Radiotherapy and Oncology 141 (2019) 321–326

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Technical Note

How to evaluate agreement between quantitative measurements

Annette Kopp-Schneider*, Thomas Hielscher

Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany

ARTICLE INFO

Article history:
Received 6 March 2019
Received in revised form 5 July 2019
Accepted 4 September 2019
Available online 25 September 2019

Keywords:
Agreement
Method comparison study
Bland–Altman plot
Limits of agreement
Concordance Correlation Coefficient

ABSTRACT

When a method comparison study is performed, the aim is to evaluate the agreement of measurements of different methods. We present the Bland–Altman plot with limits of agreement as the correct analysis methodology. We also discuss other scaled and unscaled indices of agreement and commonly used inappropriate approaches.

A frequent analysis situation in radiology is that quantitative measurements are obtained from different methods (or software tools, readers, ...) on all members of a set of items, e.g., patients. The aim of such a study is to compare the modalities and to investigate their agreement. For satisfactory agreement, the goal is to replace one measurement method with another, usually because it is easier to conduct, is less expensive or less time consuming. If measurements are made by the same method but different readers, the aim is to show the reproducibility of the measuring method. If measurements are made using the same method by the same reader, agreement reflects the repeatability of the measurement for this specific reader. It is important to note that in the situations considered here, none of the methods is considered the gold standard, hence the true value remains unknown, and the accuracy of the methods, i.e., any systematic error, cannot be assessed. An agreement study for two methods is set up such that the true (but unknown) value of each item comes from a distribution with variance σ_I^2 and this true value is observed with two methods, each associated with measurement error σ_i^2, independent of the true value of the item. Hence the variance of the measurements made by each method is $\sigma_I^2 + \sigma_i^2$, the sum of the variability of items and the precision of the method. Although methodology for assessing agreement with multiple measurements per method and item are available, we will only consider the common situation with a single measurement per method. In such a situation, precision of the methods, i.e., σ_i^2, cannot be assessed.

For method comparison studies, the focus lies on the agreement between (error-associated) measurements of the methods on each item and hence on the size of the difference between measurements of the two methods, or, as Bland and Altman [2] put it: "...the key to method comparison studies is to quantify disagreements between individual measurements". Since these studies are not aimed at analyzing population parameters (such as the population mean for each method), they further state, "It follows that we do not see a place for methods of analysis based on hypothesis testing". If the result of the comparison study is that two methods show sufficient agreement, these methods can be used interchangeably. Importantly, 'sufficient' cannot be purely defined in statistical terms. Acceptable disagreement depends on the clinical relevance of differences.

The manuscript is organized as follows. On basis of an example data set a typical analysis using correlation, linear regression and *t*-test is presented and arguments for its inappropriateness are given. Analysis with Bland–Altman plots is introduced and typical patterns of this plot are discussed. Prediction and tolerance intervals and the unscaled indices Mean Square Deviation/Root Mean Square Deviation, Total Deviation Index and Coverage Probability are considered as well. Scaled summary indices for agreement, in particular Concordance Correlation Coefficient and Intraclass Correlation Coefficient, are introduced and discussed. The manuscript closes with recommendations for a basic analysis of agreement studies.

* Corresponding author at: Division of Biostatistics, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany.

E-mail address: kopp@dkfz.de (A. Kopp-Schneider).

https://doi.org/10.1016/j.radonc.2019.09.004
0167-8140/© 2019 Elsevier B.V. All rights reserved.

Example data set

We will exploit the data set published and described in detail by Müller-Eschner et al. [10]. Briefly, preoperative computed tomography angiographies of 30 patients with thoracic aortic disease (mean age 66.8 ± 11.6 years, 23 males) were retrospectively analyzed by two blinded experts in vascular radiology. Maximum aortic diameters at different measurement positions relevant to thoracic endovascular aortic repair (TEVAR) were assessed using three measurement techniques: manual axial slices (axial), manual double-oblique multiplanar reformations (MPRs) and semiautomatic centerline analysis (CL). Two measurement positions were selected for the present analysis (P1, distal to left common carotid artery; P4, which is actually not a fixed position in the aorta but recorded the maximum diameter between the outer walls of the pathology).

