Computer-Assisted Sleep Classification according to the Standard of the American Academy of Sleep Medicine: Validation Study of the AASM Version of the Somnolyzer 24 × 7

Peter Anderer^{a, d} Arnaud Moreau^d Michael Woertz^d Marco Ross^d
Georg Gruber^d Silvia Parapatics^d Erna Loretz^d Esther Heller^e
Andrea Schmidt^e Marion Boeck^b Doris Moser^b Gerhard Kloesch^b
Bernd Saletu^a Gerda M. Saletu-Zyhlarz^a Heidi Danker-Hopfe^e Josef Zeitlhofer^b
Georg Dorffner^{c, d}

Departments of ^aPsychiatry and Psychotherapy, ^bNeurology and ^cMedical Cybernetics and Artificial Intelligence, Center for Brain Research, Medical University of Vienna, and ^dThe Siesta Group Schlafanalyse GmbH, Vienna, Austria; ^eDepartment of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Campus Benjamin Franklin (CBF), Berlin, Germany

Key Words

Automated sleep scoring · Semi-automated sleep scoring · Polysomnography · Arousal detection · American Academy of Sleep Medicine standard

Abstract

Background: In 2007, the *AASM Manual for the Scoring of Sleep and Associated Events* was published by the American Academy of Sleep Medicine (AASM). Concerning the visual classification of sleep stages, these new rules are intended to replace the rules by Rechtschaffen and Kales (R&K). **Methods:** We adapted the automatic R&K sleep scoring system Somnolyzer 24 × 7 to comply with the AASM rules and subsequently performed a validation study based on 72 polysomnographies from the Siesta database (56 healthy subjects, 16 patients, 38 females, 34 males, aged 21–86 years). Scoring according to the AASM rules were performed manually by experienced sleep scorers and semi-automatically

by the AASM version of the Somnolyzer. Manual scorings and Somnolyzer reviews were performed independently by at least 2 out of 8 experts from 4 sleep centers. **Results:** In the quality control process, sleep experts corrected 4.8 and 3.7% of the automatically assigned epochs, resulting in a reliability between 2 Somnolyzer-assisted scorings of 99% (Cohen's kappa: 0.99). In contrast, the reliability between the 2 manual scorings was 82% (kappa: 0.76). The agreement between the 2 Somnolyzer-assisted and the 2 visual scorings was between 81% (kappa: 0.75) and 82% (kappa: 0.76). **Conclusion:** The AASM version of the Somnolyzer revealed an agreement between semi-automated and human expert scoring comparable to that published for the R&K version with a validity comparable to that of human experts, but with a reliability close to 1, thereby reducing interrater variability as well as scoring time to a minimum.

Copyright © 2010 S. Karger AG, Basel

Introduction

Up to 2007, the only standard for the classification of sleep that found worldwide acceptance were the rules published by Rechtschaffen and Kales (R&K) in 1968 [1]. Since the presentation of the R&K manual for visual sleep scoring, numerous attempts at a computer-assisted identification of sleep stages have been published [2–12]. In 2005, Anderer et al. [12] presented an automatic sleep classification system (Somnolyzer 24 × 7) based on a large-scale, multicenter study performed in Europe as part of the Siesta project. Five hundred and ninety recordings, split into development and validation samples, were used. The final validation in 286 polysomnographies (PSGs) of 95 healthy subjects and 49 patients with sleep disorders revealed an overall epoch-by-epoch agreement of 80% (Cohen's kappa: 0.72) between the Somnolyzer-assisted and the human expert scoring, as compared with an interrater reliability of 77% (Cohen's kappa: 0.68) between 2 human experts scoring the same dataset. Two Somnolyzer analyses (including a structured quality control by 2 human experts) revealed an interrater reliability close to 1 (Cohen's kappa: 0.99). This study was acknowledged as 1 of only 2 evidence level 1 studies for sleep classification according to R&K by the Digital Task Force Committee of the American Academy of Sleep Medicine (AASM) [13]. First applications of the Somnolyzer in a double-blind, placebo-controlled pharmacological sleep study on the potential interaction of 20 mg paroxetine and 1 mg alprazolam in healthy young volunteers and in a single-blind, placebo-controlled clinical sleep study on the acute effects of 100 mg trazodone in patients with nonorganic insomnia related to somatoform pain disorder confirmed the validity of the classifier for scoring PSGs after treatment with psychotropic drugs [14, 15]. This validity was further confirmed by Svetnik et al. [16] in 164 PSGs of 82 subjects in a clinical trial using zolpidem in a phase advance model of transient insomnia.

In May 2007, the *AASM Manual for the Scoring of Sleep and Associated Events* was published [17]. Concerning the visual classification of sleep stages, these new rules replace the rules by R&K. As pointed out in the accompanying paper by Silber et al. [18], the adaptation of the new rules to digital scoring algorithms will again have to be tested for validity as compared with visual scoring. In the AASM classification, sleep stages S1–S4 are referred to as N1, N2 and N3, with N3 comprising the slow-wave sleep stages S3 and S4. The stage rapid eye movement (REM) is referred to as stage R,

wake as stage W and the stage 'movement time' does not exist any more. According to the AASM manual, a minimum of 3 EEG derivations sampling activity from the frontal, central, and occipital regions have to be recorded. Moreover, the minimal settings for analog bandwidths, sampling rates and amplitude resolutions are given. In addition, the AASM manual defines the start and end of the different sleep stages and describes the characteristics of various sleep/wake-related patterns such as REM, slow eye movements (SEM), alpha rhythm, vertex sharp waves, sleep spindles, K-complexes and slow-wave activity, sawtooth waves or transient muscle activity. One major change in the rules defining the start and end of the different sleep stages concerns the termination of stage N2 due to a cortical arousal. According to the AASM manual, any cortical arousal – whether or not associated with an increase in EMG muscle tone – reflects a change from N2 to N1. Changing back to N2 thereafter requires the reappearance of a K-complex unassociated with an arousal or a sleep spindle [17].

For 72 out of the 295 subjects in the Siesta database, frontal derivations had been recorded in addition to central and occipital leads. Since the sampling rates for these recordings were 200 Hz or higher, these data could be subjected to a re-analysis according to the AASM criteria. During the Siesta project, these PSGs had been visually scored according to R&K by 2 independent scorers out of a pool of 30 sleep experts from 8 European sleep labs. The R&K scoring had been based on 2 EOG, 1 (sub)mental EMG and the central EEG channels. In addition, these PSGs had been analyzed by the R&K version of Somnolyzer 24 × 7, as described by Anderer et al. [12]. For the present study, the 72 PSGs were visually scored by 2 independent scorers according to the AASM rules out of a pool of 7 sleep experts from 3 European sleep labs. AASM scoring was preceded by a 2-day training symposium with detailed discussions and 4 test scorings (not included in the present study) to ensure that all scorers correctly interpreted the new AASM standard.

The results for interrater reliabilities comparing visual expert scorings according to the AASM standard and visual expert scorings according to the R&K standard for these 72 PSGs were presented recently by Danker-Hopfe et al. [19]. The authors determined the degree of agreement on quantitative sleep parameters by means of intraclass correlations as well as by means of Cohen's kappa on the basis of epoch-by-epoch comparisons. Results showed that the overall agreement for scorings ac-

cording to the AASM standard was slightly higher than that for scorings according to the R&K standard. Cohen's kappa was 0.76 for the AASM standard and 0.74 for the R&K standard. While for wake, REM, stage 1 and slow-wave sleep interrater reliability was higher for the AASM than for the R&K standard, for sleep stage 2 it was lower. This decrease in the reliability of visually scored N2 was caused by the new rule according to which cortical arousals with or without a concurrent increase in (sub)mental EMG define the end of N2. As a consequence, N2 reliability is now closely related to arousal reliability. The authors noted that a limitation of the study was that the R&K scorings were performed without prior training [20], while the AASM scorings were preceded by a 2-day training symposium [19]. Since prior training per se may have increased interrater reliability, at least part of the observed increase in Cohen's kappa from 0.74 to 0.76 may have been caused by this training effect. In summary, the results by Danker-Hopfe et al. [19] underlined that the integration of frontal, central, and occipital leads improved interrater reliability, but that this advantage was counteracted by the decrease in interrater reliability due to the new rule according to which any cortical arousal – whether or not associated with an increase in EMG muscle tone – defines the end of N2.

In a second paper, Moser et al. [21] investigated the effects of the scoring standard on sleep parameters in the same dataset. In summary, the new standard showed only a minor influence on total sleep time, sleep efficiency and stage REM, but affected wake after sleep onset as well as the distribution of non-rapid eye movement (NREM) sleep stages (stage 1 and slow-wave sleep increased by approximately 10 min each, stage 2 sleep decreased by approximately 20 min). The authors demonstrated that the actual implications for an individual study depended on slow-wave amplitudes and their anterior-posterior gradient as well as on the density of arousals, spindles and K-complexes and might range from nearly no changes at all to increases in slow-wave sleep of up to 53 min and in stage 1 sleep of up to 44 min, with a maximal decrease in stage 2 sleep of up to 55 min [21].

In the present paper, manual scorings are used as gold standard for the validation of the AASM version of the Somnolyzer 24 × 7. The validation study includes Cohen's kappa, sensitivity, specificity, positive and negative predictive values (PPVs and NPVs) based on epoch-by-epoch comparisons as well as correlation analyses based on sleep parameters.

Subjects and Methods

Subjects

The present study is based on a subsample of the Siesta database [22]. The Siesta database comprises a total of 590 recordings of 295 subjects (198 healthy controls and 97 patients with the ICD-10 diagnoses of nonorganic insomnia related to depression or generalized anxiety disorder, Parkinson's disease, periodic limb movement disorder or sleep apnea). The recording protocol specified at least 16 channels of biosignals: 6 EEG channels with mastoid (M1, M2) as reference (Fp1-M2, C3-M2, O1-M2, Fp2-M1, C4-M1, O2-M1) and an additional EEG channel (M1-M2) for re-referencing; 2 EOG channels; 1 (sub)mental EMG; 1 EMG channel recorded from linked electrodes placed at the musculus tibialis anterior of the left and right leg; ECG and respiratory signals (airflow, movements of the chest wall and abdomen and O₂ saturation of arterial blood). Recording of additional EEG channels was recommended whenever possible. Two out of the 8 sleep labs (Department of Neurology, Medical University of Vienna and Department of Psychiatry and Psychotherapy, Charité – Campus Benjamin Franklin) contributed multichannel recordings (19 EEG channels including F3-M2 and F4-M1). Since sampling rates (200 and 250 Hz for the recordings from Berlin and Vienna, respectively) and filter settings were in agreement with the minimal requirements according to the AASM manual, this subsample of the Siesta database was usable for scoring sleep according to the new AASM rules. All recordings were done digitally. For the exchange of the digital polygraphic sleep data and the documentation of the hypnograms the EDF data format [23] was used.

The subsample used in the present study consisted of PSGs (second night after an adaptation night) of 72 subjects, i.e. 56 healthy controls and 16 patients with the ICD-10 diagnoses of nonorganic insomnia related to generalized anxiety disorder ($n = 5$), Parkinson's disease ($n = 6$), or periodic limb movement disorder ($n = 5$). The mean age of the 38 females and 34 males was 57.7 ± 18.7 years, ranging from 21 to 81 years, with a mean age of patients of 58.5 ± 19.8 years and of healthy controls of 54.9 ± 14.7 years.

Manual Scoring of Sleep Stages

For each of the 72 recordings 2 independent scorings according to the rules of R&K and a consensus scoring were available. The first scoring was always performed by the recording lab, while the labs for the second and the consensus scoring were assigned randomly [20, 24]. For the present AASM validation study, all these recordings were rescored independently by at least 2 experts according to the AASM rules. Seven scorers from 3 labs (Department of Neurology, Medical University of Vienna, The Siesta Group GmbH, Vienna, and Department of Psychiatry and Psychotherapy, Charité – Campus Benjamin Franklin) participated in the AASM scoring task. All 7 scorers have a long-standing experience and practice in sleep scoring. The number of PSGs previously scored by each scorer varied between 350 and more than 7,000. The AASM scoring was based on 2 EOG, 1 (sub)mental EMG and the 6 EEG channels (F4-A1, F3-A2, C4-A1, C3-A2, O2-A1, O1-A2).

Automated and Semi-Automated Scoring of Sleep Stages

All 72 recordings had been scored by the R&K version of the Somnolyzer 24 × 7, including the standard quality control process, as described by Anderer et al. [12]. For the present validation study, the automated scorings obtained by the AASM version of the Somnolyzer were reviewed by 2 sleep experts from 2 labs (Department of Psychiatry and Psychotherapy, Medical University of Vienna and The Siesta Group GmbH, Vienna) who independently performed a structured quality control resulting in the semi-automated scorings. While 1 expert had not been involved in the manual scoring task at all, the other had scored 22 out of the 72 PSGs manually. To facilitate the quality review process, in addition to the raw data the human experts received traces depicting preprocessed data as revealed by the automatic analysis. In detail, delta, spindle and fast beta intensity, the alpha/theta index, REM and SEM density, the EMG tonus as well as artifacts identified at frontal, central and occipital leads were displayed. The experts were advised to carefully check periods with frequent artifacts, all sleep-wake transitions (with particular emphasis on sleep onset) as well as all NREM-REM transitions (with particular emphasis on the first REM phase, which may occasionally be very short and only vaguely manifested).

The following section describes step by step the adaptations made to achieve the AASM version of the Somnolyzer. Moreover, the effect of each step on the classification and on the sleep parameters is evaluated.

Step 1: Considering Additional EEG Leads for Detecting Alpha Activity

According to the AASM manual, alpha rhythm is defined as trains of sinusoidal 8- to 13-Hz activity recorded over the occipital leads. However, during the plenary discussion in course 11 'Advanced Polysomnography 2007: Understanding and Using the New AASM Sleep Scoring Manual' at the Sleep 2007 Meeting in Minneapolis (June 10, 2007) the members of the committee stated that EEG patterns such as alpha waves, slow waves or arousals 'can be taken from any lead – i.e. from the lead in which the pattern is more evident'. Thus, both occipital and central leads were included in the 'raw data quality check' module of the Somnolyzer for artifact identification (see fig. 4 in Anderer et al. [12]) and in the subsequent 'feature extraction' module for defining density, intensity, maximal intensity and frequency of the alpha activity (for a description of the features, see table 2 in Anderer et al. [12]).

Step 2: Considering Additional EEG Leads for Detecting Slow-Wave Activity

Like in the detection of alpha activity, both frontal and central EEG leads were considered in the 'feature extraction' module for defining density and intensity of slow waves (i.e. waves with a frequency of 0.5–2 Hz and peak-to-peak amplitude above 75 µV) as well as density, intensity, maximal intensity and frequency of delta waves (i.e. waves with a frequency of 0.5–2 Hz but without amplitude criteria).

Step 3: Considering the 'Arousal Rule' for the End of N2

In contrast to the R&K rules, any cortical arousal – whether or not associated with an increase in EMG muscle tone – defines the end of stage 2 sleep until a new spindle event or K-complex occurs. If this rule is applied consequently, no distinction between K-complexes unassociated or associated with an arousal is neces-

sary. If a K-complex unassociated with an arousal occurs, it defines the start of N2 according to rule IV.5.A.1.a in the manual. If, however, this K-complex is associated with an arousal, the arousal defines the end of N2 according to rule IV.5.C.b. Thus, such an epoch is correctly assigned by applying the rule for scoring an epoch if 2 or more stages coexist during a single epoch (IV.2.B.3). The appropriate adaptations were included in the smoothing part of the 'classification' module.

The definition of a cortical arousal in the AASM manual closely follows the ASDA criteria of 1992 [25]. In addition to the minimal arousal duration (3 s), the manual also states a minimal duration of the submental EMG increase (1 s). The Siesta arousal detector was optimized on the basis of expert scorings in 12 PSG recordings and required a minimal EMG increase of 3 consecutive 0.5-second epochs for defining an arousal in REM. In addition to this small modification in respect to the minimal duration of a required concurrent EMG increase in REM (i.e. 2 instead of 3 consecutive 0.5-second epochs), the arousal detector in the AASM version of the Somnolyzer was adapted to incorporate the information from central and occipital leads. The automatic algorithm in a first step basically compares the pattern of root mean square amplitudes of 9 EEG frequency bands between a 10-second baseline and a 3-second test window by a series of linear discriminant analyses to identify abrupt and transient changes in EEG frequency with a minimal duration of 3 s separately for each EEG lead. The prior probability of an arousal event is increased in relation to the amount of increases in chin EMG amplitudes, and reduced if the EEG activity in the 10-second baseline window resembles wake EEG and/or a probable sleep spindle is identified in the 3-second test window. In a second step, arousal start and end time points are determined by combining the posterior probabilities of the EEG channels so that the arousal indices (number of arousals per total sleep time) derived from automatic and visual detections are equivalent.

Step 4: Considering the Abolition of the '3-Min Rule'

According to the AASM manual, epochs with low-amplitude mixed frequency activity even without K-complexes or spindles that follow N2 should be scored as stage N2 until a transition to stage W, N3 or R, or an arousal, or a major body movement followed by SEM occurs. That means that there is no more limitation of this 'smoothing rule' by 3 min. The appropriate adaptation was included in the smoothing part of the 'classification' module.

Step 5: Considering the Combination of S3 and S4 to Slow-Wave Sleep N3

The AASM manual combines the R&K sleep stages S3 and S4 to 1 slow-wave sleep stage N3, with the rules for scoring slow-wave sleep remaining unchanged. Thus, to comply with this rule, S3 and S4 are simply added to form N3 in the smoothing part of the 'classification' module.

Step 6: Considering the Rules for Major Body Movements

According to R&K, epochs that are obscured by muscle and movement artifacts in more than half of the epoch and that are immediately preceded and followed by sleep are scored as 'Movement Time (MT)'. MT according to R&K is not counted with either sleep or wake time but scored as a separate category. According to AASM rules, this separate category does not exist any more and epochs previously classified as MT should be scored as W if any

Fig. 1. Stage transition analysis describing the changes from scoring according to Rechtschaffen and Kales (R&K-1968) to the standard of the American Academy of Sleep Medicine (AASM-2007). On average, 1 min of MT changed to W and 0.5 min to one of the sleep stages N1, N2, N3 and R. All epochs scored as wake according to R&K were scored as W according to AASM as well. One minute of S1 changed to W if the occipital lead was considered. 1.5 min of S1 changed to N2 after abolishment of the 3-min rule. 12.5 min of S2 changed to N1 if the new arousal rule was considered and 9.5 min changed to N3 if the frontal leads were included, resulting in a reduction of 22 min from S2 to N2. All epochs scored as S3 or S4 according to R&K were scored as N3 according to AASM. And finally, 1 min of REM changed to N1 and another minute to N2 if the additional EEG leads were included in the analysis. In addition to the averaged changes for the 5 stages in minutes, the percentages of total sleep time in each stage are printed in the lower right corner of each stage box.

signs of awakening can be seen (even if wake comprises less than half of the epoch), otherwise they should be scored as the same stage as the epoch that follows them. The appropriate adaptation was added in the smoothing part of the 'classification' module.

Thus, the following scorings were included in the analysis: 144 manual scorings (2 independent manual scorings for each of the 72 recordings, labeled as 'manual scoring 1' and 'manual scoring 2'), 72 automated scorings (based on the AASM version of the Somnolyzer 24 × 7 and labeled as 'automated scoring') and 144 semi-automated scorings (2 independent manual reviews of the 72 automated scorings, labeled as 'semi-automated scoring 1' and 'semi-automated scoring 2'). Note that the experts who performed manual scoring and/or manual review were randomly assigned to 1st or 2nd scoring/review, so that the scorers were equally assigned to both groups.

Results

In the following section, the effects of the various adaptations on sleep classification and sleep parameters are described in detail (see fig. 1 for a summary):

Step 1: Considering Additional EEG Leads for Detecting Alpha Activity

This modification resulted in changes from S1 to W, specifically for epochs with alpha activity restricted to

the occipital region (see fig. 2 for an example). On average, however, this adaptation affected only 2–3 epochs per recording.

Step 2: Considering Additional EEG Leads for Detecting Slow-Wave Activity

This modification resulted in changes from S2 to N3 for epochs with delta waves just below 75 μ V at central, but above 75 μ V at frontal leads (see fig. 3 for an example with the typical frontal maximum of slow-wave amplitudes). On average, 19 epochs changed from S2 to N3 per recording due to this change. Interestingly, this adaptation also had a (slight) influence on stage REM sleep. On average, 3–4 epochs within an REM sleep phase changed from REM to N1 or N2 (due to frontally detected K-complexes). Further evaluation revealed that the time with NREM intrusions within the REM periods increased, but the number and duration of REM cycles remained unchanged.

Step 3: Considering the 'Arousal Rule' for the End of N2

The adaptations of the smoothing algorithms to comply with the 'arousal rule' resulted in an average of 25 epochs that changed from S2 to N1, since in these epochs

Fig. 2. Example of the effect of considering the occipital leads in sleep scoring (patient with insomnia related to generalized anxiety disorder, female, 44 years, 2nd night). The upper part shows the AASM hypnograms for manual scorings 1 and 2 as well as for semi-automatic scoring 1; the delta, spindle and fast beta intensity; the alpha/theta index; the REM and SEM density; the EMG tonus as well as the artifacts identified in frontal, central and oc-

cipital leads. The lower part shows a 30-second epoch in stage W with 2 EOG derivations (Pos8-M1 and Pos18-M1), a mental EMG derivation and 6 EEG derivations (F4-M1, F3-M2, C4-M1, C3-M2, O2-M1, O1-M2). Note that the alpha activity is restricted to the occipital leads. Thus, according to R&K, this epoch is scored as S1, but according to AASM, it has to be scored as stage W.

the arousals were not followed fast enough by a sleep spindle or a K-complex to keep scoring them as N2 (see fig. 4 for an example). It is evident that for recordings with frequent arousals but low K-complex and spindle densities this effect will be most prominent (up to 80 epochs).

Step 4: Considering the Abolition of the '3-Min Rule'

This modification specifically affected recordings with both low K-complex and low spindle densities. On

average, it affected only 3–4 epochs per recording that changed from S1 to N2.

Step 5: Considering the Combination of S3 and S4 to Slow-Wave Sleep N3

Since the rules for defining slow-wave sleep did not change, all stages scored as S3 or S4 were classified as N3. Note that due to the addition of the frontal EEG leads, however, N3 increased by an average of 19 epochs

Fig. 3. Example of the effect of considering the frontal leads for scoring of sleep (healthy control, female, 42 years, 2nd night). For details see figure 2. Note that the slow waves have amplitudes just

below 75 μ V at central leads, but above 75 μ V at frontal leads. Thus, according to R&K, this epoch is scored as S2, but according to AASM, it has to be scored as stage N3.

as compared with the sum of S3 and S4 (compare also step 1).

Step 6: Considering the Rules for Major Body Movements

On average, 3 epochs had been scored as MT according to the R&K rules. Approximately 2 of these epochs were classified as W according to AASM criteria, since awakening had been detected (alpha waves and/or eyes open). The other epochs were assigned to the sleep stages that followed MT. The majority of these changes were to N1 and N2, but some MT epochs changed to N3 and R as well.

Validation Based on Epoch-by-Epoch Comparisons

In the upper part of figures 2–4, the AASM hypnograms based on manual scoring by 2 independent experts are compared with those based on semi-automated scoring. As can be seen in these examples, manual and semi-automated scorings yielded comparable results.

Table 1 shows the confusion table for the comparison between the 2 independent human expert scorings, and table 2 shows the comparison between the 2 independent semi-automated scorings. In table 1, manual scoring 1 was considered as the ‘gold standard’, in table 2, semi-automated scoring 1 (i.e. Somnolyzer-assisted scoring after review 1). The values in the confusion tables give the

Fig. 4. Example for the impact of the arousal rule for scoring of sleep (healthy control, male, 81 years, 2nd night). For details see figure 2. Note the arousal in the first half of the epoch. Since the

previous epoch has been scored as sleep stage 2, according to R&K, this epoch is scored as S2, but according to AASM, it has to be scored as stage N1.

percentages of agreement and disagreement of the 2nd scoring as compared with the respective 'gold standard'. Figures 2–4 show, in addition to the raw data (lower parts), the traces depicting preprocessed data, in particular delta, spindle and fast beta intensity, the alpha/theta index, REM and SEM density, the EMG tonus as well as artifacts identified at frontal, central and occipital leads displayed for the total night (upper parts). During the 2 independent manual reviews of the automated scoring, 2,970 epochs (out of 69,506 epochs = 4.3%) and 2,420 (= 3.5%) epochs were corrected during review 1 and 2, respectively.

A comparison of table 1 and table 2 reveals that the interrater reliability increases significantly from a kappa coefficient of 0.76 for manual scorings to 0.99 for semi-automated scorings. To show the validity of the semi-automated scorings, tables 3 and 4 present the comparisons between the manual scoring 1 (= 'gold standard') and semi-automated scoring 1 and between the manual scoring 2 (= 'gold standard') and semi-automated scoring 2, respectively. The kappa coefficients for the agreements for all possible pairwise comparisons between manual, automated and semi-automated scorings are summarized in figure 5. If the disagreement between the semi-automated and the manual scorings was restricted to ep-

Table 1. Confusion table for the comparison between manual scoring 1 and manual scoring 2 (n = 72; overall agreement: 82.1%; Cohen's kappa: 0.76)

	W	N1	N2	N3	R
W	90.0	7.4	1.9	0.2	0.5
N1	12.1	59.8	22.3	0.3	5.6
N2	0.9	7.2	83.0	7.7	1.2
N3	0.1	0.2	22.7	76.9	0.0
R	0.8	3.7	3.2	0.0	92.3

The agreement/disagreement rates are given in percent of the number of epochs scored by manual scoring 1 (= 'gold standard').

Table 2. Confusion table for the comparison between semi-automated scoring 1 and semi-automated scoring 2 (n = 72; overall agreement: 99.0%; Cohen's kappa: 0.99)

	W	N1	N2	N3	R
W	98.9	0.6	0.4	0.0	0.2
N1	0.5	97.5	1.6	0.0	0.5
N2	0.0	0.1	99.2	0.6	0.1
N3	0.0	0.0	0.0	100.0	0.0
R	0.1	0.3	0.5	0.0	99.2

The agreement/disagreement rates are given in percent of the number of epochs scored by semi-automated scoring 1 (= 'gold standard').

ochs where none of the 2 manual scorers made the same decision as the Somnolyzer-assisted scoring, the overall agreement between semi-automated and manual scoring was 90.2% (Cohen's kappa: 0.87).

In tables 5 and 6, sensitivities, specificities, PPVs, NPVs as well as prevalence values are given for each sleep stage versus all other sleep stages for pairwise comparisons between manual, automated and semi-automated scorings. For all pairwise comparisons, the lowest sensitivities and PPVs were observed for N1, the lowest specificities and NPVs for N2. Moreover, manual reviews of the automated scorings resulted in improvements in all performance measures to values close to those for the comparisons between manual scorings.

The time needed for manual scoring including the marking of arousals was monitored in 55 recordings and varied between 0:24 and 2:30 h (mean 1:34). The 10th, 25th, 50th, 75th and 90th percentiles were 0:40, 1:00,

Table 3. Confusion table for the comparison between manual scoring 1 and semi-automated scoring 1 (n = 72; overall agreement: 82.2%; Cohen's kappa: 0.76)

	W	N1	N2	N3	R
W	88.6	7.3	1.5	0.1	2.4
N1	11.2	60.2	18.0	0.3	10.4
N2	1.1	7.7	82.4	6.7	2.1
N3	0.4	0.3	17.3	82.0	0.0
R	0.9	3.7	3.5	0.0	91.8

The agreement/disagreement rates are given in percent of the number of epochs scored by manual scoring 1 (= 'gold standard').

Table 4. Confusion table for the comparison between manual scoring 2 and semi-automated scoring 2 (n = 72; overall agreement: 81.1%; Cohen's kappa: 0.75)

	W	N1	N2	N3	R
W	87.4	7.8	2.0	0.2	2.6
N1	11.5	55.7	21.6	0.4	10.8
N2	0.9	8.3	81.7	6.8	2.3
N3	0.5	0.3	16.8	82.3	0.0
R	0.9	4.6	4.2	0.0	90.4

The agreement/disagreement rates are given in percent of the number of epochs scored by manual scoring 2 (= 'gold standard').

1:40, 2:10 and 2:30, respectively. It has to be noted that the experts were asked to score as accurately as possible (and not as efficiently as possible) and in addition, all scorers were aware of the fact that all recordings would be scored by at least 1 more scorer and that after completion of the scoring task a statistical comparison would be performed and the results would be discussed and evaluated in a second workshop. The time needed for epoch-by-epoch quality control of the automated scorings was below 5 min for 50% of the recordings and below 8 min for 90%. The maximal time for quality control of the automated scorings was 15 min.

Validation Based on Sleep Parameters

In table 7, Spearman rank correlation coefficients between sleep parameters derived from manual and semi-automated scorings are summarized. For all sleep parameters that should be reported according to the AASM

Fig. 5. Cohen's kappa for the overall agreement between human expert scorings (manual scoring 1 and manual scoring 2) and Somnolyzer-assisted scorings (automated scoring as well as semi-automated scoring 1 and semi-automated scoring 2). As can be seen, the semi-automated scorings have a reliability close to 1 and an agreement with manual scorings comparable to that observed between the manual scorings.

Fig. 6. Scatter plots for sleep efficiency obtained after manual scoring 1 and manual scoring 2 as well as after manual scoring 1 and Somnolyzer-assisted scoring for the validation dataset ($n = 72$; r : Spearman's rho). Note that all dots are scattered around the 45-degree line of identity. TST = Total sleep time.

Table 5. Epoch-by-epoch comparisons for each sleep stage for pairwise comparisons between manual, automated and semi-automated scorings

Stage	Sensitivity %	Specificity %	PPV %	NPV %	Prevalence %
Manual scoring 1 versus manual scoring 2					
W	90.0	97.6	88.9	97.8	17.8
N1	59.8	94.4	60.3	94.3	12.4
N2	83.0	88.9	83.7	88.4	40.7
N3	76.2	96.3	75.8	96.4	13.2
R	92.3	98.4	91.8	98.5	16.1
Manual scoring 1 versus automated scoring					
W	84.4	97.1	86.2	96.6	17.8
N1	50.6	93.6	52.6	93.1	12.4
N2	80.1	89.2	83.6	86.7	40.7
N3	80.1	96.5	77.7	96.9	13.2
R	86.4	94.7	75.8	97.3	16.1
Manual scoring 1 versus semi-automated scoring 1					
W	88.6	97.5	88.6	97.5	17.8
N1	60.2	94.2	59.5	94.4	12.4
N2	82.4	91.0	86.3	88.3	40.7
N3	81.1	96.8	79.3	97.1	13.2
R	91.8	96.7	84.4	98.4	16.1
Manual scoring 1 versus semi-automated scoring 2					
W	87.7	97.5	88.3	97.3	17.8
N1	58.4	94.1	58.4	94.1	12.4
N2	81.7	90.5	85.5	87.8	40.7
N3	81.0	96.5	77.8	97.1	13.2
R	91.1	96.6	83.7	98.3	16.1

Manual scoring 1 is defined as the gold standard (n = 72).

manual, the correlation coefficients obtained for the sleep parameters derived from manual scorings 1 and 2 are compared to the correlation coefficients obtained for the parameters derived from manual scoring 1 and semi-automated scoring 1, manual scoring 2 and semi-automated scoring 1 as well as for the average of the sleep parameters derived from the manual scorings 1 and 2 and semi-automated scoring 1. A table with very similar values is obtained if the second instead of the first semi-automated scorings are used, since the interrater reliability between the semi-automated scorings 1 and 2 is close to 1. As can be seen in table 7, the correlation coefficients between the semi-automated and the manual scoring are close to or even above the respective values for the correlation between the 2 independent manual scorings. The scatter plots for sleep efficiency (fig. 6) demonstrate as an example the high concordance between sleep parameters derived from manual and semi-automated scorings, since the values scatter closely around the 45-degree line.

Table 6. Epoch-by-epoch comparisons for each sleep stage for pairwise comparisons between manual, automated and semi-automated scorings

Stage	Sensitivity %	Specificity %	PPV %	NPV %	Prevalence %
Manual scoring 2 versus manual scoring 1					
W	88.9	97.8	90.0	97.6	18.0
N1	60.3	94.3	59.8	94.4	12.3
N2	83.7	88.4	83.0	88.9	40.4
N3	75.2	96.5	76.9	96.2	13.4
R	92.1	98.5	92.0	98.5	16.1
Manual scoring 2 versus automated scoring					
W	85.0	97.4	87.7	96.7	18.0
N1	50.3	93.5	51.8	93.1	12.3
N2	80.8	89.3	83.7	87.3	40.4
N3	80.7	96.7	79.2	97.0	13.5
R	86.4	94.8	75.9	97.3	16.1
Manual scoring 2 versus semi-automated scoring 1					
W	88.2	97.6	89.1	97.4	18.0
N1	57.3	93.8	56.2	94.0	12.3
N2	82.0	90.4	85.2	88.1	40.4
N3	80.9	96.9	80.2	97.0	13.4
R	91.0	96.7	83.9	98.3	16.1
Manual scoring 2 versus semi-automated scoring 2					
W	87.4	97.6	89.0	97.2	18.0
N1	55.7	93.7	55.2	93.8	12.3
N2	81.7	90.1	84.8	87.9	40.4
N3	81.4	96.7	79.3	97.1	13.4
R	90.4	96.5	83.3	98.1	16.1

Manual scoring 2 is defined as the gold standard (n = 72).

Discussion

In any PSG study, the first step of the analysis has to be the classification of sleep stages. Until 2007, the only worldwide accepted standard for scoring sleep stages were the rules published by Rechtschaffen and Kales in 1968 [1]. In 2007, the AASM published *The AASM Manual for the Scoring of Sleep and Associated Events*, which contained new rules for scoring sleep stages as well as arousals, cardiac, movement and respiratory events [17]. Concerning the visual classification of sleep stages, these new rules are intended to replace the rules by R&K. For commentaries on the AASM manual for scoring sleep, see Schulz [26] and Parrino et al. [27]. According to Schulz, sleep experts should concentrate on 3 major tasks related to automated sleep scoring: '(1) definition and development of an algorithm', (2) surveillance of the analyzing process, including artifact decontamination, and (3) quality control of sleep analysis' [26]. Task 1 means that

Table 7. Spearman rank correlation coefficients between sleep scoring parameters derived from manual and semi-automated scorings (n = 72)

	Manual scoring 1 vs. manual scoring 2	Manual scoring 1 vs. semi-automated scoring 1	Manual scoring 2 vs. semi-automated scoring 1	Average manual scoring vs. semi-automated scoring 1
Total sleep time	0.95	0.96	0.95	0.96
Sleep latency	0.84	0.89	0.83	0.92
Stage R latency (sleep onset to first epoch of stage R)	0.76	0.82	0.78	0.84
Wake after sleep onset (stage W after sleep onset)	0.95	0.95	0.92	0.95
Percent sleep efficiency (total sleep time/ total recording time) $\times$ 100	0.95	0.95	0.93	0.95
Time in each stage				
Stage N1	0.74	0.72	0.60	0.71
Stage N2	0.64	0.68	0.79	0.82
Stage N3	0.72	0.82	0.84	0.90
Stage R	0.90	0.86	0.85	0.87
Percent of total sleep time in each stage [(time in each stage/total sleep time) $\times$ 100]				
Stage N1	0.76	0.77	0.66	0.76
Stage N2	0.40	0.55	0.66	0.74
Stage N3	0.70	0.80	0.84	0.89
Stage R	0.88	0.85	0.84	0.86

a large network of experts in the field of both sleep and biosignal analysis have to work in close cooperation. Task 2 is part of the standardized quality control procedure, in which the sleep experts receive information from various steps of the analysis including raw data, artifacts and sleep/wake-related features. And finally, task 3 represents the epoch-by-epoch quality control of the automated scoring.

The major part of the present paper describes in detail the adaptations made to the existing automatic R&K sleep classification system to comply with the AASM rules for the classification of sleep stages in adults. Moreover, the result of a validation study in 72 recordings is presented.

A necessary requirement for the implementation of the adaptations in the software is the identification of similarities and differences between the rules in order to break down the necessary changes into individual steps. For this purpose, not only sleep experts but also experts in the field of biosignal analysis participated in the 2-day workshop in which the relationship between the two rules was discussed extensively.

As a result, the overall structure of the Somnolyzer system, as shown in figure 4 of Anderer et al. [12], remained unchanged. This structure consists of a 'raw data quality check' module including automatic artifact iden-

tification, a 'feature extraction' module defining the feature matrix per 30-second epoch, a 'feature matrix plausibility check' module, the 'classifier' module based on an expert system with subsequent smoothing algorithm for sleep stage 2 and REM as well as a 'report generation' module. The final step is a structured quality control process, where the sleep experts receive input from raw data, the 'raw data quality check' module, the 'feature matrix plausibility check' module and the 'report generation' module.

Concerning the differences between the rules, the workshop resulted in 6 necessary adaptation steps, which may also help human scorers in changing from the R&K to the AASM rules. Indeed, there is no difference in the effort it takes to train a human scorer or develop an automated system to score according to the new AASM or to the old R&K rules. Moreover, as a by-product of the software adaptation process including specification, implementation, verification and validation of each step, we were able to estimate the impact of each step on sleep stage classification. This endeavor resulted in the 'stage transition analysis', as summarized in figure 1. As expected, all adaptation steps affected the assignment of epochs to the different stages. The majority of the changes were clearly seen within NREM sleep and only few were observed between wakefulness and sleep or between

NREM and REM sleep. Moreover, some changes partially compensated each other. For instance, while approximately 25 epochs changed from S2 to N1 due to step 3 ('arousal rule' for the end of N2), approximately 3 epochs changed from S1 to N2 due to step 4 (abolition of the '3-min rule'). As shown in the paper by Moser et al. [21] based on manual scoring, the exact influence changes in the scoring rules have on the sleep stages depends on the subjects' EEG characteristics, such as the anterior-posterior gradient of slow-wave amplitudes, densities of sleep spindles, K-complexes on the one hand and arousals on the other hand. And since these characteristics may vary significantly between subjects, there is no chance to predict the effect of changing the rules for individual recordings (for further details, see Moser et al. [21]). As a consequence, results obtained with R&K scoring cannot be directly compared with results obtained with AASM scoring regardless of whether the study is based on manual scoring or on semi-automated scoring.

The present validation study based on 72 PSGs of 72 subjects showed similar results for sleep scoring according to AASM as the validation study published in 2005 based on 286 PSGs of 143 subjects for sleep scoring according to R&K. Concerning interrater reliability between human expert scorings, the epoch-by-epoch agreement was 77% (Cohen's kappa: 0.68) for the R&K scoring and 80% (Cohen's kappa: 0.75) for the AASM scoring. These figures, however, cannot be directly compared due to the different number of stages considered (6 vs. 5 stages), the different dataset (95 healthy controls and 49 sleep-disturbed patients vs. 56 healthy controls and 16 sleep-disturbed patients) and the different scoring procedure (30 experts without a prior training workshop asked to score 'as usual' according to R&K vs. 7 experts after a training workshop asked to score 'as accurately as possible' according to AASM). For detailed information on the effects of the scoring rule on interrater reliability controlling for the number of stages and recordings, see Danker-Hopfe et al. [19]. The authors summarize that on the one hand the integration of frontal, central and occipital leads improves interrater reliability, but that on the other hand this advantage is counteracted by the impairment of interrater reliability due to the new rule according to which any cortical arousal, whether or not it is associated with an EMG increase, determines the end of stage N2.

Similar to the above-described slight increase in interrater reliability between 2 independent human expert scorings, the overall epoch-by-epoch agreement between the Somnolyzer-assisted and the human expert scoring increased slightly from 80% (Cohen's kappa: 0.72) for the

R&K validation study to 82% (Cohen's kappa: 0.76) for the present AASM validation study. Thus, the validity of the Somnolyzer-assisted semi-automated and the human expert scoring is again equivalent. Most importantly, only few epochs had to be changed by the sleep experts during quality control. Indeed, the fully automated Somnolyzer analysis achieved already an epoch-by-epoch agreement of 79% (Cohen's kappa: 0.72). On average, only 4.3 and 3.5% of epochs were changed by the experts during quality control 1 and 2, respectively. Consequently, 2 automated scorings with independent quality control resulted in an interrater reliability close to 1 with an agreement of 99% (Cohen's kappa: 0.99). This means that if one asks 2 sleep experts to manually score sleep stages according to AASM based on the raw data, this will result in an agreement rate of approximately 80%. If these experts receive additional information, such as pre-assigned stages, artifacts, slow waves, sleep spindles, alpha/theta indices, high-frequency EEG and EMG muscle activities as well as SEMs and REMs, the agreement rate will increase substantially. It is obvious that this increase depends on the number of epochs changed during the quality control process and that this number again depends on the level of scrutiny of the visual control. If 2 experts have an agreement rate of 80%, the agreement rate for epochs changed during quality review will be approximately 80% as well. And thus, as long as less than 5% of epochs are changed, the interrater variability will be approximately 1% (= 20% of 5% epochs) and hence the agreement rate 99%. Even if 20% of the epochs are changed during quality control, the agreement rate will still be approximately 96% (= 100% - 20% of 20% epochs). Indeed, these results strongly support the use of semi-automated sleep scoring, specifically in clinical trials and other applications where a high reproducibility is important. For a recent debate on the pros and cons of using automated and semi-automated scoring of sleep, see Zammit [28] and Svetnik et al. [29].

Last but not least, the time required for an epoch-by-epoch review is typically reduced from 30–150 min for manual scoring to 5–8 min for semi-automated scoring, depending on the study characteristics and data quality. On average, the time required for analysis was reduced from 94 to 5 min. Even if typically less than 94 min are needed for sleep scoring by experienced and trained scorers, a reduction in scoring time by a factor of at least 5 is certainly relevant considering the increasing health care costs nowadays.

Performance measures for each sleep stage, as summarized in tables 5 and 6, confirmed that for the com-

parisons between the manual scorings sensitivity is best for stage R, followed by W, N2, N3, and finally N1, no matter whether manual scoring 1 or 2 was defined as the gold standard. A corresponding finding was described recently by Danker-Hopfe et al. [19] based on Cohen's kappa for pairwise comparisons between 2 experts as well as on Fleiss' kappa for comparisons between 6 experts. The same descending order in sensitivity is seen for the comparisons between manual and automated scorings as well as manual and semi-automated scorings. Manual review of the automated scorings changed the performance measures for N2 and N3 only marginally, but improved classification of W, N1 and R. These results demonstrate the importance of performing a review of the automated scoring, with special attention to the distinctions between wake and sleep as well as NREM and REM sleep, even though typically only few epochs are changed during the review process. Indeed, the effects of quality review of the automated scoring observed for the AASM scorings are very similar to those described by Anderer et al. [12] for R&K scorings, confirming the stability and reproducibility of computer-assisted sleep classification. Thus, the new AASM rules improved neither the variability between human and automated scorings nor that between 2 human scorings significantly.

As discussed above, automated scoring of sleep recordings is useful as a pre-evaluation for manual reviewing, since it will result in high scoring validity, high scoring reliability and high cost efficiency. Alternatively, automated scoring may be considered as a second independent evaluation, which could be used to monitor manual scoring of an expert and/or to identify problematic recordings that should be considered for further review. Since the R&K Somnolyzer, which is the basis for the AASM version, was optimized to have minimal dis-

tance to the interpretation of the scoring rules by 30 experts from 8 European sleep labs and since the optimization was based on 572 scorings from 286 PSGs, Somnolyzer-assisted scoring will mimic an 'average' scoring interpretation (e.g. for scoring N3 sleep) rather than tending to extreme positions. This behavior of avoiding extreme points of view is also reflected in the described increase in correlations between target variables derived from manual and semi-automated scorings as compared with the correlations between the average of the target variables of the 2 manual and the semi-automated scorings (table 7). The correlation for N3, for instance, increases from 0.72 for manual scoring 1 versus manual scoring 2 to 0.82 and 0.84 for semi-automated versus manual scorings 1 and 2, respectively, and is finally 0.90 for semi-automated versus the average of manual scorings 1 and 2. Consequently, within the range of 'correct' interpretations of slow-wave sleep, Somnolyzer-assisted scorings will never take the extreme positions, resulting in an either relatively low or relatively high amount of N3, but will lead to the amount of N3 sleep that is close to the mean interpretation of several scorers. And most importantly, the interpretation of the rules will not change significantly between recordings, at least as long as less than 5% of the epochs are changed during manual review. This certainly increases the comparability of scorings between experts and may increase the significance of statistical comparisons between subjects or conditions.

Acknowledgements

This study was supported by The Siesta Group Schlafanalyse GmbH, Vienna. The authors would like to express their thanks to Mag. Elisabeth Grätzhofer for her valuable editorial assistance.

References

- 1 Rechtschaffen A, Kales A: A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects. Bethesda, Public Health Service, US Government Printing Office, 1968.
- 2 Itil TM: Automatic classification of sleep stages and the discrimination of vigilance changes using digital computer methods. *Agressologie* 1969;10(suppl):603–610.
- 3 Smith JR, Karacan I: EEG sleep stage scoring by an automatic hybrid system. *Electroencephalogr Clin Neurophysiol* 1971;31:231–237.
- 4 Hoffmann R, Moffitt A, Wells R, Sussman P, Pigeau R, Shearer J: Quantitative description of sleep stage electrophysiology using digital period analytic techniques. *Sleep* 1984;7:356–364.
- 5 Kubicki S, Holler L, Berg I, Pastelak-Price C, Dorow R: Sleep EEG evaluation: a comparison of results obtained by visual scoring and automatic analysis with the Oxford sleep stager. *Sleep* 1989;12:140–149.
- 6 Prinz PN, Larsen LH, Moe KE, Dulberg EM, Vitiello MV: C STAGE, automated sleep scoring: development and comparison with human sleep scoring for healthy older men and women. *Sleep* 1994;17:711–717.
- 7 Schaltenbrand N, Lengelle R, Toussaint M, Luthringer R, Carelli G, Jacmin A, Lainey E, Muzet A, Macher JP: Sleep stage scoring using the neural network model: comparison between visual automatic analysis in normal subjects and patients. *Sleep* 1996;19:26–35.

- 8 Baumgart-Schmitt R, Herrmann WM, Eilers R: On the use of neural network techniques to analyze sleep EEG data. Third communication: robustification of the classifier by applying an algorithm obtained from 9 different networks. *Neuropsychobiology* 1998; 37:49–58.
- 9 Agarwal R, Gotman J: Computer-assisted sleep staging. *IEEE Trans Biomed Eng* 2001; 48:1412–1423.
- 10 Schwaibold M, Schochlin J, Bolz A: Automated sleep stage detection with a classical and a neural learning algorithm – Methodological aspects. *Biomed Tech (Berl)* 2002; 47(suppl 1 Pt 1):318–320.
- 11 Pittman SD, MacDonald MM, Fogel RB, Malhotra A, Todros K, Levy B, Geva AB, White DP: Assessment of automated scoring of polysomnographic recordings in a population with suspected sleep-disordered breathing. *Sleep* 2004;27:1394–1403.
- 12 Anderer P, Gruber G, Parapatics S, Woertz M, Miazhyńska T, Klösch G, Zeitlhofer J, Barbanoj M, Danker-Hopfe H, Himanen SL, Kemp B, Penzel T, Röschke J, Grözing M, Kunz D, Rappelsberger P, Saletu B, Dorffner G: An E-health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24 × 7 utilizing the Siesta database. *Neuropsychobiology* 2005;51:115–133.
- 13 Penzel T, Hirshkowitz M, Harsh J, Chervin RD, Bufkov N, Kryger M, Malow B, