Towards more reliable sleep scoring in CNS trials by way of Artificial Intelligence

Submitter Michael Lagler

Affiliation The Siesta Group

SUBMISSION DETAILS

Methodological Issue Being Addressed A review of 11 interrater reliability studies indicates substantial agreement for human expert sleep scoring, with only poor to moderate agreement in nonrapid eye movement sleep stages [1]. Consequently, some sleep endpoints are largely dependent on individual expert interpretation. We hypothesize that artificial intelligence (AI) scoring algorithms can remedy this situation without sacrificing accuracy.

Introduction The widely accepted gold standard for scoring sleep relies on human expert scoring based on neurological signals. However, there is a current move from time-consuming and error-prone visual scoring toward automated scoring of sleep stages. Recently developed AI algorithms offer consistent and reliable results and provide additional features such as estimated sleep stage probabilities. To address the increasing trend from attended in-lab polysomnography (PSG) to home sleep testing (HST), AI systems are trained to provide sleep stage information derived from cardiorespiratory signals only.

Methods In this study, we tested the hypothesis that AI-based autoscoring of the Somnolyzer algorithm (Koninklijke Philips, Netherlands) is superior to manual scoring at the group level and non-inferior to manual scoring for individual studies.
The algorithm was trained on 685 PSGs of 391 subjects from two studies and validated on four independent datasets containing 426 PSGs scored by 1 scorer and three datasets comprising up to 70 PSGs scored by up to 12 scorers. Algorithm performance was determined by calculating the Cohen’s kappa coefficient comparing AI-predicted sleep stages to manual scorings. A manual consensus scoring is derived where multiple scorings were available. Inter-scorer reliability of human experts was assessed by comparing individual scorings to an unbiased consensus of the respective other expert opinions.
The cardio-respiratory sleep staging built into Somnolyzer was validated independently on two datasets including PSGs from 592 subjects. The gold-standard sleep staging was derived from neurological signals following AASM standards, while the cardio-respiratory sleep staging was calculated using cardiac and respiratory signals, only. Algorithm performance was assessed using Cohen’s kappa coefficient for four class sleep staging (Wake, N1+N2, N3, REM).

Results Cohen’s kappa for 5-stage sleep scoring was 0.74 versus a single scorer and 0.78 versus a consensus of 6 scorers. The agreement between individual human expert scorings and unbiased consensus scorings was significantly below the algorithm performance with an average kappa coefficient of 0.69. Both hypotheses were confirmed.
Moreover, the AI-based autoscoring outputs sleep stage probabilities (hypnodensities) that quantify sleep stage ambiguity and stability while providing all the information contained in a hypnogram.

Concerning cardiorespiratory sleep staging, we observed substantial agreement to the gold-standard comparator of manual sleep staging of neurological signals (Cohen’s kappa for 4-stages: 0.68).

Conclusions Sleep scoring is moving from visual to automatic scoring, from hypnogram to hypnodensity, and opens new applications for home recordings using cardio, respiratory and/or accelerometer signals.

Co-Authors

Michael Lagler[1], Marco Ross[1], Peter Anderer[1], Georg Dorffner[1]

[1] The Siesta Group

Keywords

Keywords
polysomnography
home sleep testing
machine learning
cardiorespiratory sleep staging
hypnodensity

Guidelines I have read and understand the Poster Guidelines

Disclosures if applicable <blank>

Related tables Abstract_ISCTM_2025_08AUG25.docx