Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2019, Vol. 28(8) 2557–2565
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280218759693
journals.sagepub.com/home/smm
SAGE

MethodCompare: An R package to assess bias and precision in method comparison studies

Patrick Taffé,[1,2] Mingkai Peng,[3,4] Victoria Stagg[5] and Tyler Williamson[3,4,6]

Abstract

Bland and Altman's limits of agreement have been used in many clinical research settings to assess agreement between two methods of measuring a quantitative trait. However, when the variances of the measurement errors of the two methods are different, limits of agreement can be misleading. MethodCompare is an R package that implements a new statistical methodology, developed by Taffé in 2016. MethodCompare produces three new plots, the "bias plot", the "precision plot", and the "comparison plot" to visually evaluate the performance of the new measurement method against the reference method. The method is illustrated on three simulated examples. Note that the Taffé method assumes that there are several measurements from reference standard and possibly as few as one measurement from the new method for each individual.

Keywords

MethodCompare, biasplot, limits of agreement, differential bias, proportional bias, Bland-Altman plot, method comparison, measurement, empirical Bayes, BLUP

1 Introduction

In clinical research, Bland and Altman's limits of agreement (LoA) are frequently used to evaluate the agreement between two methods for measuring quantitative traits.[1] Often this is motivated by a new, perhaps less expensive or easier, method of measurement against an established reference standard. To evaluate the comparability of the methods, the investigator collects measurements, perhaps one or several, from each method for a set of subjects. Bland and Altman's LoA are then computed by adding and subtracting 1.96 times the estimated standard deviation to the mean differences. A scatter plot of the differences versus the means of the two variables with the LoA superimposed is used to visually appraise the degree of agreement and quantify its magnitude. Further, a regression of the differences as a function of the means is added to the plot to indicate whether there is a bias and the direction of that bias.[2]

Bland and Altman's plot may be misleading, however, in situations where the variances of the measurement errors for each method differ from one another. 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. This problem has been identified in the literature previously but, to the best of our knowledge, no simple to use and effective plots that evaluate bias and precision have been presented as an alternative.[3–7] Interested readers should look at Nawarathna and Choudhary[8] and the references therein for a recent review of the literature on measurement errors.

[1]Institute of Social and Preventive Medicine (IUMSP), University of Lausanne, Switzerland
[2]Faculty of Biology and Medicine, University of Lausanne, Switzerland
[3]Department of Community Health Sciences, University of Calgary, Canada
[4]O'Brien Institute for Public Health, University of Calgary, Canada
[5]Calgary Statistical Support, Calgary, Canada
[6]Alberta Children's Hospital Research Institute, University of Calgary, Canada

Corresponding author:
Patrick Taffé, Institute for Social and Preventive Medicine, University of Lausanne, Biopôle 2, Route de la Corniche 10, Lausanne 1010, Switzerland.
Email: patrick.taffe@chuv.ch

The MethodCompare R package implements a new method to identify and quantify the amounts of differential and proportional biases in the setting of heteroscedastic measurement errors, particularly when heteroscedasticity is a function of the latent trait. This method, first introduced by Taffé in 2016,[9] requires that several measurements are made with the reference standard for each individual (usually more than 5) and possibly only one with the new method. It allows each individual to have a different number of repeated measurements by each method, and is applicable in all circumstances with or without differential and/or proportional bias, and when the measurement errors are either homoscedastic or heteroscedastic. The implementation of the Taffé method in Stata Statistical Software has been published elsewhere.[10]

2 The measurement error model

2.1 Formulation of the model

A full presentation of the methodological theory can be found elsewhere.[9] Presented below is an abridged version of the methods. Consider the measurement error model

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

where y_{1ij} is the jth replicate measurement by method 1 on individual i, $j = 1, \dots, n_i$ and $i = 1, \dots, N$, 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, i.e. $\sigma^2_{\varepsilon_1}(x_{ij}; \boldsymbol{\theta}_1)$ 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, $\alpha_2 = 0$ 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^2_x. When method 2 is the reference standard and method 1 the new method to be evaluated, the model reduces to

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

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.[11]

2.2 Estimation of the model

The estimation process is completed in two distinct steps.

