Sample size determination in method comparison and observer variability studies

Gerke, Oke; Pedersen, Andreas Kristian; Debrabant, Birgit; Halekoh, Ulrich; Möller, Sören

Published in:
Journal of Clinical Monitoring and Computing

DOI:
10.1007/s10877-022-00853-x

Publication date:
2022

Document version:
Accepted manuscript

Citation for pulished version (APA):
Gerke, O., Pedersen, A. K., Debrabant, B., Halekoh, U., & Möller, S. (2022). Sample size determination in method comparison and observer variability studies. *Journal of Clinical Monitoring and Computing*, *36*(5), 1241-1243. https://doi.org/10.1007/s10877-022-00853-x

Sample size determination in method comparison and observer variability studies

Oke Gerke[1,2,*], Andreas Kristian Pedersen[3,4], Birgit Debrabant[5], Ulrich Halekoh[6], Sören Möller[1,7]

[1]*Department of Clinical Research, University of Southern Denmark, Odense, Denmark*

[2]*Department of Nuclear Medicine, Odense University Hospital, Odense, Denmark; ORCID: https://orcid.org/0000-0001-6335-3303*

[3]*Department of Research and Learning, Hospital of Southern Jutland, Aabenraa, Denmark; ORCID: https://orcid.org/0000-0001-5406-1675*

[4]*Department of Regional Health Research, University of Southern Denmark, Odense, Denmark*

[5]*Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark; ORCID: https://orcid.org/0000-0002-1964-3204*

[6]*Epidemiology, Biostatistics and Biodemography, University of Southern Denmark, Odense, Denmark; ORCID: https://orcid.org/0000-0003-3679-8424*

[7]*Open Patient data Explorative Network, Odense University Hospital, Odense, Denmark; ORCID: https://orcid.org/0000-0003-0858-4269*

**Correspondance to:* Oke Gerke; *e-mail: oke.gerke@rsyd.dk*

Manuscript category: Technical Note

Word count: 1113 (main text); 1966 words (all inclusive)

Sample size determination in method comparison and observer variability studies

Abstract

The comparison of two quantitative measuring devices is often performed with the Limits of Agreement proposed by Bland and Altman in their seminal Lancet paper back in 1986. Sample size considerations were rare for such agreement analyses in the past, but recently several proposals have been made depending on how agreement is to be assessed and the number of replicates to be used. We have summarized recent developments and recommendations in various situations including a distinction between method comparison and observer variability studies. These include current state-of-the-art analysis of and reporting guidelines for agreement studies. General recommendations close the paper.

Keywords Agreement · Bland-Altman analysis · Limits of Agreement · Repeatability · Reproducibility · Sample size

1 Method comparison studies

Formal sample size motivations have been scarce for agreement studies according to the *Guidelines for Reporting Reliability and Agreement Studies* [1]. Assuming normally distributed differences between two measurement methods with unknown mean μ and standard deviation σ, the paired quantiles of interest in Bland-Altman analysis are $\theta_{0.025}$ and $\theta_{0.975}$ [2]. Jan and Shieh proposed an exact 95% confidence interval $\{\widehat{\omega}_L, \widehat{\omega}_U\}$ to cover the central 95% proportion of the differences (see their Supplemental Material A and D for implementations with SAS/IML and R, respectively) [3]. They proposed to base the sample size on either the expected width of $\{\widehat{\omega}_L, \widehat{\omega}_U\}$ which must not exceed a predefined benchmark value Δ or the observed width of $\{\widehat{\omega}_L, \widehat{\omega}_U\}$ that will not exceed Δ

with an assurance probability of, say, 90% (see their Supplemental Material B, E and C, F, respectively). The former approach leads to similar sample sizes than the latter with an assurance probability of 50%. Shieh advanced this procedure into a formal hypothesis test and compared the relative performance of Type I error rate for their approach to previous two one-sided tests (TOST) procedures [4]. Shieh concluded that TOST procedures in the context of Bland-Altman analysis are too conservative, i.e. falling short of the nominal level, hence corresponding sample size formulas are problematic.

