RESEARCH ARTICLE

Bayesian linear regression model for method comparison studies

S.M.M. Lakmali*, L.S. Nawarathna and P. Wijekoon

Highlights

- The proposed Bayesian linear regression model has higher accuracy than other existing models.
- It requires less time for model fitting and performed well, specifically with small sample sizes.
- The proposed model is used to develop a methodology for agreement evaluation between two methods.
- This methodology can be used for both balanced and unbalanced method comparison data designs.

Ceylon Journal of Science 51(1) 2022: 37-42
DOI: http://doi.org/10.4038/cjs.v51i1.7977

RESEARCH ARTICLE

Bayesian linear regression model for method comparison studies

S.M.M. Lakmali[1,*], L.S. Nawarathna[2] and P. Wijekoon[2]

[1]*Postgraduate Institute of Science, University of Peradeniya, Peradeniya, Sri Lanka.*
[2]*Department of Statistics and Computer Science, University of Peradeniya, Peradeniya, Sri Lanka.*

Received: 03/08/2021; Accepted: 01/30/2022

Abstract: Method comparison is an essential area related to clinical science and study to compare a new method with an existing method to check if a new one can replace with an existing method. This study proposes an efficient methodology for homoscedastic measurements to evaluate the agreement between two methods using Bayesian inference. This proposed model introduces an accurate, model-fitting easiness, less time required, and an assumption-less model. Simulation is used to assess and compare the finite sample performance. Simulations were carried out, and the coverage probabilities and credible intervals were calculated for each trial. Coverage probabilities of model parameters, alpha, and beta, imply that those are between the credible interval with 96% and 96.5%, respectively. It is observed that the coverage probabilities are decreasing with the increase of sample size. The proposed methodology is then used to analyze the Cardiac Ejection Fraction data. The best-fitted model was selected using the minimum error values and evaluated the agreement between the two methods using that model. The proposed model was chosen as the best model, considering the two methods have good agreement. The proposed model performed well, specifically with the small sample sizes.

***Keywords*:** Agreement; Bayesian regression; homoscedastic; method comparison.

INTRODUCTION

The world is developing rapidly, and hence, with the development of health-related fields like clinical testing, nutrition, medical chemistry, and medicine, several new methods with new features have been introduced based on novel technology for the easiness of humans. (Hanneman, 2018). Due to technological advancement, new measurement methods are cheaper, faster, and easier to use, less invasive, or more reliable than the current methods became available (Haber and Barnhart, 2008). Thus, it is essential to identify the most appropriate measurement method and know whether the measurements of the new method agree enough with the existing one. If they agree sufficient, the two methods can be used interchangeably (Nawarathna and Choudhary, 2013). The meaning of interchangeability is that measurements from the new method of a subject are significantly closer or similar to a measurement from an existing method in the clinical interpretation of the measurement. Then, the new method can be replaced with the existing one. This can be achieved through statistical method comparison studies (Jensen and Kjelgaard, 2006). When considering the health-related fields, method comparison is an important factor, and hundreds of studies are published each year (Nawarathna and Choudhary, 2013). Method comparison has been developed under many concepts and novel technology. The method could be a medical device, clinical observer, or instrument used to measure continuous response such as cardiac stroke volume, blood pressure, heartbeat measurements, etc. (Bland and Altman, 1999; Choudhary, 2009). In method comparison, those measured data can be either homoscedastic or heteroscedastic measurements (Stockl *et al*., 1998; Aravind and Nawarathna, 2017; Su and Berenson, 2017). The literature-based on this topic can be found in many papers (Jensen and Kjelgaard, 2006; Haber and Barnhart, 2008; Choudhary, 2009; Nawarathna and Choudhary, 2013; Hanneman, 2018).

