The 2007 AASM Recommendations for EEG Electrode Placement in Polysomnography: Impact on Sleep and Cortical Arousal Scoring

Warren R. Ruehland, BSc(Hons)[1,2]; Fergal J. O'Donoghue, MD, PhD[1]; Robert J. Pierce, MD[1]; Andrew T. Thornton, PhD[3]; Parmjit Singh, BSc(Hons), MBA[3]; Janet M. Copland, BAppSc[4]; Bronwyn Stevens, BBNSc, PGradDip (Psych)[1]; Peter D. Rochford, BAppSc, GradDip (Bio Instr)[1]

[1]Institute for Breathing and Sleep, Austin Health, Heidelberg, Victoria, Australia; [2]Department of Medicine, University of Melbourne, Austin Health, Heidelberg, Victoria, Australia; [3]Royal Adelaide Hospital, Adelaide, South Australia, Australia; [4]Monash Medical Centre, Clayton, Victoria, Australia

Study Objectives: To examine the impact of using American Academy of Sleep Medicine (AASM) recommended EEG derivations (F4/M1, C4/M1, O2/M1) *vs.* a single derivation (C4/M1) in polysomnography (PSG) on the measurement of sleep and cortical arousals, including inter- and intra-observer variability.

Design: Prospective, non-blinded, randomized comparison.

Setting: Three Australian tertiary-care hospital clinical sleep laboratories.

Patients or Participants: 30 PSGs from consecutive patients investigated for obstructive sleep apnea (OSA) during December 2007 and January 2008.

Interventions: N/A

Measurements and Results: To examine the impact of EEG derivations on PSG summary statistics, 3 scorers from different Australian clinical sleep laboratories each scored separate sets of 10 PSGs twice, once using 3 EEG derivations and once using 1 EEG derivation. To examine the impact on inter- and intra-scorer reliability, all 3 scorers scored a subset of 10 PSGs 4 times, twice using each method. All PSGs were de-identified and scored in random order according to the 2007 *AASM Manual for the Scoring of Sleep and Associated Events.* Using 3 referential EEG derivations during PSG, as recommended in the AASM manual, instead of a single central EEG derivation, as originally suggested by Rechtschaffen and Kales (1968), resulted in a mean ± SE decrease in N1 sleep of 9.6 ± 3.9 min (P = 0.018) and an increase in N3 sleep of 10.6 ± 2.8 min (P = 0.001). No significant differences were observed for any other sleep or arousal scoring summary statistics; nor were any differences observed in inter-scorer or intra-scorer reliability for scoring sleep or cortical arousals.

Conclusions: This study provides information for those changing practice to comply with the 2007 AASM recommendations for EEG placement in PSG, for those using portable devices that are unable to comply with the recommendations due to limited channel options, and for the development of future standards for PSG scoring and recording. As the use of multiple EEG derivations only led to small changes in the distribution of derived sleep stages and no significant differences in scoring reliability, this study calls into question the need to use multiple EEG derivations in clinical PSG as suggested in the AASM manual.

Keywords: Electroencephalography, polysomnography, sleep scoring, cortical arousal scoring, sleep architecture, obstructive sleep apnea, sleep disordered breathing, inter-scorer reliability, intra-scorer reliability, kappa

Citation: Ruehland WR; O'Donoghue FJ; Pierce RJ; Thornton AT; Singh P; Copland JM; Stevens B; Rochford PD. The 2007 AASM recommendations for EEG electrode placement in polysomnography: impact on sleep and cortical arousal scoring. *SLEEP* 2011;34(1):73-81.

IN CLINICAL PRACTICE THE MOST COMMON INDICATIONS FOR POLYSOMNOGRAPHY (PSG) ARE INVESTIGATION AND TREATMENT OF OBSTRUCTIVE SLEEP apnea (OSA). For OSA diagnosis the main outcome measures from PSG are: (1) the apnea-hypopnea index (AHI), which is a measure of sleep disordered breathing events (apneas and hypopneas) per hour of sleep, (2) the arousal index (ArI) which is a measure of sleep disruption per hour of sleep and (3) various sleep scoring summary statistics describing sleep quality or sleep architecture (e.g., sleep efficiency). For these measures, scoring of arousals and sleep are not only important in their own right, but they also impact on other measures. For example, total sleep time (TST) is used as the denominator for both the AHI and ArI, and some criteria allow for hypopnea scoring if airflow reduction is accompanied by a cortical arousal.[1]

The scoring of sleep and arousals relies on visual inspection of continuous surface electroencephalography (EEG), electromyography (EMG), and electrooculography (EOG) measurements. Rechtschaffen and Kales (R&K),[2] the first consensus-based guidelines for scoring of sleep, recommended recording a minimum of one channel of central EEG (either C_3/A_2 or C_4/A_1) during PSG. The one-derivation minimum was recommended due to device limitations and because it was thought that regional differences in scalp areas were not critical for sleep scoring. Subsequently, other authors have suggested that regional differences may be important[3] and have recommended the use of more than one EEG derivation.[4]

In agreement with this view the 2007 *AASM Manual for the Scoring of Sleep and Associated Events*[1] recently recommended the use of 3 standard EEG derivations for scoring of sleep; including frontal, central, and occipital derivations. The evidence review paper underpinning the AASM manual,[5] stated the recommendations follow from the current less restrictive device limitations and from observations in healthy subjects that, although sleep spindles may be generally recorded optimally over central regions,[3,6-8] this is not the case for other

Submitted for publication March, 2010
Submitted in final revised form April, 2010
Accepted for publication May, 2010
Address correspondence to: Warren Ruehland, Institute for Breathing and Sleep, Austin Health, Ground Floor Bowen Centre, Studley Road, Heidelberg VIC 3084, Australia; Tel: +61 3 9496 3528; Fax: +61 3 9496 5128; E-mail: Warren.Ruehland@austin.org.au

sleep scoring features such as K complexes,[3,9] delta waves,[9,10] and alpha activity.[11] The review therefore suggested that the use of a single central EEG derivation may lead to sleep scoring inaccuracies.[5]

One study comparing sleep scoring using a 2-channel bipolar EEG montage, including a frontal EEG derivation, to a montage including a single central referential EEG derivation as recommended by R&K, suggested a tendency to score more deep sleep with the bipolar montage.[12] There is little data however on how increasing the number of EEG derivations in PSG may impact on sleep scoring summary statistics such as TST, sleep efficiency (SE), sleep latency (SL), REM latency (RL), or time in particular sleep stages. Also, although the above study[12] suggested equivalent inter-scorer reliability of sleep scoring between the 2 configurations, direct comparison of inter- and intra-scorer reliability using R&K *vs.* AASM recommended electrode placements was not made.