2.3 Estimation Step 1

Other methods treat x_i as a nuisance parameter and attempt to integrate it out from the joint likelihood function. The Taffé method estimates the regression model for y_{2ij} using marginal maximum likelihood allowing the variance of ε_{2ij} to be different for each decile of the empirical distribution of $\overline{y}_{2i}$ (i.e. the mean of the individual repeated measurements $\overline{y}_{2i}$ is used as a rough approximation to x_i). Then, following an empirical Bayes approach the x_i's are predicted from 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), which is the best linear unbiased prediction (BLUP) for x_i

$$\hat{x}_i = E(x_i|\mathbf{y}_{2i}) = \int x_i \frac{f_{y_2}(\mathbf{y}_{2i}|x_i) f_x(x_i)}{\int f_{y_2}(\mathbf{y}_{2i}|x_i) f_x(x_i) \mathrm{d}x_i} \mathrm{d}x_i \tag{3}$$

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 equation (3) is

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

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

It is desirable to have a smooth estimate of the heteroscedasticity, which does not depend on $\overline{y}_{2i}$ but rather on $\hat{x}_i$, the BLUP for x_i. Therefore, Taffé[9] suggests an approach similar to that of Bland and Altman[2] 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) to create a smooth estimate of the heterogeneous variance

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

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

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

Note that Taffé suggests that the form of the heterogeneity need not be a straight line and other heterogeneity structures such as nonlinear may be considered. 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. Further, 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.

2.4 Estimation Step 2

The second stage of the estimation process involves the estimation of the regression equation for y_{1ij} in equation (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. The Wald test and 95% confidence intervals for α_1 and β_1 may then be used to formally assess these biases. As before, a smooth estimate of the variance by using OLS to estimate 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 $y_{1ij} = \alpha_1^* + \beta_1^* \hat{x}_i + \varepsilon_{1ij}^*$. Then, based on the estimates $\hat{\alpha}_1^*$ and $\hat{\beta}_1^*$, the bias of the new method can be estimated as

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

To visually assess the degree of bias, the "bias plot" 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 while adding a second scale on the right showing the relationship between the estimated amount of bias and $\hat{x}_i$.

Taffé[9] shows, by simulation, that the methodology performs very well and that the estimates of the differential α_1 and proportional β_1 biases are reasonably unbiased and consistent already for sample sizes of 100 persons with 3–5 repeated measurements per individual from the reference method and with only one measurement from the new method. However, to appropriately estimate the (heterogeneous) measurement error variances, one should better have 10–15 repeated measurements per individual from the reference method and one or several measurements from the new method.

2.5 Recalibration of the new method

To remove the differential and proportional biases of the new method, we proceed to its recalibration by computing $y_{1ij}^* = (y_{1ij} - \hat{\alpha}_1^*)/\hat{\beta}_1^*$. The "comparison plot" allows the visualization of the performance of the recalibration procedure.

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. 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 the "precision plot". It is possible that after recalibration the new method turns out to be more precise (locally or globally) than the reference standard.

3 The MethodCompare package

The MethodCompare package implements the method described above and produces the bias plot, the precision plot, and comparison plot. It also allows the computation of the extended version of Bland and Altman's LoA when variances of measurement errors are possibly heteroscedastic. We illustrate the use of the MethodCompare package based on three simulated datasets. The MethodCompare package depends on the package nlme[12] to compute the BLUP of x.

3.1 Example 1

We simulated the dataset 1 with the following data generation process

$$y_{1i} = -4 + 1.2x_i + \varepsilon_{1i}, \varepsilon_{1i}|x_i \sim N\big(0, (1 + 0.1x_i)^2\big)$$
$$y_{2ij} = x_i + \varepsilon_{2ij}, \varepsilon_{2ij}|x_i \sim N\big(0, (2 + 0.2x_i)^2\big)$$
$$x_i \sim Uniform[10 - 40]$$

where $i = , 1, \ldots, 100$ and the number of repeated measurements of individual i from the new measurement method and the reference standard was $n_{1i} = 1$ and $n_{2i} \sim Uniform[10 - 15]$, respectively.

