Sleep Medicine 14 (2013) 1199–1207

Contents lists available at ScienceDirect

Sleep Medicine

journal homepage: www.elsevier.com/locate/sleep

Original Article

Scoring accuracy of automated sleep staging from a bipolar electrooocular recording compared to manual scoring by multiple raters

Carl Stepnowsky [a,b,*], Daniel Levendowski [c], Djordje Popovic [c], Indu Ayappa [d], David M. Rapoport [d]

[a] Department of Medicine, University of California, San Diego, La Jolla, CA, United States
[b] Veterans Affairs San Diego Healthcare System, San Diego, CA, United States
[c] Advanced Brain Monitoring, Inc., Carlsbad, CA, United States
[d] Department of Medicine, New York University, New York, NY, United States

ARTICLE INFO

Article history:
Received 9 January 2013
Received in revised form 25 April 2013
Accepted 26 April 2013
Available online 16 August 2013

Keywords:
Automatic sleep scoring
Electroencephalography
Electrooculography
Polysomnography
Sleep stages
Validation studies

ABSTRACT

Objectives: Electroencephalography (EEG) assessment in research and clinical studies is limited by the patient burden of multiple electrodes and the time needed to manually score records. The objective of our study was to investigate the accuracy of an automated sleep-staging algorithm which is based on a single bipolar EEG signal.

Methods: Three raters each manually scored the polysomnographic (PSG) records from 44 patients referred for sleep evaluation. Twenty-one PSG records were scored by Rechtschaffen and Kales (R&K) criteria (group 1) and 23 PSGs were scored by American Academy of Sleep Medicine (AASM) 2007 criteria (group 2). Majority agreement was present in 98.4% of epochs and was used for comparison to automated scoring from a single EEG lead derived from the left and right electrooculogram.

Results: The κ coefficients for interrater manual scoring ranged from 0.46 to 0.89. The κ coefficient for the auto algorithm vs manual scoring by rater ranged from 0.42 to 0.63 and was 0.61 (group 1, $\kappa = 0.61$ and group 2, $\kappa = 0.62$) for majority agreement for all studies. The mean positive percent agreement across subjects and stages was 72.6%, approximately 80% for stages wake (78.3%), stage 2 sleep (N2) (80.9%), and stage 3 sleep (N3) (78.1%); the percentage slightly decreased to 73.2% for rapid eye movement (REM) sleep and dropped to 31.9% for stage 1 sleep (N1). Differences in agreement were observed based on raters, obstructive sleep apnea (OSA) severity, medications, and signal quality.

Conclusions: Our study demonstrated that automated scoring of sleep obtained from a single-channel of forehead EEG results in agreement to majority manual scoring are similar to results obtained from studies of manual interrater agreement. The benefit in assessing auto-staging accuracy with consensus agreement across multiple raters is most apparent in patients with OSA; additionally, assessing auto-staging accuracy limited disagreements in patients on medications and in those with compromised signal quality.

1. Introduction

The importance of sleep on health and well-being is well-documented [1]. The challenge for the sleep field is to not only continue to increase the capacity for diagnostic sleep disorder testing, but also to improve on the ongoing long-term management of sleep disorders. Sleep disorder management might benefit from sleep studies to assess treatment efficacy, as important risk factors can change over time. If the burden of performing and scoring sleep studies was reduced, it could be used for long-term assessment and management of certain sleep and psychiatric disorders (e.g., insomnia, depression), including ongoing follow-up to monitor therapy adherence and assessing the role of therapeutic side effects and symptom resolution [2].

Historically the measurement of sleep has been accomplished with full polysomnography (PSG) in dedicated sleep laboratories. PSG provides comprehensive information about sleep architecture in a controlled laboratory environment. PSG will continue to be the standard against which other methods can be evaluated. However, full PSG is difficult to do on a repeated basis due to its complexity, effort, and costs. The attempt to obtain the same sleep information from more limited electroencephalography (EEG) montages, which could be automatically scored, would greatly contribute to the ease of including sleep analyses in multiple clinical or research settings.

* Corresponding author. Address: Veterans Affairs San Diego Healthcare System, 3350 La Jolla Village Drive (111n-1), San Diego, CA 92161, United States. Tel.: +1 858 642 1240; fax: +1 858 552 4321.
E-mail address: cstepnowsky@ucsd.edu (C. Stepnowsky).

1389-9457 Published by Elsevier B.V. Open access under CC BY-NC-ND license.
http://dx.doi.org/10.1016/j.sleep.2013.04.022

Manual sleep scoring is the gold standard, requiring trained sleep technicians to apply visual pattern recognition to the signals. In the best of circumstances, interrater reliability among scores approaches 0.90 and direct percent agreement approaches 80% to 85%. In typical clinical settings, these agreement metrics typically are less even with quality oversight. Within clinical research, the effects of lowered scoring reliability are that correlation coefficients are less robust, sample size requirements are increased, statistical power is reduced, and ultimately clinical trial costs are higher [3].

Computerized or automated scoring is one way to overcome some of these issues [4,5]. A previous review addressed the question of whether or not computerized polysomnographic analysis can reliably and accurately score sleep stages. Concerning sleep stage validation, the literature provided evidence that computerized scoring is reliable and accurate, relative to human scoring but with some caveats. In particular, the findings are not necessarily generalizable but are specific to the systems, algorithms, and specific human scoring training that are employed [4]. The review also suggests that the classification accuracy of any given system must be evaluated in both normal and sleep-disordered samples of patients. In addition, age-related changes need to be considered, and the need for high-quality recordings is critical.

