spirit of Bland and Altman's LoA. The methodology requires several measurements with the reference standard and possibly only one with the new method for each individual. Actually, each individual may have a different number of repeated measurements by each method. It is applicable in all circumstances with or without differential and/or proportional bias and when the measurement errors are either homoscedastic or heteroscedastic.

2 The measurement error model

2.1 Formulation of the model

Consider the measurement error model:

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

$$x_{ij} \sim f_x(\mu_x, \sigma_x^2)$$

where y_{1ij} be the *j*th replicate measurement by method 1 on individual *i*, $j = 1,...,n_i$ and $i = 1,...,N$, 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 method 1 and 2. It is assumed that the variances of these errors, i.e. $\sigma_{\varepsilon_1}^2(x_{ij};\boldsymbol{\theta}_1)$ and $\sigma_{\varepsilon_2}^2(x_{ij};\boldsymbol{\theta}_2)$, 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 represent the true unknown but constant value of the trait for individual *i* and, therefore, $x_{ij} \equiv x_i$ (this assumption may be relaxed, see discussion).

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

$$\begin{aligned} & y_{1ij} = \alpha_1 + \beta_1 x_i + \varepsilon_{1ij}, \quad \varepsilon_{1ij} \mid x_i \sim N(0, \sigma_{\varepsilon_1}^2(x_i;\boldsymbol{\theta}_1)) \\ & y_{2ij} = x_i + \varepsilon_{2ij}, \;\; \varepsilon_{2ij} \mid x_i \sim N(0, \sigma_{\varepsilon_2}^2(x_i;\boldsymbol{\theta}_2)) \\ & x_i \sim f_x(\mu_x, \sigma_x^2) \end{aligned} \tag{2}$$

Nawarathna and Choudhary[9] have considered a slightly more general model with method by subject interactions. This refinement is not necessary in our setting, as the focus is on identifying differential and proportional biases in order to recalibrate the new method, and these interactions are absorbed into the measurement error terms. Note that this measurement error model is slightly different from the classical measurement error model in that the heteroscedasticity depends on the latent trait and not on an observed average.[10]

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.

Nawarathna and Choudhary[9] estimate the parameters of this model by bivariate maximum likelihood. Their approach is complicated by the evaluation of the integrals in the marginal likelihood function and requires special numerical methods such as Laplace approximation or Gauss-Hermite quadrature. We have developed another more simple and expeditious way to estimate this model by a two-stage procedure, which performs effectively as demonstrated by the simulation study (see below).

2.2 Estimation of the model

In the first stage, instead of treating x_i as a nuisance parameter and integrating it out from joint likelihood function we estimate the regression model for y_{2ij} by marginal maximum likelihood accounting non-parametrically for the heteroscedasticity by allowing the variance of ε_{2ij} to be different for each decile of the empirical distribution of $\bar{y}_{2i}$ (i.e. the mean of the individual repeated measurements $\overline{y}_{2i}$ is used as a rough approximation to x_i). Then, we adopt an empirical Bayes approach to predict x_i by the mean of its posterior distribution (i.e. the mean of the conditional distribution of x_i given the vector $\mathbf{y}_{2i}$ of observations for individual *i* by method 2),[11] which is the best linear unbiased prediction (BLUP) for x_i:

$$\begin{aligned}\hat{x}_i &= E(x_i \mid \mathbf{y}_{2i}) \\ &= \int x_i \frac{f_{y_2}(\mathbf{y}_{2i} \mid x_i) f_x(x_i)}{\int f_{y_2}(\mathbf{y}_{2i} \mid x_i) f_x(x_i)\, dx_i} dx_i\end{aligned} \tag{3}$$

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.

When f_x is the normal density, then (3) is:

$$\hat{x}_i = \sigma_x^2 \boldsymbol{\iota}' \mathbf{V}_i^{-1} (\mathbf{y}_{2i} - \boldsymbol{\iota}\hat{\mu}_x) + \hat{\mu}_x \tag{4}$$

where $\boldsymbol{\iota}$ is a n_i vector of ones and $\mathbf{V}_i = \sigma_x^2 \boldsymbol{\iota\iota}' + diag(\sigma_{\varepsilon_2}^2(x_i;\boldsymbol{\theta}_2))$ is the variance covariance matrix of $\mathbf{y}_{2i}$.

Our estimate of the heteroscedasticity, however, is rough and it is desirable to get a smooth estimate to be able to compare the precision of each method, which does not depend on $\overline{y}_{2i}$ but rather on $\hat{x}_i$ the BLUP for x_i. Therefore, following a similar approach to that of Bland and Altman[2] we compute a smooth estimate of the (heterogeneous) variance of the measurement errors by regressing the absolute values of the residuals $\hat{\varepsilon}^*_{2ij}$, from the linear regression model $y_{2ij} = \alpha_2^* + \beta_2^* \hat{x}_i + \varepsilon^*_{2ij}$, on $\hat{x}_i$ by ordinary least squares (OLS):

$$|\hat{\varepsilon}^*_{2ij}| = \theta_2^{(0)} + \theta_2^{(1)} \hat{x}_i + \nu_{ij} \tag{5}$$

Under the normality assumption $|\varepsilon^*_{2ij}|$ follows a half-normal distribution with mean $E(|\varepsilon^*_{2ij}|) = \sigma_{\varepsilon_2}(\hat{x}_i;\theta_2)\sqrt{2/\pi}$. Therefore, a smooth standard deviation estimate is obtained as:

$$\hat{\sigma}_{\varepsilon_2}(\hat{x}_i;\hat{\theta}_2) = \hat{E}(|\hat{\varepsilon}^*_{2ij}|)\sqrt{\pi/2} = (\hat{\theta}_2^{(0)} + \hat{\theta}_2^{(1)} \hat{x}_i)\sqrt{\pi/2} \tag{6}$$

The form of the heterogeneity need not be a straight line and a fractional polynomial may be used instead if the investigator believes that the straight line model is too restrictive.[11] In any case, a graphical representation of $|\hat{\varepsilon}^*_{2ij}|$ versus $\hat{x}_i$ provides a good start to visually check the plausibility of the straight line model. Finally, a scatter plot of y_{2ij} versus $\hat{x}_i$ with the estimated regression line and the 95% prediction limits computed as $\hat{\alpha}_2^* + \hat{\beta}_2^* \hat{x}_i \pm 2\hat{\sigma}_{\varepsilon_2}(\hat{x}_i;\boldsymbol{\theta}_2)$ may also be useful to assess the fit.

