Editorial

eISSN 2005-8330
https://doi.org/10.3348/kjr.2023.0965
Korean J Radiol 2024;25(4):325-327

Korean Journal of Radiology
KJR

Statistical Methods for the Analysis of Inter-Reader Agreement Among Three or More Readers

Kyunghwa Han[1], Leeha Ryu[2]

[1]Department of Radiology, Research Institute of Radiological Science, and Center for Clinical Imaging Data Science, Yonsei University College of Medicine, Seoul, Republic of Korea
[2]Department of Biostatistics and Computing, Yonsei University Graduate School, Seoul, Republic of Korea

Keywords: Agreement; Reliability; Reproducibility; Repeatability; Statistical method; Statistical analysis; Multiple; Reader; Rater; Observer

The inter-reader agreement is a key imaging interpretation-related outcome parameter in radiological research. The interpretation of medical images can be affected by the subjective assessment of readers. Therefore, the measure of inter-reader agreement is crucial. A substantial degree of reliability is required in clinical research involving image interpretation.

To evaluate the extent of inter-reader agreement, measures of agreement such as kappa, intraclass correlation coefficient (ICC), and concordance correlation coefficient (CCC) are commonly employed [1]. In a study with only two readers, statistical methods for analyzing inter-reader agreement can be easily applied and interpreted. Nevertheless, research studies may necessitate the involvement of three or more readers to enhance the generalizability of results across diverse clinical practices [2]. Researchers maybe less acquainted with statistical methods for analyzing inter-reader agreements involving three or more readers compared to methods for two readers. Therefore, this article provides a brief guide on statistical methods for analyzing inter-reader agreement among three or more readers. The recommended methods are listed in Table 1. Statistical methods were classified according to the scale of the readers' interpretations: binary (e.g., presence vs. absence of a finding/disease) or nominal scale (e.g., category of imaging findings), ordinal scale (e.g., a 5-point Likert scale for image quality from 1 for poor quality to 5 for good quality), and continuous scale (e.g., size measurement of a lesion).

Received: October 1, 2023 **Revised:** November 19, 2023
Accepted: December 5, 2023
Corresponding author: Kyunghwa Han, PhD, Department of Radiology, Research Institute of Radiological Science, and Center for Clinical Imaging Data Science, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea
• E-mail: khhan@yuhs.ac

Binary or Nominal Scale

Cohen's kappa [3] was used by only two readers. However, for three or more readers, statistical analysis is more complex because all possible combinations between multiple readers are included in calculating the statistics of agreement. Fleiss's kappa [4] and Conger's kappa [5] are well-known alternatives to Cohen's kappa for three or more readers. Light's kappa [6] represents the average Cohen's kappa value calculated for all two-reader combinations. Additionally, to overcome some drawbacks of these chance-corrected measures such as prevalence paradoxes [7], other measures such as Gwet's agreement coefficient 1 (AC1) [8], Brennan–Prediger's BP [9], and Krippendorff's α [10] are used.

Ordinal Scale

Statistical methods for evaluating the agreement of ordinal scales among three or more readers have not been

Table 1. Recommended statistical methods for analysis of interreader agreement among three or more readers

Type of ratings	Statistical method
Binary or nominal	Brennan-Prediger's BP
	Conger's Kappa
	Fleiss' Kappa
	Gwet's AC1
	Krippendorff's α
	Light's Kappa
Ordinal	Generalized weighted Kappa
	Gwet's AC2
	Intraclass correlation coefficient
	Light's Kappa
Continuous	Concordance correlation coefficient
	Intraclass correlation coefficient

The methods are presented in alphabetical order rather than by frequency of use or order of recommendation.
AC = agreement coefficient

firmly established, and these methods are generally not available in user-friendly statistical software programs. Therefore, published radiological studies often inadequately use the Fleiss Kappa statistic, either neglecting the ordinal nature of the data or applying Cohen's weighted kappa to all possible reader pairs. Weighted Kappa [11] is another version of Cohen's kappa that incorporates the weights of pairs of categories in cases with ordinal ratings. The weights can be assigned differently when calculating the kappa to account for varying degrees of agreement between the ratings. However, they can only be applied to data from two readers because they are based on the cross-tabulation of the ratings between the two readers.

