Pediatric RESEARCH

CORRESPONDENCE

Methodological issues on interrater reliability of the modified Sarnat examination in preterm infants

Pediatric Research (2020) 87:614; https://doi.org/10.1038/s41390-019-0741-9

Dear Editor,

We carefully read the influential article by Pavageau and colleagues published in the journal Pediatric Research in 2019.[1] To determine the reliability of the modified Sarnat neurologic examination in late and moderately preterm neonates born at 32–36 weeks' gestational age was the purpose of this study. Kappa values were used to evaluate and analyze the agreement between examiners.[1] Based on the results, the reliability of the neurologic exam between the gold standard (GS) study investigator and groups of attending neonatologists was good to excellent ($k > 0.72$) in most categories except for Moro and tone. While the agreement was poor/fair for both tone and Moro categories in infants born at 32–34 weeks' gestation ($k = 0.20$–0.60), at 35–36 weeks' gestation, in contrast, the agreement was perfect for the tone and Moro categories ($k = 1.0$). However, when the GS examiner was compared to groups of attending examiners, the agreement in the Moro and tone categories was fair, $k = 0.46$.[1]

Depending on the type of variables, reliability analysis can be performed in different ways, one of which is the kappa coefficient, which has been used to assess agreement for qualitative variables. However, applying kappa for such a situation in particular circumstances can provide misleading results. These conditions are as follows: when the prevalence difference in each group can significantly change the kappa value. The second condition is when there are more than two categories.[2–6] Finally, the last critical situation occurs when the marginal distribution of voters' responses is different.[2–6] In such a situation, we strongly recommend using weighted kappa. Table 1 illustrates these circumstances with a hypothetical example and shows how much kappa (0.44 as moderate and 0.80 as very good) can change across circumstances with different prevalence rates and number of categories.[3–6]

The authors concluded that there was strong reliability with the exception of Moro and tone for the modified Sarnat in preterm infants. However, the experience of the examiners can influence these results and can improve reliability in tone and Moro agreement after 35 weeks. Eventually, their institution adopted these two goals: first, by providing education that targets the assessment of tone in preterm infants, and second, by testing a new neurological examination form that omits the Moro and provides details for evaluating the tone adapted from the Dubowitz and Hammersmith infant neurological examinations.[1] Such a conclusion may have been due to the inappropriate use of the statistical test, which can ultimately lead to a misleading message. In this letter, we pointed out the disadvantages of using kappa to assess agreement.

Table 1. The kappa and weighted kappa values for calculating reliability between two examiners for more than two categories and depend on prevalence.

		Examiner 1			Sum
	Grade	**1**	**2**	**3**	
Examiner 2	**1**	60	20	1	81
	2	2	12	4	18
	3	3	11	11	25
Sum		65	43	16	124
		Estimate			
Kappa		0.43			
Weighted kappa		0.63			

AUTHOR CONTRIBUTIONS

M.N. and S.M. conceptualized and designed the study. M.N. and S.M. participated in the writing of the first draft of the manuscript, reviewed the revisions and approved the final manuscript as submitted.

ADDITIONAL INFORMATION

Competing interests: The authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Shokofeh Maleki[1] and Mehdi Naderi[1]
[1]*Clinical Research Development Centre, Taleghani and Imam Ali Hospital, Kermanshah University of Medical Sciences, Kermanshah, I.R., Iran*
Correspondence: Mehdi Naderi (m.naderi51@yahoo.com)

REFERENCES

1. Pavageau, L., Sánchez, P.J., Steven Brown, L. & Chalak, L. F. Inter-rater reliability of the modified Sarnat examination in preterm infants at 32-36 weeks' gestation. *Pediatr. Res.* https://doi.org/10.1038/s41390-019-0562-x (2019).
2. Szklo, M. & Nieto, F. J. *Epidemiology Beyond the Basics* 3rd edn (Jones and Bartlett Publisher, Manhattan, New York, 2014).
3. Sabour, S. & Dastjerdi, E. V. Reliability of four different computerized cephalometric analysis programs: a methodological error. *Eur. J. Orthod.* **35**, 848 (2013).
4. Naderi, M. & Sabour, S. Reproducibility of diagnostic criteria associated with atypical breast cytology: a methodological issue. *Cytopathology* **29**, 396 (2018).
5. Sabour, S. & Ghassemi, F. The validity and reliability of a signal impact assessment tool: statistical issue to avoid misinterpretation. *Pharmacoepidemiol Drug Saf.* **25**, 1215–1216 (2016).
6. Naderi, M. & Sabour, S. Inter and intraobserver reliability and critical analysis of the FFP classification of osteoporotic pelvic ring injuries: methodological issue. *Injury* **50**, 1261–1262 (2019).

Received: 24 September 2019 Accepted: 7 October 2019
Published online: 2 January 2020