BMC Nephrology

MATTERS ARISING

Open Access

Evaluating inter-rater reliability in the context of "Sysmex UN2000 detection of protein/creatinine ratio and of renal tubular epithelial cells can be used for screening lupus nephritis": a statistical examination

Ming Li[1,2], Qian Gao[2], Jing Yang[1*] and Tianfei Yu[3,4*]

Abstract

Background The evaluation of inter-rater reliability (IRR) is integral to research designs involving the assessment of observational ratings by two raters. However, existing literature is often heterogeneous in reporting statistical procedures and the evaluation of IRR, although such information can impact subsequent hypothesis testing analyses.

Methods This paper evaluates a recent publication by Chen et al., featured in *BMC Nephrology*, aiming to introduce an alternative statistical approach to assessing IRR and discuss its statistical properties. The study underscores the crucial need for selecting appropriate Kappa statistics, emphasizing the accurate computation, interpretation, and reporting of commonly used IRR statistics between two raters.

Results The Cohen's Kappa statistic is typically used for two raters dealing with two categories or for unordered categorical variables having three or more categories. On the other hand, when assessing the concordance between two raters for ordered categorical variables with three or more categories, the commonly employed measure is the weighted Kappa.

Conclusion Chen and colleagues might have underestimated the agreement between AU5800 and UN2000. Although the statistical approach adopted in Chen et al.'s research did not alter their findings, it is important to underscore the importance of researchers being discerning in their choice of statistical techniques to address their specific research inquiries.

Keywords Lupus nephritis, AU5800, UN2000, Inter-rater reliability, Kappa statistic, Cohen's Kappa, Weighted Kappa

*Correspondence:
Jing Yang
yangjing@hrbeu.edu.cn
Tianfei Yu
yutianfei2001@163.com
[1] Department of Software Engineering, College of Computer Science and Technology, Harbin Engineering University, 150001 Harbin, China
[2] Department of Computer Science and Technology, College of Computer and Control Engineering, Qiqihar University, 161006 Qiqihar, China
[3] Heilongjiang Provincial Key Laboratory of Resistance Gene Engineering and Protection of Bioaffiliationersity in Cold Areas, Qiqihar University, 161006 Qiqihar, China
[4] Department of Biotechnology, College of Life Science and Agriculture Forestry, Qiqihar University, 161006 Qiqihar, China

The evaluation of inter-rater reliability (IRR) is integral to research designs involving the assessment of observational ratings by two raters. However, existing literature is often heterogeneous in reporting statistical procedures and the evaluation of IRR, although such information can impact subsequent hypothesis testing analyses [1, 2]. This commentary evaluates a recent publication by Chen et al. [3], featured in *BMC Nephrology*, aiming to introduce an alternative statistical approach to assessing IRR and discuss its statistical properties. The study underscores the crucial need for selecting appropriate Kappa statistics, emphasizing the accurate computation, interpretation, and reporting of commonly used IRR statistics between two raters. The analysis focuses on the raised issues, contributing to the discourse on IRR assessment methodology. We are committed to addressing these concerns comprehensively, providing insights into the selection of suitable statistical measures for robust IRR evaluation. This effort aims to elevate the quality of research practices in IRR assessment, thereby fostering a more accurate and reliable foundation for scientific investigations across diverse fields.

Kappa statistic

When delving into the examination of nominal and categorical data, researchers commonly turn to classical statistical techniques such as the Kappa statistic and its iterations like Cohen's Kappa. These metrics are instrumental in gauging the agreement among diverse observers, a pivotal aspect for ensuring methodological rigor and reliability in research. Through their application, researchers aim to precisely evaluate the consistency of ratings, establishing a robust foundation for the integrity of their findings and the overall validity of their study.

Cohen's Kappa

Cohen's Kappa, a widely utilized statistical method for assessing IRR, presents certain limitations [4]. Primarily designed for fully-crossed designs with precisely two raters, Cohen's Kappa may exhibit biases and limitations in certain scenarios. Unlike simple percent agreement, it accounts for chance agreement, emphasizing its utility for two raters dealing with two categories or for unordered categorical variables having three or more categories [5–7]. Recognizing these constraints is crucial when employing Cohen's Kappa, urging researchers to consider alternative approaches tailored to their specific study designs and categorical data structures.

Cohen's Kappa is calculated as follows:

$$k_C = \frac{\sum_{j=1}^{n} u_{jj}\left(ii'\right) - \sum_{j=1}^{n} p_{ij}p_{i'j}}{1 - \sum_{j=1}^{n} p_{ij}p_{i'j}} \tag{1}$$

The value of $u_{jj}\left(ii'\right)$ is the proportion of objects put in the same category *j* by both raters i and i'. The value of p_{ij} is the proportion of objects that rater i assigned to category j .