Why correlation, linear regression analysis and *t*-test do not answer the agreement question

A typical evaluation in this situation is the scatterplot of measurements obtained from one method versus the other, together with the line of identity. Fig. 1A shows this for the analysis of reproducibility of the axial measurements at one location (P1) for the two readers. Often this plot is complemented with Pearson's correlation coefficient (here $r = 0.86$), the *p*-value for the test whether the coefficient is equal to 0 (here $p \ll 0.001$) and sometimes by the 95%-Confidence Interval (95%-CI) for *r* (here 0.72–0.93). While the plot alone is informative and fulfills the recommendation to always plot raw data, the correlation coefficient only indicates a strong linear association, but by itself does not provide information about agreement. To stress the point, consider comparing axial measurements of two readers where one reader obtains the same measurement as the other, but with an offset of exactly 1 mm; these two values are perfectly correlated and what is sometimes called "consistent", but they lack absolute agreement. Another disadvantage of the correlation coefficient is that it depends on the variability of items. Considering methods with equal precision, i.e., equal variance of the measurement error σ^2, then the correlation is the ratio of item variability σ_I^2 and the sum of item and measurement variability $\sigma_I^2 + \sigma^2$ (see Supplement S1 for details) and thus will increase with increasing item variability. In the present example, correlation between both readers drops to 0.49 if we decrease the item variability by excluding 20% of measurement at both the lower and upper end (Supplementary Table 1).

Sometimes a linear regression line with intercept 0 is fitted to the scatterplot and it is tested whether the slope is equal to one. In the current data set, this test yields a *p*-value of 0.36, and the 95%-CI for the slope ranges from 0.99 to 1.03: Would this suggest relevant agreement? Since the power to reject the null hypothesis depends also on the sample size, not rejecting the null hypothesis cannot be taken as evidence for agreement.

Another approach sometimes taken is to compare the measurements by paired *t*-test, i.e., investigating the difference in measurements. This answers the question whether the population means of the first and the second rater are different. In our example, the *p*-value is $p = 0.28$ and the 95%-CI for the differences between Reader 1 and Reader 2 ranges from -0.92 to 0.27 with a mean of -0.33. This shows that there is no significant bias between the two readers. This finding by itself is interesting, but it only addresses the population mean difference and does not answer the question to which extent the measurements on an individual item (dis)agree. In addition, the same issue with power as for the slope testing applies here, i.e., more samples might yield a significant bias here. Note, that a statistically significant bias might be considered negligible if the measurement error itself is larger, a quantity that cannot be estimated with a single measurement per item and method.

Graphical analysis method: Bland–Altman plot

A simple graphical method to assess agreement is the Bland–Altman plot (e.g., [1,2]. This plot addresses the quantity of interest, namely the differences *d* between the two measurements. The brilliant idea proposed by Bland and Altman was to plot the difference of the measurements versus their average. It would certainly be best to plot the difference against the true value, but this is unfortunately unknown, and the best surrogate is the average of the two measurements. Since the difference and each one of the measurements are statistically dependent, it would not be a good idea to plot the difference versus one of the measurements. If both measurements are associated with similar measurement error, they are uncorrelated and no slope in the scatterplot should be observed (for details see Supplement S2). Arguing that the differences typi-