In the context of variance component analysis, the *repeatability coefficient* (RC) is derived from within-subject variance σ_w^2 and estimated as $1.96\sqrt{2} \cdot \hat{\sigma}_w^2$. The RC is an estimate for the limit within which 95% of differences are expected to lie [5,6]. The RC can be used with multiple assessments from each subject; in case of single measurements from each subject by each method, the RC coincides with half the width of the Bland-Altman Limits of Agreement [7]. Yi and colleagues proposed an equivalence test for agreement in case of *k* repeated measurements from each subject ($k \geq 2$), using analysis of variance [8]. The test aims at confirming that $1.96\sqrt{2} \cdot \hat{\sigma}_w^2$ is small enough to be acceptable. It can be formulated as testing H_0: $\sigma_w^2 \geq \sigma_U^2$ against H_1: $\sigma_w^2 < \sigma_U^2$, with predefined unacceptable within-subject variance σ_U^2. Denoting the difference between σ_U^2 and the assumed population within-subject variance σ^2 as Δ ($\Delta = \sigma_U^2 - \sigma^2$, with $\sigma^2 < \sigma_U^2$), the sample size is derived from determining the degrees of freedom (*df*) that make both sides of the following equation equal:

$$\frac{\chi^2_{df,1-\beta}}{\chi^2_{df,\alpha}} = \frac{\sigma_U^2}{\sigma_U^2 - \Delta}.$$

Here, α and $1-\beta$ represent the significance level and the power, respectively. The number of subjects to be included is derived from *df* and depends on the number of

repeated measurements *k*. For k=2, the number of subjects is equal to the *df*; for *k*>2, it is equal to *df*/(*k*−1).

Employing repeated measurements enables the investigation of variance parameters that describe the uncertainty between and within subjects. Increasing the planning complexity with repeated measurements suggests determining sample sizes on simulation-based methodology [9,10].

Carstensen pointed out that many observations are required to produce stable variance estimates [10]. He assumed scaled Chi-squared distributions for variance estimates, assessed approximate 95% confidence intervals for these, and compared the widths of the 95% confidence intervals for 20 to 500 *df*. Based on these assessments, he made a rough, general recommendation of 50 subjects with three repeated measurements on each method.

2 Observer variability analysis

Though, technically, Bland-Altman Limits of Agreement can be equally applied in methods comparison and observer variability studies, usually only very few (most often two) fixed methods are compared with each other, whereas interrater variability assessments will ultimately aim at generalizability of, say, clinical readings, independent of a specific set of raters employed. Christensen and colleagues provided sample size motivations when using the Limits of Agreement with the mean (LOAM) for multiple observers [11]. The measurements are assumed to follow an additive two-way random effects model, and sample size considerations are based on the width of confidence intervals for the proposed LOAM. They ascertained that a higher precision for the confidence intervals is obtained by increasing the number of observers while increasing the number of subjects is not sufficient. This underlines the inherent difference between method and observer comparisons, and it mirrors the need to

illuminate interrater variability in multicenter studies. Christensen and colleagues made an R-package, R-scripts, and their example for the LOAM calculations available in a GitHub repository.

3 Analysis and reporting

Olofsen and colleagues presented a formal description of more advanced Bland-Altman analysis models employing repeated measurements and provided a freely available online implementation of it [12,13]. These methods are based on analysis of variance and make, therefore, use of normality assumptions. Taffé asserted that Bland-Altman Limits of Agreement may be misleading when the variances of the measurement errors of the two methods are different [14,15]. In this case, he proposed a set of graphs that support the investigator to assess bias, precision, and agreement between two measurement methods. Corresponding sample size considerations would have to be based on simulation studies.

Abu-Arafeh and colleagues reviewed the reporting of Bland-Altman analysis across five anesthetic journals and derived a list of 13 key features for adequate presentation of a Bland and Altman analysis (see their Table 1 in [16]). Likewise, the *Guidelines for Reporting Reliability and Agreement Studies* comprised 15 items to keep in mind for transparent reporting (see their Table 1 in [1]).

4 Recommendations

The necessary assumptions for any sample size rationale and the targeted level of precision require careful planning in light of the research context and the pre-specified study goal [3]. One general advice is, though, that the sampling procedure should result in the inclusion of study subjects contributing with measurements across the whole measurement range of clinical interest and relevance [10]. The Preiss-Fisher procedure

[17] is a tool for the visual assessment to this end (for an exemplification, see, for instance [18]). Using the *Guidelines for Reporting Reliability and Agreement Studies* [1] already in the planning phase of a study will support purposive rigor.

