Interobserver agreement issues in radiology

Mehdi Benchoufi, E. Matzner-Lober, Nicolas Molinari, A.-S. Jannot, P. Soyer

Title: **Interobserver agreement issues in radiology**

Short title: Interobserver agreement

Authors

M. Benchoufi [a,b]
E. Matzner-Lober [c]
N. Molinari [d]
A.-S. Jannot [b,e, f]
P. Soyer [b, g*]

Affiliations

[a] Center of Research in Epidemiology and Statistics (CRESS), French Institute of Health and Medical Research (INSERM), National Institute of Agricultural Research (INRA), Paris, France.

[b] Université de Paris, 75006 Paris, France

[c] CREST ENSAE, UMR 9194, 91120 Palaiseau, France

[d] IMAG, CNRS, Univ Montpellier, Department of Statistics, CHU Montpellier, 34000 Montpellier, France

[e] Centre de Recherche des Cordeliers, Sorbonne Université, INSERM, 75006, Paris, France.

[f] Assistance Publique-Hôpitaux de Paris (AP-HP), Department of Biostatistics, Medical Informatics and Public Health, Hôpital Européen Georges Pompidou, 75015, Paris, France
[g] Assistance Publique-Hôpitaux de Paris (AP-HP), Department of Radiology, Hôpital Cochin, 75014, Paris, France

* **Corresponding author**: philippe.soyer@aphp.fr

Abstract: Agreement between observers (*i.e.*, inter-rater agreement) can be quantified with various criterion but their appropriate selections are critical. When the measure is qualitative (nominal or ordinal), the proportion of agreement or the kappa coefficient should be used to evaluate inter-rater consistency (*i.e.*, inter-rater reliability). The kappa coefficient is more meaningful that the raw percentage of agreement, because the latter does not account for agreements due to chance alone. When the measures are quantitative, the intraclass correlation coefficient (ICC) should be used to assess agreement but this should be done with care because there are different ICCs so that it is important to describe the model and type of ICC. The Bland-Altman method can be used to assess consistency and conformity but its use should be restricted to comparison of two raters.

Index terms: Reproducibility of results; Interobserver agreement; Radiology; Kappa test; Intraclass correlation coefficient

1. Introduction

For decades, radiology was restricted to data acquisition and image interpretation, with a final diagnosis made by a radiologist. However, with progress in software and computerized image analysis, radiology is now used to predict response to treatment, tumor mutation, disease-free survival in patients with cancer and provide myriad other information [1, 2, 3]. However, as a fundamental pillar, the human eye (and brain) is still the first step of any process but the skill of a radiologist depends on many variables and two different radiologists with different experiences may give different interpretations or different ratings. This variability in perception and interpretation is a critical issue in radiology. In this regard, the results of an imaging technique often depend on the intrinsic qualities of the observer. It is now recommended to use different observers and independent readings to have a clear idea of the possible variations of a given imaging test across radiologists [4]. This is because a new imaging procedure must be robust and have a good interobserver agreement to have utility in everyday practice.

The purpose of this article is to stress the importance of interobserver variability assessment and provide simple rules on how to use the most appropriate statistical test to assess interobserver variability, depending on the specific setting.

2. Why interobserver agreement is important?

Collected data are meaningful when they are fully representative of a given variable. Many causes may produce no representative or erroneously considered data. As an example, ill-defined contour of a tumor on imaging may induce errors in dimension and volume measurement that may further affect the calculation of the apparent diffusion coefficient of the volume in question on diffusion-weighted magnetic resonance imaging. Another problem in imaging is when the data (or imaging criteria) are subjective. This applies to criteria such as lesion homogeneity *vs.* heterogeneity or degree of enhancement (i.*e*., no enhancement, mild, moderate, marked and intense enhancement). It is clear that such criteria greatly depend on observer perception.

In radiology, agreement between readers can be analyzed when comparing different imaging techniques or determining presence of a given variable (*e.g.*, imaging criterion). Agreement between readers (inter-rater agreement) can be quantified with various settings but their appropriate selection is critical and depends on the nature of the measurements.

3. Agreement for qualitative measurements

When the measure is qualitative (nominal or ordinal), the proportion of agreement or Cohen's kappa coefficient should be used to evaluate inter-rater consistency (*i.e.*, inter-rater reliability). However, the kappa coefficient is more meaningful than the raw proportion of agreement that does not account for agreements due to chance alone. When more than two readers are available, the Fleiss's kappa coefficient should be used [5]. A common error is the use of a chi-square test (χ^2) that evaluates statistical association between two readers but not a measure of agreement. Another error is to perform a McNemar test to take into account the fact that data are paired, but this test is not relevant in this situation as paired analysis is appropriate when observations are paired but not when two observers rate the same observation.