In line with previous recommendations,[13] the 2007 AASM manual[1] recommended that arousal scoring should include information from occipital and central EEG derivations. The associated evidence review paper[14] stated that the number of arousals scored would be expected to be larger using occipital EEG derivations in addition to central EEG derivations. Although to the best of our knowledge, this statement has not been formally evaluated, it has been previously reported that the addition of frontal EEG increases the detection of respiratory-related arousals.[15] In addition, it is yet to be examined whether using multiple compared to a single EEG derivation would lead to any differences in intra- or inter scorer reliability in arousal scoring.

The primary aim of this study was to assess the impact of using a montage including 3 EEG derivations (frontal, central, and occipital) *vs*. a montage including 1 central EEG derivation on PSG sleep and arousal scoring summary statistics, in a cohort of patients presenting for diagnosis or exclusion of obstructive sleep apnea. A secondary aim was to assess the impact of using 3 *vs*. 1 EEG derivations on both the intra- and inter-scorer reliability of sleep and arousal scoring.

METHODS

Patient Selection

This study utilized 30 single-night PSGs sourced during December 2007 and January 2008 from the Royal Adelaide Hospital sleep laboratory in Adelaide, South Australia, from consecutive patients being investigated for clinically suspected OSA. PSGs were not considered if they were being primarily conducted for investigation of other sleep disorders, if they were for research purposes, or if they involved implementation or review of treatment.

PSG Recordings

PSGs were recorded using Compumedics E-series monitoring equipment (Abbotsford, Victoria, Australia), according to the recommendations of the AASM Manual.[1] For EEG, electrodes were placed according to the manual's "recommended" rather than "alternative" derivations. Specifically, the following EEG derivations were recorded: F4/M1, C4/M1, O2/M1, as well as back-up derivations: F3/M2, C3/M2, and O1/M2. The recording configuration also consisted of left and right EOG (alternative AASM placement), ECG, chin EMG, nasal pressure, thermistor, body position, thoracic and abdominal excursion (inductance plethysmography), oxygen saturation via finger pulse oximetry (MasimoSET Radical; Masimo, Irvine, CA), left and right leg movement (piezoelectric sensors), and sound (in-room decibel meter).

PSG Scoring

All PSG scoring was performed manually using Profusion PSG 2 software (Compumedics, Abbotsford, Victoria, Australia), following the recommendations outlined in the 2007 AASM manual,[1] with one exception for arousal scoring outlined below. Sleep and arousal scoring occurred in a single pass of the recorded data. PSGs were either configured to display a montage containing 3 EEG derivations (M_{3EEG}) including F4/M1, C4/M1, and O2/M1, or a montage containing 1 EEG derivation (M_{1EEG}) at C4/M1. Care was taken to ensure that the display size of all EEG channels was identical regardless of the number of channels displayed, and use of back-up channels was permitted for sections of poor recording quality. The exception to the arousal scoring recommendation in the AASM manual,[1] which states that arousals scoring should incorporate information from central and occipital EEG derivations, was that an arousal could be scored from any of the EEG channels displayed. Additionally, scorers were instructed to mark arousal length accurately and to mark arousals to wakefulness as a 15-second event. To allow characterization of the patient sample, respiratory events were marked in a separate pass of the data, using the 2007 manual's alternative hypopnea definition, which is important to note given the impact hypopnea definition may have on AHI.[16]

Scorers

Three scorers, from 3 separate Australian clinical sleep laboratories participated in this study; one from the Royal Adelaide Hospital, Adelaide, one from the Austin Hospital, Melbourne, and one from Monash Medical Centre, Melbourne. All scorers participated in intra and inter-laboratory scoring concordance programs and were of varying experience (2 with > 10 y experience; 1 with approximately 1 y experience). Prior to commencement of the study, 2 PSGs were scored by each scorer, and were subjected to epoch-by-epoch and event-by-event inter-scorer reliability analysis. The results of this analysis were presented to the scorers and were used to help identify and correct major discrepancies in scoring interpretation.

Protocol

Each scorer analyzed PSGs 1-10 four times each, twice using M_{3EEG} and twice using M_{1EEG}. In addition scorer 2 analyzed PSGs 11-20 twice, once using each method, and scorer 3 analyzed PSGs 21-30 twice, once using each method. Thus overall, scorer 1 performed a total of 40 scorings, and scorers 2 and 3 performed 60 each. For each scorer all PSGs and versions were de-identified and presented in random order to eliminate any order effect, with the exception that no 2 versions of the same PSG were ever presented consecutively. In addition, to avoid study recognition, scorers were instructed not to score more than 5 of these PSGs per week. The time taken to score each PSG was recorded by each scorer.

This study was approved by the Austin Health Human Research Ethics Committee.

Analysis

All 30 PSGs were used in examining the impact of EEG derivations on PSG summary statistics. For this analysis, each scorer scored a unique set of 10 PSGs twice, once using M_{3EEG} and once using M_{1EEG}. Although in total, PSGs 1-10 were scored 4 times by each scorer to allow assessment of scoring reliability, summary statistics analysis only utilized scorings undertaken by scorer 1, and only 1 version of the 2 available using each method; the version used was determined prior to randomization.

Sleep and arousal scoring summary statistic differences between M_{3EEG} *vs.* M_{1EEG} were tested for significance using paired-sample *t*-tests. The net sleep stage specific changes in sleep scoring were also reported to improve understanding of differences in sleep summary statistics.

PSGs 1-10 were used in assessment of inter- and intra-scorer reliability. Epoch-by-epoch inter-scorer reliability for sleep scoring was assessed using Fleiss' multi-scorer kappa.[17,18] Event-by-event inter-scorer reliability for arousals was also assessed using multi-scorer kappa modified for continuous measurements.[19] The modification utilizes the proportions of time spent in agreement and disagreement in the continuous time series rather than examining the presence or absence of events in an arbitrarily defined interval.[19] Raw agreement, expressed as percentage agreement[17] for sleep and as proportion of specific agreement for positive ratings (PSA)[18] for arousals, was also presented for comparison. The calculation of PSA takes into account measurement of agreement on the presence but not the absence of an event, and events were considered to match if they overlapped at any time. PSA is interpreted as the probability that if a randomly chosen scorer detects an event, a second randomly chosen scorer would also detect the event.[18]

Inter-scorer reliability differences between M_{3EEG} *vs.* M_{1EEG} were compared using paired-sample *t*-tests. The 2 versions using each method were averaged prior to analysis.