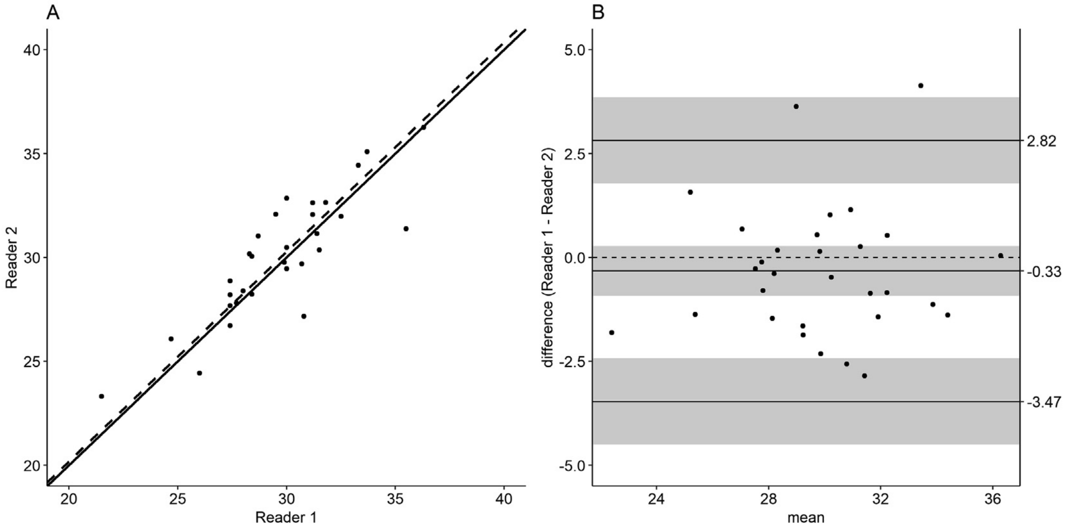

Fig. 1. Comparison of axial P1 measurements for Reader 2 and Reader 1. A: Scatterplot with line of identity (=concordance) given as solid line and the regression line through the origin (with estimated slope of 1.01) shown as dashed line. B: Bland–Altman Plot with solid lines for bias and LoAs, with 95%-Confidence Intervals shown as grey zones.

cally follow a normal (Gaussian) distribution, Bland and Altman proposed to add three lines to the scatterplot based on the mean difference or bias ($\bar{d}$) and its standard deviation (s). One line is drawn for the bias $\bar{d}$, one for the so-called upper Limit of Agreement (LoA) at $\bar{d} + 1.96s$ and one for the lower LoA at $\bar{d} - 1.96s$, using the 0.975-quantile of the standard normal distribution, 1.96. Due to the normal distribution, 95% of the data points are expected to lie between the upper and the lower LoA. A Bland–Altman plot for agreement in axial diameter measurement at P1 of Reader 1 and Reader 2 is shown in Fig. 1B. The plot indicates that there is only a slight systematic shift between the two readers of −0.3 mm (as noted above) and that approximately 95% of the differences lie between −3.5 and almost +2.8 mm. Indeed, in this data set a total of 2 points, i.e. 7%, fall outside the LoAs. It should be noted that the LoAs and the bias are not fixed quantities but are also associated with variability since they are derived from the data, and their variability can again be assessed by reporting the respective 95%-CIs (for details see [2]), shown as grey zones in Fig. 1B. It is only sensible to interpret LoAs if there is no systematic relationship between mean and difference. In contrast to the correlation coefficient, the LoAs only depend on the mean and standard deviation of the differences between measurements and hence are independent of the variability of the items, see Supplementary Table 1.

Clinicians perform a method comparison study to investigate whether the methods show good or, even better, "significant" agreement. They typically expect the statistician to provide an answer to their question and are disappointed that s/he cannot make this assessment without background from the clinician on what constitutes an acceptable agreement in that specific situation, i.e., to which extent measurements can be considered interchangeable. The clinician must judge whether the range of deviations covered by the LoAs is acceptable by putting it into the clinical context and assessing the impact of discrepancies. Ideally, the range of clinically tolerated deviations should be defined upfront.

Regarding agreement between measurement methods, one question in the exemplary radiological study was whether the measurements obtained with the time consuming manual method MPR agree with the semiautomatic CL analysis which can also be performed by non-expert readers. Fig. 2A shows the respective Bland–Altman plot for Reader 1 for P4. The Bland–Altman plot for Reader 2 looks very similar, also showing a significant bias (as assessed by the 95%-CI) of MPR measurement being about

