View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

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.

Published by Elsevier B.V. Open access under CC BY-NC-ND license.

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

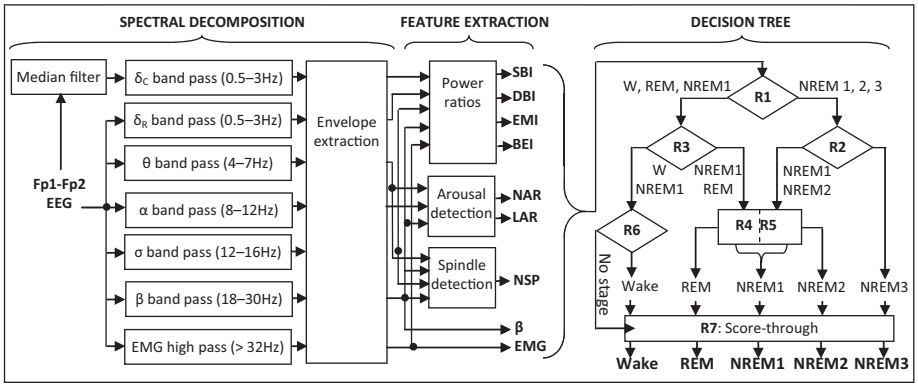

Fig. 1. Block diagram of the FP-STAGER algorithm.

(NREM3), whereas the beta-dominated cluster was divided into REM/NREM1 and wake/NREM1 subclusters (nodes R2, R3). REM sleep was identified in two steps which resembled the AASM rules for initiation and continuation of REM scoring: seed epochs were first identified with high precision using one set of thresholds, followed by examination of the 3-min segments around each seed against another set of thresholds. At the next level, nodes R4 and R5 separated NREM1 epochs with arousals from the NREM1/2 and REM/NREM1 clusters and node R6 identified wake- and arousal-free NREM1 epochs. The epochs unclassified at nodes R1–R6 were assigned a stage using a simple score-through rule (node R7).

2.5. Data analysis

The reliability of manual sleep staging was determined by the interrater κ coefficients and the percentage epochs in agreement between one rater vs concurrent agreement across the other two raters and by the percentage of epochs of which at least two raters agreed. Differences among raters were further elucidated by comparing each of their manual staging vs the auto-staging algorithm to derive sensitivity and positive predictive values (PPV) by stage and across stage κ coefficients (all measures based on pooled epochs across subjects).

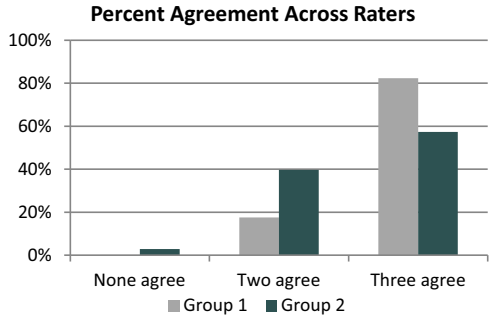

Fig. 2. Percentage of epochs with no agreement, majority agreement (at least two agree), and consensus agreement (all three agree) among raters.

Further assessment was made by comparison of the automated algorithm to majority agreement with κ coefficients and by the percent of epochs in agreement based on pooled epochs across subjects and on averaging of by-subject sensitivity and PPV values.

All epochs were manually staged, submitted for automated staging, and used in the analyses. To compute the by-subject sensitivity and PPV, epochs with no consensus agreement were dropped and at least 11 epochs with consensus agreement were included in mean values reported by and across stages and groups.

Box plots were used to assess interrater variability and auto-staging performance by stage and group. The box represents the distributions of the second and third quartile about the median, the whiskers represent the 10% and 90%, and the Δ identifies the outliers. Group data were combined for presentation of Bland–Altman plots and interclass correlations to assess auto-staging performance related to total sleep time, sleep efficiency, sleep latency, wake after sleep onset, Stage N3 and REM onset, and percentage of sleep time sleep stages 1 (N1), 2 (N2), N3 and REM.

3. Results

3.1. Agreement among raters

Fig. 2 reports on the percentage of epochs with no agreement, majority agreements (at least two agreed), and consensus agreement (all three agreed) among raters by group. Table 1 shows the all-stage agreement among raters as measured by the pooled κ coefficients (left side of Table) and percent agreements between one rater and consensus agreement with the other two raters (right side of Table). The agreements within group 1 (raters 1 vs 2 and 1 vs 3) and within group 2 (raters 1 vs 2) were strong. There was relatively less agreement within group 1 (raters 1 vs 3) and group 2 (rater 1 vs 3 and 2 vs 3). The percentage of epochs in which there was majority agreement was slightly better for group 1 (99.9%) compared to group 2 (97.1%), with the overall majority agreement being 98.4%. The strong agreement between raters 1 vs 2 of group 2 masked their substantial disagreements with rater 3. Table 2, which compares individual rater agreement vs the auto-staging algorithm, further highlights differences among the raters

Table 1
All stage agreement for manual scoring among raters by and across groups.