In example 1, there are between 10 and 15 repeated measurements by the reference standard and only one by the new measurement method for each individual. The differential and proportional biases are -4 and 1.2, respectively. The standard deviation of the measurement errors is heteroscedastic for both measurement methods and increases with the level of the underlying true latent trait. However, the dispersion of the reference standard is twice that of the new measurement method.

```
> library(MethodCompare)
> data(data1)
> head(data1)

  id   y1         y2
1 1    14.39119   6.81835
2 1    NA         17.48666
3 1    NA         13.69647
4 1    NA         22.77729
5 1    NA         11.54489
6 1    NA         13.51292
```

First, we estimate the differential and proportional biases.

```
> measure_model <- measure_compare(data1)
> measure_model$Bias

                   Estimate    2.5%       97.5%
Differential bias  -3.166945  -5.305945   -1.027940
Proportional bias  1.197453    1.108928   1.285978
```

The estimated differential bias is -3.17 95% CI = $[-5.31, -1.03]$ and proportional bias 1.2 95% CI = [1.11, 1.29]. These values are not too far and the confidence intervals cover the true values.

The LoA plot, bias plot, precision plot and comparison plot can be produced with the following commands.

```
> bland_altman_plot(data1,new="y1",Ref="y2",ID="id",fill=TRUE)
> bias_plot(measure_model)
> precision_plot(measure_model)
> compare_plot(measure_model)
```

In Figure 1, the LoA plot indicates that there is a slight positive bias of the new measurement method for low values of the estimated latent trait level (i.e. BLUP of x) and a negative bias for high values. On the contrary,