In the second stage, we proceed 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. A Wald test as well as 95% confidence intervals for α_1 and β_1 may be used to formally assess these biases. Again, we can compute a smooth estimate of the (heterogeneous) variance of the measurement errors by proceeding like before and estimating by OLS the model $|\hat{\varepsilon}^*_{1ij}| = \theta_1^{(0)} + \theta_1^{(1)}\hat{x}_i + \omega_{ij}$, where $|\hat{\varepsilon}^*_{1ij}|$ is the absolute value of the residuals $\hat{\varepsilon}^*_{1ij}$ from the linear regression model $y_{1ij} = \alpha_1^* + \beta_1^*\hat{x}_i + \varepsilon^*_{1ij}$.

Based on the estimates $\hat{\alpha}_1^*$ and $\hat{\beta}_1^*$ of the differential and proportional biases one can compute an estimate of the bias of the new method:

$$bias_i = \hat{\alpha}_1^* + \hat{x}_i(\hat{\beta}_1^* - 1) \tag{7}$$

A very useful figure to visualize the bias of the new method (i.e. method 1) is obtained by graphing a scatter plot of y_{1ij} and y_{2ij} versus the BLUP $\hat{x}_i$ along with the two regression lines and add a second scale on the right showing the relationship between the estimated amount of bias and $\hat{x}_i$, which we call a “bias plot”.

Simulations show that our methodology performs very well and one may obtain reasonably unbiased and consistent estimates of the differential α_1 and proportional β_1 biases, as well as of the (heterogeneous) measurement error variances already with sample sizes of 100 persons and 10 to 15 repeated measurements per individual from the reference method and only 1 measurement from the new method.

2.3 Recalibration of the new method

To remove the differential and proportional biases of the new method we proceed to its recalibration by computing $y^*_{1ij} = (y_{1ij} - \hat{\alpha}_1^*)/\hat{\beta}_1^*$. Now that y_{2ij} and y^*_{1ij} are on the same scale we can compare the variances of the measurement errors to determine which method is more precise. Before proceeding, one can check the quality of the recalibration by checking that the estimated intercept and slope of the linear regression model $y^*_{1ij} = \alpha_1^* + \beta_1^*\hat{x}_i + \varepsilon^*_{1ij}$ are zero and 1. As we would like to compare y_{2ij} with y^*_{1ij} (and not with y_{1ij}) it is advisable to recalculate a smooth estimate of the measurement errors variance of y^*_{1ij} by proceeding like before.

One can then proceed to the comparison of the variances by making a scatter plot of the estimated standard deviations $\hat{\sigma}_{\varepsilon_1}(\hat{x}_i;\boldsymbol{\theta}_1)$ and $\hat{\sigma}_{\varepsilon_2}(\hat{x}_i;\boldsymbol{\theta}_2)$ versus $\hat{x}_i$, which we call “precision plot”. It is possible that after recalibration the new method turns out to be more precise (locally or globally) than the reference standard.

2.4 Why Bland and Altman’s plot may be misleading

Bland and Altman have suggested to plot the differences $D_{ij} = y_{1ij} - y_{2ij}$ versus the averages $A_{ij} = (y_{1ij} + y_{2ij})/2$, and add to the plot the regression line of the relationship between D_{ij} and A_{ij} in addition to the LoA. The problem is that the regression line may show a positive or negative slope when there is no bias or have a zero slope in the presence of a bias. To see the reason, consider the linear regression of D_{ij} on A_{ij}, where from (1) we have

$D_{ij}=(\alpha_1-\alpha_2)+(\beta_1-\beta_2)x_i+\varepsilon_{1ij}-\varepsilon_{2ij}$ and $A_{ij}=(\alpha_1+\alpha_2)/2+(\beta_1+\beta_2)x_i/2+(\varepsilon_{1ij}+\varepsilon_{2ij})/2$ and after substitution of x_i:

$$D_{ij}=\underbrace{(\alpha_1-\alpha_2)-(\alpha_1+\alpha_2)\frac{(\beta_1-\beta_2)}{(\beta_1+\beta_2)}}_{a}+\underbrace{2\frac{(\beta_1-\beta_2)}{(\beta_1+\beta_2)}}_{b}A_{ij}\underbrace{-(\varepsilon_{1ij}+\varepsilon_{2ij})\frac{(\beta_1-\beta_2)}{(\beta_1+\beta_2)}+\varepsilon_{1ij}-\varepsilon_{2ij}}_{\varepsilon_{ij}} \tag{8}$$