Epoch-by-epoch intra-scorer reliability of sleep scoring was assessed using Cohen's pairwise κ.[20] Event-by-event intra-scorer reliability of arousal scoring was also assessed using pairwise κ modified for continuous measurements as described above. Raw agreement was also presented for comparison as described above.

Intra-scorer reliability differences between M_{3EEG} *vs.* M_{1EEG}, were tested for significance using a general linear model. As this analysis was based on the same set of 10 PSGs scored by all 3 scorers using both methods, *Scorer* and *PSG* were specified in the model as additional explanatory variables, ensuring that these potential sources of variation were correctly accounted for in the analysis.

Inter-scorer reliability of sleep and arousal summary statistics was assessed by determining the maximal absolute difference between the 3 scorers (i.e., the range) and intra-scorer reliability of sleep and arousal summary statistics was assessed by determining the absolute difference between paired scorings. As it was not possible to assume normality or find appropriate transformations for all absolute difference outcomes, differences between M_{3EEG} *vs.* M_{1EEG} were tested for statistical significance using the Wilcoxon signed ranks test.

The difference in time taken to score PSGs between all scorings of M_{3EEG} *vs.* M_{1EEG}, was tested for significance using a paired-sample *t*-test.

Data were transformed to satisfy distributional assumptions of normality prior to analysis where appropriate, and results are expressed as mean ± standard error unless otherwise stated. P values < 0.05 were accepted as statistically significant.

Table 1—Summary of patient characteristics for the patients studied in the PSG summary statistics comparison and the subset of patients studied in the inter- and intra-scorer reliability comparison

	Comparison	
Parameter	PSG summary statistics	Reliability
n	30	10
Age	51 (37, 63)	52 (43, 63)
Gender M/F	19/11	7/3
BMI (kg/m^2)	34.3 (30.9, 38.1)	34.6 (31.2, 40.2)
ESS	9.0 (7.0, 11.0)	9.0 (7.3, 12.0)
AHI(/h)	15.8 (5.7, 52.1)	17.3 (10.2, 46.4)
TDT (min)	431 (414, 471)	416 (410, 436)

Values are median (interquartile range). PSG, polysomnography; BMI, body mass index; ESS, Epworth Sleepiness Scale; AHI, apnea-hypopnea index; TDT, total dark time. AHI derived using AASM alternative hypopnea definition and using an average of all available scorings.

RESULTS

Patient Characteristics

Patient characteristics for all 30 patients studied, and for the subset of 10 patients studied in the inter- and intra-scorer reliability comparison, are shown in Table 1.

PSG Summary Statistics

When scoring with M_{3EEG} *vs.* M_{1EEG} there was a statistically significant reduction in stage N1 sleep and a significant increase in stage N3 sleep (Table 2). No other statistically significant differences were found for any other sleep or arousal scoring summary statistics. For individual sleep stages, the same pattern of results was observed if they were expressed as total time (minutes) or as a percentage of total sleep time spent in the particular sleep stage.

When examining the distribution of sleep stage specific changes when using M_{3EEG} *vs.* M_{1EEG} (Table 3), the most noteworthy changes were a net increase in N2 of 9.7 ± 3.0 min at the expense of N1 and net increase in N3 of 10.0 ± 2.7 min at the expense of N2.

Inter-Scorer Reliability

There were no statistically significant differences observed in epoch-by-epoch sleep scoring inter-scorer reliability when scoring with M_{3EEG} *vs.* M_{1EEG}, despite a trend to higher mean values of Fleiss' κ when scoring with M_{3EEG} for sleep scoring overall, and for specific sleep stages N2, N3, and W (Table 4; Figure 1). For sleep scoring overall, the equivalent raw per-

Table 2—Sleep and arousal scoring summary statistics derived using M_{3EEG} *vs.* M_{1EEG} (n = 30 observations; 10 per scorer)

Parameter	Montage		Difference		
	M_{3EEG}	M_{1EEG}	$M_{3EEG} - M_{1EEG}$	95% CI	P-Value
Total sleep time (min)	326.1 (11.9)	323.1 (12.4)	3.0 (3.3)	-3.8, 9.8	0.369
Sleep efficiency (%)	75.1 (2.4)	74.4 (2.5)	0.7 (0.8)	-0.8, 2.3	0.340
Sleep latency (min)*	26.3 (3.7)	27.8 (3.9)	-1.5 (1.3)	-4.2, 1.2	0.368
Stage R latency (min)*	136.2 (13.1)	146.6 (16.6)	-10.4 (15.8)	-43.1, 22.2	0.662
Wake after sleep onset (min)	82.2 (9.1)	83.7 (9.4)	-1.5 (2.8)	-7.3, 4.2	0.590
Time in each sleep stage (min)					
N1	70.7 (11.6)	80.3 (12.7)	-9.6 (3.9)	-17.5, -1.7	0.018
N2	146.9 (10.5)	145.0 (10.5)	1.9 (3.8)	-5.9, 9.6	0.623
N3	63.5 (6.3)	52.8 (6.2)	10.6 (2.8)	4.8, 16.4	0.001
NR - Total	281.0 (10.4)	278.1 (11.2)	2.9 (3.6)	-4.4, 10.1	0.426
R	45.1 (4.5)	45.0 (4.6)	0.2 (1.6)	-3.3, 3.5	0.919
Percent of TST in each sleep stage					
N1	21.5 (3.2)	24.6 (3.6)	-3.2 (1.1)	-5.5, -0.8	0.010
N2	45.3 (2.8)	45.1 (2.7)	0.2 (1.1)	-2.1, 2.5	0.868
N3	19.8 (1.9)	16.6 (1.9)	3.1 (0.9)	1.3, 5.0	0.002
NR - Total	86.5 (1.3)	86.3 (1.4)	0.2 (0.5)	-0.8, 1.2	0.733
R	13.5 (1.3)	13.7 (1.4)	-0.2 (0.5)	-1.2, 0.8	0.773
Arousal index (/h)	26.6 (3.3)	28.3 (3.6)	-1.8 (1.2)	-4.2, 0.6	0.139
Arousal count	144.9 (19.9)	153.5 (20.5)	-8.6 (6.2)	-21.3, 4.0	0.174

All data reported as mean (standard error) on original scale. M_{3EEG}: Montage including 3 EEG derivations, from frontal, central, occipital regions; M_{1EEG}: Montage including 1 central EEG derivation only; 95%CI, 95% confidence intervals; R, REM sleep; NR, NREM sleep; N1, stage 1 sleep; N2, stage 2 sleep; N3, stage 3 sleep. *Data log transformed to normalize distribution prior to analysis.

