SLEEP, 2022, 1–3

https://doi.org/10.1093/sleep/zsac190
Advance Access Publication Date: 11 August 2022
Editorial

EDITORIAL

Sleep scoring moving from visual scoring towards automated scoring

Thomas Penzel*

Interdisciplinary Sleep Medicine Center, Charite center for Pneumology CC12, Charite Universitatsmedizin Berlin, Berlin, Germany

*Corresponding author. Thomas Penzel, Interdisciplinary Sleep Medicine Center, Charite center for Pneumology CC12, Charite University Hospital, Chariteplatz 1, 10117 Berlin, Germany. Email: thomas.penzel@charite.de.

Introduction

For many years we have been using visual sleep scoring to quantify sleep stages and all events occurring during sleep. This required a long path with initial standardization of rules for visual sleep staging starting with the sleep of healthy young volunteers and continuing with all the sleep abnormalities we know today. Currently, rules for sleep staging are laid out in the AASM manual version 2.6 which provides definitions for sleep stages and the most commonly observed events related to the sleep disorders with the highest prevalence [1]. We recognize the high variability in sleep scoring results achieved by expert sleep scorers [2, 3]. Visual sleep scoring is still, however, a very valid task because we may observe unexpected events during sleep and this teaches us much about the abnormalities observed during sleep. Visual sleep scoring is also very important for newcomers to the field of sleep medicine, so that they learn and understand how sleep changes across the night, how much sleep varies from person to person, and how to identify unusual and abnormal events during sleep.

The Problem

Sleep recording and sleep scoring have become part of regular diagnostic procedures in sleep medicine and have become a kind of "mass production" for diagnosing sleep disorders given the high prevalence of these disorders. In the process of "mass production," sleep scoring, as far as routine clinical work, can be tedious, can be boring, and hence can be sensitive to errors in boring work, related to the normal variation of vigilance in the sleep scorer when doing his/her job. Today we have validated computer software, which helps to make the boring work of scoring raw medical data more robust against human errors. The best example for making use of modern computer software to make human scoring of medical data more robust is cancer diagnosis and especially scoring of mammography x-ray images [4]. Errors in scoring mammography x-ray images should be lower than 5%. And they are in the 5% range with automated approaches according to quality control studies [5]. For sleep stage scoring, we still accept error rates of 15% or more if counted on an epoch-by-epoch comparison. If the agreement between sleep stage scorers is 85%, this is acceptable, and start to be worried if the agreement drops below 70%. As a side note, these numbers are a simplified estimate, because quantitative reliability studies use many different methods, like Cohen's *K*, Fleiss *K*, Pearson product-moment correlation coefficient, and intraclass correlation coefficient (ICC) which reflect different statistical properties of differences [3, 6]. Making use of algorithms and software approaches as used for cancer diagnosis (e.g. machine learning, artificial intelligence, artificial neural networks, big data analysis, and deep learning) can help us improve the accuracy of sleep staging and make it more robust. How could this be achieved? First, I propose we need a common reference database on which sleep staging software can be tested and against which sleep staging software needs to be validated.

We know that a lot of training scorers in different sleep centers will reduce the variability in scoring between centers [7]. The AASM Interscorer reliability program is one effort to decrease variability in scoring by encouraging the training of sleep scorers [8]. We also know that improving definitions for events helps decrease the variability in scoring. This has worked well for the scoring of apnea and hypopnea events [6, 9]. Agreement between scorers became moderate and clinically acceptable

with this approach [10]. This is helpful for sleep apnea diagnosis. It means that sleep apnea diagnosis achieves reliable results. The approach of redefining sleep stages and making definitions more precise and simple was one of the initial aims when creating the AASM manual for the scoring of sleep. Some fuzziness in the definitions of Rechtschaffen and Kales was removed and the definitions became more precise. Without a doubt, this reduced the variability in the results of expert sleep scoring sleep stages to some extent [11]. However, too much variability is still observed for the manual scoring of sleep stages [2, 3].