	κ Agreement across all stages			% Epochs in agreement			
	Raters 1 vs 2	Raters 1 vs 3	Raters 2 vs 3	Rater 1 vs 2 and 3 agree	Rater 2 vs 1 and 3 agree	Rater 3 vs 1 and 2 agree	At least two raters agree
Group 1	0.85	0.89	0.77	83.8	92.2	89.2	99.9
Group 2	0.80	0.46	0.49	64.7	61.1	86.0	97.1
Overall	0.83	0.68	0.64	73.2	75.2	87.4	98.4

Table 2
Mean agreements between auto-staging and individual human raters by stage.

Rater	Wake		N1		N2		N3		REM		Overall	
	Sen	PPV	Sen	PPV	Sen	PPV	Sen	PPV	Sen	PPV	Agree	κ
Group 1: rater agreements based on pooled epochs across subjects												
1	78.6	82.2	43.5	51.1	79.1	75.3	82.3	75.4	79.2	71.5	72.2	0.63
2	78.9	80.4	40.3	53.7	75.9	74.5	90.1	55.4	78.6	69.3	69.6	0.59
3	77.2	82.7	42.0	47.0	78.3	74.3	81.6	78.2	78.5	71.5	71.4	0.62
Group 2: rater agreements based on pooled epochs across subjects												
1	77.3	80.6	24.5	21.6	80.7	64.3	62.9	86.6	65.8	82.0	69.5	0.60
2	77.3	81.6	30.1	17.9	78.9	75.1	71.4	77.0	65.4	82.6	72.6	0.63
3	69.8	71.4	19.0	34.5	62.5	63.2	73.4	26.4	58.8	65.9	57.1	0.42

Abbreviations: N1, sleep stage 1; N2, sleep stage 2; N3, sleep stage 3; REM, rapid eye movement sleep; PPV, positive predictive values; Sen, sensitivity.

Table 3
Contingency table between auto-staging and majority scoring by human raters.

	Auto-staging					
	Wake (%)	N1 (%)	N2 (%)	N3 (%)	REM (%)	No. epochs
Group 1: based on pooled epochs						
Wake	82.5	16.2	2.6	0.3	6.1	3847
N1	12.9	51.5	16.0	1.5	16.7	3567
N2	3.9	25.4	75.5	19.8	4.9	6839
N3	0.2	0.1	3.6	78.1	0.3	1687
REM	0.5	6.8	2.1	0.2	72.0	1855
No cons.	0.1	0.2	0.1	0.0	0.0	14
No. epochs	3682	3032	7240	1807	2048	17,809
Group 2: based on pooled epochs						
Wake	81.5	23.5	4.2	0.7	6.6	4727
N1	7.7	21.0	5.5	0.5	5.8	1416
N2	4.5	33.5	71.1	19.9	4.1	8015
N3	0.6	1.0	9.5	76.5	0.0	3389
REM	3.5	16.9	5.8	0.9	81.8	3205
No cons.	2.2	4.0	4.0	1.6	1.7	630
No. epochs	4541	1922	9016	3248	2655	21,382

Abbreviations: N1, sleep stage 1; N2, sleep stage 2; N3, sleep stage 3; REM, rapid eye movement sleep; cons., consensus; No., number.

in their staging of stage N2, N3, and REM as well as consistent discrepancies in performance for rater 3 of group 2 vs all other raters.

3.2. Majority agreement vs automated staging

Table 3 presents a contingency matrix of majority agreements across subjects by sleep stage and group. Based on the pooling of epochs, the overall agreement and all stage κ coefficient were similar between groups 1 (72.8% and 0.63%) and 2 (73.1% and 0.64%).

Table 4 presents the sensitivity and positive predictive values by and across stages and groups. Based on individual results averaged across subjects, the overall agreement from groups 1 and 2 also were similar (72.2 vs 73.0). The sensitivities for each of the stages were above 0.73 with the exception of N1 when averaged across groups. The PPV for stage N2 was equivalent to the overall agreement but was slightly lower compared to stages wake, N3, and REM. The two groups were comparable on most measures of sensitivity and PPV; however, group 2 had a lower sensitivity compared to group 1 for stage N3 and a lower sensitivity and higher PPV for stage REM. When agreements were stratified by OSA severity, patients with moderate to severe OSA had superior sensitivities and PPV for REM and PPV for wake and inferior sensitivities and PPV for N3 and PPV for N2, compared to those with an AHI <15.