3.1. Agreement for binary data

The simplest way of calculating agreement for binary data is the proportion of agreement that is obtained by summing all agreements (*i.e.*, the numbers in the diagonal of the confusion matrix) and dividing the sum by the numbers of all observations usually denoted by po. It is also called accuracy. Although this method is intuitive, it does not take into account agreements that can be obtained by chance alone and that is the essence of the Kappa statistics. It takes into account the percentage of agreement and the percentage of agreement expected by chance commonly denoted by pc. Kappa is obtained as the ratio of po-pc divided by 1-pc. So kappa could be negative (disagreement) but is upper bounded by 1. The interpretation of kappa values can be made using various classifications. Landis & Koch classification is the most commonly used (0.00, poor; 0.00-0.20, slight; 0.21-0.40, fair; 0.41-0.60, moderate; 0.61-0.80, substantial and 0.81-1.00, almost perfect) [6]. However, other semi-quantitative classifications exist so that it is recommended that the authors specifically indicate the classification they use in the Materials & Methods section of their article [7].

3.2. Agreement for multiclass data (nominal and ordinal data)

The limitation of the Cohen's kappa coefficient is that it does not account for the degree of disagreement. The unweighted Kappa test takes into account the percentage of agreement and the percentage of agreement expected by chance. However, it does not take into account the degree of disagreement; meaning that all disagreements are considered equally with zero weight to all disagreement cells of the matrix and it is helpful to provide a measure that gives also the extent of disagreement.

Although the use of the weighted Kappa test for nominal data is debated because the Kappa value depends on the prevalence in each category and the number of categories, it is then more appropriate to use the weighted Kappa coefficient for ordinal data. Classically the weights are evaluated according to its distance to the diagonal in the confusion table and are increasing when the cell is away from the diagonal. This choice is practically relevant for ordinal data. We recommend to discuss these issues with a statistician or to use dedicated statistical software. Because the Kappa statistic is affected by the prevalence of a variable, similar to predictive values, its use is not recommended for rare findings [8].

Several statistical software give a corresponding *P* values but, a significant *P* value only indicates that the kappa value is different from 0 and not the strength of agreement [9], so that the *P* value is irrelevant to the evaluation of agreement with the Kappa test [10]. On the opposite, it is recommended to provide a 95% confidence interval (CI) of kappa value that is more meaningful. After calculating the standard error of kappa (SE_K), the 95% CI is Kappa ± 1.96 SE_K.

4. Agreement for quantitative data

When the measures are quantitative such as scores or volumes, the ICC should be used to assess agreement. Many different ICCs exist and one should describe the model, type and definition used in the calculation in the appropriate setting. It is important to differentiate between absolute agreement and consistency. As an example, when one reader constantly grades a score higher than the other reader, there is a high consistency but poor agreement. The difficulty here is that ICC can measure both agreement and consistency (*i.e.*, association between the measurements performed by each observer). By definition the ICC for absolute agreement is always smaller than the ICC for consistency. Consistency may be considered to search for a systemic bias between independent readers. The absolute agreement definition should be selected for test-retest and intra-rater reliability assessment.

4.1. Intraclass correlation coefficient (ICC)

One important concern is that there are different ICCs that must be carefully selected to obtain meaningful information. The choice of ICC depends on the design of the study and the priority for type of agreement (*i. e.*, consistency or absolute agreement). A clear and detailed description of the criteria used for best selection of ICC is discussed elsewhere [11].

The appropriate model of ICC can be selected among a one-way random-effect, a two-way random effect or a two-way mixed-effects model [11, 12, 13]. The one-way random-effects model should be used in imaging studies when two different imaging sets are rated by two different sets of raters, which is a rare scenario. The two-way random effects model should be used when the same set of imaging sets are rate by two (or more) different raters which are selected at random from a pool of raters. This latter model should be preferred because it provides results that can be generalized to any raters with the same characteristics (*i.e.*, expertise, years of experience or specialization). The two-way mixed-effects model should be selected when the same set of imaging examinations are done by selected raters that are the only raters of interest. This is the case when the goal of the study is to determine the reliability of the raters involved in the study because the results cannot be generalized to other raters. The two or more raters are considered as fixed. The two-way mixed effect should be also selected to test intra rater reliability and test-retest reliability [12].