Table 3—Net sleep stage specific changes in sleep scoring (minutes) when using M_{3EEG} *vs.* M_{1EEG} (n = 30 PSGs)

		M_{3EEG}				
	Stage	W	N1	N2	N3	R
M_{1EEG}	W	—	1.9 (2.5)	0.6 (0.7)	0.2 (0.1)	0.3 (0.4)
	N1		—	9.7 (3.0)	0.5 (0.2)	1.4 (0.9)
	N2			—	10.0 (2.7)	-1.5 (1.0)
	N3				—	0.0 (0.0)
	R					—

All data reported as mean (standard error). The net 10-minute N2 - N3 change represents a net increase in N3 when scoring with M_{3EEG} at the expense of N2 when scoring with M_{1EEG}. M_{3EEG}: Montage including 3 EEG derivations, from frontal, central, occipital regions; M_{1EEG}: Montage including 1 central EEG derivation only; W, Wake; N1, stage 1 sleep; N2, stage 2 sleep; N3, stage 3 sleep; R, REM sleep.

centage agreement was 77% ± 3% and 76% ± 3% using M_{3EEG} and M_{1EEG}, respectively. Likewise, there was no statistically significant difference in inter-scorer reliability for arousal scoring between the 2 methods (Table 4; Figure 1). For arousal scoring raw agreement, identical PSA of 0.58 ± 0.03 was found when using both M_{3EEG} and M_{1EEG}.

When examining the inter-scorer reliability of PSG summary statistics using the maximal absolute difference between scorers (range), there was no statistically significant differences using M_{3EEG} *vs.* M_{1EEG} for any of the summary statistics examined (Table 5).

Intra-Scorer Reliability

No statistically significant differences were observed in epoch-by-epoch intra-scorer reliability, measured using Cohen's κ, comparing M_{3EEG} with M_{1EEG}, for sleep scoring overall, or for sleep stages considered separately; nor was there a significant difference observed for event-by-event arousal scoring reliability (Table 6; Figure 2). For sleep scoring overall, the equivalent mean percentage agreement was 83% ± 2% and 83% ± 1% using M_{3EEG} and M_{1EEG}, respectively. For arousal scoring, the PSA was 0.70 ± 0.03 and 0.72 ± 0.03 using M_{3EEG} and M_{1EEG}, respectively.

When examining the intra-scorer reliability of PSG summary statistics using the absolute difference between PSG pairs, there was no statistically significant difference using M_{3EEG} *vs.* M_{1EEG} for any of the summary statistics examined (Table 7).

Time Taken to Score PSGs

Data on time taken to score PSGs were available for 2 of the 3 scorers. There was no statistically significant difference in time taken to score PSGs when using M_{3EEG} *vs.* M_{1EEG} for PSG scoring (P = 0.306); the mean time to score PSGs using M_{3EEG} was 1 h 19 m ± 5 m compared to 1 h 15 m ± 6 m using M_{1EEG}. The mean difference between methods (M_{3EEG} - M_{1EEG}) was 3 minutes (95%CI: −3 m, 11 m).

DISCUSSION

The main findings of this study were that increasing the number of EEG derivations from 1 central derivation to 3 derivations (frontal, central, and occipital), in a cohort of suspected OSA patients, resulted in a small decrease in stage N1 sleep and a small increase in stage N3 sleep. No other significant differences were observed in other sleep or arousal scoring summary statistics; nor were there differences in inter- or intra-scorer scoring reliability.

The decrease in stage N1 sleep with 3 derivations is consistent with observations that K-complexes, which are a key scoring feature for stage N2 sleep, are more prominently observed in frontal regions of the cortex.[3,9] Thus a portion of epochs scored as stage N1 with a single central derivation are scored as stage N2 with montages that include a frontal derivation. The increase in stage N3 sleep is also consistent with observations that slow waves, a key feature of stage N3 sleep, are also more prominent in frontal regions of the cortex[9,10]; thus shifting epochs of stage N2 sleep scored with a single EEG derivation to stage N3 with multiple derivations. The lack of difference in stage N2 sleep using the two methods is then explained by the offsetting effect of these two subtle changes.

Moser et al.[21] recently compared sleep classification according to the newer AASM standard,[1] utilizing multiple EEG derivations, compared to R&K criteria,[2] utilizing only a central EEG derivation. Although both the present study and that of Moser et al.[21] found an increase in stage N3 sleep when scoring with multiple EEG derivations, there were some notable differences in the design and the results between that study and the present study. That study found an increase in stage N1 when scoring with multiple EEG derivations[21] using the new AASM standard rather than a decrease as found in the present study; and that study also found a decrease in stage N2 and an increase in WASO,[21] whereas the present study found no differences. Moser et al.,[21] however, did not distinguish differences related to rule changes from those related to EEG derivation differences. Indeed, the increase in N1 was thought to be explained by the scoring rule[1] which states a period of stage N2 is terminated by a cortical arousal.[21] The decrease in stage N2 was thought to be due to a combination of this same rule and the increase in slow wave detection with frontal leads[21]; the increase in WASO was suggested to be due to rule changes[1] relating to sleep onset and the scoring of movement time.[21] Scoring rules were not a factor in the present study as all PSGs were scored according to the 2007 AASM manual.[1] Another point of difference worth noting is that the majority of PSGs examined in the study of Moser et al.[21] were not from a clinical population and none were from patients suspected of OSA, whereas all the PSGs analyzed in the present study were from a clinical population being investigated for OSA.

Table 4—Multi-scorer (3 scorers) inter-scorer reliability (epoch-by-epoch / event-by-event multi-scorer κ) for polysomnography (n = 10) derived using M_{3EEG} *vs.* M_{1EEG}

Parameter	Montage		Difference		
	M_{3EEG}	M_{1EEG}	$M_{3EEG} - M_{1EEG}$	95% CI	P-Value
Sleep scoring—overall	0.67 (0.04)	0.65 (0.05)	0.024 (0.020)	-0.020, 0.069	0.246
Sleep scoring—stage specific					
N1	0.40 (0.03)	0.40 (0.04)	-0.001 (0.023)	-0.052, 0.050	0.960
N2	0.61 (0.04)	0.59 (0.04)	0.023 (0.025)	-0.033, 0.079	0.370
N3	0.60 (0.06)	0.50 (0.09)	0.104 (0.048)	-0.005, 0.212	0.060
R	0.88 (0.02)	0.88 (0.02)	-0.002 (0.015)	-0.038, 0.035	0.921
W	0.80 (0.04)	0.77 (0.05)	0.034 (0.020	-0.011, 0.080	0.121
Arousal scoring	0.42 (0.03)	0.41 (0.03)	0.001 (0.015)	-0.033, 0.035	0.948

Data are represented as mean (standard error). M_{3EEG}: Montage including 3 EEG derivations, from frontal, central, occipital regions; M_{1EEG}: Montage including 1 central EEG derivation only; 95%CI, 95% confidence intervals; R, REM sleep; N1, stage 1 sleep; N2, stage 2 sleep; N3, stage 3 sleep; W, wake.