Estimation of (8) by OLS generally provides biased estimates of a and b as:

$$\operatorname{cov}(A_{ij},\varepsilon_{ij})=\frac{1}{(\beta_1+\beta_2)}\left[\sigma^2_{\varepsilon_1}(x_i;\boldsymbol{\theta}_1)\beta_2-\sigma^2_{\varepsilon_2}(x_i;\boldsymbol{\theta}_2)\beta_1\right] \tag{9}$$

which is generally different from 0 and, therefore, A_{ij} cannot be considered as being exogenous it is, rather, endogenous.
OLS provides unbiased estimates only when:

$$\operatorname{cov}(A_{ij},\varepsilon_{ij})=0 \quad \Leftrightarrow \quad \frac{\sigma^2_{\varepsilon_1}(x_i;\boldsymbol{\theta}_1)}{\sigma^2_{\varepsilon_2}(x_i;\boldsymbol{\theta}_2)}=\frac{\beta_1}{\beta_2} \tag{10}$$

i.e. there is no bias whenever the variances of the measurement errors are strictly proportional to the proportional bias, a special condition that has little chance to truly hold in practice.
One can show that the OLS estimates of a and b are given by:

$$\begin{aligned} b_{OLS}&=2\frac{(\beta_1^2-\beta_2^2)\sigma_x^2+[\sigma^2_{\varepsilon_1}(x_i;\boldsymbol{\theta}_1)-\sigma^2_{\varepsilon_2}(x_i;\boldsymbol{\theta}_2)]}{(\beta_1+\beta_2)^2\sigma_x^2+[\sigma^2_{\varepsilon_1}(x_i;\boldsymbol{\theta}_1)+\sigma^2_{\varepsilon_2}(x_i;\boldsymbol{\theta}_2)]} \\ a_{OLS}&=(\alpha_1-\alpha_2)+(\beta_1-\beta_2)\mu_x-b_{OLS}\left(\frac{\alpha_1+\alpha_2}{2}+\frac{(\beta_1+\beta_2)}{2}\mu_x\right) \end{aligned} \tag{11}$$

Therefore, a zero slope occurs when there is no proportional bias and the measurement errors variances are strictly equal. However, whenever the variances are not equal a zero slope is possible in presence of a differential bias. Conversely, a non-zero slope is difficult to interpret in general and may mislead the investigator into believing that there is a proportional bias when actually the measurement error variances are different but there is truly no such bias. The same luckless situations may also occur with the differential bias.
Fortunately, the methodology we have developed does not suffer from these limitations and allows us to correctly identify and quantify the bias. We, therefore, suggest the use of a "bias plot" (see 2.2) to visualize the performance and the bias of the new method of measurement, as well as of a "precision plot" (see 2.3) to assess the precision of the new method relative to that of the standard.

3 A simulation study

We will demonstrate in this simulation study that our methodology to assess the biases, recalibrate the new method, and compare the precision of the two measurement methods performs very well for sample sizes of 100 individuals and between 10 to 15 measurements per individual by the reference standard and only one by the new method. We have deliberately decided to focus on the setting where one has only one measurement from the new method, which is an unfavorable data setting. However, the conclusions drawn carry over naturally to the more favorable case with repeated measurements from the new method to be evaluated.

For our simulations we considered the following data generating process:

$$\begin{aligned} & y_{1i} = -4 + 1.2x_i + \varepsilon_{1i}, \quad \varepsilon_{1i} \mid x_i \sim N(0,(1+0.1x_i)^2) \\ & y_{2ij} = x_i + \varepsilon_{2ij}, \;\; \varepsilon_{2ij} \mid x_i \sim N(0,(2+0.2x_i)^2) \\ & x_i \sim Uniform[10-40] \end{aligned} \tag{12}$$

where $i =,1,...,100$ and the number of repeated measurements of individual i from the reference standard was $n_i \sim Uniform[10-15]$. The new method has differential bias of -4 and a proportional bias of 1.2 . However, the variance of the measurement errors from method 1 is smaller than that of the reference method 2.

The Bland and Altman' LoA plot extended to the setting where there is heteroscedasticity of the measurement errors does not seem to indicate any bias (Figure 1):

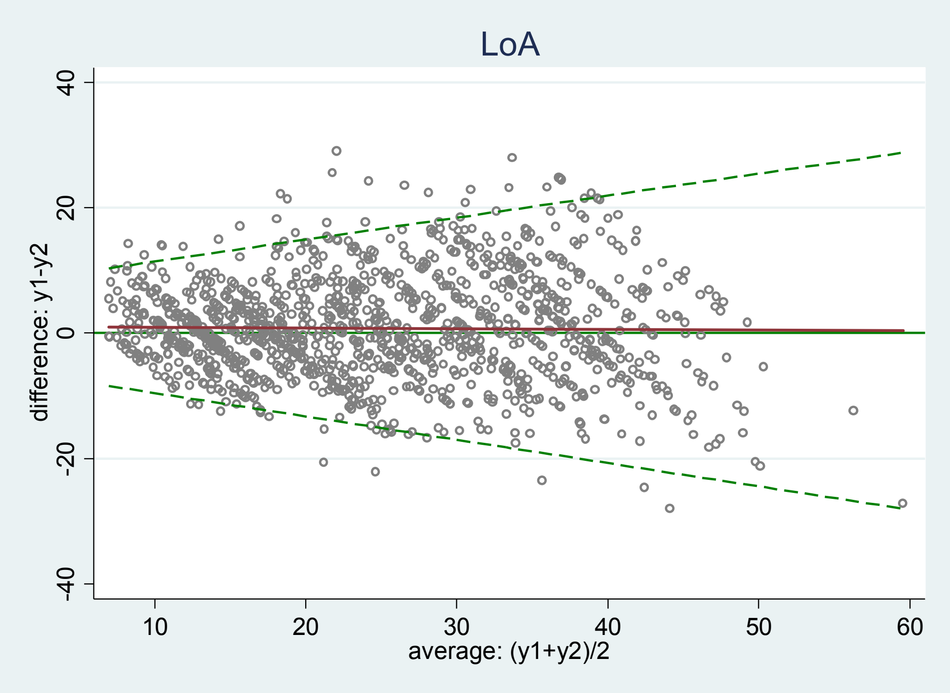

Figure 1. Bland and Altman' LoA plot when there is heteroscedasticity. The regression line does not seem to indicate any bias.

