SPECIAL ARTICLES

JCSM
Journal of Clinical Sleep Medicine

http://dx.doi.org/10.5664/jcsm.2352

COMMENTARY

Inter-scorer Reliability between Sleep Centers Can Teach Us What to Improve in the Scoring Rules

Commentary on Rosenberg et al. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. J Clin Sleep Med 2013;9:81-87.

Thomas Penzel, Ph.D.; Xiaozhe Zhang, M.D.; Ingo Fietze, M.D.

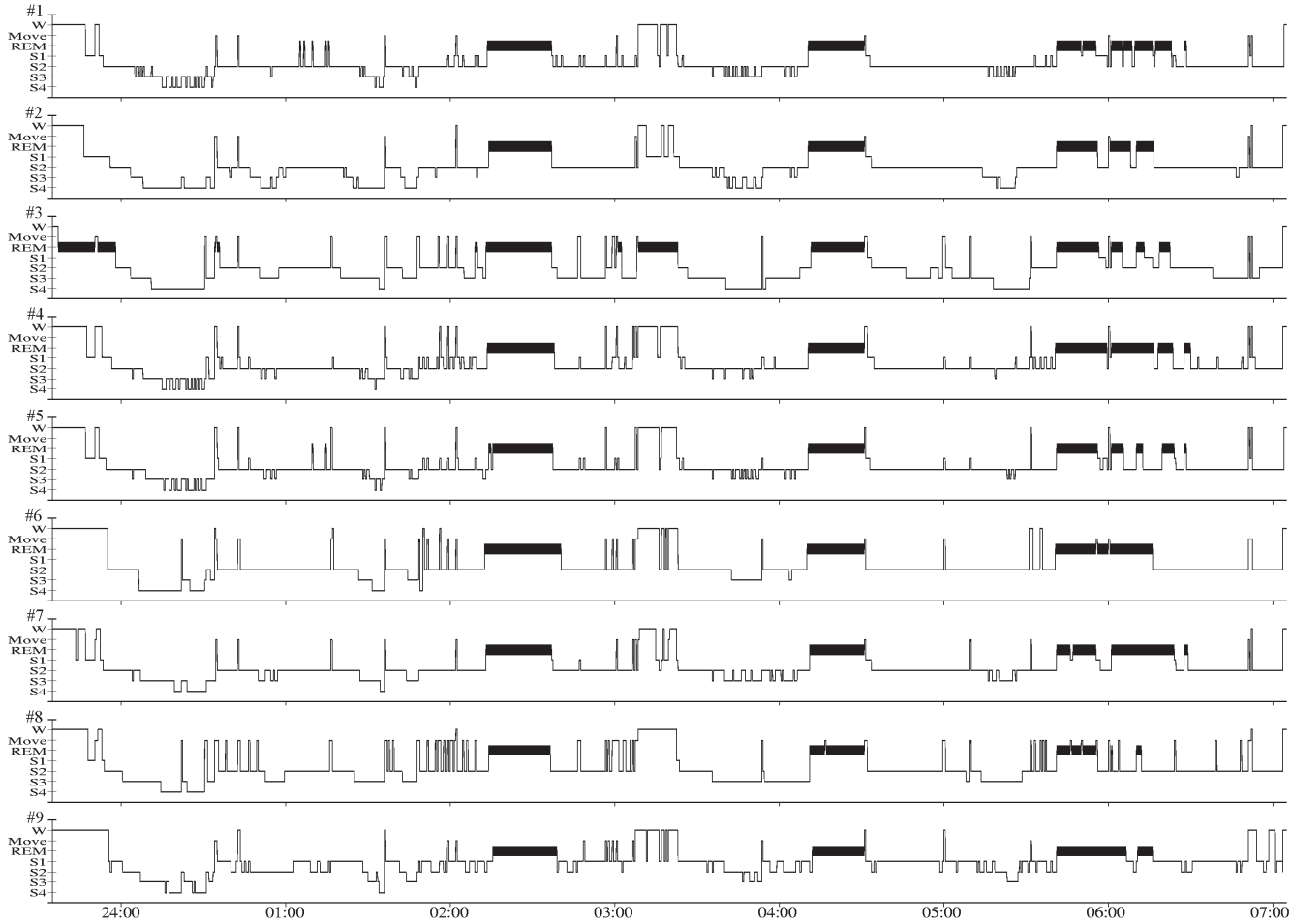
Sleep Medicine Center, Charité Universitätsmedizin Berlin, Berlin, Germany

With the first revision of the scoring manual for scoring sleep and associated events,[1] the discussion on reliability of sleep scoring is in the focus again. The revision of the scoring manual keeps the visual scoring of sleep as the reference method following the previous manual of Rechtschaffen and Kales.[2] As long as visual scoring of sleep is the reference method, we need to know the reliability of this procedure. Visual scoring of sleep (or any other signal of image in medicine) by its nature has an inherent subjective component. Therefore to assess the reliability of visual sleep scoring is a very important task in sleep medicine. A high reliability is needed for creditability of the sleep scoring procedure. For the purpose of comparing sleep scoring, the AASM Inter-scorer Reliability Program was developed and offered to sleep centers.[3] This program started in April 2010. The program has been increasingly used since then. Now, based on an experience with this program, an evaluation of the inter-scorer reliability can be presented on a very large number of scorers and has been documented in a remarkable paper.[3] Statistics about agreement of scoring sleep stages and details on sleep stage transitions are provided. The paper demonstrates a high reliability for R sleep (with 90.5% agreement) and lowest agreement for N1 (with 63.0%). Furthermore the paper focuses on the transitions between sleep stages.

Whenever performing a systematic study on the inter-scorer reliability of sleep the question about the truth comes up. In all dubious cases we would like to know the "true" state. This would be the "gold standard." Since the truth is not known, one democratic compromise is to take the majority vote as the truth. Another truth could be a consensus scoring. For a consensus scoring, a group of scorers is sitting together on all dubious epochs, and they discuss the various arguments for their different opinion until they find a consensus. This approach was selected for the Siesta database of sleep recordings, being a typical European approach.[4] Another historical approach is "truth by age," which refers to the scoring performed by one experienced expert scorer, usually a respected and elderly person. The AASM inter-scorer reliability program uses two board certified sleep specialists as scorers to provide a gold standard. Differences between them are resolved, and thus this can be regarded as a consensus scoring. This is a fair approach to this fundamental problem. It is strongly recommend to avoid the term "gold standard," even for a sleep center director. The reason is obvious: the center director is just another well-experienced scorer and not necessarily the true reference. If there is a need to keep track of the position or the experience of a scorer then it is recommended to use the term "director scorer" or "experienced scorer" to annotate the credentials. This makes clear the respect for the achievements and definitely is better than suggesting that this particular scorer always scores the truth.

It is important to note that the reliability in scoring was improved from the initial Rechtschaffen and Kales scoring to the AASM scoring.[4] In order to show the limitations of the reliability in scoring according to Rechtschaffen and Kales a number of studies have investigated the reliability. Here we reference one study on reliability of sleep scoring according to Rechtschaffen and Kales.[5] In Germany, with the accreditation of sleep centers in the mid-1990s, an effort to compare sleep scoring in the sleep centers was started. This continued an even earlier attempt, where a comparison of sleep scoring was started by sending a classical paper record with a sleep EEG from one center to the next. This effort ceased without success because the sleep EEG book was lost in the mail after it passed five centers. In addition, this would have been one single recording only. In 2000, eight digital sleep recordings were stored on one CD-ROM, together with viewing and scoring software.[5] Copies of the CD-ROM were distributed to all accredited sleep centers in Germany. The eight recordings just filled the capacity of the CD-ROM. The eight recordings consisted of three healthy subject recordings, three sleep apnea recordings, and two recordings with periodic leg movement syndrome. The sleep centers were asked to send back the scorings as digital files to the German Sleep Society. Finally, nine centers sent back their scorings, and these were compared statistically and visually. Two examples of the visual comparison are presented in **Figures 1** and **2**. Although the results of this study were disappointing because so few scorings were received back, the comparison of the scorings made some essential points clear. These points are presented here. There is high reliability in scoring REM sleep. Some scorings show very obvious errors (see scorer #3 with his error in scoring REM sleep at the beginning of the night). Reliability between scorers in healthy sleepers (see **Figure 1**) is much higher than in patients, particularly sleep apnea patients (see **Figure 2**). A

Figure 1—A comparison of sleep scorings of nine scorers from different sleep centers across Germany based on the same recording[5]



The recording was performed in a subject without sleep disorders being part of a control study. The comparison shows well some uncertainties between sleep stages and a major agreement in general. It shows also a remarkable error by scorer #3 with REM sleep in the beginning of the night.

previous study showed the lower agreement of scoring sleep stages in sleep apnea subjects as well.[6] There is a certain flavor in scoring when the rules are not obvious. Some scorers like to score sleep in apnea patients with a lot of MOVE; others like to score a lot of WAKE or they ignore all the arousals and score sleep stage 2. Some scorers say, even if this is sleep stage 2 with sleep spindles, the strongly disturbed sleep would justify this scoring as sleep stage 1 only. Transitions between sleep stages present the biggest source of differences between sleep scorers. There is no gold standard scorer. This requires appropriate statistical tools for calculation of reliability. Here the Fleiss kappa coefficient was applied.

One motivation for the AASM manual was to remove the ambiguity in scoring sleep stages in patients with sleep disorders such as sleep apnea and simplify sleep scoring. Today much of this aim has been achieved, but the ultimate goal has not been reached. For sleep stage N1, the agreement is as low as 63.0%.[3]

One open issue to be clarified with additional studies is the issue of using multiple EEG derivations when scoring sleep and investigating inter-observer variability.[7]

As a result of this, the comparison as presented by the Inter-scorer Reliability Program can teach us where there are remaining weak issues that need to addressed in future improvements of the scoring rules. An in-depth analysis of the deviations is a definite help to the AASM to improve reliability in scoring.

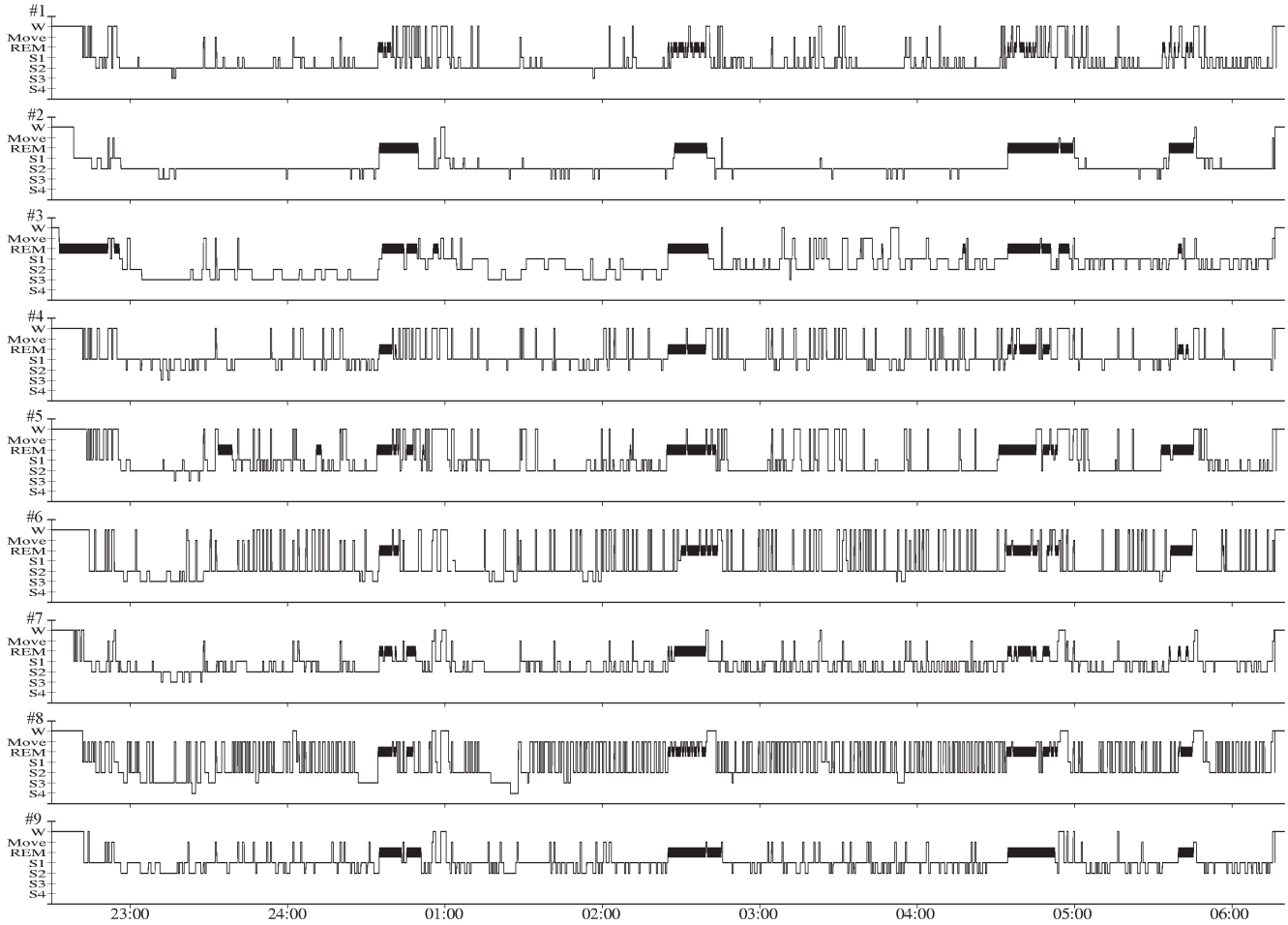
One lesson learned from the previous comparisons is that the subjective component in visual sleep scoring can be minimized by extensive training of sleep scorers. Optimal training for experienced sleep scorers is participation in consensus scoring rounds.

