Pediatric RESEARCH

www.nature.com/pr

CORRESPONDENCE

Statistical rigor and kappa considerations: which, when and clinical context matters

Pediatric Research (2020) 88:5; https://doi.org/10.1038/s41390-020-0890-x

We thank Maleki and Naderi[1] for their correspondence letter highlighting the methodological kappa optimal use in the March issue of *Pediatric Research* in response to "Interrater reliability of the modified Sarnat examination in preterm infants".[2]

The key findings in late preterm infants 33−36 weeks of age who were evaluated in the first 6 h after birth with the modified Sarnat exam were the following: (1) The reliability between the gold standard study investigator and the group of attending neonatologists was good to excellent ($k > 0.72$) in most categories except for Moro and tone. (2) While the agreement was poor for both tone and Moro categories in infants born at 32–34 weeks' gestation ($k = 0.20$–0.60), it dramatically improved at 35−36 weeks suggesting an important maturation effect.

Unweighted and weighted kappa are both widely used to measure the degree of agreement between two independent raters. While we agree with the general value of weighted kappa statistics and the importance of monitoring partial agreement, we have purposefully selected to use the conservative nonweighted Kappa statistic in this particular clinical situation. The decision is based clinically on the need to completely agree on what constitutes moderate−severe encephalopathy. In such a situation, a near miss is not acceptable; you either categorically initiate hypothermia or you do not. This determined the use of a statistical approach of complete agreement among gold standard and other examiners.

As a review, Cohen first introduced unweighted kappa in 1960,[3] as a chance-corrected index of agreement for categorical variables with Kappa = 1 representing a perfect agreement between two raters. Subsequently, weighted kappa[4] was introduced in 1968 to find the agreement of two raters when using nominal scores. Whereas unweighted kappa does not distinguish among degrees of disagreement, weighted kappa incorporates the magnitude of each disagreement and provides partial credit for disagreements when agreement is not complete.[5] The usual approach is to assign weights to each disagreement pair. While linear and quadratic agreement weights are common in the literature and available in statistical packages, other weights can be used depending upon the impact of disagreement.

We have now further performed the weighted kappa statistics for both linear and quadratic agreement to respond to the correspondence. We report, as one would expect, improvement in the kappa statistics when comparing the unweighted to linear and quadratic weighted statistics respectively for each of tone (kappa improved from 0.46 to 0.48 to 0.54) and Moro (kappa improved from 0.51 to 0.63 to 0.73). The conclusions and overall message of the study remain that the strong inter-reliability agreement does not include tone/Moro, which are significantly influenced by the gestational maturity and the experience of the examiner.

In conclusion, we emphasize that rigor and using the correct statistical approach is of essence for any scientific publication. This correspondence highlights the need for collaborative effort from both clinical and statistical perspectives. The clinical context dictates the best statistical approach. Regarding the modified Sarnat Exam, the clinical relevance is to report the complete agreement that affects decisions to initiate urgent hypothermia therapy.

ACKNOWLEDGEMENTS

NIH Grants K23HD069521 and 1R01NS102617-01 support L.F.C. The funding organization had no role in design and conduct of the study; collection, management, analysis and interpretation of the data; preparation, review and approval of the manuscript, and decision to submit the manuscript for publication.

AUTHOR CONTRIBUTIONS

L.F.C. drafted the letter and L.H. performed the statistics. All authors reviewed and approved the final version.

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.

Lina F. Chalak[1], Lara Pavageau[1], Beverly Huet[2] and Linda Hynan[2]
[1]*Neonatal-Perinatal Medicine, UT Southwestern Medical Center, Dallas, TX, USA and* [2]*Population Health and Statistics, UT Southwestern Medical Center, Dallas, TX, USA*
Correspondence: Lina F. Chalak (lina.chalak@utsouthwestern.edu)

REFERENCES

1. Maleki, S. & Naderi, M. Methodological issues on interrater reliability of the modified Sarnat examination in preterm infants. *Pediatr. Res.* **87**, 614 (2020) https://doi.org/10.1038/s41390-019-0741-9.
2. Pavageau, L., Sanchez, 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.* **87**, 697–702 (2020). https://doi.org/10.1038/s41390-019-0562-x.
3. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychological Meas.* **20**, 37–46 (1960).
4. Cohen, J. Weighed kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bull.* **70**, 213–220 (1968).
5. Maclure, M. & Willett, W. C. Misinterpretation and misuse of the kappa statistic. *Am. J. Epidemiol.* **126**, 161–169 (1987).

Received: 14 February 2020 Accepted: 7 March 2020
Published online: 10 April 2020