On the other hand, the bias plot (Figure 2) illustrates that the new method underestimates the trait up to 20 and then overestimate it gradually more and more, thereby clearly illustrating the occurrence of differential and proportional biases:

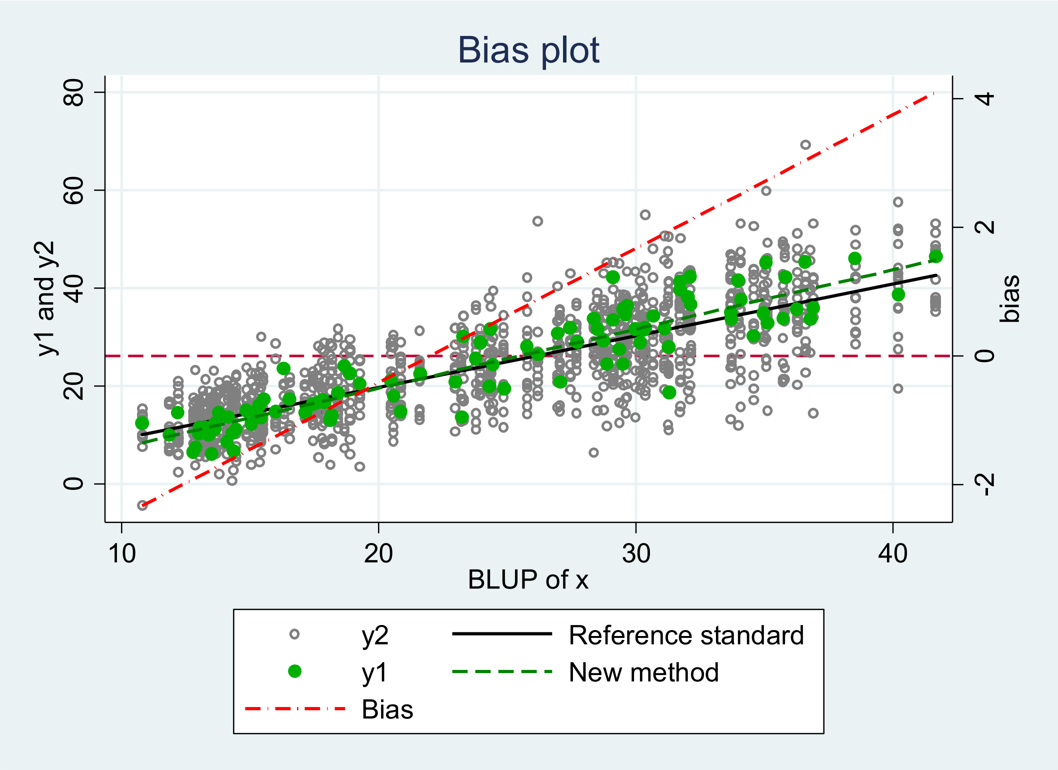

Figure 2. The bias plot shows the scatter plot of the two measurement methods y_{1ij} and y_{2ij} versus the BLUP $\hat{x}_i$ with the two regression lines added. The second scale on the right shows the relationship between the estimated amount of bias and the predicted value $\hat{x}_i$ (i.e. BLUP of x_i, the latent trait).

Estimation of the regression equation for y_{1ij} by OLS after having substituted the BLUP $\hat{x}_i$ for the true unmeasured trait x_i allowed us to identify a differential bias of -3.85 95%CI = [-6.81; -0.88] (true value is -4) and a proportional bias of 1.19 95%CI = [1.08; 1.29] (true value is 1.2). Based on 1000 simulations we found with our sample size coverage rates very close to nominal value for both parameters (97% for the differential bias and 95% for the proportional bias).

The precision plot before (Figure 3a) and after (Figure 3b) recalibration of the new method allows the comparison of the standard errors of the measurement errors of the two methods:

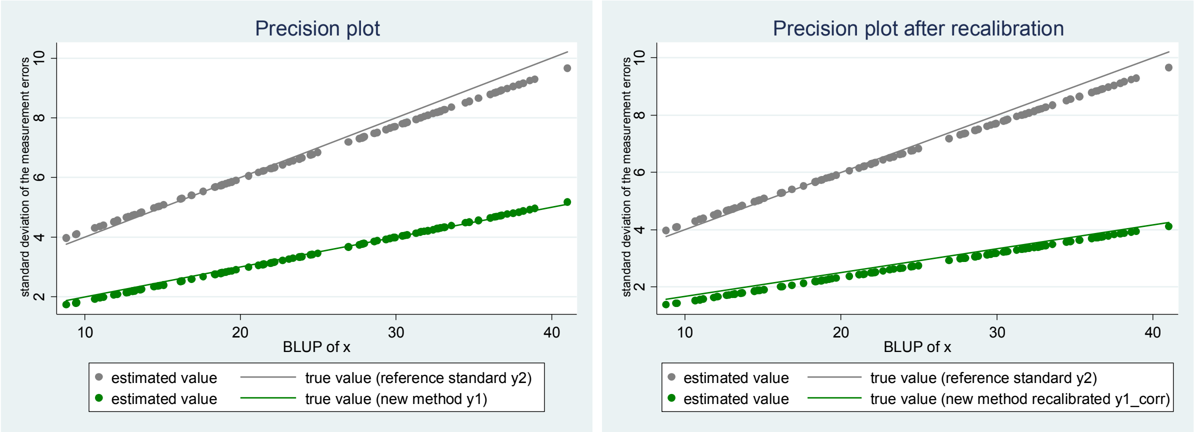

Figure 3. In these precision plots we have illustrated the true and estimated standard deviation of the measurement errors of the two methods before recalibration (left) and after (right).