Table 5 provides a comparison of averaged sensitivity and PPV values stratified by factors known to affect EEG sleep staging. The first analysis compares the agreements between patients on medications (i.e., benzodiazepines [$n = 3$], antidepressants [$n = 2$], benzodiazepines and antidepressants [$n = 2$], Adderall [$n = 1$]) to those not on these medications. As expected, the overall agreement was greater for patients not on medications (74.9 vs 70.5). The EEG from patients on medications contributed to reduced sensitivity for stage N3 and REM, reduced PPV for wake and stage N3, increased sensitivity for stage N1, and increased PPV for stage N2 and REM. The second analysis in Table 5 compares the agreements across sites using different recording equipment. The equipment used at site 1 contributed to superior agreement between consensus manual staging and the overall auto-staging (75.5 vs 71.1) and virtually all stage-related sensitivity and PPV values compared to site 2.

3.3. Distribution of interrater and auto-staging performance

Fig. 3 shows the box plots of the distribution of agreement by and across stage among raters auto-staging vs majority agreement, stratified by group. In the all stages box plot for group 2, one can see that the range of distribution for rater 3 was wider than for the other raters, but that this increased variability only had a minor impact on the majority agreement vs auto-staging percent agreement, which remained comparable to that of group 1. It would appear that much of the variability for rater 3 of group 2 concerned the scoring of stage 3 and REM to a slightly lesser degree.

Table 4
Mean agreements between auto-staging and majority scoring by human raters by stage.

Condition	Wake		N1		N2		N3		REM		Overall
	Sen	PPV	Sen	PPV	Sen	PPV	Sen	PPV	Sen	PPV	Agree
Agreements based on average of individual results by and across groups											
Group 1	79.9	83.1	36.5	47.7	81.9	72.6	86.1	79.4	81.1	74.3	72.2
Group 2	76.6	79.5	26.8	20.5	79.8	73.2	72.0	78.7	66.5	81.5	73.0
Overall	78.3	81.4	31.9	34.7	80.9	72.9	78.1	79.0	73.2	78.2	72.6
Agreements based on average of individual results by OSA severity											
AHI <15 ($n = 32$)	78.0	80.0	31.2	33.7	80.9	76.6	80.1	79.8	70.9	76.7	73.2
AHI >15 ($n = 12$)	79.3	85.1	33.6	37.5	81.0	63.0	70.7	76.2	81.1	83.4	71.0

Abbreviations: N1, sleep stage 1; N2, sleep stage 2; N3, sleep stage 3; REM, rapid eye movement sleep; PPV, positive predictive values; Sen, sensitivity; AHI, apnea–hypopnea index; OSA, obstructive sleep apnea.

Table 5
Mean agreements between auto-staging and human raters by stage for group 2.

Condition	Wake		N1		N2		N3		REM		Overall
	Sen	PPV	Sen	PPV	Sen	PPV	Sen	PPV	Sen	PPV	Agree
Agreement stratified by use of medications											
Meds (n = 9)	77.2	72.5	33.6	20.9	77.1	76.8	67.8	74.3	62.4	86.0	70.5
No meds (n = 12)	76.2	84.9	21.7	20.2	81.8	70.5	75.1	82.0	69.5	78.2	74.9
Agreement stratified by differences in signal quality											
Site 1 (n = 9)	78.2	86.5	33.5	21.2	82.2	77.4	69.5	85.1	69.9	83.5	75.5
Site 2 (n = 12)	75.3	74.4	21.8	20.0	78.0	70.0	73.8	73.9	63.9	80.0	71.1

Abbreviations: N1, sleep stage 1; N2, sleep stage 2; N3, sleep stage 3; REM, rapid eye movement sleep; PPV, positive predictive values; Sen, sensitivity; Meds, medicine.