Several models are introduced for comparison with different conditions and assumptions for the homoscedastic measurements when considering the past literature. The three most popular Regression Models for method comparison data are Mixed Effect Model, Deming Regression (DR), and Passing Boblok Regression (PBR) (Linnet, 1998; Jensen and Kjelgaard, 2006; Bilić-Zulle, 2011). Besides, the mixed-effect model can be considered the most frequently used model for method comparison studies (Jensen and Kjelgaard, 2006; Roy *et al*., 2015; Parker *et al*., 2016). Moreover, these mixed-effects models are used in various disciplines like physical, biological, and social sciences when repeated measurements are used for the modelling. In DR, the orthogonal least square estimates are used to minimize the error of observed measurements of the test method and estimate the slope and intercept in DR (Linnet, 1998). When both the test method and the reference method are measured with an error, DR provides unbiased regression estimates. DR is often used in method comparison to look for systematic differences between the two measurement methods. This model is used under the assumption of the ratio of measurement error as a constant (Linnet, 1998; Magari, 2002). The PBR method can be considered a regression type with no particular assumptions regarding data distribution (Bilić-Zulle,

**Corresponding Author's Email:* manusmml91@gmail.com
https://orcid.org/0000-0002-7570-6986

2011). This method is a non-parametric robust method used in two-dimensional data where both variables, X and Y, are measured with an error. It fits the linear regression and checks the agreement whether the intercept is near to zero and the slope is near to one (Stockl *et al*., 1998; Haber and Barnhart, 2008). However, the accuracy of the model is an issue of this method (Bilić-Zulle, 2011). The method comparison models in previous studies (Yin *et al*., 2008; Choudhary and Kunshan, 2010) are working well with moderate or large sample sizes. However, a method comparison model with small sample sizes is vital for clinical trials. Therefore, this study aims to develop a new method that works well with small sample sizes with high accuracy.

To achieve this goal, a Bayesian Linear Regression (BLR) model (Andrew *et al*., 2004; O'Hagan *et al*., 2004), which can deal with the small sample sizes, is introduced for method comparison studies. The homoscedastic measurements, which have equal error variance for all the measurements, are considered in the proposed model. Bayesian inference accomplished this objective by giving novel experiences of the method comparison. When considering the Bayesian viewpoint, it suggests probability as a more general concept to represent the uncertainty of the event. The Bayesian inference allows us to use some belief that we already have, called prior data. This Prior (Kass and Wasserman, 1996) data help to calculate the probability of a given event with good impact and accuracy. There are no particular assumptions about the parameter distribution under the Bayesian inference. Further, this method can be used for both replicated and non-replicated measurements.

MATERIALS AND METHODS

Bayesian homoscedasticity model for method comparison

The proposed methodology of this study can be described in three steps. The first step is to introduce the Bayesian Linear Regression Model for method comparison studies that satisfy the homoscedastic condition. A simulation study to verify the model performance is considered as the second step. The simulation study was used to verify the performance of the new model and the accuracy compared with the existing related models. The third step evaluates the agreement between the two methods under the proposed model. A good agreement implies that the two methods can be interchangeable when the difference between the measurements of the two methods is relatively low (Tim, 2019). These steps are discussed step by step according to the order to describe the methodology more clearly.

Consider a method comparison study with two homoscedastic measurement methods for n subjects. The first method is considered the reference method, and the other is the test method. Let y_i; $i = 1,2,\ldots,n$ be the observed test response, and x_i; $i = 1,2,\ldots,n$ be the observed reference response on the ith measurement. Besides, y_i' s is assumed to be normally distributed with mean μ_i and variance σ^2, where μ_i has a linear relationship with x_i which is not estimated as a single value. Then, the proposed Bayesian model is defined as,

$$Y_i = \mu_i + \varepsilon_i \;\; ; \qquad i = 1, 2, \ldots, n,$$

$$\text{where } \mu_i = \alpha + \beta * X_i \;\; and \;\; \varepsilon_i \sim N(0, \sigma^2)$$

$$Y_i \,|\alpha, \beta, \sigma^2 \sim N(\alpha + \beta * X_i, \sigma^2),$$

where α and β are constants. The precision of this model can be expressed as, $\tau = 1/\sigma^2$. The proposed model is a simple linear regression model fitted under the homoscedastic measurements, and we then convert it into a Bayesian model using Bayesian inference. Selecting the priors for parameters is the next step, and informative priors are considered in this study. Due to the lack of details about the distribution, Normal and Gamma distributions introduced in the literature with different parameter settings were checked to select the best priors (Punt and Hilborn, 1997; Gelman, 2006; Choudhary, 2007; Choudhary and Yin, 2010). Selected prior distributions can be expressed as

$$\alpha \sim N(0,1)\,, \beta \sim N(0,1) \quad \text{and} \quad \sigma^2 \sim gamma(0.001, 0.001)$$

The Shapiro-Wilk normality test (Rocho *et al*., 2012) was used to check the normality of the residuals, and the Breush Pagan Test (Halunga *et al*., 2017) and the residual scatter plot were used to check the homoscedasticity of variance. Moreover, prior distributions for α and β parameters were used to add the Bayesian concepts to the simple linear regression model.