Clearly, the estimation procedure for the variance of the measurement errors performs very well and despite the relatively small sample size the estimated standard deviations are very close to their true values. As is apparent, the recalibration slightly modifies the standard deviation of the new method. Again, based on 1000 simulations we found with our sample size coverage rates for the parameters of the heteroscedastic variances very close to nominal value.

We computed Bland and Altman' LoA plot for the recalibrated method (Figure 4) to illustrate that in the absence of bias the figure may mislead the reader into believing that there is a bias:

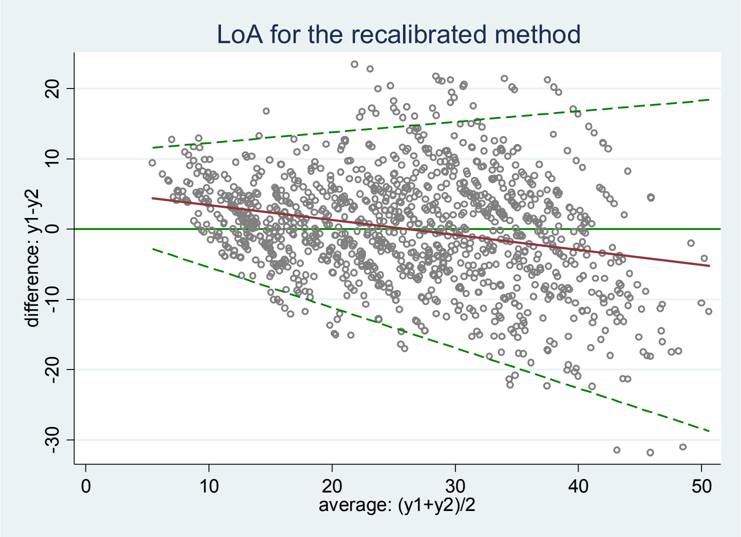

Figure 4. Bland and Altman' LoA plot after recalibration. The regression line seems to indicate a proportional bias when there is none.

Finally, to visualize the performance of our recalibration procedure we have represented the reference standard y_2, the new method y_1, and the recalibrated version y_{1_corr} in the following figure (Figure 5):

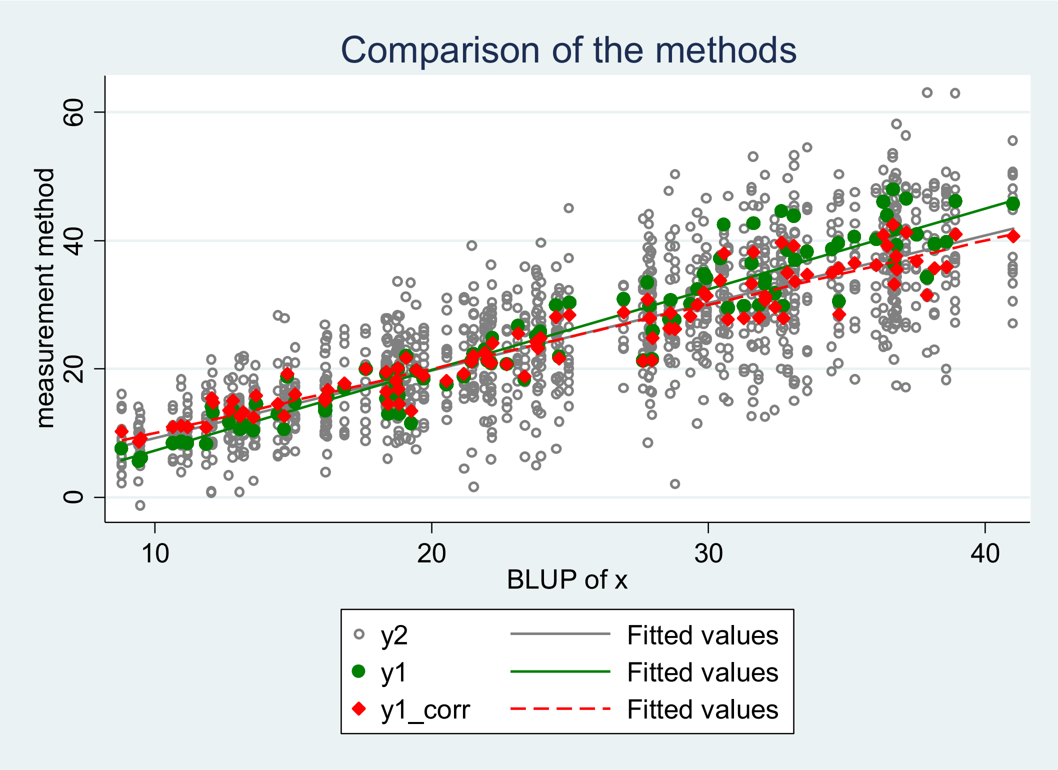

Figure 5. Comparison of the reference standard y_2 with the new method y_1 and its recalibrated version y_{1_corr}. Clearly, after recalibration the bias of method 1 has been eliminated.

4 A worked example

To illustrate our methodology we used the same data set on systolic blood pressure measurements as Bland & Altman in their 1999 paper[2]. Very briefly, three systolic blood pressure measurements were simultaneously made on 85 individuals by two observers (J and R) and an automatic blood pressure measuring machine (S). The measurements were repeated three times to provide three repeated values on each individual by each method. For our illustration, we will consider the measurements made by observer J as the reference standard and assess the performance of the automatic blood pressure measuring machine S.