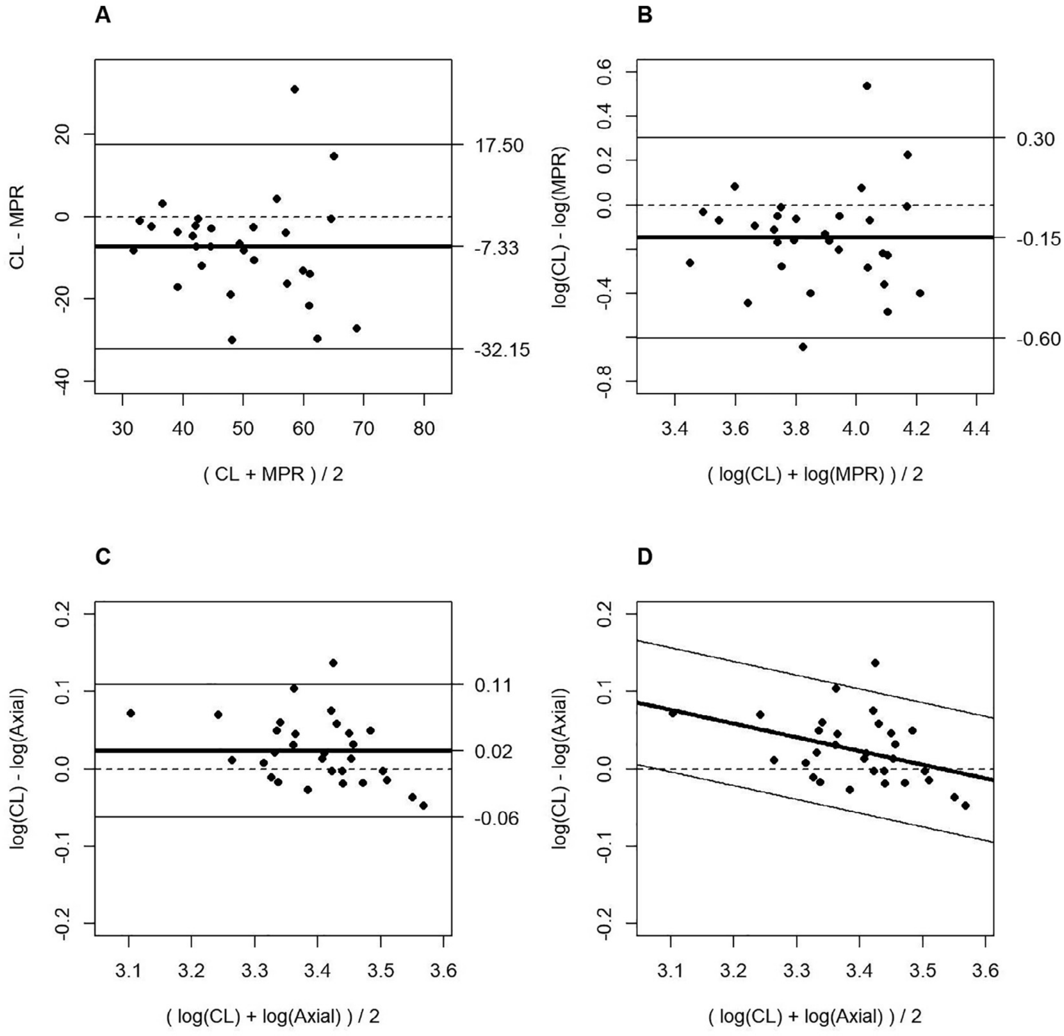

Fig. 2. Different shapes of point clouds in Bland–Altman plot for CL and MPR measurements for P4, and for CL versus axial measurements at P1. A: Bland–Altman plot illustrating that the variability of the difference increases with the value. B: Bland–Altman plot of log-transformed values of CL versus MPR measurement. Variability of the difference of log-values (i.e., log of the ratio) is constant over the range. C: Bland–Altman plot of log-transformed values of CL versus axial measurement with horizontal LoAs, showing a negative slope in the point cloud. D: Bland–Altman plot of log-transformed values of CL versus axial measurement with LoAs derived from regressing the difference versus the average of log-transformed values. All measurements for Reader 1.

7 mm larger than CL measurement. If the clinician would be willing to tolerate the deviation between the two methods as assessed by the LoAs, then the less expensive method could be used for replacement.