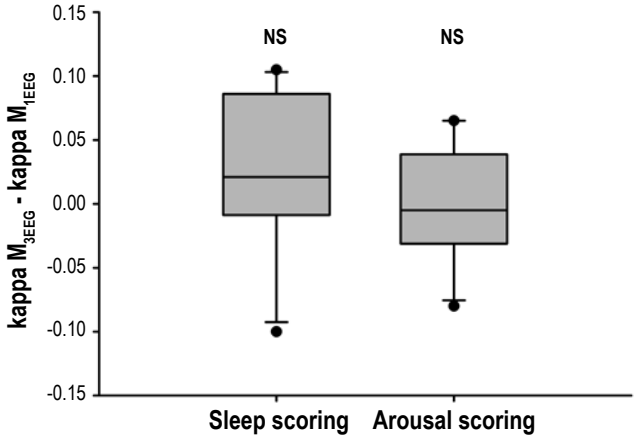

Figure 1—Box plots showing paired multi-scorer (3 scorers) inter-scorer reliability differences, measured using Fleiss' κ, when using M_{3EEG} *vs.* M_{1EEG}, for both sleep and arousal scoring (n = 10 PSGs). M_{3EEG}: Montage including 3 EEG derivations, from frontal, central, occipital regions; M_{1EEG}: Montage including 1 central EEG derivation only. Central line = median; Box boundary = 25th and 75th percentiles; Error bars = 10th and 90th percentiles; ● = outliers. NS = differences between M_{3EEG} and M_{1EEG} not significant at $P < 0.05$.

In the present study we chose to score arousals from all three EEG derivations rather than two derivations (central and occipital) as recommended in the AASM manual,[1] mainly based on previous reports of an increase in respiratory-related arousals with the addition of frontal leads.[15] The evidence review paper underpinning the AASM manual[14] suggested one might expect the number of arousals scored to be greater with an increase in EEG derivations. Interestingly we found no significant difference in arousal index or arousal count when using 3 compared to 1 EEG derivation. Although not statistically different, the present study shows a trend to a reduction in the number of arousals scored when using multiple compared to a single EEG derivation. One might conceive that the added complexity of scoring from multiple channels may lead to arousals being overlooked on one channel when attention is focused on others; alternatively additional channels may have been used to discount a possible arousal in one channel rather than all EEGs being treated independently. However, arousal scoring findings should be considered with the knowledge that arousal scoring reliability, in this and other studies,[22] is generally lower compared to other PSG measures, thus reducing the chances of finding statistically significant differences between methods.[23]

The present study is the first to describe intra-scorer in addition to inter-scorer reliability for sleep and arousal scoring, when comparing multiple to single EEG derivations. In summary, no significant differences were found in inter- or intra-scorer sleep or arousal scoring reliability when using 3 compared to 1 EEG derivations for PSG analysis. These findings are in contrast to the study of Danker-Hopfe et al.[24] which compared inter-scorer sleep scoring reliability of the 2007 AASM criteria,[1] using multiple referential EEG derivations, to R&K criteria using only a central EEG derivation. That study found a small but significant ($P = 0.02$) increase in pairwise inter-scorer reliability measured using Cohen's κ when scoring according to the AASM standard (mean ± SD: 0.75 ± 0.11) compared to R&K (0.72 ± 0.10); a finding mainly related to an increase in inter-scorer reliability for scoring of stage R, wake, and N1.[24] In the present study, although similarly small improvements were observed for multi-scorer inter-scorer sleep scoring reliability, measured

Table 5—Inter-scorer reliability (maximal difference/range) of PSG sleep and arousal scoring summary statistics derived using M_{3EEG} *vs.* M_{1EEG} (n = 10 PSGs)

	Montage		Difference	
Parameter	**M_{3EEG}**	**M_{1EEG}**	**$M_{3EEG} - M_{1EEG}$**	**P-Value**
Total sleep time (min)	15.6 (10.4, 38.6)	19.5 (5.1, 46.9)	-1.4 (-24.7, 8.8)	0.445
Sleep efficiency (%)	3.7 (2.5, 8.7)	4.6 (1.3, 13.3)	-0.4 (-5.8, 2.2)	0.333
Sleep latency (min)	6.8 (0.0, 9.3)	5.9 (0.9, 7.5)	0.4 (-1.3, 2.9)	0.474
Stage R latency (min)	9.3 (1.5, 75.8)	3.5 (1.8, 7.5)	1.8 (-0.8, 6.8)	0.150
Wake after sleep onset (min)	10.8 (6.3, 35.3)	15.3 (3.9, 42.1)	-2.8 (-21.9, 7.3)	0.333
Time in each stage (min)				
N1	29.0 (14.0, 63.5)	27.3 (15.9, 54.6)	2.8 (-10.4, 19.9)	0.575
N2	22.1 (19.3, 51.1)	23.4 (19.6, 61.4)	-5.0 (-18.9, 2.9)	0.153
N3	22.8 (11.9, 27.3)	18.4 (7.6, 37.2)	2.5 (-12.6, 6.6)	0.878
NR - Total	19.3 (11.6, 28.3)	24.4 (6.1, 45.6)	-5.6 (-19.4, 11.2)	0.575
R	4.3 (0.0, 13.8)	5.3 (0.0, 11.9)	0.0 (-5.1, 2.9)	0.612
Arousal index (/h)	5.5 (3.6, 9.0)	6.6 (4.1, 16.0)	-0.8 (-4.5, 0.8)	0.308
Arousal count	24.5 (10.0, 48.4)	31.3 (13.6, 31.3)	-6.5 (-15.3, 4.4)	0.374

All data reported as median (inter-quartile range). M_{3EEG}: Montage including 3 EEG derivations, from frontal, central, occipital regions; M_{1EEG}: Montage including 1 central EEG derivation only; R, REM sleep; NR, NREM sleep; N1, stage 1 sleep; N2, stage 2 sleep; N3, stage 3 sleep.