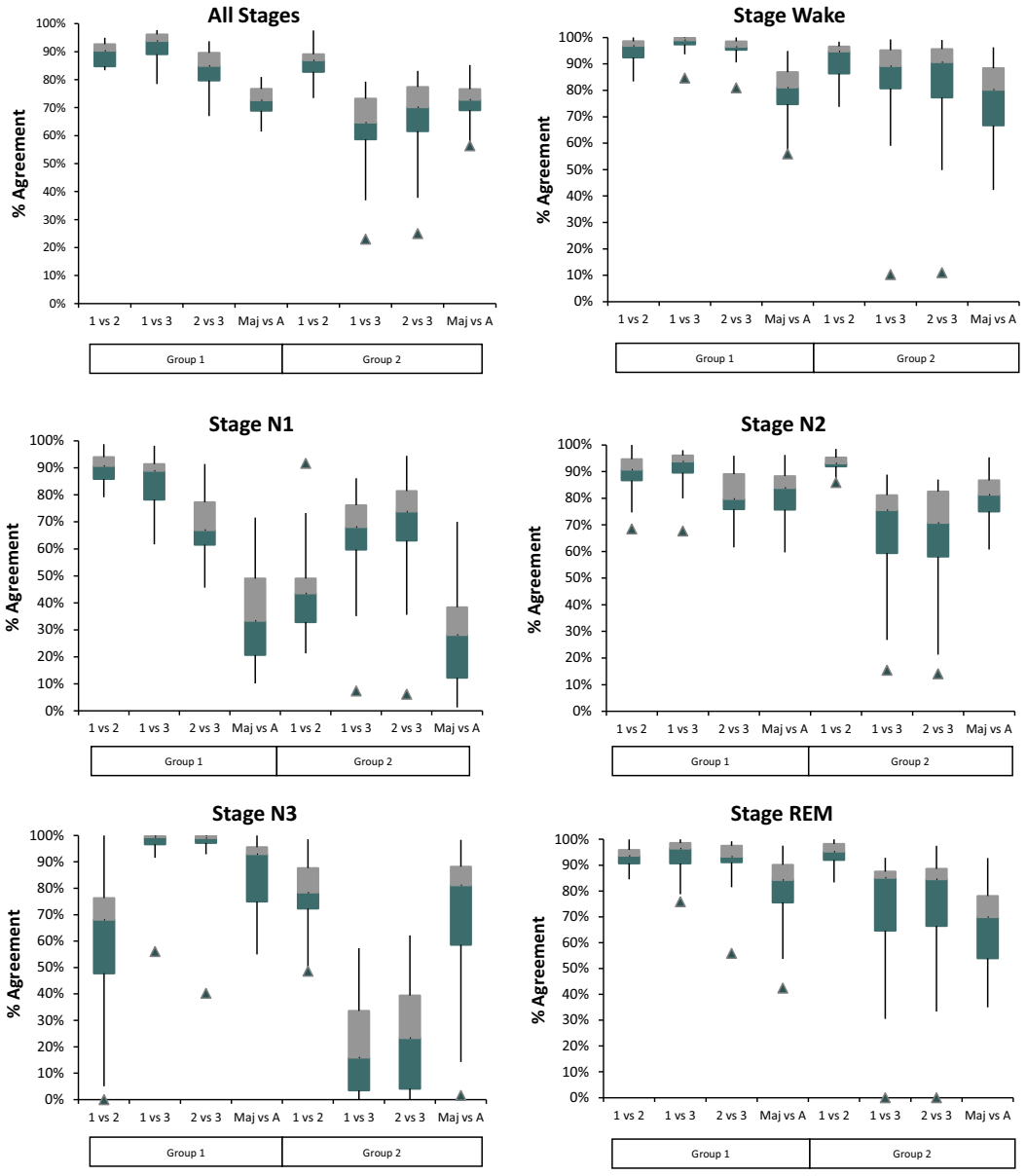

Fig. 3. Box plots showing distribution of agreement (sensitivity) across and by stage among raters and majority agreement vs auto-staging, stratified by group.

When observing the auto-staging vs majority agreement across the sleep stages, it was apparent that the percent agreement was strong for all combined stages, wake, and N2, and N3; the percent agreement slightly decreased for REM and was the lowest for N1. Correspondingly, the interrater reliability was the highest for all combined stages, wake, and N2; it slightly decreased for N3 and REM and was the lowest for N1.

3.4. Bland–Altman plots

Figs. 4 and 5 show the Bland–Altman plots between auto-staging and majority agreement pooled across groups. Fig. 3 provides the plots for total sleep time, sleep efficiency, sleep latency, wake after sleep onset, N3 onset, and REM onset. The plots show that the interclass correlation coefficients were all above 0.70 and were

