Reporting Standards for a Bland–Altman Agreement Analysis

Created by: Oke Gerke [1]

1, Department of Clinical Research, University of Southern Denmark, Odense, DENMARK Department of Nuclear Medicine, Odense University Hospital, Odense, DENMARK; oke.gerke@rsyd.dk

Version received: 11 June 2020

The Bland–Altman Limits of Agreement is a popular and widespread means of analyzing the agreement of two methods, instruments, or raters in quantitative outcomes. An agreement analysis could be reported as a stand-alone research article but it is more often conducted as a minor quality assurance project in a subgroup of patients, as a part of a larger diagnostic accuracy study, clinical trial, or epidemiological survey. Consequently, such an analysis is often limited to brief descriptions in the main report. Therefore, in several medical fields, it has been recommended to report specific items related to the Bland–Altman analysis. Seven proposals were identified from a MEDLINE/PubMed search on March 03, 2020, three of which were derived by reviewing anesthesia journals. Broad consensus was seen for the a priori establishment of acceptability benchmarks, estimation of repeatability of measurements, description of the data structure, visual assessment of the normality and homogeneity assumption, and plotting and numerically reporting both bias and the Bland–Altman Limits of Agreement, including respective 95% confidence intervals. Abu-Arafeh et al. provided the most comprehensive and prudent list, identifying 13 key items for reporting (Br. J. Anaesth. 2016, 117, 569–575). The 13 key items should be applied by researchers, journal editors, and reviewers in the future, to increase the quality of reporting Bland–Altman agreement analyses.

Background

The Bland–Altman Limits of Agreement (BA LoA), or simply Bland–Altman plots, are used widely in method comparison studies with quantitative outcomes, as evidenced by more than 34,753 citations of the seminal Lancet paper to date.[1] In this analysis, a pair of observations is made from the same subject, with two different methods. Subsequently, the means and differences of these pairs of values for each subject are displayed in a scatter plot. The plot usually also shows a line for the estimated mean difference between the two methods (a measure of the bias between the two methods), and lines indicating the BA LoA (within which approximately 95% of all population differences would lie).[1][2] Use of the BA LoA assumes that the differences are normally distributed.

Kottner et al. pointed out that agreement and reliability assessment is either conducted in dedicated studies with a respective primary focus or as a part of larger diagnostic accuracy studies, clinical trials, or epidemiological surveys that report agreement and reliability as a quality control.[3] The latter is often done in subsamples, resulting in small to moderate sample sizes; sample sizes as small as 10 are, by no means, an exception.[4] Such a supplementary agreement or reliability analysis is often limited to brief descriptions in the main report, lacking details for sufficient transparency.

The *Guidelines for Reporting Reliability and Agreement Studies* (GRRAS) comprise a comprehensive checklist of 15 items that support the transparent reporting of agreement and reliability studies.[3] Item no. 10 and 13 relate to the description of the statistical analysis and reporting of estimates of reliability and agreement, including measures of statistical uncertainty. However, agreement and reliability studies can easily become complex investigations when considering different sources for variation in the data, leading naturally to repeatability coefficients based on variance components analyses.[5][6] This is why Item no. 10 and 13 are neither specific with respect to agreement analysis performed by means of BA

LoA.

During the past two decades, researchers have attempted to establish reporting standards for BA plots in various fields. In a review of methodological reviews, Gerke identified reporting standards for BA agreement analyses and singled out the most comprehensive and appropriate list.[7]

Proposals of reporting standards for agreement analyses with BA LoA

Seven publications, published before March 03, 2020, were identified.[8][9][10][11][12][13][14] Three out of seven studies were published in anesthesia journals, while the remaining stemmed from various fields (Table 1).

Table 1: Characteristics of studies proposing reporting items for BA analysis. N/A: not applicable. This table was reproduced from Gerke.[7]

Publication	Field/Area	Search Approach or Target Journals	Time Frame	Evidence Base
Flegal (2019)[8]	Self-reported vs. measured weight and height	Unrestricted; reference lists of systematic reviews, repetition of 2 PubMed searches of these, “related articles” in PubMed	1986–May 2019	N = 394 published articles
Abu-Arafeh (2016)[9]	Anesthesiology	Anaesthesia, Anesthesiology, Anesthesia & Analgesia, British Journal of Anaesthesia, Canadian Journal of Anesthesia	2013–2014	N = 111 papers
Montenij (2016)[10]	Cardiac output monitors	N/A	N/A	Expert opinion
Olofsen (2015)[11]	Unrestricted	N/A	N/A	Narrative literature review and Monte Carlo simulations
Chhapola (2015)[12]	Laboratory analytes	PubMed and Google Scholar	2012 and later	N = 50 clinical studies
Berthelsen (2006)[13]	Anesthesiology	Acta Anaesthesiologica Scandinavica	1989–2005	N = 50
Mantha (2000)[14]	Anesthesiology	Seven anesthesia journals	1996–1998	N = 44