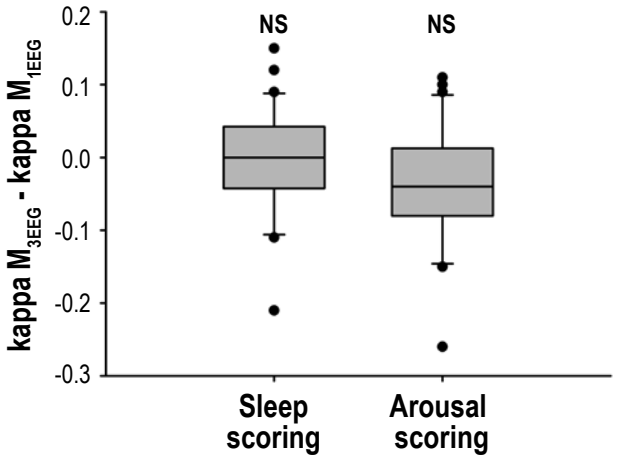

Figure 2—Box plots showing paired intra-scorer reliability differences, measured using Cohen's κ, when using M_{3EEG} *vs.* M_{1EEG}, for both sleep and arousal scoring (n = 30 observations; 10 PSGs x 3 scorers). M_{3EEG}: Montage including 3 EEG derivations, from frontal, central, occipital regions; M_{1EEG}: Montage including 1 central EEG derivation only. Central line = median; Box boundary = 25th and 75th percentiles; Error bars = 10th and 90th percentiles; ● = outliers. NS = differences between M_{3EEG} and M_{1EEG} not significant at $P < 0.05$.

with Fleiss' κ, using 3 EEG derivations compared to 1 EEG, these differences were not statistically significant. Additionally, no statistically significant differences were found when considering sleep stages separately.

There are a number of possible explanations for the differences in findings between the present study and that of Danker-Hopfe et al.[24]: (1) The present study examined inter-scorer reliability on a relatively small number of PSGs and may be under-powered to find differences. Despite this it appears safe to say that any differences between methods are not large and/or consistent; (2) The differences found in the Danker-Hopfe[24] study may be related to rule differences between AASM and R&K criteria rather than differences in EEG derivations; (3) The Danker-Hopfe study[24] examined a similar data set to that of Moser[21] discussed previously; thus, in contrast to the present study, the majority of the PSGs were not from a clinical population; (4) As discussed by the authors the AASM sleep scoring reliability may be inflated in the Danker-Hopfe study[24] by a smaller number of experts used to score the AASM PSGs compared to the R&K PSGs, and the fact that the AASM scoring was conducted after a training symposium but the R&K scoring was not; (5) There may be an order effect in the Danker-Hopfe study,[24] with all the AASM PSG scoring occurring approximately 6 years later than the R&K PSG scoring; in the present study PSGs were scored in random order.

A benefit in arousal scoring reliability might be expected due to the localization of the alpha rhythm over the occipital cortex.[11] In the present study, there was no indication of an improvement in reliability when using multiple EEGs compared to a single EEG; in fact there was a trend observed for a reduction in intra-scorer reliability when using multiple EEG derivations. As discussed earlier, the added complexity of scoring from multiple channels may negate any advantage of scoring PSG with occipital EEG derivations.

A novel aspect of this study is the methodology utilized to examine arousal scoring reliability, where we used κ to examine event-by-event reliability in a continuous time series. It is based on a method mentioned but not utilized by Ayappa et al.[25] in examining scoring reliability of respiratory-effort related arousals and originally described by Conger.[19] Traditionally, studies have not examined event-by-event arousal scoring reliability in a continuous time series but have examined reliability of arousal indices[22,26-28] or have examined event-by-event reliability on pre-selected sections of record.[29,30] Although the κ statistic for arousal scoring has the theoretical advantage over raw agreement of incorporating chance agreement in the assessment of reliability,[18,20] another difference between κ and the raw agreement index (PSA) presented in the present study for arousal scoring must be noted. In contrast to κ, PSA calculated in the present study is independent of scored event length or the degree of overlap between events; events scored by different scorers or methods are considered to agree if they overlap at any time. This is likely to inflate any differences observed between kappa and raw agreement, and it is for this reason that scorers were specifically instructed to mark event length accurately in the present study.

We were unable to ascertain from the present study whether using the AASM manual's alternative EEG placement, which includes bipolar derivations, would produce different results. This is plausible, given that sleep spindles, for example, have shown differences in their topographic distribution between referential and bipolar recordings.[10] However, a study by van

Sweden et al.[12] suggests the results may be similar. That study evaluated a two-channel bipolar montage compared to a single central R&K montage and found a tendency to score more deep sleep with the bipolar montage, and little difference in sleep scoring inter-rater reliability. Nevertheless, studies similar to ours directly comparing the AASM recommended and alternative recordings are warranted so that future standards can be evidence based, and able to furnish a single recommendation for electrode placement in PSG. Replication of the present study in normal and other clinical populations would also be beneficial to increase the evidence base on which future recommendations for electrode number and placements are made. Additionally, studies larger than ours are required to examine whether factors such as gender, age, or OSA severity may influence the impact of EEG derivations on PSG summary statistics and scoring reliability. For example, aging may impact the findings due to the different distribution of sleep stages observed in elderly compared to younger populations.[31] However the studies of Moser[21] and Danker-Hopfe[24] suggest the impact of age and gender may be small. When comparing the AASM (multiple EEG derivations) vs. R&K (single EEG derivation) standards, age effects were only observed on the amount of REM sleep scored[21] and the reliability of scoring stage 1 sleep[24]; no other age or gender effects were observed.[21,24]

A limitation that must be recognized in the present study is that scorers could not be blinded to scoring method and therefore could be subject to bias. We tried to limit this by using multiple scorers from different sleep centers, thus there was no discussion of expectations between scorers. Furthermore, the observation that there was no difference in time taken to score using either method, suggested that scorers were not being more diligent using one method over the other. Using scorers from different sleep centers also had the advantage of increasing the generalizability of the present study to clinical practice.

The fact that using additional EEG derivations in PSG scoring results in only small changes in sleep scoring summary statistics, and no improvement in scoring reliability, brings into question the need to use multiple EEG derivations for PSG in a clinical setting. In fact when commenting on the 2007 *AASM Manual for the Scoring of Sleep and Associated Events*,[1] the Italian Association of Sleep Medicine[32] suggested that without significant improvement there is no reason to justify the, "technical, economic and scientific sacrifices" required to adopt the new recommendations. Our study suggests however, that at least there is no extra burden in the time required to analyze PSGs when using additional EEG derivations.