We have previously published the accuracy of an auto-staging algorithm applied to a single channel based on the differential recording from left and right electrooocular (EOG) signals, compared to manual sleep staging based on a full PSG montage [6]. This single-electrode montage takes advantage of the information encoded in the left and right EOG signals as well as the frontal EEG. The previous cross-validation was limited, as only one rater per record was used.

Our study was designed to cross-validate our auto-staging algorithm on the single EEG/EOG lead in a new test dataset using agreement of 3 raters who scored each record as a reference. The use of multiple human scorers in our study helped to assess interrater reliability and also to improve the assessment of accuracy by minimizing scorer bias. Comparisons were made between two subgroups to highlight between-laboratory differences in the interpretation of the same rules applied to visual staging.

2. Methods

2.1. Study design

Our cross-sectional study was designed to compare interrater staging across three raters and then to compare the automated sleep-staging algorithm with majority scoring interrater agreement.

2.2. Data selection

The entire dataset included 44 studies in subjects with a mean age of 43 years (minimum, 22 years and maximum, 69 years) with 32% women, all undergoing full laboratory PSG. The dataset was developed by pooling the data from two projects by the similarity of methods, which included the use of three raters. The data used in our study were not used to train the algorithm or previously used in any way related to the algorithm; these data represent a new and independent test dataset.

Group 1 records were acquired at the New York University (NYU) School of Medicine using Sandman PSG equipment. Across the 23 records, the average apnea–hypopnea index (AHI) was 1 + 22 events per hour and included six healthy controls, five patients with an AHI <5, five patients with mild obstructive sleep apnea (OSA)(AHI, 5–15/hours), and seven patients with moderate to severe OSA. For the sleep staging, rater 1 was an expert in sleep staging unaffiliated with NYU (Mayo Clinic) and raters 2 and 3 were registered polysomnographic technicians (RPSGT) from NYU with expertise in staging sleep for research studies.

Group 2 consisted of a subset of 21 records from a separate group of 46 PSGs based on inclusion criteria requiring a minimum of 20 epochs of REM and stage 3 sleep (N3) from the initial diagnostic sleep staging and an AHI <30 events per hour. Of the 21 records, nine were acquired at NYU School of Medicine using Sandman PSG equipment and 12 were acquired at the Sleep Medicine Associates of New York City using Compumedics E series PSG equipment. The combined average AHI was 8 + 7.8 events per hour with 10 patients having an AHI <5, six patients having mild OSA, and five patients having moderate to severe OSA. Rater 1 (boarded in sleep) and rater 2 (RPSGT) were from University Services, Philadelphia, PA, and rater 3 was a RPSGT from NYU.

2.3. Manual scoring

The full PSG montage used for manual sleep staging provided electroencephalographic recordings from C3, C4, O1, O2, and Fz (referenced to the linked mastoids), left and right electrooculography (EOG-L and EOG-R), and submental electromyography (EMG). Group 1 data were scored using the criteria developed by Rechtschaffen and Kales (R&K) [2], as incorporated into their clinical scoring protocols. Group 2 data were scored according to the 2007 American Academy of Sleep Medicine (AASM) scoring rules [3]. The AHI for both groups was based on 10-s cessation in breathing or a 30% reduction in airflow coupled to a 4% decrease in oxyhemoglobin saturation. Raters were blind to the automated scoring.

2.4. Automated scoring

Three major steps were applied to the auto-staging algorithm: spectral decomposition of the input signal, computation of descriptors of sleep macro- and microstructure, and classification of 30-s epochs into one of the five stages (wake, REM, nonrapid eye movement sleep stage 1 [NREM1], NREM sleep stage 2 [NREM2] or NREM sleep stage 3 [NREM3]) (Fig. 1). The input signal is decomposed into delta, theta, alpha, sigma, beta, and EMG bands using digital filters. Two signals were derived in the delta band, one from the raw signal, and one after removal of ocular artifacts with a median filter. The other bands were extracted directly from the raw signal (eye movements had little impact on the signal power >4 Hz). Descendant signals in each band were integrated and fed to the feature extraction block.

Six descriptors of sleep macrostructure (SBI, DBI, EMI, BEI, $\overline{EMG}$, and $\bar{\beta}$) were derived for each 30-s epoch; their selection was guided by the literature [7] and attempts to mitigate between-subject variability of the envelopes in each band. Three descriptors of microstructure also were determined: number of spindles, number of arousals, and total length of all arousals in the epoch. Spindles and arousals were detected by contrasting short-term fluctuations to long-term trends in the signal [8]. Spindles were identified as 0.5- to 2-s segments of the signal during which the sigma envelope was larger than the theta, alpha, and beta envelopes and its instantaneous value exceeded the median value of the sigma envelope calculated over the preceding 30 s by a factor of 2. Cortical arousals during NREM sleep were detected as 3- to 15-s segments during which the instantaneous alpha envelope exceeded the respective median values calculated over the preceding 90 s by a factor of 2.

The macro- and microstructure descriptors were fed to a hierarchical decision tree with seven nodes. Node R1 classified epochs into NREM cluster (NREM2, NREM3, or some NREM1) or beta-dominated cluster (wake, REM, or most of NREM1). The NREM cluster was further separated into light (NREM1/2) and deep sleep