Towards proving the validity of computer-supported sleep scoring for pediatric trials

Georg Dorffner[1,2], Michael Lagler[2] ,Georg Gruber[2], Patrick Erber[1], Lukas Pinka[1], Ruth Luigart[3], Barbara Schneider[3]

[1:] Medical University of Vienna, Section for Artificial Intelligence, Vienna, Austria

[2:] The Siesta Group, Vienna, Austria, [3:] Children's Hospital St. Marien, Landshut, Germany

METHODOLOGICAL QUESTION

Is it possible to prove statistical equivalence of computer supported sleep scoring in a pediatric population, as compared to visual expert scoring?

INTRODUCTION

In previous work (Dorffner et al., 2021, ICSTM 2021 fall conference) we have shown that existing algorithms for scoring sleep based on polysomnographic recordings can be adapted to the characteristics of infant and child recordings, especially that of the electroencephalogram (EEG) which is known to change during maturation (Berry et al., 2017). Here, we took this observation a step further and ventured to derive a validated version of the adapted algorithm that would be fit to be used in pediatric clinical trials involving sleep as an endpoint..

REFERENCES

(1) Berry, R. B., et al. (2017). The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications, version 2.4. American Academy of Sleep Medicine, Darien.

(2) Anderer, P., et al. (2010). "Computer-assisted sleep classification according to the standard of the American Academy of Sleep Medicine: validation study of the AASM version of the Somnolyzer 24 x 7."

Neuropsychobiology 62(4): 250-264.

(3) Rogers, J. L.; Howard, K. I.; Vessey, J. T. (1993). "Using significance tests to evaluate equivalence between two experimental groups". Psychological Bulletin. 113 (3): 553–565.

MEDICAL UNIVERSITY OF VIENNA

METHODS

Based on a dataset from a children's sleep laboratory (reported in Dorffner et al., 2021), we focused on the two age groups, 5-9 and 9-14 years. In the previous study, it was found that for both age groups, a 50% EEG amplitude scaling factor and a cut-off frequency between alpha and theta bands of 5Hz (as opposed to 7Hz for adults), represent optimal scoring parameters applied to the fully validated sleep scoring tool, Somnolyzer (Anderer et al. 2010) – see fig.1 for an example.

Here, Somnolyzer, with those adaptations, was applied to a randomly selected validation data set, that was not part of parameter adaptation. For each of those recordings, two independent expert scorings were available. Their average difference in any of the main sleep endpoints considered – percentage in each sleep stage (N1P, N2P, N3P, and RP) – was seen as a tolerable deviation in a statistical equivalence test. For such a test, the 90%-confidence interval of differences between the adapted Somnolyzer and the primary expert scoring would need to be fully within the tolerance to be considered statistically equivalent (Rogers et al. 1993).

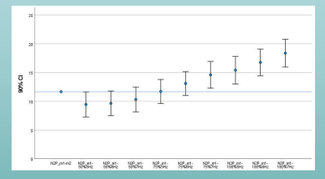


Fig. 1. The influence of different parameter settings on the confidence intervals of the resulting endpoint variables. Percentage of N2 sleep is shown as an example.

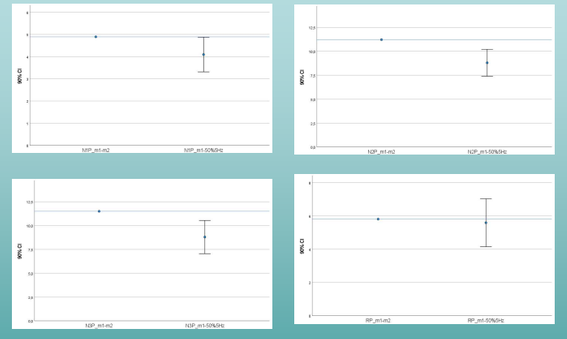
DISCLOSURES

Michael Lagler is an employee, Georg Dorffner and Georg Gruber are employees and shareholders, of The Siesta Group, a service provider for measuring electrophysiological signals including sleep in clinical trials.

RESULTS

For both age groups, most of the endpoint variables could be proven statistically equivalent when compared to visual expert scoring. For age group 9-14, the 90%-confidence intervals of N1P [3.31 4.88], N2P [7.39 10.22], and N3P [7.05 10.55] were entirely below the tolerance intervals of 4.89, 11.22, and 11.50, respectively. Only RP [4.15 7.03] did not reach equivalence as compared to the tolerance of 5.81. For age group 5-9, equivalence was proven for N2P ([5.79 9.09], as compared to 11.99), and N3P ([6.49 10.67] as compared to 12.67). RP ([2.56 4.41] as compared to 4.36) was slightly outside the significant equivalence, while N1P ([3.55 5.45] as compared to 3.4) clearly missed equivalence.

Figures 2 and 3 highlight the main validation results for the two age groups. Confidence intervals are shown vis-à-vis the acceptable tolerance threshold given by the deviation between two expert scorers.



THE siesta GROUP

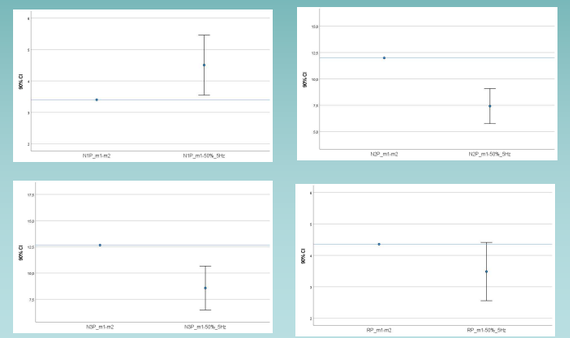


Fig. 2 (on the right): 90% confidence intervals for the percentage of the four sleep stages, as compared to the mean difference between experts, for age group 9-14 years.

Fig. 3 (above): Same as in fig. 2, for age group 5-9 years.

CONCLUSION

With only a few exceptions (N1P for age group 5-9 and RP for age group 9-14) the main sleep endpoint variables passed the statistical equivalence test, as compared to the average deviation of two human experts. Since only two such expert scorings were available, a more truthful estimate of the acceptable tolerance interval could reveal that even those exceptions are within acceptable limits. In conclusion, the study has largely proven that adapted sleep scoring algorithms can be considered validated to children as young as 5 years old.

Contact: georg.dorffner@meduniwien.ac.at