Table 6—Intra-scorer reliability (epoch-by-epoch / event-by-event pairwise κ) of PSG analyzed using M_{3EEG} *vs.* M_{1EEG} (n = 30 observations; 10 PSGs x 3 scorers)

	Montage		Difference		
Parameter	**M_{3EEG}**	**M_{1EEG}**	**$M_{3EEG} - M_{1EEG}$**	**95% CI**	**P-Value**
Sleep scoring—overall*	0.75 (0.02)	0.75 (0.02)	-0.004 (0.013)	-0.031, 0.022	0.691
Sleep scoring—stage specific					
N1	0.54 (0.03)	0.56 (0.03)	-0.015 (0.023)	-0.062, 0.033	0.615
N2*	0.70 (0.03)	0.71 (0.02)	-0.010 (0.017)	-0.045, 0.026	0.440
N3*	0.70 (0.03)	0.66 (0.05)	0.043 (0.050)	-0.059, 0.145	0.507
R	0.90 (0.01)	0.92 (0.01)	-0.018 (0.013)	-0.044, 0.010	0.219
W*	0.86 (0.02)	0.85 (0.02)	0.009 (0.009)	-0.010, 0.028	0.506
Arousal scoring†	0.56 (0.03)	0.60 (0.02)	-0.037 (0.014)	-0.066, -0.008	0.082

All data reported as mean (standard error) on original scale. M_{3EEG}: Montage including 3 EEG derivations, from frontal, central, occipital regions; M_{1EEG}: Montage including 1 central EEG derivation only; 95%CI, 95% confidence intervals; R, REM sleep; N1, stage 1 sleep; N2, stage 2 sleep; N3, stage 3 sleep; W, wake. *Parameter raised to the power of 3 to normalize distribution of differences prior to analysis; †Parameter raised to the power of 2 to normalize distribution of differences prior to analysis.

Table 7—Intra-scorer reliability (pair-wise absolute difference) of PSG sleep and arousal scoring summary statistics derived using M_{3EEG} *vs.* M_{1EEG} (n = 30 observations; 10 PSGs x 3 scorers)

	Montage		Difference	
Parameter	**M_{3EEG}**	**M_{1EEG}**	**$M_{3EEG} - M_{1EEG}$**	**P-Value**
Total sleep time (min)	5.0 (1.9, 15.5)	5.0 (2.0, 10.5)	0.3 (-1.8, 7.1)	0.301
Sleep efficiency (%)	1.3 (0.5, 3.8)	1.3 (0.5, 2.9)	0.1 (-4.3, 1.8)	0.302
Sleep latency (min)	0.5 (0.0, 7.5)	1.0 (0.4, 3.0)	0.0 (-0.6, 5.4)	0.438
Stage R latency (min)	1.5 (0.5, 5.0)	1.0 (0.5, 3.0)	0.5 (-1.0, 3.3)	0.613
Wake after sleep onset (min)	4.5 (1.9, 12.6)	4.5 (1.0, 9.6)	0.5 (-2.4, 7.1)	0.656
Time in each stage (min)				
N1	7.8 (2.8, 17.6)	7.0 (3.0, 15.1)	0.0 (-2.3, 5.1)	0.700
N2	15.3 (6.5, 22.3)	11.3 (2.9, 19.3)	2.3 (-6.9, 9.0)	0.600
N3	11.3 (5.9, 20.9)	7.3 (2.4, 18.4)	2.5 (-5.1, 10.5)	0.150
NR - Total	6.5 (2.4, 13.5)	6.8 (2.5, 12.5)	0.8 (-2.5, 6.6)	0.338
R	1.5 (0.5, 5.0)	1.0 (0.0, 4.5)	0.0 (-0.4, 3.0)	0.275
Arousal index (/h)	3.1 (1.3, 6.6)	2.8 (1.4, 5.7)	-0.3 (-2.5, 2.7)	0.854
Arousal count	11.5 (7.0, 33.8)	19.0 (6.8, 40.5)	-5.0 (-18.3, 8.3)	0.252

All data reported as median (inter-quartile range); Abbreviations: M_{3EEG}: Montage including 3 EEG derivations, from frontal, central, occipital regions; M_{1EEG}: Montage including 1 central EEG derivation only; R, REM sleep; NR, NREM sleep; N1, stage 1 sleep; N2, stage 2 sleep; N3, stage 3 sleep.

CONCLUSION

The main findings of this study were that using three EEG derivations during PSG, as recommended in the *AASM Manual for the Scoring of Sleep and Associated Events*,[1] instead of a single central EEG derivation, resulted in a small decrease in N1 sleep and small increase in N3 sleep, in a cohort of patients being investigated for OSA. No other significant differences were observed for any other sleep or arousal scoring summary statistics; nor were any differences observed

in inter-scorer and intra-scorer reliability for scoring sleep or cortical arousals. This information is valuable for those changing practice to comply with the AASM recommendations, for those using portable devices that are unable to comply with the recommendations due to limited channel options, and for development of future standards for PSG scoring and recording. Given that the use of multiple EEG derivations resulted in small changes in the distribution of derived sleep stages, and no significant difference in scoring reliability, this study implies there is no significant advantage in using multiple over single EEG derivations for scoring sleep and arousals in suspected OSA patients undergoing clinical polysomnography.

ABBREVIATIONS

AASM, American Academy of Sleep Medicine;
EEG, electroencephalogram;
PSG, Polysomnography;
OSA, Obstructive sleep Apnoea;
AHI, Apnoea Hypopnoea Index;
ArI, Arousal Index;
TST, Total sleep time;
R&K, Rechtschaffen & Kales;
SE, Sleep efficiency;
SL, Sleep latency;
RL, REM latency;
EOG, electrooculogram;
ECG, electrocardiogram;
EMG, electromyogram;
M_{3EEG}: Montage including 3 EEG derivations, from frontal, central, occipital regions;
M_{1EEG}: Montage including 1 central EEG derivation only;
PSA, proportion of specific agreement;
N1, stage 1 sleep;
N2, stage 2 sleep;
N3, stage 3 sleep;
R, rapid eye movement sleep;
NR, non-rapid eye movement sleep;
W, wake;
ESS, Epworth sleepiness scale

ACKNOWLEDGMENTS

We would like to thank Marnie L. Collins for assistance with statistical analysis.

