An important step toward automation of polysomnography analyses

SLEEPJ, 2025, 48, 1–2
https://doi.org/10.1093/sleep/zsaf147
Advance Access publication 19 June 2025
Editorial

Editorial

An important step toward automation of polysomnography analyses

Matteo Cesari[1,*], Andreas Brink-Kjaer[2] and Irene Rechichi[3]

[1]Department of Neurology, Medical University of Innsbruck, Innsbruck, Austria,
[2]Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark and
[3]Department of Control and Computer Engineering, Politecnico di Torino, Turin, Italy

*Corresponding author. Matteo Cesari, Department of Neurology, Medical University of Innsbruck, Anichstrasse 35, 6020 Innsbruck, Austria. Email: matteo.cesari@i-med.ac.at.

Manual polysomnography (PSG) scoring is a highly specialized, labor-intensive, time-consuming task, and additionally suffers from inter- and intra-rater variability [1]. All these limitations potentially affect patient care: the bottleneck created by manual PSG scoring might cause delays in the diagnosis and treatment of sleep disorders, and the mentioned variability might even lead to inconsistent diagnoses and treatment plans. Automated algorithms that can perform PSG scoring quickly and objectively hold a high potential in overcoming these limitations, ultimately improving patient care and reducing the economical burdens of sleep examinations.

The efforts toward developing automatic systems started in the 1960s [2], but significant improvements in their accuracy and reliability were obtained in the last years with novel artificial intelligence (AI) approaches [3]. In particular, the availability of large datasets and of increased computation power allowed researchers to develop highly performing system for scoring sleep stages [4, 5], identifying respiratory events [6], limb movements [7], and arousals [8]. Until now, however, researchers have focused on isolated tasks or a subset of tasks [9], while a unique system integrating all the mentioned tasks is missing.

In the work of Nasiri et al., the Complete Artificial Intelligence Sleep Report (CAISR) system is proposed [10]. CAISR integrates sleep staging, arousal detection, identification of respiratory events (obstructive, central and mixed apneas, hypopneas, and respiratory effort-related arousals), and limb movement in a unique and unified tool. CAISR was therefore developed with the aim of performing all routine tasks carried out by technicians in their manual scoring of PSG.

The authors of this work are to be commended for the rigorous methodology and the strong performance of their system compared to both human raters and competing automatic methods. The excellent methodology and results make CAISR a milestone toward full automation of PSG analysis. In particular, concerning performance on three different multi-scored test datasets, it is worth highlighting that CAISR had performances that were matching or even surpassing the human ones, making it a reliable and precise tool. In addition to the test performance, which is reported in detail in the manuscript [10], we would like to highlight and discuss other aspects of the study.

First, the authors employed a large and heterogeneous dataset of over 25,000 PSGs coming from seven independent datasets, thus reinsuring the generalizability of the developed system, although for one of these databases, fine-tuning was applied to improve performance. In particular, the datasets included high heterogeneity across sex, ethnicity, cohort types, in-lab, and at-home recordings. The authors correctly mention the possibility of a regional bias, due to the fact that the databases were mainly originating from North America. This aspect should be indeed investigated with the perspective of using CAISR in the scoring of PSGs coming from different world regions. Furthermore, it is fundamental that future works investigate the performance of CAISR in populations with REM sleep behavior disorder and neurodegenerative diseases, who present abnormalities in their sleep micro-structure, which typically challenge automatic scoring systems [11].

Second, the authors generated exhaustive platinum labels in one dataset, which were obtained by multiple rounds of checks of annotations across authors. The generation of these high-quality labels allowed the authors to identify systematic biases and errors that are performed in regular clinical scorings. As an example, the exact timing of the start and end of an arousal is typically not important in clinical scoring, as only the presence of the arousals is relevant for clinical evaluations. Thereby, this creates biases and errors in human scoring, which might be reflected in poor agreement performance between automated systems and humans. To avoid inheriting this bias, the authors refined labels for training purposes, which allowed their model to learn from the labels while minimizing bias. This is an important message of the work of Nasiri et al., as it underlines the importance of high-quality scoring that is needed to both correctly optimize and validate automated systems. Sleep societies should strive

to provide more and larger datasets with high-quality standard scorings so that automated scoring systems can become more and more precise in executing tasks.