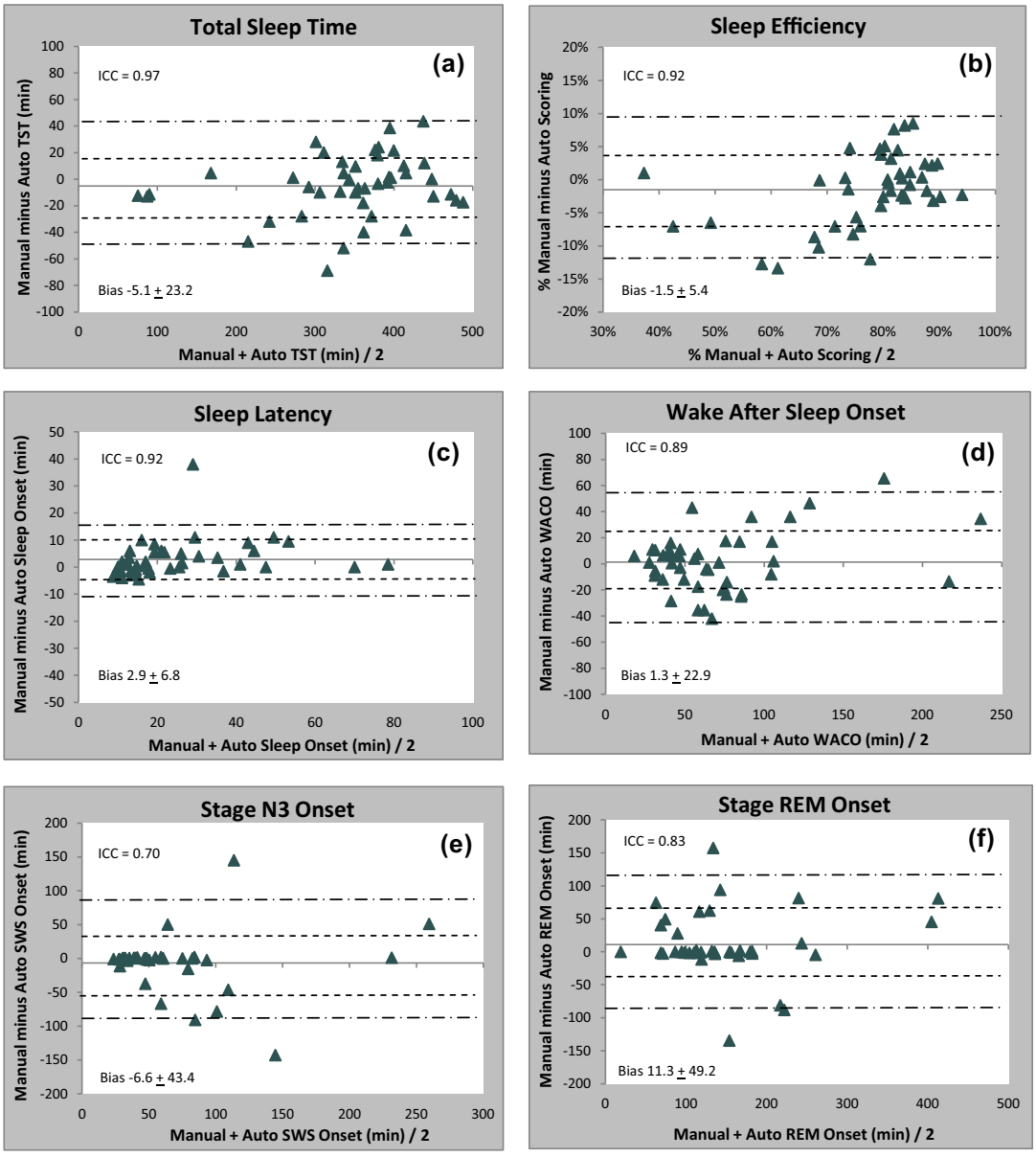

Fig. 4. Interclass correlations and Bland–Altman plot comparisons between majority agreement and auto-staging for detection of (a) total sleep time, (b) sleep efficiency, (c) sleep latency, (d) wake after sleep onset, (e) sleep stage 3 (N3) onset, and (f) rapid eye movement (REM) sleep onset.

the highest for total sleep time, sleep efficiency, sleep latency, and wake after sleep onset. Fig. 4 provides the plots for percentage of N1, N2, N3, and REM. The interclass correlation coefficients were 0.78, 0.53, 0.80, and 0.76, respectively. Because the interrater variability was primarily limited to stages N3 and REM, the data between the two groups (and scoring rules) were pooled to assess the standard measures of sleep architecture presented in Figs. 4 and 5.

4. Discussion

Our study evaluated the accuracy of an auto-staging algorithm used to stage sleep compared to manual scoring based on multiple raters with majority agreement between raters used to assess the accuracy of the auto-staging algorithm. The main finding of our study was that the auto-staging algorithm was comparable to manual scoring for all stages combined, wake, N2, and N3, while the agreement slightly decreased for REM. The similarly performed algorithm of whether or not R&K (group 1) or AASM criteria (group 2) for scoring was applied. The κ score between the automated algorithm and pooled majority agreement across stages was comparable for groups 1 and 2 (0.63 vs 0.64, respectively) and the overall agreement, whether or not it was pooled across epochs (72.8 vs 73.1) or averaged across subjects (72.2 and 73.0), was strong. The overall agreement between the majority agreement and the auto-staging across subjects as well as for those with sleep-disordered breathing was equivalent to the epoch-by-epoch agreement for 10 pairs of scorers [9].