One other lesson learned from all previous studies on this subject is that inter-scorer reliability is not only an issue of one national sleep society but of all national sleep societies worldwide and between them. International inter-scorer reliability results need to be presented next.

CITATION

Penzel T; Zhang X; Fietze I. 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.

Figure 2—A comparison of sleep scorings of nine scorers from different sleep centers across Germany based on the same recording[5]



The recording was performed in a subject with obstructive sleep apnea. All the difficulties applying Rechtschaffen and Kales to sleep apnea EEG are obvious. Some scorers preferred to score arousal events as WAKE, others as MOVE, others as stage 1, and others as stage 2 sleep. REM sleep remained most consistent. Even the error by scorer #3 remained to be the same.

REFERENCES

1. Berry RB, Brooks R, Gamaldo CE, Harding SM, Marcus CL and Vaughn BV for the American Academy of Sleep Medicine. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.0. www.aasmnet.org, Darien, Illinois: American Academy of Sleep Medicine, 2012.
2. Rechtschaffen A, Kales A. *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. Washington, DC: US Department of health, Education and Welfare Public Health Service – NIH/NIND, 1968.
3. Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine Inter-scorer reliability program: sleep stage scoring. *J Clin Sleep Med* 2013;9:81-7.
4. Danker-Hopfe H, Anderer P, Zeitlhofer J, et al. Interrater reliability for sleep scoring according to Rechtschaffen & Kales and the new AASM standard. *J Sleep Res* 2009;32:139-49.
5. Penzel T, Behler PG, von Buttlar M, et al. Reliability of visual evaluation of sleep stages according to Rechtschaffen and Kales from eight polysomnographs by nine sleep centers. *Somnologie* 2003;7:49-58.
6. Normal RG, Pal I, Stewart C, et al. Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep* 2000;23:901-908.
7. Ruehland WR, O'Donoghue FJ, Pierce RJ, et al. The 2007 AASM recommendations for EEG electrode placement in polysomnography: impact on sleep and cortical arousal scoring. *Sleep* 2011;34:73-81.

ACKNOWLEDGMENTS

The authors want to thank the participants of the trial to compare sleep scorings between sleep centers in Germany as referred in Penzel et al.[5] Support was received for a research visit of Xiaozhe Zang by the Sino-German Science Center in Beijing (grant No GZ 598).

SUBMISSION & CORRESPONDENCE INFORMATION

Submitted for publication December, 2012
Accepted for publication December, 2012

Address correspondence to: Thomas Penzel, Ph.D., Charite - Universitatsmedizin Berlin, Department of Cardiology, Sleep Center, Chariteplatz 1, CCM 11, Berlin, D-10117, Germany; Tel: (49) 30-450513013; Fax: (49) 30-450513906; E-mail: thomas.penzel@charite.de

DISCLOSURE STATMENT

This was not an industry supported study. The authors have indicated no financial conflicts of interest.