The ICC estimate is only an expected value so that ICCs should be reported with their corresponding 95% CIs. It is commonly admitted that ICC < 0.5 indicates poor reliability, ICC between 0.5 and 0.75 indicates moderate reliability, ICC between 0.75 and 0.9 indicates good reliability and ICC > 0.90 indicates excellent reliability [11]. As a limitation, ICC does not give information about the variability of data with the magnitude of measurements. Alternative approaches include the jackknife methods, the corrected F test, the ordinal regression fixed-effects method, and the random effects test [14]. Regression methods may also be used to assess consistency, however ICC is more commonly used and is needed.

4.2. Bland-Altman method

The Bland-Altman approach can be used to assess consistency and conformity but its use should be restricted to comparison of two raters. The Bland-Altman method provides a visual assessment of differences between quantitative evaluations by showing a residual like plot of the difference of the observed pairs of readings against their mean values [15]. This method defines a limit of agreement by combining the mean (m) and the standard deviation (SD) of the difference as m ± 1.96 SD. A scatterplot is drawn to understand dispersion of variables using x-axis (mean) and y-axis (difference) of the two measures. Good agreement is displayed as a diminished scattering of points with points lying close to the mean bias line. This type of plot allows identification of outliers as well as an examination for trend using linear regression analysis. The plot of difference against mean in Bland-Altman analysis may allow investigating for possible correlation between measurement error (difference between two methods) and the assumed true value (average value of two methods) [16].

The Bland-Altman approach can be used to compare two distinct methods or one method to a standard of reference [17]. However, before using the Bland Altman, the variables must be checked for normal distribution. Another limitation of the Bland-Altman analysis is that it is not appropriate for comparing repeated measurements, unless a random effect model is added to the analysis [18, 19].

5. Conclusion

Intra- and inter observer agreement is a critical issue in imaging [20, 21, 22]. This can be assessed using different settings, depending on the study design and the types of data. When categorical data are reported, agreement should be corrected for chance by using κ statistics. The ICC is an acceptable summary measure for agreement between continuous data, but it does not give information about the variability of data with the magnitude of measurements.

Disclosure of interest

The authors declare that they have no competing interest in relation with this article.

Funding

This study received no funding.

CRediT authorship contribution statement

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

References

1. Yamashita K, Hatae R, Hiwatashi A, Togao O, Kikuchi K, Momosaka D. Predicting TERT promoter mutation using MR images in patients with wild-type IDH1 glioblastoma. Diagn Interv Imaging 2019;100:411–9.
2. Vourtsis A. Three-dimensional automated breast ultrasound: technical aspects and first results. Diagn Interv Imaging 2019;100:579–92.
3. Madico C, Herpe G, Vesselle G, Boucebci S, Tougeron D, Sylvain C, et al. Intra peritoneal abdominal fat area measured from computed tomography is an independent factor of severe acute pancreatitis. Diagn Interv Imaging 2019;100:421–6.
4. Soyer P. Agreement and observer variability. Diagn Interv Imaging 2018;99:53–4.
5. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378.
6. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
7. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012;22:276–82.

8. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problem of two paradoxes. J Clin Epidemiol 1990;43:543–9.
9. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37:360–3.
10. Jakobsson U, Westergren A. Statistical methods for assessing agreement for ordinal data. Scand J Caring Sci 2005;19:427–31.
11. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15:155–163.
12. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.
13. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1:30–46.
14. Obuchowski NA, Zepp RC. Simple steps for improving multiple-reader studies in radiology. AJR Am J Roentgenol 1996; 166:517–21.
15. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–10.
16. Mantha S, Roizen MF, Fleisher LA, Thisted R, Foss J. Comparing methods of clinical measurement: reporting standards for Bland and Altman analysis. Anesth Analg 2000;90:593–602.
17. Doğan NO. Bland-Altman analysis: a paradigm to understand correlation and Agreement. Turk J Emerg Med 2018;18:139–41.
18. Woodman RJ. Bland-Altman beyond the basics: creating confidence with badly behaved data. Clin Exp Pharmacol Physiol 2010;37:141–2.
19. Tipton E, Shuster J. A framework for the meta-analysis of Bland-Altman studies based on a limits of agreement approach. Stat Med 2017;36:3621–35.
20. Bankier AA, Levine D, Halpern EF, Kressel HY. Consensus interpretation in imaging research: is there a better way? Radiology 2010;257:14–17.
21. Choi MH, Oh SN, Park GE, Yeo DM, Jung SE. Dynamic contrast-enhanced MR imaging of the rectum: Correlations between single-section and whole-tumor histogram analyses. Diagn Interv Imaging 2018;99:537–45.
22. Mulé S, Hoeffel C. Evaluation of measurement variability in quantitative analyses: Application to dynamic contrast-enhanced MRI histogram analysis in rectal cancer. Diagn Interv Imaging 2018;99:421–2.