Weighted Kappa

When assessing the concordance between two raters for ordered categorical variables with three or more categories, the commonly employed measure is the weighted Kappa [8–10]. Two variations of weighted Kappa exist: the linear weighted Kappa (LWK) [11] and the quadratic weighted Kappa (QWK) [12]. LWK assigns weights based on linear distances between categories, while QWK uses quadratic distances. Both LWK and QWK offer more nuanced insights into IRR compared to Cohen's Kappa. The selection between LWK and QWK hinges on the nature of the data. To ensure a comprehensive understanding of disagreements, it is advisable to report both coefficients, particularly in situations where not all disagreements hold equal significance [13]. This dual reporting strategy contributes to a more thorough evaluation of the distribution of disagreements [14], enhancing the accuracy and depth of the assessment of consistency and reliability in intricate datasets.

Weighted Kappa is calculated as follows:

$$w_{ij}^{(m)} = 1 - \left(\frac{|i-j|}{n-1}\right)^{m} \tag{2}$$

$$k_m = 1 - \frac{1 - \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^{(m)} p_{ij}}{1 - \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^{(m)} p_i q_j} \tag{3}$$

Where m≥1, p and q are relative frequencies, which reflect the proportion of frequency to the number of samples. $p_i = \sum_{j=1}^{n} p_{ij}$ and $q_i = \sum_{j=1}^{n} p_{ji}$. In special cases, k_1 is the LWK and k_2 is the QWK.

The interpretation of Kappa value

Cohen devised a classification system for interpreting Kappa values, indicating various levels of agreement [4]. However, McHugh [15] highlighted practical concerns, arguing that labeling a 61% agreement rate as "substantial" might be misleading, especially in critical settings like clinical laboratories where a 40% error rate would be significant. He emphasized the need for a higher standard, with many sources recommending a minimum interrater agreement of 80%. McHugh proposed an alternative interpretation of Kappa values, categorizing≤0.20 as no agreement, 0.21 to 0.39 as minimal agreement, 0.40 to 0.59 as weak agreement, 0.60 to 0.79 as moderate agreement, 0.80 to 0.90 as strong agreement, and values exceeding 0.90 as almost perfect agreement. This alternative approach considers the practical implications

Table 1 The Kappa coefficient between the AU5800 and UN2000

Test parameters		UN2000*						
		-	1+	2+	Total	κ_c	κ_{lW}	κ_{qW}
AU5800(mg/g)	< 150	427	21	0	448	0.858 ($p < 0.001$, 95% CI = 0.816–0.899)	0.909 ($p < 0.001$, 95% CI = 0.882–0.937)	0.947 ($p < 0.001$, 95% CI = 0.931–0.964)
	150 ~ 490	4	48	8	60			
	≧ 500	0	7	105	112			
	Total	431	76	113	620			

The data has been cited from the article published by Chen et al. [3]. κ_c: Cohen's Kappa value; κ_{lw}: linear weighted Kappa value; κ_{qw}: quadratic weighted Kappa value; CI: confidence interval.

of different Kappa values, offering a nuanced perspective on agreement levels in situations where accuracy has substantial real-world consequences. By addressing these concerns, McHugh's interpretation provides a more contextually relevant framework for understanding and applying Kappa values, particularly in critical decision-making environments.

Comparative Kappa statistics between Cohen's and weighted approaches

Chen et al. assessed the viability of utilizing automated urine sediment analysis (UN2000) for lupus nephritis screening. 284 urine samples from systemic lupus erythematosus patients were examined with UN2000, evaluating protein/creatinine ratio (P/C) and renal tubular epithelial cells. Employing biochemical analysis and microscopy as the gold standard, the Kappa consistency test demonstrated strong and good agreement for P/C and renal tubular epithelial cells (RTEC), respectively (Cohen's Kappa, 0.858). Setting P/C≥2+as the sole screening standard yielded the highest specificity, positive predictive value, and coincidence for lupus nephritis. Combining P/C≥2+or RTEC>2.8 cells/μl as the standard maximized sensitivity and negative predictive value. UN2000 proves effective in lupus nephritis screening by detecting P/C and RTEC. Yet, as mentioned earlier, in the context of three-category ordinal variables, opting for weighted Kappa is often a more suitable approach for evaluating IRR compared to Cohen's Kappa.