Proposed Solution

In this issue of *SLEEP*, there is a report of a new computer-supported sleep scoring software that has been compared against sleep scoring ambiguity across expert sleep scorers with their visual scoring results [12]. A new and very promising approach of this newly published work is, that the authors decided to apply their sleep scoring software on three databases. Choosing different datasets really includes a representative variety of sleep recordings and of expertise in sleep scorers. We know that there are "decision flavors" among groups of human expert sleep scorers. With "flavors" I mean group differences in scoring, as observed between nations, possibly schools educating sleep scoring, and similar differences [13]. Based on this concept, it is impossible to find a perfect agreement among human scorers and as a consequence, agreement with a sleep scoring software will be challenging. It is time to overcome exactly this limitation. Taking several datasets together with a variety of healthy sleepers and pathological sleep, and with a variety of expert sleep scorers, all really experienced in their scoring of sleep stages and events, is the way forward and reach a better agreement for tasks like sleep scoring. An exciting result of the published work is, that there is no reference sleep expert truth. The hypnodensity-based chart presents probabilities for sleep stages and thus treats all expert sleep scorers equally [14]. There is no longer a gold reference sleep scorer expert, but the reference is built from probabilities of sleep stage scoring. This is a fair and adequate way to build a reference dataset for testing sleep scoring software. The influence of different "flavors" of different sleep scorers is averaged out to some extent. Is it possible to create a database of sleep stage scoring accepted by all sleep centers and all sleep stage scoring experts? The new reference dataset finally should be held by a professional scientific society like AASM or by an academic institution, possibly linked to a generally accessible sleep resource database. But this is a different discussion.

The strength of the study presented here is, that it presents a new approach to auto-scoring, which despite using up-to-date computer software algorithms, reaches a limited accuracy. As this report explains, accepting the limited accuracy is a very fair way, by showing the ambiguity of multiple expert scorers with the hypnodensity-based approach. The hypnodensity-based approach does not favor one sleep scorer, but is based on probabilities. We can learn from this study, that the time is now to create computer-based automated sleep scoring. The computational tools are out there. We can learn, that we need to create a reference database for the validation of computer-based automated sleep scoring. We can learn that we need to treat human expert sleep scorers equally by using hypnodensity-based probabilities. And finally, we should come to a better sleep stage accuracy than 85% by using the help of modern computer algorithms. Of course, we can never forget that visual sleep scoring will remain part of education in sleep medicine.

Funding

No support was obtained for this work.

Disclosure Statement

The author receives institutional grant support from Cidelec and the European Union funds. The author received speaker fees and travel compensation from Jazz Pharmaceutical, Löwenstein Medical, Neuwirth Medical Technology, and Philips. He is a consultant to Bayer Healthcare, Idorsia, Cerebra, Jazz Pharmaceutical, and Philips paid to the institution. He holds shares at Advanced Sleep Research GmbH, The Siestagroup GmbH, and Nukute Oy.

References

1. Berry RB, *et al.*; for the American Academy of Sleep Medicine. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. Version 2.6*. Darien, IL: American Academy of Sleep Medicine; 2020.
2. Younes M, *et al*. Reliability of the American Academy of Sleep Medicine rules for assessing sleep depth in clinical practice. *J Clin Sleep Med*. 2018;**14**(2):205–213. doi:10.5664/jcsm.6934
3. Lee YJ, *et al*. Interrater reliability of sleep stage scoring: a meta-analysis. *J Clin Sleep Med*. 2022;**18**(1):193–202. doi:10.5664/jcsm.9538
4. Gilbert FJ, *et al.*; CADET II Group. Single reading with computer-aided detection for screening mammography. *N Engl J Med*. 2008;**359**(16):1675–1684.
5. Porzio M, *et al*. MAMMO_QC: free software for quality control (QC) analysis in digital mammography and digital breast tomosynthesis compliant with the European guidelines and EUREF/EFOMP protocols. *Biomed Phys Eng Express*. 2021;**7**(6):1–8. doi:10.1088/2057-1976/ac2076
6. Punjabi NM, *et al*. Computer-assisted automated scoring of polysomnograms using the Somnolyzer system. *Sleep*. 2015;**38**(10):1555–1566. doi:10.5665/sleep.5046
7. Biswal S, *et al*. Expert-level sleep scoring with deep neural networks. *J Am Med Inform Assoc*. 2018;**25**(12):1643–1650.
8. Rosenberg RS, *et al*. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. *J Clin Sleep Med*. 2013;**9**(1):81–87. doi:10.5664/jcsm.2350
9. Kuna ST, *et al*. Agreement in computer-assisted manual scoring of polysomnograms across sleep centers. *Sleep*. 2013;**36**(4):583–589. doi:10.5665/sleep.2550
10. Malhotra A, *et al*. Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. *Sleep*. 2013;**36**(4):573–582. doi:10.5665/sleep.2548
11. Danker-Hopfe H, *et al*. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *J Sleep Res*. 2009;**18**(1):74–84.
12. Bakker JP, *et al*. Scoring sleep with artificial intelligence enables quantification of sleep stage ambiguity: hypnodensity based on multiple expert scorers and auto-scoring. *Sleep*. 2022;**45**(10). doi:10.1093/sleep/zsac154.

13. Penzel T, *et al*. Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules. *J Clin Sleep Med*. 2013;**9**(1):89–91. doi:10.5664/jcsm.2352

14. Olesen AN, *et al*. Automatic sleep stage classification with deep residual networks in a mixed-cohort setting. *Sleep*. 2021;**44**(1). doi:10.1093/sleep/zsaa161