Simulation study

The primary motivation of this section is to evaluate the finite sample model performance of the proposed Bayesian Linear Regression Model to identify the characteristics of the model in a better way. The simulation work is used for parameter estimation and to calculate the coverage probability of the measurements measured by two methods. Just Another Gibbs Sampler (JAGS) language was used for simulation as JAGS provides a helpful platform for Bayesian modelling (Choudhary and Yin, 2010).

First, the accuracy of the model was evaluated by calculating coverage probability and the credible intervals (CI) (Gardner and Altman, 1986) of each run. All data points were used with 1000 simulation runs. True initial values for parameters; α, β, and are taken as 0.002, 0.804, and 1.424, respectively. With these values, the BLR model provides enough evidence that the model fitted well to the data with acceptable higher coverage probabilities. Next, the simulation was carried out to verify the characteristics of the proposed model using a different number of trials with different initial values. Initial values were taken between zero to one since the most suitable values of parameters lie in this range. Moreover, in linear regression, the test method to the reference method should be yielded in a straight line that goes through the origin, and if measurements deviate from that, it means the lack of agreement. Hence, the alpha range is 0 to 1 for simulation settings (Magari, 2002). Also, various settings were used to represent the agreement as High, Moderate, and Low. Each of the following settings of initial values indicated in Table 1 was simulated with 20, 50, 70, 100, 500, and 1000 sample sizes.

Table 1: Initial values of two parameters for different settings in simulation.

Parameter	Setting 1 (Low Agreement)	Setting 2 (Moderate Agreement)	Setting 3 (High Agreement)
(α , β)	(0.9, 0.1)	(0.5, 0.5)	(0.2, 0.9)

For each setting of coverage probabilities, the upper and lower limits of the credible interval were calculated.

Evaluation of the agreement under the Bayesian regression model

In method comparison, evaluating the agreement of measurements is essential since it helps in decision-making. This study used the best-fitted model among all the comparable models to measure the agreement between the two methods. The agreement of the two methods in clinical observation can be graphically represented by the Limit of Agreement (LOA) plot. The key idea about LOA is that the large proportion of differences in the measurements are within the acceptable margin (say $\pm\ \delta_0$), then those measurements agree on enough to use interchangeably (Bland and Altman, 1999; Choudhary, 2009; Carstensen, 2010). That margin can be expressed as the (μ - 1.96σ, μ + 1.96σ) (when the measurements follow the normal distribution with mean μ and variance σ^2.

In addition to the LOA, three statistical measures, namely Total Deviation Index (TDI), Mean Squared Deviation (MSD), and Concordance Correlation Coefficient (CCC) values (Andrew *et al*., 2004; Choudhary, 2009) were calculated to evaluate the agreement. Next, all three agreement measurements with the specific equations considering the proposed model are described. The TDI below has non-negative values that indicate the smaller values as better agreement.

$$\mathrm{TDI}(\pi_0) = \tau\left\{X_1^2(\pi_0, \frac{\mu^2}{\tau^2})\right\}^{1/2} \quad (1)$$

Equation (1), $X_1^2(\pi_0, \Delta)$ denotes the π_0 the percentile of a non-centralized chi-squared distribution with one degree of freedom and non-centrality parameter Δ. Also, the mean and the standard deviation are denoted by μ and τ respectively. Its homoscedastic version with the proposed model can be obtained as equation (2).

$$\mathrm{TDI}(\pi_0) = \tau\left\{X_1^2(\pi_0, \frac{(\beta_0 + \beta_1 X_i)^2}{\sigma^2})\right\}^{1/2} \quad (2)$$