Applying the proposed methodology to assess bias and precision of the automatic blood pressure measuring machine, with respect to the measurements made by observer J, we found a differential bias of 34.0 [mmHG], 95%CI = [23.5, 44.6], and a proportional bias of 0.86, 95%CI = [0.77, 0.94], whereas as the LoA plot seems to indicate only a differential bias (Figure 6):

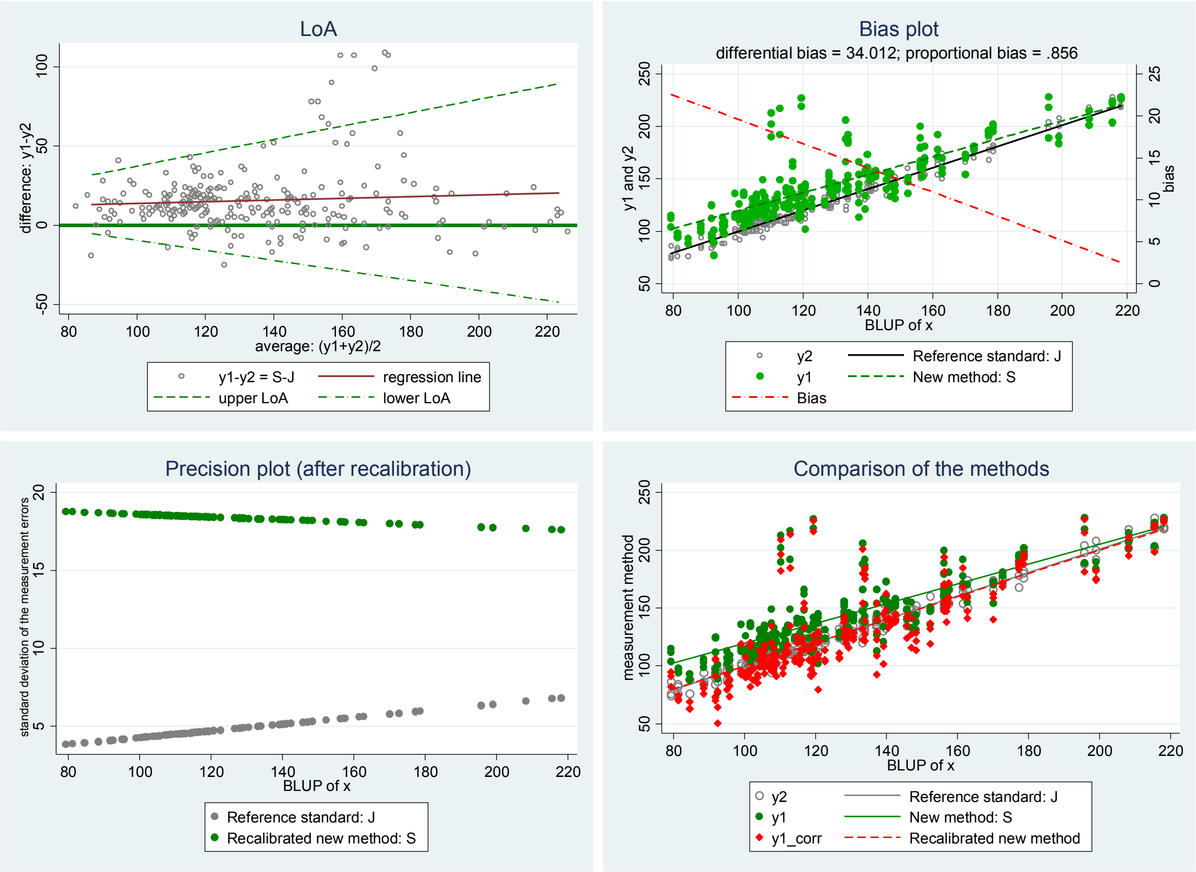

Figure 6. (top left) Bland and Altman' LoA plot, (top right) bias plot showing the amount of differential and proportional bias of the automatic blood pressure measuring machine, (bottom left) precision plot showing that the dispersion of the blood pressure measuring machine is much larger than that of observer J, (bottom right) scatter plot illustrating that the recalibration of the automatic machine's measurements (i.e. y_{1_corr}) is effective in removing bias.

The precision and comparison plots show that despite effective recalibration the blood pressure measuring machine still performs poorly in terms of precision (the measurement errors from machine S are much larger than those from observer J).

5 Discussion

We have developed a new estimation procedure to compare two quantitative measurement methods (one of which is the reference standard), which is widely applicable both when one has repeated measurements from the reference standard and possibly only one measurement per individual from the new method to be evaluated, and when measurement errors are homoscedastic or heteroscedastic. Our methodology circumvents the limitations of Bland and Altman's LoA methodology and allows to consistently quantify the amount of differential and proportional biases. We have proposed two new plots, the bias plot and the precision plot, which allow for the first to visualize the performance of the two methods as well as the bias of the new method on the second scale, and for the second to compare the precision of the two methods.

Our model of measurement is not new and was inspired by Nawarathna and Choudhary[9]. However, our estimation procedure is different from that adopted by these authors. They treated the latent variable as a nuisance parameter and integrated it out from the likelihood function. As a result their estimation procedure is complex and its implementation daunting.

We have, on the other hand, developed an estimation procedure based on an empirical Bayes approach, which proceeds in two steps. It has the advantage of being easier to implement, and works very well. Another advantage of our approach is that the BLUP of x leads naturally to the construction of the bias and precision plots. We have shown by simulations that it performs very well already with sample sizes of 100 individuals and 10 to 15 repeated measurements from the reference standard and only one measurement from the new method; estimated parameters as well as 95% prediction limits for both measurement methods all have proper coverage rates. When the sample size is increased to 300 individuals with 30 ~ 40 repeated measurements the estimated biases and curves are almost indistinguishable from their true values.