Third, from a methodological point of view, it is interesting that, while for sleep stage scoring and arousal detection deep learning architectures were employed, simpler rule-based algorithms were instead used for the identification of respiratory events and limb movements. Nevertheless, rule-based algorithms still showed high performance for their tasks. While rule-based systems are less likely to inherit scoring biases, based on our experience, they are typically more sensitive to artifacts, equipment choice, and recording settings. We therefore think that further external validation of CAISR is necessary to prove their robustness across different setups.

Fourth, it is worth highlighting that the authors shared their code and trained models in an open-access repository so that researchers can freely download it and test it on their own data. While this has to be appreciated, it must be mentioned that the usage of CAISR is still meant for people with coding experience and not directly for more clinical researchers. Furthermore, the authors mention that CAISR might need to be fine-tuned for specific applications and settings, but a clear fine-tuning workflow is not provided. These limitations affect many of the methods and algorithms that have been developed until now. It is fundamental that future research will focus on how to better integrate these research tools in sleep medicine environments. We think that sleep societies should define clear frameworks on how tools like CAIRS can be shared with the community, to further improve usability and standardization.

In conclusion, CAISR is the first well-validated, automated system for the complete scoring of sleep stages, arousals, respiratory events, and limb movements. CAIRS has the potential to overcome the limitations of manual sleep scoring and standardized evaluation of PSGs. We encourage researchers worldwide to perform independent validations of CAISR to further prove its robustness and identify potential regional or other types of biases.

Disclosure Statement

Financial disclosure: None.

Non-financial disclosure: None.

References

1. Cesari M, Stefani A, Penzel T, *et al*. Interrater sleep stage scoring reliability between manual scoring from two European sleep centers and automatic scoring performed by the artificial intelligence-based Stanford-STAGES algorithm. *J Clin Sleep Med*. 2021;**17**(6):1237–1247. doi:10.5664/jcsm.9174
2. Itil TM, Shapiro DM, Fink M, Kassebaum D. Digital computer classifications of EEG sleep stages. *Electroencephalogr Clin Neurophysiol*. 1969;**27**(1):76–83. doi:10.1016/0013-4694(69)90112-6
3. Bandyopadhyay A, Oks M, Sun H, *et al*. Strengths, weaknesses, opportunities, and threats of using AI-enabled technology in sleep medicine: a commentary. *J Clin Sleep Med*. 2024;**20**(7):1183–1191. doi:10.5664/jcsm.11132
4. Hermans LWA, Huijben IAM, van Gorp H, *et al*. Representations of temporal sleep dynamics: Review and synthesis of the literature. *Sleep Med Rev*. 2022;**63**:101611. doi:10.1016/j.smrv.2022.101611
5. Fiorillo L, Puiatti A, Papandrea M, *et al*. Automated sleep scoring: a review of the latest approaches. *Sleep Med Rev*. 2019;**48**:101204. doi:10.1016/j.smrv.2019.07.007
6. Mostafa SS, Mendonça F, Ravelo-García A G, Morgado-Dias F. A systematic review of detecting sleep apnea using deep learning. *Sensors*. 2019;**19**(22):4934. doi:10.3390/s19224934
7. Carvelli L, Olesen AN, Brink-Kjær A, *et al*. Design of a deep learning model for automatic scoring of periodic and non-periodic leg movements during sleep validated against multiple human experts. *Sleep Med*. 2020;**69**:109–119. doi:10.1016/j.sleep.2019.12.032
8. Qian X, Qiu Y, He Q, *et al*. A review of methods for sleep arousal detection using polysomnographic signals. *Brain Sci*. 2021;**11**(10):1274. doi:10.3390/brainsci11101274
9. Zahid AN, Jennum P, Mignot E, Sorensen HBD. MSED: a multi-modal sleep event detection model for clinical sleep analysis. *IEEE Trans Biomed Eng*. 2023;**70**(9):2508–2518. doi:10.1109/TBME.2023.3252368
10. Nasiri S, Ganglberger W, Nassi T, *et al*. CAISR: achieving human-level performance in automated sleep analysis across all clinical sleep metrics. *Sleep*. 2025;**48**(8). doi: 10.1093/sleep/zsaf134
11. Cesari M, Christensen JAE, Sixel-Doring F, *et al*. A clinically applicable interactive micro and macro-sleep staging algorithm for elderly and patients with neurodegeneration. *Annu Int Conf IEEE Eng Med Biol Soc*. 2019;**2019**:3649–3652. doi:10.1109/EMBC.2019.8856705