The MSD measure is given by the below equation (3).

$$\mathrm{E}(D^2) = \mu^2 + \sigma^2 = (\beta_0 + \beta_1 X_i)^2 + \sigma^2 \quad (3)$$

MSD has non-negative values, where D represents the difference between the two measurements. Also, smaller values give better agreement. The agreement measurement CCC ranges between -1 and 1 and is defined as equation (4).

$$\mathrm{CCC} = \frac{2\rho\sigma_1\sigma_2}{(\mu_1 - \mu_2)^2 + \sigma_1^2 + \sigma_2^2} \quad (4)$$

Here μ_1, μ_2 are the means of two measurements and σ_1, σ_2 are the respective variances, while ρ is the correlation coefficient of the two measurements.

CI is used to move the single value estimate to the range of values with more confidence which assures the precise of the data. The calculation of the CI of the sample statistic is given by CI = sample statistics ± margin of error. The margin of the error can be calculated as the product of the standard error of the point estimates and the critical value (z) derived from the standard normal distribution. The standard error differs according to the sample statistics, mean, odds ratio, or proportion. According to the proposed model, CI's for α and β is calculated by considering the mean as a sample statistic using the following equations:

$$CI = \hat{\alpha} \pm Z_{\frac{\alpha}{2}}\left(\frac{\sigma_\alpha}{\sqrt{n}}\right) \quad (5)$$

$$CI = \hat{\beta} \pm Z_{\frac{\alpha}{2}}\left(\frac{\sigma_\beta}{\sqrt{n}}\right) \quad (6)$$

where α_α and β_β are the standard deviation of alpha and beta estimators and $Z_{\alpha/2}$ is taken from the normal distribution while α is the confidence coefficient. Here, the standard error can be taken as $\sigma_j/\sqrt{n}$ where σ_j is the standard deviation of the point estimate where $j = \sigma, \beta$ and n is the sample size.

Application to cardiac ejection fraction dataset

The Cardiac ejection fraction dataset (Bland and Altman, 1999) has been used to illustrate the proposed method. This data comes from cardiac ejection fraction trials measurements involving 60 pairs of measurements measured by twelve individuals, with 3 - 7 replicates per individual. Two methods were used in this study. One is the standard method, where measurements are taken by considering radionuclide ventriculography (RV), and the other newly introduced method is impedance cardiography (IC). A Mercury column sphygmomanometer is used to measure both measurements. We are interested in quantifying the agreement between the two methods using a Bayesian Regression Model to apply a proposed model and check whether those can be used interchangeably.

RESULTS AND DISCUSSION

The simulation summary of the proposed model, including calculated coverage probabilities and credible intervals, is given in Table 2.

According to the above results, the average value of the parameter alpha is 0.798, which lies between the credible interval of 0.01 and 1.72. In the same way, the parameter beta had 0.733 as the average value, and the credible interval is (0.56, 0.88). The and coverage

probability imply that the alpha value and the beta value should be in given credible intervals of 0.96 and 0.965 of actual probabilities, respectively.

According to the settings in Table 1, the second simulation was carried out with different sample sizes. The average estimated values, their credible intervals, and coverage probabilities for α and β values are summarised in Table 3.

According to the above results, it can be identified that when the sample size increases, the coverage probability gets low. When both the sample size and the settings are considered, the highest coverage probability for both α and β was found in setting (0.5, 0.5) in a sample size of twenty. This simulation summary implies that the best results from the fitted regression model for the homoscedastic measurements in method comparison data were obtained for the small sample sizes. Moreover, the moderate agreement gives better results for the proposed model in different sample sizes according to the coverage probability. Also, according to the coverage probabilities, the proposed model fits well for Setting 2 for all sample sizes. Further, according to the simulation results, all settings perform well when the sample is less than 500.

Practical evaluation of the proposed model was done by using the Cardiac ejection fraction dataset. First, the assumptions were checked, and then, the accuracy of the proposed model was evaluated with other existing models. The *p*-value of the Shapiro-Wilk normality test 0.074 (> 0.05) indicates that the data set is normally distributed. Also, the assumption of homoscedasticity is checked using the Breush Pagan Test and graphical methods. Then, the existing model and the newly introduced model are compared. Table 4 shows the estimated Error Sum of Squares (SSE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Root Mean Squares Error (RMSE), and Mean Absolute Error (MAE) values computed for the three existing models.

Table 2: Simulation Summary for Dataset.

Alpha Value	Lower Limit	Upper Limit	Coverage Probability of Alpha	Beta Value	Lower Limit	Upper Limit	Coverage Probability of Beta
0.798	0.01	1.72	96%	0.733	0.56	0.88	96.5%

Table 3: Summary results of simulation for estimated values for α and β, their lower and upper 95% credible intervals, and coverage probabilities computed for different settings.

Setting		α				β			
		Avg. value	Lower Limit	Upper Limit	Coverage Probability	Avg. value	Lower Limit	Upper Limit	Coverage Probability
N = 20	1	1.2800	0.2501	3.7536	0.9533	0.4469	0.1277	0.7661	0.9667
	2	0.8497	-0.8703	2.5697	0.9765	0.6503	0.2836	1.0169	0.9765
	3	0.5656	0.3042	2.5212	0.9513	0.5991	0.4043	0.7939	0.9511
N = 50	1	1.2951	0.4878	2.1023	0.9531	0.6243	0.3887	0.8600	0.9553
	2	1.2654	0.4612	2.0696	0.9645	0.6255	0.4789	0.7820	0.9665
	3	1.3115	0.5029	2.1202	0.9530	0.6207	0.4752	0.7961	0.9520
N = 70	1	1.4605	0.7849	2.1361	0.9474	0.6408	0.0746	0.7917	0.9487
	2	1.4603	0.7839	2.1367	0.9565	0.5951	0.4749	0.7553	0.9491
	3	1.4591	0.7820	2.1363	0.9375	0.5954	0.4754	0.7544	0.9367
N = 100	1	1.2581	0.4576	2.0586	0.9356	0.6306	0.4867	0.7745	0.9034
	2	1.3764	0.7554	1.9973	0.9617	0.6218	0.5095	0.7341	0.9545
	3	1.3664	0.7852	1.9476	0.9434	0.6120	0.5080	0.7342	0.9112
N = 500	1	1.4875	0.8025	1.7207	0.4534	0.5973	0.5557	0.7164	0.3233
	2	1.3829	0.7577	2.0081	0.9090	0.6221	0.4993	0.7449	0.9014
	3	1.5030	0.8021	1.7892	0.4110	0.6000	0.5490	0.6510	0.4323
N = 1000	1	1.4619	0.8020	1.7267	0.3221	0.5947	0.5477	0.6418	0.2230
	2	1.1234	0.9027	1.3440	0.4194	0.6600	0.6199	0.7001	0.4194
	3	1.4630	0.7897	1.7953	0.4660	0.5960	0.5480	0.6490	0.3770

Table 4: Estimated SSE, MAPE, MAE, RMSE, and MSE values computed using five fitted models.

Model	SSE	MAPE	MAE	RMSE	MSE
BLR	**6.6659**	0.1569	**0.6573**	**0.7784**	**0.6059**
DR	6.8262	**0.1508**	0.7079	0.7877	0.6205
PB	6.9309	0.1511	0.7123	0.7937	0.6300
Mixed	362.1872	0.2923	1.4381	1.7519	3.0693

The lowest values of the MAE, MSE, SSE, and RMSE imply the best-fitted model since all estimated values measure the error of the proposed model. Due to the fact that Bayesian inference gives more relative information through the prior distribution, it has comparatively low errors than other models. Table 4 shows that the lower errors support building the high accuracy of the Bayesian model. Hence, the proposed Bayesian linear regression model is the best compared to the other fitted models.

By considering the difference between the two methods of measurements, LOA (Choudhary, 2007) was obtained between the range of (-1.2815, 2.4858), and Figure 1 shows its plot. Since all data, except one point, is between the limits, there is a good agreement between the two methods.

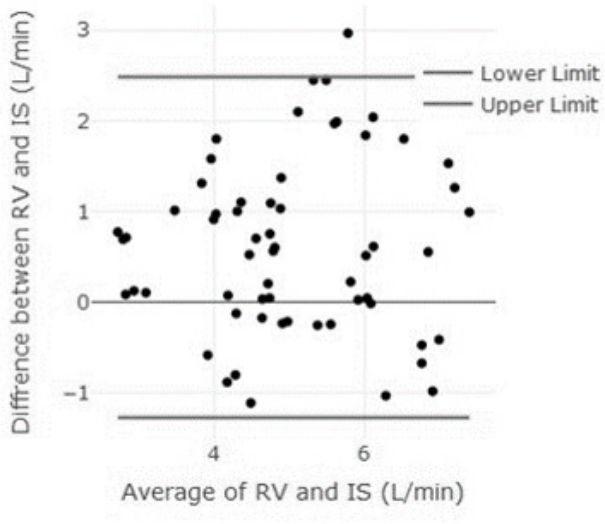

Figure 1: Limit of agreement of the two measurements of the dataset.

Besides, to evaluate the agreement between the two methods, MSD, TDI, and CCC measurements were calculated. The estimated values for the TDI, MSD, and CCC are 1.2862, 0.478, and 0.66248, respectively. When calculating the TDI, π_0 is taken as 0.95, and the CI of alpha and the CI of beta are taken as (-2.069, 1.862) and (0.326, 1.169), respectively. The above results clearly visualize the proposed Bayesian linear regression model and verify the model characteristics more accurately for the Cardiac ejection fraction dataset. According to the three measures, MSD, TDI, and CCC, all the agreement measures verify that the methods agree well.

CONCLUSION

This paper introduced a Bayesian Regression Model for method comparison data for homoscedastic measurements. The proposed Bayesian Regression Model performed well for homoscedastic measurements in method comparison data because of the higher accuracy than other existing models, model fitting easiness, less time required, and assumption-less model. Besides, it is performed well with sample sizes up to 500. According to the sample sizes, the highest coverage probability gives by the sample size 20. The proposed model shows the higher coverage probability for a moderate agreement setting in each sample size. Moreover, the proposed model is used to develop a methodology for agreement evaluation in two measurement methods. Agreement measurements and graphical results imply a good agreement between the two methods. Further, the proposed methodology can be used for homoscedastic measurements in method comparison for both balanced and unbalanced data designs. Although the proposed model is limited to homoscedastic measurements of method comparison data, this model can be easily extended to heteroscedastic measurements.

ACKNOWLEDGEMENT

The authors are grateful to anonymous reviewers for their comments that significantly improved this article. They are also thankful to Professor D. G. Altman and Professor J. M. Bland for providing the Cardiac ejection fraction dataset used in this study.

DECLARATION OF CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

Andrew, G., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004). *Bayesian Data Analysis*. Chapman & Hall/ CRC, A CRC Press Company, Boca Raton London New York Washington, D.C, 290-394. DOI: https://doi.org/10.1201/9780429258411.

Aravind, K.R. and Nawarathna, L.S. (2017). A statistical method for assessing agreement between two methods of heteroscedastic clinical measurements using the copula method. *Journal of Medical Statistics and Informatics* **5**(1): 3. DOI: https://doi.org/10.7243/2053-7662-5-3.

Bilić-Zulle, L. (2011) Comparison of methods: passing and bablok regression. *Biochemia Medica* **21**(1): 49-52. DOI: https://doi.org/10.11613/BM.2011.010.

Bland, J.M. and Altman, D.G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* **8**(2): 135-160. DOI: https://doi.org/10.1177/096228029900800204.

Carstensen, B. (2010). Comparing methods of measurement: extending the LoA by regression. *Statistics in Medicine* **29**: 401-410. DOI: https://doi.org/10.1002/sim.3769

Choudhary, P.K. (2007). Semiparametric regression for assessing agreement using tolerance bands. *Computational Statistics and Data Analysis* **51**(12): 6229-6241. DOI: https://doi.org/10.1016/j.csda.2007.01.006.

Choudhary, P.K. (2009). Interrater agreement. In *Methods and Applications of Statistics in the Life and Health Sciences*, Balakrishnan N et al. (Ed.), John Wiley: New York, 461- 480.