The scorers included in our study were all highly trained and experienced sleep scorers. Despite this level of scoring expertise, there were still prominent human scoring discrepancies. In group 1, the κ coefficients suggested that raters 2 and 3 had a lower rate of agreement (Table 1); Table 2 suggests that their disagreement primarily occurred with stage N3. For group 2, rater 3 scored grossly different than raters 1 and 2 across all stages. This finding was apparent based on the κ scores and across stage agreements (Table 1) and the differences in agreement with the auto-staging algorithm (Table 2). In group 2, raters 1 and 2 also staged N3 differ-

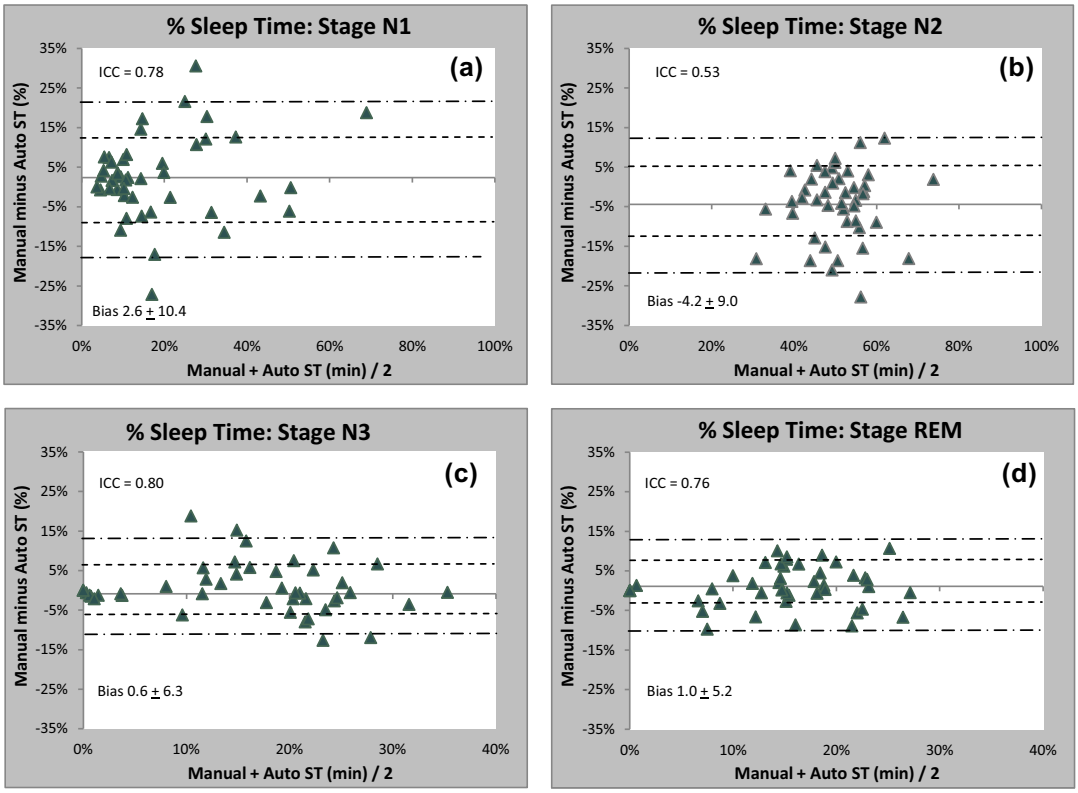

Fig. 5. Interclass correlations and Bland–Altman plot comparisons between majority agreement and auto-staging for respective percentages of sleep time (a) stage 1 sleep (N1), (b) stage 2 sleep (N2), (c) stage 3 sleep (N3), and (d) rapid eye movement (REM) sleep.

ently based on their sensitivity and PPV values presented in Table 2. The benefit of using multiple raters [10] is highlighted by the fact that the overall agreement for group 2 was greater than group 1 (see Tables 3 and 4); the use of majority agreement coupled with similar scoring styles for two of the three raters in group 2 masked their substantial disagreement with rater 3.

Our study utilized two approaches to measure the agreement between manual scoring and auto-staging. The pooling of epochs (Table 3) tends to hide wide variances in agreement across subjects, while the averaging of individual agreement metrics across subjects (Table 4) applies equivalent weight to results based on a limited number of epochs. The distinction between these two approaches is highlighted by between-group comparisons. For group 1, the agreements were stronger when averaged across subjects vs pooling for stages N3 and REM. For group 2, the agreements were weaker when averaged across subjects vs pooling for stages N3 and REM. These findings are further reflected in the box plot presentation of sensitivity distributions (Fig. 3), demonstrating more consistent agreement among raters (as occurred with group 1) contributing to higher and more consistent levels of agreement vs the auto-staging.

OSA is a factor that disrupts sleep architecture and influences scoring reliability, yet it had a limited impact on the sensitivity and PPV for stages wake, N1, and N2. Stage N3 sensitivities decreased as OSA severity increased, likely resulting from suppressed slow-wave sleep (SWS) in those with moderate to severe OSA. The improved sensitivity and PPV in auto-staged detection of REM for those with moderate to severe OSA may be explained by more consolidated and easier-to-recognize signal patterns. When comparing the results from our study to a previously published comparison of the auto-staging algorithm in patients with OSA, the benefit of multiple raters becomes obvious. When comparing the auto-staging algorithm to a single rater vs majority agreement in patients with an AHI <15, the sensitivities were similar for wake (78.0 vs 79.7), N1 (31.2 vs 25.1), N2 (80.9 vs 77.7), N3 (80.1 vs 83.9), and REM (70.9 vs 74.6). For patients with moderate to severe OSA, the sensitivities markedly improved from a single rater to a majority agreement for wake (71.1 vs 79.3), N2 (70.5 vs 81.0), N3 (65.0 vs 70.7), and REM (67.4 vs 81.1).