Institutions at which work was performed: Institute for Breathing and Sleep, Austin Health, Heidelberg, Victoria, Australia; Royal Adelaide Hospital, Adelaide, South Australia, Australia; Monash Medical Centre, Clayton, Victoria, Australia.

Financial Support: This study was supported by research grants from the Institute for Breathing and Sleep and the National Health and Medical Research Council of Australia (grant numbers 430300 and 430302).

DISCLOSURE STATEMENT

This was not an industry supported study. Mr. Ruehland has received research support from ResMed, Fisher and Paykel Healthcare, and Philips Respironics. Dr. O'Donoghue has participated in research studies sponsored by Sanofi-Aventis, Boehringer Ingelheim, GlaxoSmith Kline, Bayer and Novartis, and has received research support from ResMed and Phillips Respironics. Dr. Thornton has participated in research studies sponsored by Actelion. Mr. Rochford has received research support from Compumedics. The other authors have indicated no financial conflicts of interest. Prof. Pierce is deceased.

REFERENCES

1. Iber C, Ancoli-Israel S, Chesson A, Quan S, for the American Academy of Sleep Medicine. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. 1st ed. Westchester, IL: American Academy of Sleep Medicine, 2007.
2. Rechtschaffen A, Kales A. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Los Angeles: Brain Information Service, Brain Research Institute, UCLA, 1968.
3. McCormick L, Nielsen T, Nicolas A, Ptito M, Montplaisir J. Topographical distribution of spindles and K-complexes in normal subjects. Sleep 1997;20:939-41.
4. Himanen SL, Hasan J. Limitations of Rechtschaffen and Kales. Sleep Med Rev 2000;4:149-67.
5. Silber MH, Ancoli-Israel S, Bonnet MH, et al. The visual scoring of sleep in adults. J Clin Sleep Med 2007;3:121-31.
6. De Gennaro L, Ferrara M, Bertini M. Topographical distribution of spindles: variations between and within NREM sleep cycles. Sleep Res Online 2000;3:155-60.
7. De Gennaro L, Ferrara M, Bertini M. Effect of slow-wave sleep deprivation on topographical distribution of spindles. Behav Brain Res 2000;116:55-9.
8. Zeitlhofer J, Gruber G, Anderer P, Asenbaum S, Schimicek P, Saletu B. Topographic distribution of sleep spindles in young healthy subjects. J Sleep Res 1997;6:149-55.
9. Happe S, Anderer P, Gruber G, Klosch G, Saletu B, Zeitlhofer J. Scalp topography of the spontaneous K-complex and of delta-waves in human sleep. Brain Topogr 2002;15:43-9.
10. Werth E, Achermann P, Borbely AA. Fronto-occipital EEG power gradients in human sleep. J Sleep Res 1997;6:102-12.
11. Adrian ED, Matthews BHC. The Berger rhythm potential changes from the occipital lobes in man. Brain 1934;57:355-85.
12. van Sweden B, Kemp B, Kamphuisen HA, Van der Velde EA. Alternative electrode placement in (automatic) sleep scoring (Fpz-Cz/Pz-Oz versus C4-A1). Sleep 1990;13:279-83.
13. EEG arousals: scoring rules and examples: a preliminary report from the Sleep Disorders Atlas Task Force of the American Sleep Disorders Association. Sleep 1992;15:173-84.
14. Bonnet MH, Doghramji K, Roehrs T, et al. The scoring of arousal in sleep: reliability, validity, and alternatives. J Clin Sleep Med 2007;3:133-45.
15. O'Malley EB, Norman RG, Farkas D, Rapoport DM, Walsleben JA. The addition of frontal EEG leads improves detection of cortical arousal following obstructive respiratory events. Sleep 2003;26:435-9.
16. Ruehland WR, Rochford PD, O'Donoghue FJ, Pierce RJ, Singh P, Thornton AT. The new AASM criteria for scoring hypopneas: impact on the apnea hypopnea index. Sleep 2009;32:150-7.
17. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378-82.
18. Fleiss JL. The measurement of interrater agreement. In: Statistical methods for rates and proportions. 2nd ed. New York: John Wiley & Sons, 1981:212-36.
19. Conger AJ. Kappa reliabilities for continuous behaviors and events. Educ Psychol Meas 1985;45:861.
20. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37-46.
21. Moser D, Anderer P, Gruber G, et al. Sleep classification according to AASM and Rechtschaffen & Kales: effects on sleep scoring parameters. Sleep 2009;32:139.
22. Whitney CW, Gottlieb DJ, Redline S, et al. Reliability of scoring respiratory disturbance indices and sleep staging. Sleep 1998;21:749-57.
23. Stepnowsky CJ, Jr., Berry C, Dimsdale JE. The effect of measurement unreliability on sleep and respiratory variables. Sleep 2004;27:990-5.

24. Danker-Hopfe H, Anderer P, Zeitlhofer J, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. J Sleep Res 2009;18:74-84.
25. Ayappa I, Norman RG, Krieger AC, Rosen A, O'Malley RL, Rapoport DM. Non-invasive detection of respiratory effort-related arousals (RERAs) by a nasal cannula/pressure transducer system. Sleep 2000;23:763-71.
26. Loredo JS, Clausen JL, Ancoli-Israel S, Dimsdale JE. Night-to-night arousal variability and interscorer reliability of arousal measurements. Sleep 1999;22:916-20.
27. Wong TK, Galster P, Lau TS, Lutz JM, Marcus CL. Reliability of scoring arousals in normal children and children with obstructive sleep apnea syndrome. Sleep 2004;27:1139-45.
28. Smurra MV, Dury M, Aubert G, Rodenstein DO, Liistro G. Sleep fragmentation: comparison of two definitions of short arousals during sleep in OSAS patients. Eur Respir J 2001;17:723-7.
29. Drinnan MJ, Murray A, Griffiths CJ, Gibson GJ. Interobserver variability in recognizing arousal in respiratory sleep disorders. Am J Respir Crit Care Med 1998;158:358-62.
30. Crowell DH, Kulp TD, Kapuniai LE, et al. Infant polysomnography: reliability and validity of infant arousal assessment. J Clin Neurophysiol 2002;19:469.
31. Ohayon MM, Carskadon MA, Guilleminault C, Vitiello MV. Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: Developing normative sleep values across the human lifespan. Sleep 2004;27:1255-73.
32. Parrino L, Ferri R, Zucconi M, Fanfulla F. Commentary from the Italian Association of Sleep Medicine on the AASM manual for the scoring of sleep and associated events: For debate and discussion. Sleep Med 2009;10:799-808.