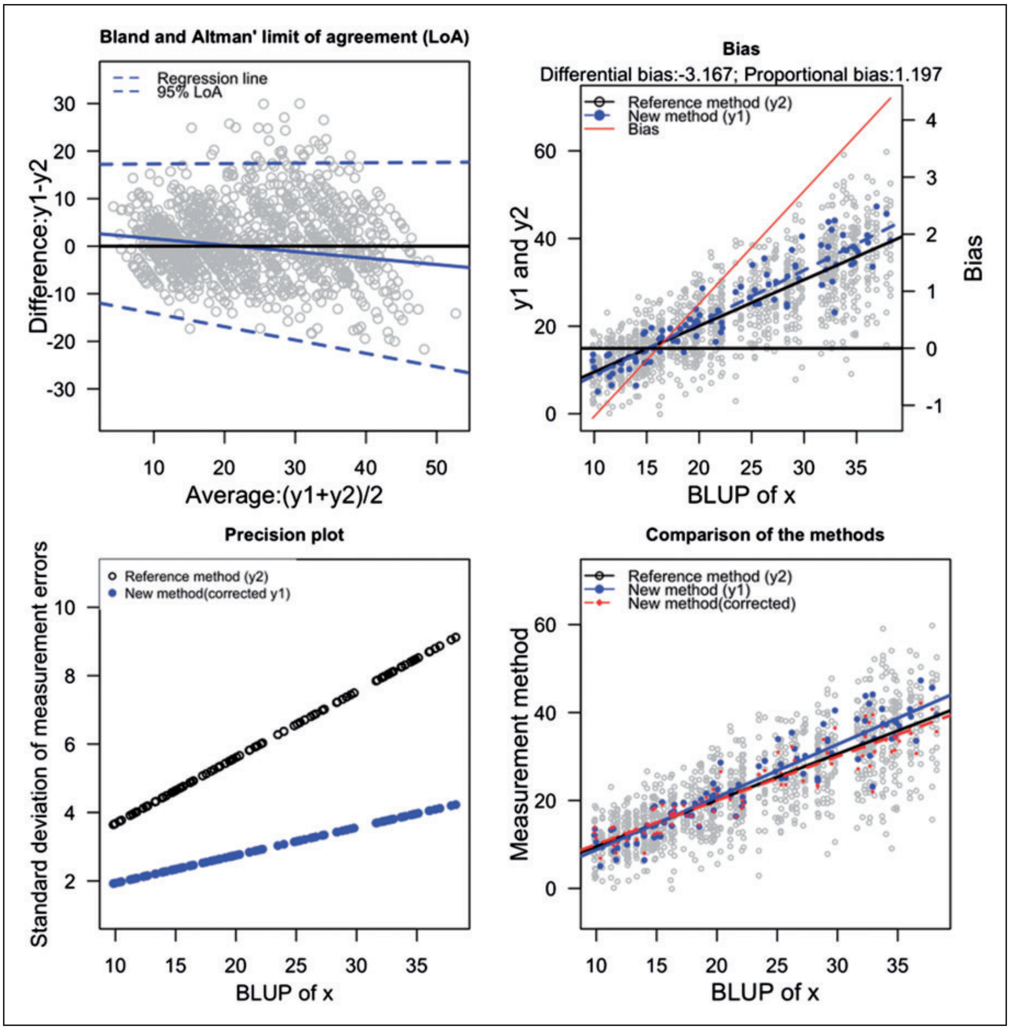

Figure 1. (Top left) Bland and Altman's LoA plot, (top right) bias plot showing the amount of bias of the new measurement method, (bottom left) precision plot showing the precision (i.e. standard deviation of the measurement error) of each measurement method, (bottom right) scatter plot illustrating that the recalibration of the new measurement method (i.e. corrected) was effective.

the bias plot illustrates that the bias is negative for low values and positive for high values. The precision plot shows that after recalibration the new measurement method is about twice as precise as the reference standard, and that both measurement methods are more precise for lower values than for higher values of the true latent trait. One can see on the comparison plot that recalibration of the new measurement method was effective in removing bias.

3.2 Example 2

Example 2 is similar to example 1 except that there are between 1 and 5 repeated measurements (instead of just 1) by the new measurement method for each individual. The dataset was simulated based on the following data generation process

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

where $i = ,1,\ldots,100$ and the number of repeated measurements of individual i from the new measurement method and the reference standard was $n_{1i} \sim Uniform[1-5]$ and $n_{2i} \sim Uniform[10-15]$, respectively.

```
> measure_model <- measure_compare(data2)
> measure_model$Bias

                   Estimate       2.5%       97.5%
Differential bias  -3.423592   -5.015038   -1.832140
Proportional bias   1.187037    1.127484    1.246590
```

The estimated differential bias is -3.42 95% CI $= [-5.01, -1.83]$ and proportional bias to 1.19 95% CI $= [1.13, 1.25]$. Compared with example 1 with only measurement by the new method, estimate of differential bias in example 2 is closer to the truth and their confidence intervals are narrower.

```
> bland_altman_plot(data2,new = "y1",Ref = "y2",ID = "id",fill = TRUE)
> bias_plot(measure_model)
```