Additional simulations have shown that even with only 3 to 5 repeated measurements from the reference standard and only one measurement from the new method coverage rates are still very close to nominal value for the proportional and differential biases. However, estimation of the heteroscedasticity deteriorates. We would recommend having at least 8 to 12 repeated measurements from the reference standard to reliably assess the precision of the two measurement methods. Actually, it is important to have repeated measurements from the reference standard as our methodology relies essentially on the BLUP of x_i, whereas repeated measurements from the new method will increase precision of the estimated heteroscedastic relationship.

It is also interesting to note that using the mean $\bar{y}_{2i}$ of the repeated measurements instead of the BLUP $\hat{x}_i$ for predicting x_i the coverage rate of the confidence intervals for both the proportional and differential biases deteriorates; with our sample size it was only 80% for the differential bias and 75% for the proportional bias. With less than 10 to 15 repeated measurements it deteriorates even more dramatically (with 5 to 8 repeated measurements it was only 45% for the differential bias and 41% for the proportional bias, whereas it was still 95% respectively 93% with the BLUP $\hat{x}_i$). Simulations show that at least 45 to 50 repeated measurements per individual are required to have approximatively proper coverage rates when using the mean $\bar{y}_{2i}$ for predicting x_i. This is not surprising, particularly with unbalanced data, given that the BLUP methodology “borrows information” from the whole data set and not only from one individual.[12]

We have shown that the major drawback of Bland and Altman’s LoA methodology is that the regression of D versus A by OLS provides generally biased estimates as the variances of the measurement errors usually differ between the two methods and are unlikely to be strictly proportional to the proportional bias of the new method; that is A is endogenous and not exogenous. Therefore, the LoA plot may mislead the researcher into believing that there is a bias whenever there is none and conversely believe that there is no bias when actually there is truly a bias. The great advantage of our methodology is that the regression of y_{1ij} on the BLUP $\hat{x}_i$, as well as that of y_{2ij}, provide consistent estimates. Also, if the investigator believes that the straight line model (i.e. the regression of y_{1ij} on the BLUP $\hat{x}_i$) is not appropriate or wants to formally assess this assumption, a fractional polynomial may be used instead.[13]

Cartensen[8] argued that OLS prediction of method 2 from method 1 by the linear regression model $y_{2ij} = \alpha + \beta y_{1ij} + \varepsilon_{ij}$ was not appropriate since y_{1ij} was endogenous and the regression parameters α and β are biasedly estimated. For the purpose of identifying the biases we perfectly agree with him. However, when the goal is prediction it is not clear if the endogeneity of y_{1ij} really poses a problem. To investigate this issue we performed the following simulation. We generated a sample of 50000 observations according to equations

(2) with $\varepsilon_{1ij} \sim N(0,64)$, $\varepsilon_{2ij} \sim N(0,16)$, $\alpha_1 = 0$, and $\beta_1 = 1$. We used the first 25000 observations to estimate by OLS the linear regression model $y_{2ij} = \alpha + \beta y_{1ij} + \varepsilon_{ij}$. Then, we computed the predicted values $\hat{y}_{2ij}$ (OLS prediction). Following Cartensen[8], we also estimated the regression of differences on averages by OLS and used the estimated coefficients to predict y_{2ij} (LoA prediction) according to equation (3) in his paper. Then, for the last 25000 observations, we computed the coverage rates of the two prediction methods and found nominal coverage rates of 95% for both methods. However, when restricting the prediction to the subsample of values of y_{1ij} smaller than 0, the coverage rate of OLS prediction was still 95%, whereas that of Cartensen's method (i.e. LoA prediction) dropped down to 82% (Figure 7 left). This result is not entirely surprising as the condition specified by equation (10) for the OLS to provide unbiased estimates when regressing the differences on averages was violated. Now, by simulating with $\beta_1 = 64/16$ condition (10) is verified and OLS regression of differences on averages provides unbiased estimates. In that case, the coverage rate is approximately 95% for both OLS prediction and LoA prediction even when restricting to the subsample of values of y_{1ij} smaller than 40 (Figure 7 right):

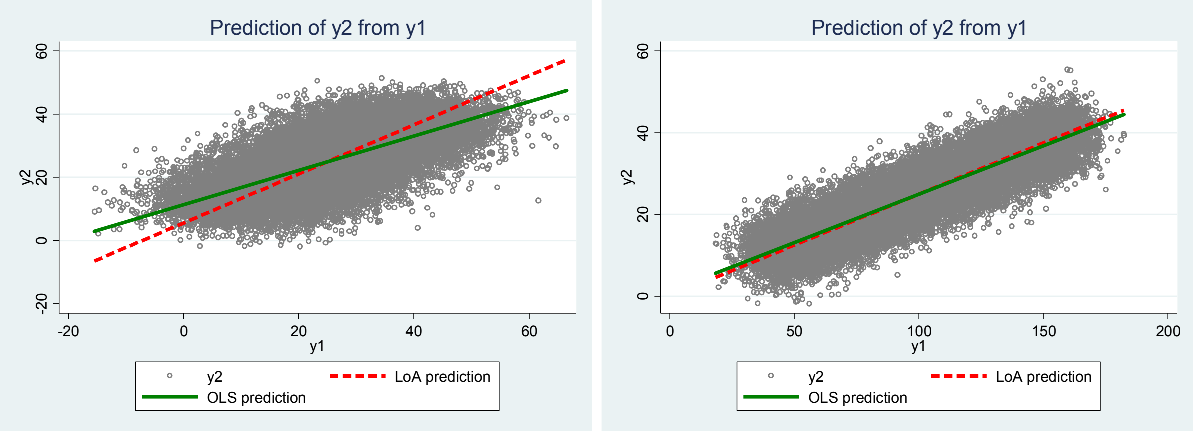

Figure 7. OLS prediction of method 2 from method 1 by simple linear regression (OLS prediction) and by Cartensen's method (LoA prediction). Clearly the two methods differ when condition (10) is not verified (left), whereas this is no more the case when it holds (right).