Medications used by patients with insomnia and psychiatric disorders suppress SWS and REM, contributing to more difficult recognition of the emergence of these stages and likely contributed to the reduced overall agreement between majority agreement and auto-staging. Patients on medications that suppress stage N3 were observed with inferior sensitivity and PPV values. These signal patterns likely contributed to the superior PPV in patients on medications during stage N2. The combination of improved sensitivity in detection of stage N1 with improved PPV during REM in patients on medications suggests that when REM occurred, it was consolidated and more easily recognized. Because the auto-scoring algorithm is based on power spectral density ratios, it applies a broader, more consistent approach to stage these signal patterns. Rather than relying solely on EEG amplitude (which is influenced by skull thickness, age and/or medications), the auto-staging algorithm assesses magnitude differences in the delta and theta power to stage N3, applying a level of precision that is not possible with visual scoring. The disadvantage of the current algorithm is its lack of a time-series application to the individually staged epochs. Human scoring incorporates decision making that reduces the likelihood that stage N1 and REM are confused due to known timing (i.e., at the onset of sleep), and humans adjust the scoring to assign epochs the same stage to create uninterrupted blocks of REM or SWS. Investigations are underway to incorporate higher-level temporal-based analyses to our staging algorithm.

Signal quality also is a factor that influences manual staging; a noisy signal makes it more difficult to differentiate wake from REM and to visually identify subtle sleep spindles used to differentiate stages N2 and N3. The two group 2 raters who scored similarly and who were unaffiliated with either site noted difficulty in manually staging the site 2 studies, due to signal quality. This difficulty became apparent when the agreements were stratified by site. The data acquired at site 1 resulted in superior overall agreement and superior PPV values in every stage as compared to site 2. The sensitivities also were superior from site 1 for every stage except N3. Auto-staging has the ability to isolate the power spectra associated with arousals and sleep spindles or cause subtle increases in beta and decreases in EMG, which are not visually apparent when signal quality is inconsistent.

As is common with sleep scoring, the agreement in our study was lowest for N1, regardless if it was manual or computerized. The agreements between the raters in group 1 vs the auto-staging algorithm was equivalent to the previously reported between rater agreement from different sleep centers [9] (i.e., approximately 0.40). In a previous report that compared manual scoring by R&K vs AASM standard criteria, N1 was found to have the lowest κ coefficient score of any sleep stage (0.41 and 0.46, respectively) [11]. Group 2 included raters from different sleep centers and their agreement was substantially lower than group 1. Both signal quality and lower sleep-disordered breathing severity, which is inversely proportional to N1 accuracy in manual staging and with auto-staging, may have contributed to this outcome. Part of the issue with scoring N1 is the difficulty in identifying the transition from wake to N1, which may in part be explained by the finding that upwards of 20% of the general population generate little or no alpha [12]. Detection of alpha activity is even more problematic when frontopolar sensor placements are used. Despite the well-known difficulty in identifying N1, it typically accounts for a small percentage of the night and the overall percent agreement between manual scoring, and auto-staging was still high in our study. In our study the primary misclassification of stage N1 was into the bordering stages N2 and wake. A substantially greater percentage of epochs manually staged as N1 were auto-staged as REM for group 2 vs 1 (16.9% vs 6.8%, respectively).

Previous research has examined auto-staging algorithms of a subset of physiologic signals from polysomnography that was collected in a controlled environment [13–15]. These kinds of evaluations provide a robust comparison of the auto-scoring algorithm relative to manual scoring because of the use of the same physiologic signals. However, generalizability to portable systems is limited. Auto-scoring algorithms that are based on signals from novel sleep recording devices require that the data be collected by those systems simultaneously with PSG. Our analysis was solely based on censoring the PSG signals down to one pair of frontal/eye electrodes. Our study evaluated an algorithm which can subsequently be used by a portable sleep recording system, but evaluation of the algorithm on PSG signals is considered the first step in the validation process. A previous paper showed that the algorithm is reliable and valid in those with and without OSA; however, the authors utilized a single rater [6]. Our study provides further evidence on the efficacy of the algorithm by utilizing three raters and majority agreement. One of the limitations of our study was that there were insufficient data to statistically evaluate the interaction between OSA severity, medication, and signal quality on the obvious differences in manual scoring reliability and auto-scoring accuracy.