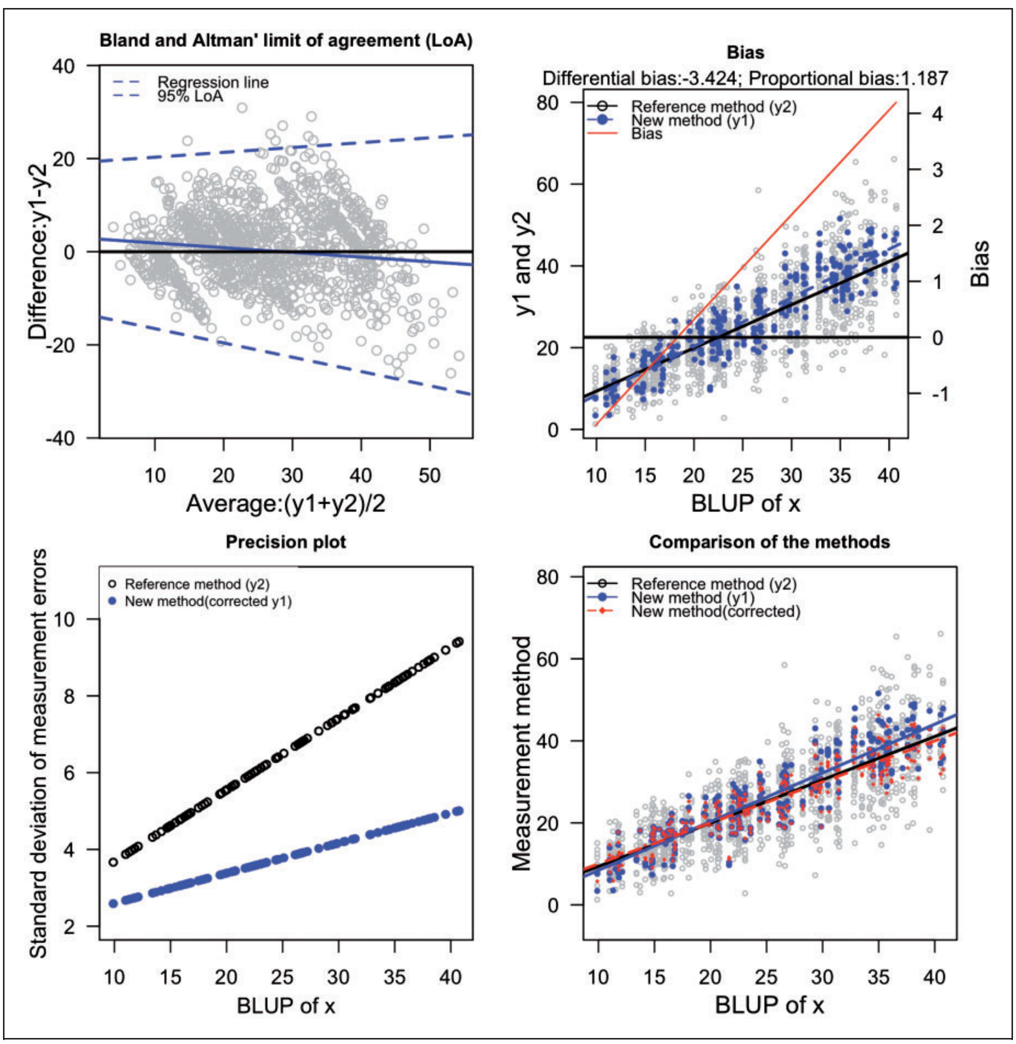

Figure 2. (Top left) Bland and Altman's LoA plot, (top right) bias plot showing the amount of bias of the new measurement method, (bottom left) precision plot showing the precision (i.e. standard deviation of the measurement error) of each measurement method, (bottom right) scatter plot illustrating that the recalibration of the new measurement method (i.e. corrected) was effective.

```
> precision_plot(measure_model)
> compare_plot(measure_model)
```

Consistent with the results for example 1, the LoA plot indicates that there is a slight positive bias of the new measurement method for low values of the estimated latent trait level (i.e. BLUP of x) and a negative bias for high values. In contrast, the bias plot (correctly) illustrates that the bias is negative for low values and positive for high values (Figure 2). The conclusions for the precision and comparison plots are the same as for example 1.

3.3 Example 3

In example 3, there are between 10 and 15 repeated measurements by the reference standard and between 1 and 5 repeated measurements by the new measurement method for each individual. The differential bias amounts to 3 and proportional bias to 0.9. The standard deviation of the measurement errors is heteroscedastic for both measurement methods and increases with the level of the underlying true latent trait. However, the dispersion of the reference standard is much lower than that of the new measurement method.