In sum, when condition (10) is verified both methods seem to be equivalent as they provide almost undistinguishable predictions, whereas when this is not the case OLS prediction seems to perform best, thereby challenging Cartensen's conclusions.

We have investigated the sensitivity of the BLUP for x_i to the heteroscedasticity by pretending that the variance of ε_{2ij} was constant. It turned out that not accounting for the heteroscedasticity of ε_{2ij} did not affect the BLUP of x_i as its value was only very slightly modified. One possible explanation is that the first term in V_i dominates the second when computing the inverse. Also, the empirical Bayes approach seems to be quite robust to distributional assumptions regarding the latent trait. Indeed, in our simulations we computed the BLUP of x_i under the assumption that f_x was the normal density, whereas x_i was actually drawn in a uniform distribution. Nevertheless, a scatter plot of x_i versus $\hat{x}_i$ as well as the results of our simulations illustrate that this does not to introduce any important bias. Actually, Jiang[14] has shown that with sufficient repeated observations per individual and sufficient number of individuals the empirical distribution of the BLUP of x_i (computed under the multivariate normal distribution of the random effects and errors) will converge to its true distribution.

We have made the simplifying assumption of a constant latent trait value for each individual, i.e. $x_{ij} \equiv x_i$. This assumption may easily be relaxed if, for example, one expects a trend in the latent trait, as may be the case in a longitudinal design. In that case, one may specify $\mu_x(t) = \alpha + \beta t$ and $x_{it} = (\alpha + u_i^{(1)}) + (\beta + u_i^{(2)})t$ with $u_i^{(j)} \sim N(0, \sigma_j^2)$. The measurements from the new method need not be taken at the same time as for the reference standard as the trend depends on the follow-up time *t*.

We have implemented in gllamm the approach advocated by Dunn[10] when error variances are heteroscedastic (chapter 4.8 and appendix 3). Unfortunately, in our setting gllamm was extremely slow and failed to converge when heteroscedastic errors where allowed (despite allowing up to 60 quadrature points). Also, we were unable to specify a model which allowed the heteroscedasticity to depend on the latent trait instead of the average of the repeated measurements. Dunn[10] assumes the normality of the distribution of the latent trait, whereas with our methodology it is not necessary, as discussed above. It is unclear, however, what are the consequences on the estimates with gllamm when the normality assumption is not met. We have, therefore, proposed a different estimation procedure which seems to be quite robust with repeated measurements from the reference standard whatever the distribution of the latent trait.

Finally, extensive simulations show that our methodology still performs very well when the amount of differential and proportional biases, as well as the form of the heterogeneity, are varied.

In summary, we have developed a new estimation procedure to assess bias and precision of a quantitative measurement method relative to the reference standard, which is simple to implement and performs very well even when the measurement errors are heteroscedastic. We also have proposed two new plots, the bias and precision plots, to help the investigator visually and clinically appraise the performance of the new method. These plots do not have the shortcomings of Bland and Altman's LoA and still are in spirit of the original paper. We are currently developing a Stata package (as well as an R package), which will be submitted to the Stata Journal (resp. R Journal), that implements this methodology (as well as Bland and Altman's LoA extended to the case of repeated measurements and heteroscedasticity[2]).

References

1. Bland JM and Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1**: 307-310.
2. Bland JM and Altman DG. Measuring agreement in method comparison studies. *Stat Meth Med Res* 1999; **8**: 135-160.
3. Ludbrook J. Confidence in Altman-Bland plots: a critical review of the method of differences. *Clin Exp Pharmacol P* 2010; **73**: 143-149.
4. Hopkins WG. Bias in Bland-Altman but not regression validity analyses. *Sportscience* 2004; **8**: 42-46.
5. Krouwer JS. Why Bland-Altman plots should use X, not (Y+X)/2 when X is a reference method. *Stat Med* 2008; **27**: 778-780.
6. Ludbrook J. Linear regression analysis for comparing two measurers or methods of measurement: but which regression ? *Clin Exp Pharmacol P* 2010; **37**: 692-699.

7. Carstensen B, Simpson J, and Gurrin LC. Statistical models for assessing agreement in method comparison studies with replicate measurements. *Int J Biostat* 2008; **4**: Article 16.
8. Carstensen B. Comparing methods of measurement: Extending the LoA by regression. *Stat Med* 2010, **29**: 401-410.
9. Nawarathna LS and Choudhary PK. A heteroscedastic measurement error model for method comparison data with replicate measurements. *Stat Med* 2015; **34**: 1242-1258.
10. Dunn G. *Statistical evaluation of measurement errors: design and analysis of reliability studies* 2nd ed. Arnold, London, 2004.
11. Verbeke G and Molenberghs G. *Linear mixed models in practice*. Lecture Notes in Statistics 126, Springer, 1997.
12. Robinson JK. That BLUP is a Good Thing: The Estimation of Random Effects. *Stat Sci* 1991; **6**: 15-32.
13. Royston P and Altman DG. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. *J Roy Stat Soc C-App* 1994; **43**: 429-467.
14. Jiang J. Asymptotic properties of the empirical BLUP and BLUE in mixed linear models. *Stat Sinica* 1998; **8**: 861-885.