The scatterplot in Fig. 2A suggests a funnel-shaped form: For a larger average the difference is larger too. This suggests that a logarithmic transformation of measurements and assessing the proportional deviation might be appropriate. Fig. 2B shows the resulting Bland–Altman plot for the natural logarithm with no further obvious pattern. The interpretation of the LoAs involves back transformation by taking the antilog and addresses the ratio of the two measurements: 95% of the ratio of the manual and the semi-automated measurements will lie between $e^{-0.6} = 0.55$ and $e^{0.3} = 1.35$, i.e., CL measurements are between 45% smaller and 35% larger than MPR measurements, and in the mean, axial measurements are 14% smaller than MPR measurements.

To illustrate a different frequently occurring type of pattern, Fig. 2C shows the Bland–Altman plot for logarithmic transformation of axial and CL measurements for Reader 1 at P1. The scatterplot has an obvious negative slope. One reason for this shape may be that the variance of both methods is different, and indeed, sd (log(axial)) = 0.11, whereas sd(log(CL)) = 0.09. Another reason for a slanted scatterplot may be a difference in scale between the two measurements, a situation that we do not consider in the present manuscript. For details why different precisions lead to a slope in the scatterplot, see Supplement S2 with Supplementary Figs. 1 and 2. Horizontal LoAs in such a situation are too far apart and might be particularly imprecise for very low or high values. Bland and Altman [2] provide a method to derive LoAs based on regressing differences on averages, shown in Fig. 2D. The central line is given by the linear regression and accompanying LoAs parallel to the regression line are shifted by the 0.975-quantile of the standard normal distribution times the square root of the residual variance in the regression analysis. LoAs can also be understood as a prediction interval for measurements by one method based on measurements by a second method. Here, LoAs with slope better characterize this relationship between methods. Supplementary Fig. 3 shows the corresponding plot on the original scale of the measurements with LoAs derived from the analysis of log-transformed data, and Supplementary Fig. 4 for the ratio of measurements versus the average. The original scale might facilitate the interpretation, but the complex shape of the LoAs only emphasizes that there is no straightforward interchangeability of methods, at least not for the whole measurement range. While a simple linear regression based on averages and differences is easy to compute and to interpret, it ignores the measurement error in predictors and the correlation of errors in the Bland–Altman plot. More complex regression approaches [5] exist to correctly account for that but require some knowledge of the measurement error, either based on replicated measurements or some prior information.

Prediction and tolerance intervals

Although Bland and Altman introduce the LoAs using the 0.975-quantile of the standard normal distribution, their interpretation as prediction interval, i.e., as interval for a future observation, would necessitate increasing the quantile by a factor of $\sqrt{(n+1)/n}$, n being the number of items. For small n, the 0.975-quantile of the t-distribution with $n-1$ degrees of freedom should be used instead of 1.96, because s is estimated from the data. In our example in Fig. 1, this would lead to slightly wider 95% agreements limits of -3.7 to 3.0 mm. That is, only with large n do the standard LoAs have the interpretation of a prediction interval, in which case the 95%-CIs of the LoAs also become very narrow. An even more stringent interval is the so-called tolerance interval, which guarantees that, e.g., 95% of differences for future measurements lie within its limits with a pre-specified confidence, i.e., probability, of, e.g., 95% (see, e.g., Francq and Govaerts [5]). The estimation of tolerance intervals is complex and usually solved approximately. In our example the 95% tolerance interval (at 95% confidence) is -4.4 to 3.8 mm, i.e., 2 mm wider than the standard LoAs but very close to the lower and upper limit of their corresponding 95%-CIs. Supplementary Fig. 5 illustrates the different types of intervals for the agreement between readers.

Other unscaled approaches: MSD/RMSD, TDI and CP

A common method to aggregate accuracy and precision between two raters/methods into a single number is the Mean Square Deviation (MSD) of the pairwise differences d. The MSD is the average of the squared differences of the pairwise measurements and corresponds to the sum of the squared bias $\overline{d}$ and the variance of d. A MSD of 0 represents perfect agreement while the MSD becomes larger if bias $\overline{d}$ and/or variance of d increases. The square root of the MSD (RMSD) gives the agreement as assessed by the MSD on the original measurement scale, which is 1.6 mm in our example of the agreement between readers.

Lin and colleagues [9] have suggested reporting the Total Deviation Index (TDI) and/or the Coverage Probability (CP). For a given probability π, typically $\pi = 0.95$, the probability of the absolute difference being smaller than TDI_{π} equals π, hence it provides a symmetrical band around the horizontal axis in the Bland–Altman plot that is (theoretically) expected to contain $\pi = 95\%$ of the differences and is in the unit of the measurements. In contrast to the LoAs, the TDI does not cover the 'middle' π% of the population because it is symmetrical around 0 and not $\overline{d}$. Consequently, for increasing bias but same variance both MSD and TDI increase while the LoAs would be shifted, retaining the same width. Lin [8] showed that an upper bound for TDI_{π} is proportional to the RMSD. A confidence interval for the TDI has the interpretation of a tolerance interval, as it is an interval within which a specified proportion of the differences fall with, e.g., 95% confidence. The CP is equivalent to the TDI but defined conversely: Given an absolute maximum difference $d_{\max}$ (again in the unit of measurement), CP_{d} is the probability of observing a difference smaller or equal to $d_{\max}$. In our example, the $\mathrm{TDI}_{0.95}$ yields an absolute maximum difference of 3.2 mm, and correspondingly the CP is 92% at a maximal absolute difference of 3 mm.

Summarizing agreement in a scaled index: CCC and ICC

Clinicians often wish to summarize data into an easily interpretable index that can be used for assessing the goodness of agreement between studies. Lin [7] introduced the Concordance Correlation Coefficient (CCC), which is a scaled index, showing 0 if no correlation is observed, 1 for total concordance and -1 for perfect negative concordance. The CCC is best explained by going back to Fig. 1A. It can be shown that

$$CCC = 1 - \frac{\text{expected squared perpendicular(ESP)deviations from the identity line}}{\text{ESP deviations from the identity line when measurements are uncorrelated}}$$

If the measurements of Reader 1 and Reader 2 are uncorrelated, one would expect the points in Fig. 1A to be randomly scattered instead of clustering around the identity line. Supplement S5 gives a more formal representation of the CCC and some further insights about its value. Due to its definition, the CCC can be understood as a scaled and chance-corrected version of the MSD, for details see Supplement S5.

The CCC for the agreement of axial measurements at P1 between the readers is 0.85 with 95%-CI from 0.72 to 0.93. Although this index corrects some of the unwanted features of the correlation coefficient when applied to agreement questions, it still is dependent on the range, i.e., the variability of items, and hence on the design of the study, as shown in Supplement S5 and Supplementary Table 1. Since the selection of items controls the CCC in a similar way as the correlation coefficient, there is no universally valid assessment of which CCC value indicates good agreement. Hence, it is not well suited to compare agreement between studies. The CCC is closely related and often very similar in value to the Intraclass Correlation Coefficient (ICC). There are different definitions of the ICC for different designs and aims [12]. All of these are based on ANOVA models and assume identical precision between methods. Since we are interested in the interchangeability of measurements, we refer to types of ICC measuring the so-called absolute agreement rather than just consistency, which ignores systematic shifts between raters. In our example of the agreement between the two raters that are considered a random selection of radiologists (ICC type 2 in [12] terminology), the corresponding ICC is 0.86.

Sometimes the ICC is also used to assess agreement between categorical measurements. While the ICC and the weighted kappa coefficient are equivalent under certain conditions [13], it is generally recommended to use the (weighted) kappa coefficient for categorical data. In analogy to CCC, the kappa coefficient quantifies observed agreement in relationship to chance agreement. Measures of agreement for categorical data are as manifold as those for continuous outcomes and are not covered in this short note.