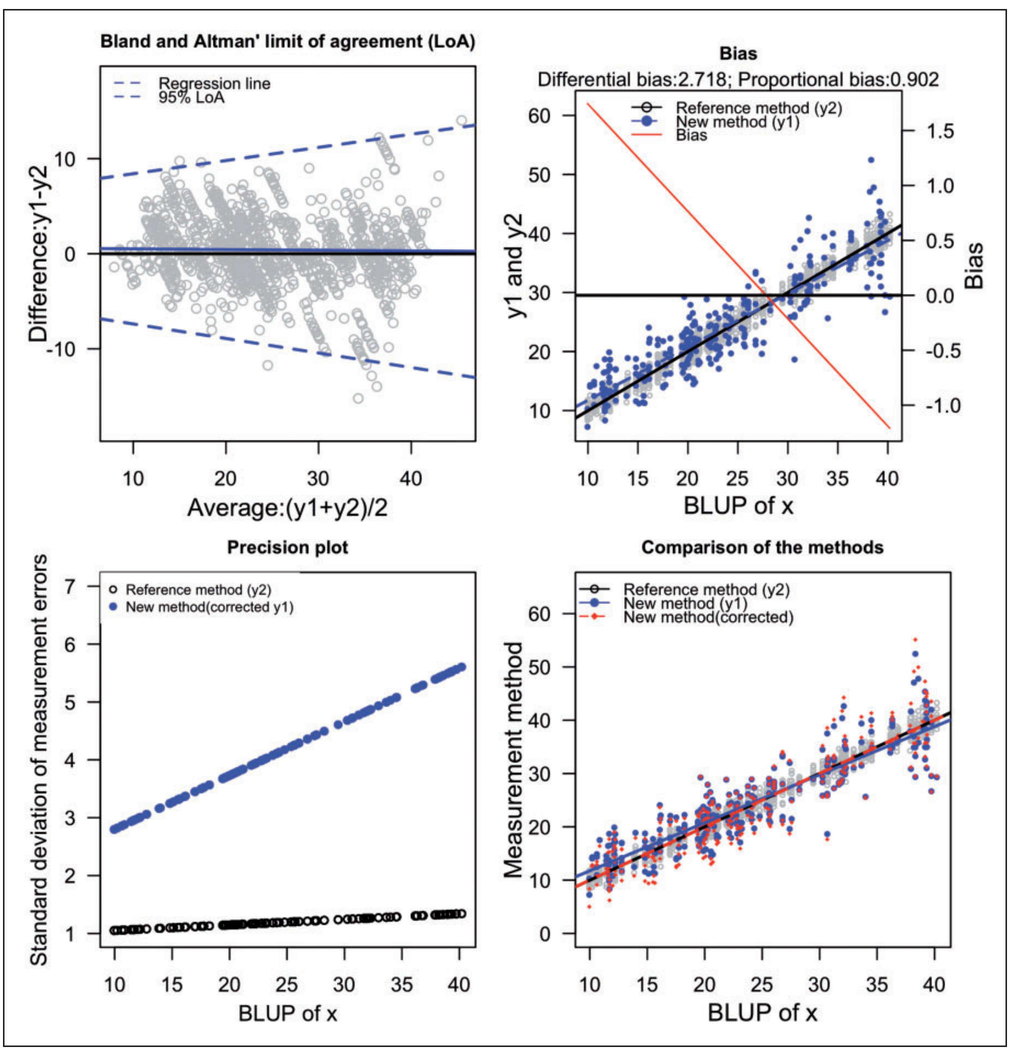

Figure 3. (Top left) Bland and Altman's LoA plot, (top right) bias plot showing the amount of bias of the new measurement method, (bottom left) precision plot showing the precision (i.e. standard deviation of the measurement error) of each measurement method, (bottom right) scatter plot illustrating that the recalibration of the new measurement method (i.e. corrected) was effective.