Choudhary, P.K. and Kunshan Y. (2010). Bayesian and frequentist methodologies for analyzing method comparison studies with multiple methods. *Statistics in Biopharmaceutical Researc*h **2**(1): 122-132. DOI: https://doi.org/10.1198/sbr.2010.08096.

Choudhary, P.K. and Yin, K. (2010). Bayesian and frequentist methodologies for analyzing method comparison studies with multiple methods. *Statistics in Biopharmaceutical Research* **2**: 122-132. DOI: https://doi.org/10.1198/sbr.2010.08096.

Gardner, M.J. and Altman, D.G. (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal (Clinical Research Ed.)* **292**(6522): 746-750. DOI: https://doi.org/10.1136/bmj.292.6522.746.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper). *Bayesian Analysis* **1**(3): 515-534. DOI: https://doi.org/10.1214/06-BA117A.

Haber, M.J. and Barnhart, H.X. (2008). A general approach to evaluating agreement between two observers or methods of measurement from quantitative data with replicated measurements. *Statistical Methods in Medical Research* **17**(2): 151-169. DOI: https://doi.org/10.1177/0962280206075527.

analytical error ratio in method comparison studies. *Clinical Chemistry* **44**(5): 1024-1031. DOI: https://doi.org/10.1093/clinchem/44.5.1024.

Magari, R.T. (2002). Statistics for laboratory method comparison studies. *BioPharm* **15**: 28-32.

Nawarathna, L.S. and Choudhary, P.K. (2013). Measuring agreement in method comparison studies with heteroscedastic measurements. *Statistics in Medicine* **32**(29): 5156-5171. DOI: https://doi.org/10.1002/sim.5955.

O'Hagan, A., Forster, J.J. (2004). Bayesian inference.In A. O'Hagan (Eds.) *Volume 2b of Kendall's Advanced Theory of Statistics*, Arnold, London, UK. Arnold, 496pp.

Parker, R.A., Weir, C.J., Rubio, N., Rabinovich, R., Pinnock, H., Hanley, J., McCloughan, L., Drost, E.M., Mantoani, L.C., William MacNee, W. and McKinstry B. (2016). Application of mixed effects limits of agreement in the presence of multiple sources of variability: exemplar from the comparison of several devices to measure respiratory rate in COPD patients. *PLoS ONE*. **11**(12): e0168321. DOI: https://doi.org/10.1371/journal.pone.0168321.

Punt, A.E. and Hilborn, R. (1997). Fisheries stock assessment and decision analysis: the bayesian approach. *Reviews in Fish Biology and Fisheries* **7**: 35-63. DOI: https://doi.org/10.1023/A:1018419207494.

Rochon, J., Gondan, M. and Kieser, M. (2012). To test or not to test: preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology* **12**: 81. DOI: https://doi.org/10.1186/1471-2288-12-81.

Roy, A., Fuller, C.D., Rosenthal, D.I, Thomas, C.R. and Jr, C.R.T. (2015). Comparison of measurement methods with a mixed effects procedure accounting for replicated evaluations (COM3PARE): method comparison algorithm implementation for head and neck IGRT positional verification. *BMC Medical Imaging* **15**: 35. DOI: https://doi.org/10.1186/s12880-015-0074-z.

Stöckl, D., Dewitte, K. and Thienpont, L.M. (1998). Validity of linear regression in method comparison studies: is it limited by the statistical model or the quality of the analytical input data *Clinical Chemistry* **44**(11): 2340-2346. https://pubmed.ncbi.nlm.nih.gov/9799762/

Su, H. and Berenson, M.L. (2017). Comparing Tests of homoscedasticity in simple linear regression. *JSM Mathematics and Statistics* **4**(1): 1017-1027.

Tim, P.M., Ian, R.W. and Michael, J.C. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine* **38**: 2074-2102. DOI: https://doi.org/10.1002/sim.8086.

Yin, K., Choudhary, P.K., Varghese, D. and Goodman, S.R. (2008). Goodman SR. A Bayesian approach for sample size determination in method comparison studies. *Statistics in Medicine* **27**(13): 2273-2289. DOI: https://doi.org/10.1002/sim.3124.