

LETTER TO THE EDITOR



Reply to the letter by Siamak Sabour

We thank the author of the letter for reading our paper carefully and emphasizing a very important statistical methodology which is indeed often applied in a wrong way. The author stresses two points about our analysis: first, it is stated that the ICC instead of the ICC should be used and, second, that the κ statistic is not a valid tool to assess reliability but a weighted κ should be calculated.

We want to present our point of view and show the validity of our analysing strategy:

The first issue refers to the ICC. Obviously, we did not apply Pearson's correlation coefficient (r) as stated by the author of the letter (which would not be a valid approach to assess reliability but is only a measure of correlation). Instead, we used the intra-class correlation coefficient (ICC), which is distinctly different from Pearson's r and did not calculate an 'inter-class correlation' (this term was never used in our article). The author of the letter states that we should not use the 'inter-class correlation' but the 'intra-class correlation coefficient' using the acronym 'ICCC' and the article by Lawrence [1] as reference. However, Lawrence introduced an alternative approach to the ICC: the concordance correlation coefficient. The acronym 'CCC' is commonly used for this measure (not 'ICCC') [2]. This does not equal the ICC as clearly described in Lawrence's article. Secondly, the ICC is a standard approach to analyse reliability [3–5], and as stated above, not equal to Pearson's r . Furthermore, the CCC and the ICC both lead to very similar results. An overview can be found in the article by Watson and Petrie [3].

The second issue about the κ -statistic is a very interesting and important concern. It is true that there is a lot of discussion about that measure. The author mentioned two of the common criticisms: The dependency of the κ -statistic on the marginal prevalence and the on the number of categories. However, these criticisms cannot be applied to all situations. For example, Vach (2005) [6] states that we should stop criticizing the κ -statistic: Besides other issues, Vach addresses the problem of the dependency on the prevalence (and the main criticism stated by the author of the letter) and explains that the κ -statistic is a good measure of reliability able to adjust for imbalances in marginal prevalence. Regarding the dependency of the κ -statistic on the observed prevalence, Vach states that 'this type of dependence does not, however, indicate any drawback or weakness of the κ -statistic. It merely reflects the basic aim of the κ -statistic' (p. 656). We certainly agree with this statement as the statistic measures the amount of agreement observed on top of the agreement by chance. Furthermore, Vach says that 'the expected amount of agreement by chance depends on the marginal prevalences, and so the correction depends on it' (p. 656). He summarizes 'it is the aim of the κ -statistic that identical agreement rates should be judged differently in the light of the marginal prevalences, which determine the expected amount of chance agreement. Hence, it makes no sense to criticize κ for exactly fulfilling this property' (p. 659).

The comparability of the results of a reliability analysis between different variables with different numbers of categories might be questionable in general. In our article, we exclusively check variables with only two categories for reliability. Therefore, we think that this issue does not apply to our article. Furthermore, the weighted κ does not solve this problem because it is only applicable to ordinal data and it also depends on the number of categories as described by Benner and Kliebisch [7]. Obviously, different analysts have different opinions about the use of the κ -statistic. However, we agree with Vach on the properties of κ . We aimed at correcting the measure of agreement for the marginal prevalence and, hence, the κ -statistic is a valid tool to analyse the data.

We thank the author of the letter for stressing this important topic and do not wish to criticize his opinion. Indeed we are thankful for the opportunity to discuss the characteristics of the applied statistical approach. We hope that we were able to show the validity of our analysing strategy.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

- [1] Li LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 1989;45:255–268.
- [2] Liu J, Tang W, Chen G, et al. Correlation and agreement: overview and clarification of competing concepts and measures. *Shanghai Arch Psychiatry*. 2016;28:115–120.
- [3] Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. *Theriogenology*. 2010;73:1167–1179.
- [4] Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep*. 1966;19:3–11.
- [5] Stolarova M, Wolf C, Rinker T, et al. How to assess and compare inter-rater reliability, agreement and correlation of ratings: an exemplary analysis of mother-father and parent-teacher expressive vocabulary rating pairs. *Front Psychol*. 2014;5:509.
- [6] Vach W. The dependence of Cohen's kappa on the prevalence does not matter. *J Clin Epidemiol*. 2005;58:655–661.
- [7] Brenner H, Kliebsch U. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*. 1996;7:199–202.

Bernadette Pretzl

Section of Periodontology, Department of Conservative
Dentistry Clinic for Oral, Dental and Maxillofacial Diseases,
University Hospital Heidelberg, Heidelberg, Germany
✉ bernadette.pretzl@med.uni-heidelberg.de

Lorenz Uhlmann

Institute of Medical Biometry and Informatics (IMBI), University
Hospital Heidelberg, Heidelberg, Germany

Received 21 May 2017; accepted 22 May 2017

© 2017 Acta Odontologica Scandinavica Society