All of the epochs in each of the records were scored by the raters; none of those scored by R&K criteria included epochs classified as movement time. The lack of artifact in these records was likely attributed to vigilant technicians who were aware that the data were to be included in a clinical study. Acquisition under less optimal conditions can result in data that should be identified and rejected from analysis by the auto-staging algorithm. The auto-staging algorithm used in our study rejects signals characterized by high to peak amplitudes (i.e., ±150 μV) or dramatic amplitude changes which are highly correlated across the two referential signals. One benefit of using a differential recording for sleep staging is that artifact common to both sensors (i.e., in phase) is rejected. Although it was not required for this dataset, the algorithm was designed to be applied to one of the two referential channels when artifact was detected in one but not both of the inputs to the differential signal.

Future sleep studies may be best served by consideration of a hybrid model, whereby a validated scoring algorithm provides the initial scoring and then the study is manually reviewed and scored. One frequently advocated approach to using an auto-staging algorithm is to manually review and overscore its output. However, as historically many discrepancies to human scoring resulted when auto-staging was performed, this resulted in minimal gain of time over solely doing human scoring. If an auto-staging algorithm can result in a high direct percentage agreement, it can both save time for the manual review and result in more accurately scored studies. Our study showed a direct percent agreement between algorithm and majority agreement of 98.4%. Using such an effective algorithm also can have the benefit of allowing a scorer to focus more time on those notoriously difficult stage transitions, such as wake/S1 and S1/REM.

Conflict of interest

The ICMJE Uniform Disclosure Form for Potential Conflicts of Interest associated with this article can be viewed by clicking on the following link: http://dx.doi.org/10.1016/j.sleep.2013.04.022.

Acknowledgments

This study was funded in part by National Institutes of Health grants MH088282-1 and HL68463-04. The authors wish to thank Peter Hauri, PhD, Chandra Matadeen-Ali, MD, Michael Simms, and Rakhil Kanevskaya for their assistance with scoring the polysomnography data.

References

[1] Colten HR, Altevogt BM, editors. Committee on sleep medicine and research, sleep disorders and sleep deprivation: an unmet public health problem. Washington (DC): The National Academies Press; 2006.
[2] Epstein LJ, Kristo D, Strollo Jr PJ, Friedman N, Malhotra A, Patil SP, et al. Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults. J Clin Sleep Med 2009;5:263–76.
[3] Stepnowsky Jr CJ, Berry C, Dimsdale JE. The effect of measurement unreliability on sleep and respiratory variables. Sleep 2004;27:990–5.
[4] Penzel T, Hirshkowitz M, Harsh J, Chervin RD, Butkov N, Kryger M, et al. Digital analysis and technical specifications. J Clin Sleep Med 2007;3:109–20.
[5] Hirshkowitz M, Moore CA. Computers in sleep medicine. In: Kryger MH, Roth T, Dement WC, editors. Philadelphia (PA): Saunders; 2000.
[6] Levendowski DJ, Popvic D, Berka C, Westbrook PR. Retrospective cross-validation of automated sleep staging using electroocular recording in patients with and without sleep disordered breathing. Int Arch Med 2012;5:21.
[7] Uchida S, Maloney T, Feinberg I. Σ (12–16Hz) and β (20–28Hz) EEG discriminate NREM and REM sleep. Brain Res 1994;659:243–8.
[8] De Carli F, Nobili L, Gelcich P, Ferrillo F. A method for the automatic detection of arousals during sleep. Sleep 1999;22:561–72.
[9] Norman R, Pal I, Stewart C, Rapoport DM. Interobserver agreement among sleep scorers from different centers in a large dataset. Sleep 2000;23:901–8.
[10] Penzel T, Glos M, Schobel C, Sebert M, Diecker B, Fietze I. Revised recommendations for computer-based sleep recording and analysis. Conf Proc IEEE Eng Med Biol Soc 2009;2009:7099–101.
[11] Danker-Hopfe H, Anderer P, Zeitlhofer J, Boeck M, Dorn H, Gruber G, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. J Sleep Res 2009;18:74–84.

[12] Silber MH, Ancoli-Israel S, Bonnet MH, Chokroverty S, Grigg-Damberger MM, Hirshkowitz M, et al. The visual scoring of sleep in adults. J Clin Sleep Med 2007;3:121–31.
[13] Berthomier C, Drouot X, Herman-Stoïca M, Berthomier P, Prado J, Bokar-Thire D, et al. Automatic analysis of single-channel sleep EEG: validation in healthy individuals. Sleep 2007;30:1587–95.
[14] Krakovska A, Mezeiova K. Automatic sleep scoring: a search for an optimal combination of measures. Artif Intell Med 2011;53:25–33.
[15] Fraiwan L, Lweesy K, Khasawneh N, Wenz H, Dickhaus H. Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier. Comput Methods Programs Biomed 2012;108:10–9 [published online ahead of print December 16, 2011].