Improved methods are available for this purpose (Table 1). Light's kappa [6], which uses the average of the weighted kappa values obtained for all possible reader pairs, maybe a potentially useful strategy; however, the method does not consider the agreement among all readers. Generalized weighted Kappa including Gwet's AC2 (Table 1) that incorporate different types of weights to the Kappa for binary or nominal scale is suggested in the literature and implemented in R package 'irrCAC' [12]. Although the statistical properties of these approaches have not been fully demonstrated, researchers have attempted to demonstrate them.

Continuous Scale

Inter-reader agreement is generally assessed based on reliability statistics such as ICC [13] or CCC (Table 1) [14]. Additionally, relevant statistical methods for quantitative imaging parameters, which are measured on a continuous scale, are present in the Radiological Society of North America-Quantitative Imaging Biomarkers Alliance (RSNA-QIBA), and COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) initiative [15,16]. In addition, a graphical presentation through the Bland-Altman plot accommodated by multiple readers [17] can be created by calculating the points of differences and averages of multiple measurements on the x- and y-axes, respectively, for each reader with different symbols/colors. Although the modified Bland-Altman plot cannot provide a measure of inter-reader agreement, the limits of agreement between the two methods can be estimated by considering multiple ratings by multiple readers.

Further Consideration

To evaluate inter-reader agreement on an ordinal scale, various statistics have been proposed and comparative studies [18] have been published. In particular, statistical methods for ordinal scales among three or more raters are not well known, and most of the statistics are based on nominal ratings with some weights. However, whether authors used weighted statistics often remains unclear. Chance-corrected measures of agreement are limited by the prevalence effects, imbalances among categories, and missing values. To overcome these issues, presenting the analytical results alongside the proportion of observed agreements may help readers understand inter-reader rating data. Researchers should present the statistical methods and software used to analyze the inter-reader agreement data to reflect the rating nature and promote high quality and transparency of the reporting.

Conflicts of Interest

Kyunghwa Han who is on the Statistical Consultant of the *Korean Journal of Radiology* was not involved in the editorial evaluation or decision to publish this article. The remaining author has declared no conflicts of interest.

Author Contributions

Conceptualization: Kyunghwa Han. Writing—original draft: Kyunghwa Han. Writing—review & editing: all authors.

ORCID IDs

Kyunghwa Han
https://orcid.org/0000-0002-5687-7237

https://doi.org/10.3348/kjr.2023.0965

Leeha Ryu
https://orcid.org/0000-0002-6575-9531

Funding Statement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1I1A1A01059893).

REFERENCES

1. Park JE, Han K, Sung YS, Chung MS, Koo HJ, Yoon HM, et al. Selection and reporting of statistical methods to assess reliability of a diagnostic test: conformity to recommended methods in a peer-reviewed journal. *Korean J Radiol* 2017;18:888-897
2. Atzen SL, Bluemke DA. Top 10 tips for writing your scientific paper: the radiology scientific style guide. *Radiology* 2022;304:1-2
3. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37-46
4. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971;76:378
5. Conger AJ. Integration and generalization of kappas for multiple raters. *Psychol Bull* 1980;88:322
6. Light RJ. Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychol Bull* 1971;76:365
7. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990;43:543-549
8. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol* 2008;61(Pt 1):29-48
9. Brennan RL, Prediger DJ. Coefficient kappa: some uses, misuses, and alternatives. *Educ Psychol Meas* 1981;41:687-699
10. Krippendorff K. Bivariate agreement coefficients for reliability of data. *Sociol Methodol* 1970;2:139-150
11. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;70:213-220
12. Gwet KL. *Handbook of inter-rater reliability*. 4th ed. Gaithersburg, MD: Advanced Analytics, LLC, 2014
13. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420-428
14. Carrasco JL, Phillips BR, Puig-Martinez J, King TS, Chinchilli VM. Estimation of the concordance correlation coefficient for repeated measures using SAS and R. *Comput Methods Programs Biomed* 2013;109:293-304
15. Hernaez R. Reliability and agreement studies: a guide for clinical investigators. *Gut* 2015;64:1018-1027
16. Raunig DL, McShane LM, Pennello G, Gatsonis C, Carson PL, Voyvodic JT, et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. *Stat Methods Med Res* 2015;24:27-67
17. Jones M, Dobson A, O'Brian S. A graphical method for assessing agreement with the mean between multiple observers using continuous measures. *Int J Epidemiol* 2011;40:1308-1313
18. Mitani AA, Freer PE, Nelson KP. Summary measures of agreement and association between many raters' ordinal classifications. *Ann Epidemiol* 2017;27:677-685.e4