COMPARISON OF CONFIDENCE INTERVAL METHODS FOR INTER-RATER RELIABILITY COEFFICIENT (FLEISS' KAPPA).

Tagliaferri Sara [1], Neri Serena[2], Ghi Tullio[2], Malvezzi Matteo[1]

[1] *Dipartimento di Medicina e Chirurgia, Università di Parma, Parma, Italia*
[2] *Dipartimento di Medicina e Chirurgia, Unità di Scienze Chirurgiche, Ostetricia e Ginecologia, Università di Parma, Parma, Italia*

Introduction

The inter-rater reliability of measurements can be performed by several approaches. The Kappa coefficient was firstly introduced by Cohen (1960) to evaluate the agreement between two raters, and then generalized and revised by different researchers [1]. For qualitative data (two or more categories) and two or more raters, the Fleiss' Kappa (Fleiss'K), based on the concept that observed agreement is adjusted for the chance agreement, is spite of it not allowing for missing data, is largely used [2 - 4]. In the peer-reviewed literature, some doubts on assessment of the uncertainties of Fleiss'K emerged. In particular, it has been stated that the standard error given in Fleiss et al. (1979) [2] should be used for quantifying precision of Fleiss' K only for testing the hypothesis of zero agreement among raters. If this assumption of no agreement is not satisfied, a different variance formula has been proposed to be used in any statistical inference procedure [1]. In the peer-reviewed literature, several studies do not report confidence intervals of inter-rater reliability coefficients [5, 6] or Authors did not consider the assumption on the variance equations implemented in commercial statistical softwares, causing a misuse of standard error by Fleiss to quantify the precision of Fleiss' K [1].

Aims

To compare 95% confidence interval methods for inter-rater reliability according to Fleiss' K variance formulas proposed by Fleiss et al. (1979) [2] and Gwet (2021) [1] among a group of obstetricians' visual interpretation of intrapartum cardiotocographs (CTGs).

Methods

This study evaluates the performance of two different methods for constructing a 95% confidence interval according to different variance formulas available in the literature: the equation introduced by Fleiss et al. (1979) and a newer one proposed by Gwet (2021). The first one has been implemented in the R package "rel" and in SPSS software, the second one in the newer R package "irrCAC". These approaches were used to retrospectively analyse the overall agreement in a real world dataset, including fifty-three intrapartum fetal CTG records evaluated by four independent obstetricians at the University of Parma. CTGs classification was performed according to fetal CTG guidelines proposed by the International Federation of Gynecology and Obstretrics (FIGO) [7], which aim to predict neonatal acidemia of the patterns and categorizes CTGs as normal, suspicious and pathological. The classification of Fleiss' K coefficients was based on cut-off values provided for Cohen's kappa (from poor to excellent agreement) [8].

	Fleiss' K coefficient	se (K)	95%CI
Gwet (2021) $variance_{unweighthed}$	0.429	0.023	0.382-0.476
Gwet (2021) $variance_{linear\ weights}$	0.523	0.023	0.505-0.600
Fleiss (1979) variance	0.429	0.042	0.347-0.511

Table 1. *Fleiss' Kappa coefficients, standard error (se) and 95% confidence interval constructed by using the variance formulas, unweighted and linear-weighted, proposed by Gwet (2021) and by Fleiss et al. (1979).*

Results

The percentage of CTGs classified as normal was 11.1%, 13.2%, 17% and 18.5% for rater 1, 2, 3 and 4, respectively. Number of records indexed as suspicious and pathological showed a larger variability among wives, ranging between 14% and 30% for suspicious CTGs and between 52.8% and 69.8% for pathological cardiotocographs. According to cut-off values proposed for Cohen's kappa and transferred to Fleiss' K, the inter-rater agreement obtained by the three models was moderate (Table 1). The standard error computed by the equation proposed by Gwet (2021) was 0.023 for both, the unweighted and the linear-weights-adjusted se (K), while it was 0.042 according to multi-rater Fleiss' equation (1979). A smaller 95% confidence interval width was obtained by performing K statistics through Gwet's equation (0.382-0.476 and 0.505-0.600, respectively for raw and weighthed coefficients), when compared to the one computed by the Fleiss' variance formula (0.347-0.511).

Conclusions

The performance of two approaches to measure the Fleiss' Kappa (according to Fleiss, 1979, and Gwet, 2021) showed inter-rater reliability coefficients of moderate agreement. A smaller 95% confidence interval width was obtained by performing K statistics through Gwet's equation, when compared to the one computed by the Fleiss' variance formula. In order to provide valid confidence intervals, that consider the assumption on hypothesis testing, and are comparable with other studies in the literature, researchers should consider specific characteristics and formulas of commercial softwares.

Bibliografia

1. Gwet K.L. Large-Sample Variance of Fleiss Generalized Kappa. Educ Psychol Meas, 2021 Aug;81(4):781-790.
2. Fleiss J.L., Nee J.C. and Landis J.R. The large sample variance of kappa in the case of different sets of raters. Physiological Bullettin, 1979;86(5):974-977.
3. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions (3rd ed.). John Wiley & Sons, Inc. Hoboken, New Jersey; 2003.
4. Zapf A., Castell S., Morawietz L. et al. Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? BMC Med Res Methodol, 2016 Aug 5;16:93.
5. Martí Gamboa S., Giménez O.R., Mancho J.P. et al. Diagnostic Accuracy of the FIGO and the 5-Tier Fetal Heart Rate Classification Systems in the Detection of Neonatal Acidemia. Am J Perinatol. 2017 Apr;34(5):508-514.
6. Devane D., Lalor J. Midwives' visual interpretation of intrapartum cardiotocographs: intra- and inter-observer agreement. J Adv Nurs. 2005 Oct;52(2):133-41.
7. Ayres-de-Campos D., Spong C.Y., Chandraharan E et al. FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography. Int J Gynaecol Obstet. 2015 Oct;131(1):13-24.
8. Landis J.R., Koch G.G. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159-74.