In method comparison studies with single measurements by each method, the sample size calculations can be liberally based on the expected width for an exact 95% confidence interval to cover the central 95% proportion of the differences [3]. A more conservative approach, resulting in larger sample sizes, would be to require that the observed width of above exact 95% confidence interval will not exceed a predefined benchmark value Δ with an assurance probability exceeding 50% [3]. In case of k repeated measurements from each subject ($k \geq 2$), the equivalence test for agreement proposed by Yi and colleagues can be used [8]. In observer variability analysis with multiple observers, sample size considerations can be based on the width of confidence intervals for the proposed LOAM [11]. R-scripts are readily available for all sample size calculations, especially those that have to be solved iteratively [3,4,11].

Acknowledgments The authors would like to thank research librarian Mette Brandt Eriksen, PhD (University Library of Southern Denmark), for assisting with reviewing the literature. Moreover, the authors would like to express their gratitude to an anonymous reviewer and an associate editor for their helpful comments on an earlier version that improved the manuscript.

Statements & Declarations

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

Author Contributions OG contributed to the conception of the work, and AKP acquired all materials. All authors assessed and interpreted research articles for this work, and OG drafted the manuscript. All authors revised it critically for important intellectual content, approved the final version to be published, and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL (2011) Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 64:96–106. https://doi.org/10.1016/j.jclinepi.2010.03.002

2. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307-310. https://doi.org/10.1016/S0140-6736(86)90837-8

3. Jan SL, Shieh G (2018) The Bland-Altman range of agreement: Exact interval procedure and sample size determination. Comput Biol Med 100:247–252. https://doi.org/10.1016/j.compbiomed.2018.06.020

4. Shieh G (2020) Assessing agreement between two methods of quantitative

measurements: Exact test procedure and sample size calculation. Stat Biopharm Res 12:352-359. https://doi.org/10.1080/19466315.2019.1677495

5. Bland JM, Altman DG (1999) Measuring agreement in method comparison studies. Stat Methods Med Res 8:135–160. https://doi.org/10.1177/096228029900800204

6. Bland JM (2015) Frequently asked questions on the design and analysis of measurement studies. https://www-users.york.ac.uk/~mb55/meas/comfaq.htm. Accessed 08 Feb 2022

7. Gerke O, Vilstrup MH, Segtnan EA, Halekoh U, Høilund-Carlsen PF (2016) How to assess intra- and inter-observer agreement with quantitative PET using variance component analysis: a proposal for standardization. BMC Med Imaging 16:54. https://doi.org/10.1186/s12880-016-0159-3

8. Yi Q, Wang PP, He Y (2008) Reliability analysis for continuous measurements: Equivalence test for agreement. Stat Med 27:2816–2825. https://doi.org/10.1002/sim.3110

9. Choudhary PK, Nagaraja HN (2017) Measuring Agreement. Models, Methods, and Applications. Wiley, Hoboken, NJ, pp 279-287

10. Carstensen B (2010) Comparing Clinical Measurement Methods. Wiley, Chichester, UK, pp 127-131

11. Christensen HS, Borgbjerg J, Børty L, Bøgsted M (2020) On Jones et al.'s method for extending Bland-Altman plots to limits of agreement with the mean for multiple observers. BMC Med Res Methodol 20:304. https://doi.org/10.1186/s12874-020-01182-w

12. Olofsen E, Dahan A, Borsboom G, Drummond G (2015) Improvements in the application and reporting of advanced Bland-Altman methods of comparison. J Clin Monit Comput 29:127–139. https://doi.org/10.1007/s10877-014-9577-3

13. Olofsen E (2021) Webpage for Bland-Altman Analysis. https://sec.lumc.nl/method_agreement_analysis. Accessed 11 Nov 2021

14. Taffé P (2020) Assessing bias, precision, and agreement in method comparison studies. Stat Methods Med Res 29:778-796. https://doi.org/10.1177/0962280219844535

15. Taffé P (2021) When can the Bland & Altman limits of agreement method be used and when it should not be used. J Clin Epidemiol 137:176-181. https://doi.org/10.1016/j.jclinepi.2021.04.004

16. Abu-Arafeh A, Jordan H, Drummond G (2016) Reporting of method comparison studies: A review of advice, an assessment of current practice, and specific suggestions for future reports. Br J Anaesth 117:569–575. https://doi.org/10.1093/bja/aew320

17. Preiss D, Fisher J (2008) A measure of confidence in Bland-Altman analysis for the interchangeability of two methods of measurement. J Clin Monit Comput 22:257–259. https://doi.org/10.1007/s10877-008-9127-y

18. Gerke O (2020) Reporting Standards for a Bland-Altman Agreement Analysis: A Review of Methodological Reviews. Diagnostics (Basel) 10:334. https://doi.org/10.3390/diagnostics10050334