The dataset was generated based on following process

$$y_{1i} = 3 + 0.9x_i + \varepsilon_{1i}, \varepsilon_{1i}|x_i \sim N\big(0, (2 + 0.06x_i)^2\big)$$
$$y_{2ij} = x_i + \varepsilon_{2ij}, \varepsilon_{2ij}|x_i \sim N\big(0, (1 + 0.01x_i)^2\big)$$
$$x_i \sim Uniform[10 - 40]$$

where $i = , 1, \ldots, 100$ and the number of repeated measurements of individual i from the new measurement method and the reference standard was $n_{1i} \sim Uniform[1 - 5]$ and $n_{2i} \sim Uniform[10 - 15]$, respectively.

```
> measure_model <- measure_compare(data3)
> measure_model$Bias

                      Estimate      2.5%        97.5%
Differential bias     2.7175510    1.3701808    4.064921
Proportional bias     0.9023004    0.8491638    0.955437
```

The estimated differential bias is 2.72 95% CI = [1.37, 4.06] and the estimated proportional bias is 0.90 95% CI = [0.85, 0.96]. The estimated values for differential and proportional bias are close to the true values.

```
> bland_altman_plot(data3,new = "y1",Ref = "y2",ID = "id",fill = TRUE)
> bias_plot(measure_model)
> precision_plot(measure_model)
> compare_plot(measure_model)
```

The plots are shown in Figure 3. The LoA plot does not indicate any bias of the new measurement method, whereas the bias plot illustrates that the bias is positive for low values and negative for high values. The precision plot shows that after recalibration the new measurement method is clearly less precise than the reference standard. Of interest is to note that the dispersion of measurement errors of the reference standard is almost constant throughout the whole range of the latent trait, whereas that of the new method is, by comparison, sharply increasing. Again, the recalibration of the new measurement method was very effective in removing bias, as illustrated by the comparison plot.

4 Discussion

Using simulated data where the relationship between the true latent trait and the two measurement methods is known, we have illustrated that the new method was effective in removing existing bias of the new measurement method, and in assessing after recalibration the precision of the two measurement methods. We also have illustrated that there are settings where Bland and Altman's LoA methodology is misleading, whereas MethodCompare package allows one to properly identify, quantify, and correct for any biases.

5 Summary

This article describes the MethodCompare package to assess the bias and precision in method comparison studies. It implemented a new method developed by Taffé. The package includes the functions for estimation of the differential and proportion bias and plots for comparison between the new method and reference method.

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.

ORCID iD

Tyler Williamson http://orcid.org/0000-0001-5029-2345

References

1. Bland JM and Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **327**: 307–310.
2. Bland JM and Altman DG. Measuring agreement in method comparison studies. *Stat Meth Med Res* 1999; **8**: 135–160.
3. Hopkins WG. Bias in Bland–Altman but not regression validity analyses. *Sportscience* 2004; **8**: 42–46.
4. Krouwer JS. Why Bland–Altman plots should use x, not $(y+x)/2$ when x is a reference method. *Statistics in medicine* 2008; **27**: 778–780.
5. 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.
6. Ludbrook J. Confidence in Altman–Bland plots: a critical review of the method of differences. *Clin Exp Pharmacol Physiol* 2010; **37**: 143–149.
7. Carstensen B. Comparing methods of measurement: extending the LoA by regression. *Stat Med* 2010; **29**: 401–410.
8. Nawarathna LS and Choudhary PK. A heteroscedastic measurement error model for method comparison data with replicate measurements. *Stat Med* 2015; **34**: 1242–1258.
9. Taffé P. Effective plots to assess bias and precision in method comparison studies. *Stat Meth Med Res* 2016 (in press). DOI 101177/0962280216666667.
10. Taffé P, Peng M, Stagg V, et al. biasplot: a package to effective plots to assess bias and precision in method comparison studies. *Stata J* 2017; **17**: 208–221.
11. Dunn G. *Statistical evaluation of measurement errors: design and analysis of reliability studies*, 2nd ed. London: Arnold, 2004.
12. Pinheiro J, Bates D, DebRoy S, et al. *nlme: linear and nonlinear mixed effects models*, 2016. URL http://CRAN.R-project.org/package=nlme. R package version 3.1-128.