Upon examining the data provided by the authors, the agreement between AU5800 and UN2000 was assessed using three Kappa values with SPSSAU (https://spssau.com/) (Table 1). There was strong agreement between the two categories, with Cohen's Kappa, and almost complete agreement with LWK and QWK. As a result, LWK and QWK are the preferred measures for more sensitive evaluation that emphasizes larger differences in judgment when assessing agreement.

Conclusion

In conclusion, Chen and colleagues might have underestimated the agreement between AU5800 and UN2000. When choosing an IRR statistical test, researchers should consider variable coding, study design, and the purpose of the estimate. It's crucial to assess the statistic's suitability and explore alternatives. Although the statistical approach adopted in Chen et al.'s research did not alter their findings, it is important to underscore the importance of researchers being discerning in their choice of statistical techniques to address their specific research inquiries.

Abbreviations

IRR	inter rater reliability
LWK	linear weighted Kappa
QWK	quadratic weighted Kappa
P/C	protein/creatinine ratio
RTEC	renal tubular epithelial cells
κ_c	Cohen's Kappa value
κ_{lw}	linear weighted Kappa Value
κ_{qw}	quadratic weighted Kappa value
CI	confidence interval

Acknowledgements

Not applicable.

Author contributions

M.L. wrote the original draft of the manuscript. Q.G. was involved in the analysis and interpretation of the data. J.Y. and T.-F.Y. contributed to the conception and design of the study. All authors reviewed the manuscript.

Funding

This work was supported by the Fundamental Research Funds in Heilongjiang Provincial Universities (145209121) and the Heilongjiang Province Leading Talent Echelon Reserve Leader Funding Project. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Data availability

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 22 January 2024 / Accepted: 8 March 2024
Published online: 13 March 2024

References

1. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol. 2012;8(1):23.
2. Hughes J. Sklar's omega: a gaussian copula-based framework for assessing agreement. Stat Comput. 2022;32(3):46.
3. Chen Y, Zhao Y, Zhang Z, Cheng X, Lin J, Li J, et al. Sysmex UN2000 detection of protein/creatinine ratio and of renal tubular epithelial cells can be used for screening lupus nephritis. BMC Nephrol. 2022;23(1):328.
4. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.
5. Gao P, He W, Jin Y, Zhou C, Zhang P, Wang W, Hu J, Liu J. Acute kidney injury after infant cardiac surgery: a comparison of pRIFLE, KDIGO, and pROCK definitions. BMC Nephrol. 2023;24(1):251.
6. Liu C, He Y, Venn AJ, Jose MD, Tian J. Childhood modifiable risk factors and later life chronic kidney disease: a systematic review. BMC Nephrol. 2023;24(1):184.
7. Anderson BM, Qasim M, Correa G, Evison F, Gallier S, Ferro CJ, et al. A clinical frailty scale obtained from MDT discussion performs poorly in assessing frailty in haemodialysis recipients. BMC Nephrol. 2023;24(1):80.
8. Stojanowski J, Konieczny A, Rydzyńska K, Kasenberg I, Mikołajczak A, Gołębiowski T, et al. Artificial neural network - an effective tool for predicting the lupus nephritis outcome. BMC Nephrol. 2022;23(1):381.
9. Keane D, Glyde M, Dasgupta I, Gardiner C, Lindley E, Mitra S, et al. Patient perspectives of target weight management and ultrafiltration in haemodialysis: a multi-center survey. BMC Nephrol. 2021;22(1):188.
10. Alsalemi N, Sadowski CA, Elftouh N, Louis M, Kilpatrick K, Houle SKD, et al. The effect of renin-angiotensin-aldosterone system inhibitors on continuous and binary kidney outcomes in subgroups of patients with diabetes: a meta-analysis of randomized clinical trials. BMC Nephrol. 2022;23(1):161.
11. Cicchetti DV, Allison T. A New Procedure for assessing reliability of Scoring EEG Sleep recordings. Am J EEG Technol. 1971;11(3):101–10.
12. Fleiss JL, Cohen J. The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas. 1973;33(3):613–9.
13. Vanbelle S. A new interpretation of the weighted kappa coefficients. Psychometrika. 2016;81(2):399–410.
14. Bayram KB, Şengül İ, Aşkin A, Tosun A. Inter-rater reliability of the Australian spasticity Assessment Scale in Poststroke spasticity. Int J Rehabil Res. 2022;45(1):86–92.
15. McHugh ML. Interrater reliability: the Kappa statistic. Biochemia Med (Zagreb). 2012;22(3):276–82.

Publisher's Note