Discussion

We have presented the Bland–Altman plot with LoAs as a simple and intuitive way to analyze data from a quantitative agreement study. Such an analysis does not result in a scaled index that is transferable between studies, and it leaves the physician to assess whether agreement is sufficient in the context of the medical application. Insofar, no final answer is obtained based on pure statistical reasoning, and the physician remains responsible for judging the extent of agreement. Scaled indices such as the CCC and the ICC seemingly absolve the physician from the responsibility of defining acceptable disagreement but they depend on the variability of the items and thus on the study design.

It should be noted that the discussed methods are valid only if the (in case of the Bland–Altman plot: difference of) quantitative measurements are approximately normally distributed (in case of the ICC with equal precision). If the variables cannot be transformed to an approximate normal distribution, these methods will be inappropriate. For the Bland–Altman plot, a nonparametric alternative based on empirical quantiles of the distribution of measurement differences can be used instead [2].

Vast literature exists for extensions of the methods presented here. Bland and Altman [2] advocate the collection of repeated measurements per item to be able to estimate precision of the methods and they suggest extensions for various scenarios. A monograph published by Carstensen [3] thoroughly discussing the analysis of method comparison studies based on LoA for a large variety of situations is recommended. The monographs by Lin et al. [9], Shoukri [11] and Choudhary and Nagaraja [4] address various unscaled and scaled approaches to agreement and also cover the analysis of agreement for categorical data. Furthermore, all of these books also deal with more general statistical models for agreement allowing not only for bias between methods but also for potential scale difference.

Recently, guidelines for reporting reliability and agreement studies (GRRAS) were proposed [6]. The authors acknowledge that due to inadequate reporting, interpretation and synthesis of study results are often difficult, and they developed a set of fifteen issues to be addressed when reporting such studies.

Conclusions

In summary, the following approach should be taken for the analysis of an agreement study with two methods and no replicate measurements:

(1) Check measurements for plausibility/data entry errors, e.g., admissible/plausible range.
(2) Generate a Bland–Altman plot and inspect whether the shape of the scatterplot indicates a slope or a funnel.
(3) If points are randomly scattered over the range of the horizontal axis, add lines for LoAs and bias, possibly with 95%-CIs.
(4) The LoAs indicate the range in which 95% of differences between the two methods are expected.
(5) Based on the scientific background, assess from the LoAs whether the methods show sufficient agreement.
(6) In case of a funnel-shaped form, use log-transformation for the measurements and proceed from (1).
(7) In case of a slope
    - inspect the variance of each method separately as this may be the cause.
    - derive LoAs by regression approach.
(8) If possible, collect repeated measurements for every method to assess precision of the measurement methods. For evaluation, consult Bland and Altman [2] or Carstensen [3].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We thank Dr. Müller-Eschner for letting us use the data for our various analyses. We are very grateful to the referees for their remarks on an earlier version of this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2019.09.004.

References

[1] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–10.
[2] Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8:135–60.
[3] Carstensen B. Comparing clinical measurement methods: a practical guide. John Wiley & Sons Ltd.; 2010.
[4] Choudhary PK, Nagaraja HN. Measuring agreement: models, methods, and applications. John Wiley & Sons Ltd.; 2017.

[5] Francq BG, Govaerts B. How to regress and predict in a Bland–Altman plot? Review and contribution based on tolerance intervals and correlated-errors-in-variables models. Statistics Med 2016;35:2328–58.

[6] Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 2011;64:96–106.

[7] Lin L. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255–68.

[8] Lin L. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. Stat Med 2000;19:255–70.

[9] Lin L, Hedayat AS, Wu W. Statistical tools for measuring agreement. New York, NY, USA: Springer; 2012.

[10] Müller-Eschner M, Rengier F, Partovi S, Weber TF, Kopp-Schneider A, Geisbüsch P, et al. Accuracy and variability of semiautomatic centerline analysis versus manual aortic measurement techniques for TEVAR. Eur J Vasc Endovasc Surg 2013;45:241–7.

[11] Shoukri MM. Measures of interobserver agreement. Boca Raton, FL, USA: Chapman&Hall/CRC; 2004.

[12] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–3428.

[13] Fleiss JL, Cohen J. The equivalence of weighted kappa and the Intraclass Correlation Coefficient as measures of reliability. Educational and Psychological Measurement 1973;33:613–9. In this issue.