Sixteen reporting items were proposed across these seven studies:

1. Pre-established acceptable limit of agreement
2. Description of the data structure (e.g., number of raters, replicates, block design)
3. Estimation of repeatability of measurements if possible (mean of differences between replicates and respective standard deviations)
4. Plot of the data, and visual inspection for normality, absence of trend, and constant variance across the measurement range (e.g., histogram, scatter plot)
5. Transformation of the data (e.g., ratio, log) according to 4), if necessary
6. Plotting and numerically reporting the mean of the differences (bias)
7. Estimation of the precision, i.e., standard deviation of the differences or 95% confidence interval for the mean difference
8. Plotting and numerically reporting the BA LoA
9. Estimation of the precision of the BA LoA by means of 95% confidence intervals
10. Indication of whether the measurement range is sufficiently wide (e.g., apply the Preiss-Fisher

procedure[15])
11. Between- and within-subject variance or stating that the confidence intervals of the BA LoA were derived by taking the data structure into account
12. Software package or computing processes used
13. Distributional assumptions made (e.g., normal distribution of the differences)
14. Sample size considerations
15. Correct representation of the x-axis
16. Upfront declaration of conflicts of interest

Broad consensus was seen for the a priori establishment of acceptable LoA (Item #1); estimation of repeatability of measurements in case of available replicates within subjects (#3); visual assessment of a normal distribution of differences and homogeneity of variances across the measurement range (#4); and plotting and numerically reporting both bias and the BA LoA, including respective 95% confidence intervals (#6–9). A description of the data structure (#2), between- and within-subject variance (or stating that confidence intervals for the BA LoA were derived by accounting for the inherent data structure; #11), and distributional assumptions (#13) followed. Only one review raised the issue of a sufficiently wide measurement range (#10), sample size determination (#14), or correct representation of the x-axis (#15). Upfront declaration of conflicts of interest (#16) also appeared only once, but this can generally be presumed to be covered by the ethics of authorship. Besides, there seems to be a tacit consensus of the fact that the x-axis must show average values of the two methods compared (#15), as also discussed by Bland and Altman.[16] The issue of sample size determination (#14) was discussed in more detail by Gerke.[7]

The list of reporting items proposed by Abu-Arafeh et al.[9] was the most comprehensive (13 out of 16 items), followed by those proposed by Montenij et al.[10] (10 out of 16 items) and Olofsen et al.[11] (9 out of 16 items). The latter two lists were complete subsets of Abu-Arafeh et al.'s list,[9] with the exception of Item #14 on the list by Montenij et al.[10] The most recently published list by Flegal et al.[8] comprised items that were derived as a modified version of those suggested by Abu-Arafeh et al.[9] Specifically, they omitted items related to statistical software and repeated measurements, as the latter are rarely applied in studies entailing self-reported weight and height.[8]

A worked example for the reporting items proposed by Abu-Arafeh et al.[9] can be found elsewhere.[7] Such an extended analysis can, generally speaking, easily accompany the report of the main study as *Online Supplemental Material*. Journal space restrictions are no longer a valid argument for reducing the reporting of agreement or reliability to a few lines of the main report.

Conclusions

The work of Abu-Arafeh et al.[9] represents the most comprehensive and prudent list of reporting items for BA analysis, identifying 13 key items. Considering GRRAS[3] as a broad reporting framework for agreement and reliability studies, Abu-Arafeh et al.[9] concretized its Item 10 (statistical analysis) and Item 13 (estimates of reliability and agreement including measures of statistical uncertainty) in the context of the Bland–Altman analysis in method comparison studies. A rigorous application of and compliance with the 13 key items recommended by Abu-Arafeh et al.[9] will increase both the transparency and quality of such agreement analyses. Researchers, journal editors, and reviewers are obligated to employ more diligence in producing and assessing BA analyses in the future.

References

1. Bland, J.M.; Altman, D.G.; Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1986**, *1*, 307-310.
2. Bland, J.M.; Altman, D.G.; Measuring agreement in method comparison studies. *Stat. Methods Med. Res.* **1999**, *8*, 135-160, 10.1177/096228029900800204.
3. Kottner, J.; Audigé, L.; Brorson, S.; Donner, A.; Gajewski, B.J.; Hróbjartsson, A.; Roberts, C.; Shoukri, M.; Streiner, D.L.; Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J. Clin. Epidemiol.* **2011**, *64*, 96-

106, 10.1016/j.jclinepi.2010.03.002.
4. Rojulpote, C.; Borja, A.J.; Zhang, V.; Aly, M.; Koa, B.; Seraj, S.M.; Raynor, W.Y.; Kothekar, E.; Kaghazchi, F.; Werner, T.J.; et al. Role of 18F-NaF- PET in assessing aortic valve calcification with age. *Am. J. Nucl. Med. Mol. Imaging* **2020**, *10*, 47-56.
5. Gerke, O.; Möller, S.; Debrabant, B.; Halekoh, U.; Odense Agreement Working Group; Experience applying the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) indicated five questions should be addressed in the planning phase from a statistical point of view. *Diagnostics* **2018**, *8*, 69, 10.3390/diagnostics8040069.
6. Gerke, O.; Vilstrup, M.H.; Segtnan, E.A.; Halekoh, U.; Høilund-Carlsen, P.F.; How to assess intra- and inter-observer agreement with quantitative PET using variance component analysis: A proposal for standardisation. *BMC Med. Imaging* **2016**, *16*, 54, 10.1186/s12880-016-0159-3.
7. Gerke, O.; Reporting Standards for a Bland-Altman Agreement Analysis: A Review of Methodological Reviews. *Diagnostics* **2020**, *10*, E334, 10.3390/diagnostics10050334.
8. Flegal, K.M.; Graubard, B.; Ioannidis, J.P.A.; Use and reporting of Bland-Altman analyses in studies of self-reported versus measured weight and height. *Int. J. Obes. (Lond.)* **2020**, *44*, 1311-1318, 10.1038/s41366-019-0499-5.
9. Abu-Arafeh, A.; Jordan, H.; Drummond, G.; Reporting of method comparison studies: A review of advice, an assessment of current practice, and specific suggestions for future reports. *Br. J. Anaesth.* **2016**, *117*, 569-575, 10.1093/bja/aew320.
10. Montenij, L.J.; Buhre, W.F.; Jansen, J.R.; Kruitwagen, C.L.; de Waal, E.E.; Methodology of method comparison studies evaluating the validity of cardiac output monitors: A stepwise approach and checklist. *Br. J. Anaesth.* **2016**, *116*, 750-758, 10.1093/bja/aew094.
11. Olofsen, E.; Dahan, A.; Borsboom, G.; Drummond, G.; Improvements in the application and reporting of advanced Bland-Altman methods of comparison. *J. Clin. Monit. Comput.* **2015**, *29*, 127-139, 10.1007/s10877-014-9577-3.
12. Chhapola, V.; Kanwal, S.K.; Brar, R.; Reporting standards for Bland-Altman agreement analysis in laboratory research: A cross-sectional survey of current practice. *Ann. Clin. Biochem.* **2015**, *52 Pt 3*, 382-386, 10.1177/0004563214553438.
13. Berthelsen, P.G.; Nilsson, L.B.; Researcher bias and generalization of results in bias and limits of agreement analyses: A commentary based on the review of 50 Acta Anaesthesiologica Scandinavica papers using the Altman-Bland approach. *Acta. Anaesthesiol. Scand.* **2006**, *50*, 1111-1113, 10.1111/j.1399-6576.2006.01109.x.
14. Mantha, S.; Roizen, M.F.; Fleisher, L.A.; Thisted, R.; Foss, J.; Comparing methods of clinical measurement: Reporting standards for Bland and Altman analysis. *Anesth. Analg.* **2000**, *90*, 593-602.
15. Preiss, D.; Fisher, J.; A measure of confidence in Bland-Altman analysis for the interchangeability of two methods of measurement. *J. Clin. Monit. Comput.* **2008**, *22*, 257-259, 10.1007/s10877-008-9127-y.
16. Bland, J.M.; Altman, D.G.; Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* **1995**, *346*, 1085-1087.

Keywords

agreement; Bland–Altman plot; confidence interval; interrater; Limits of Agreement; method comparison; repeatability; reporting; reproducibility; Tukey mean-difference plot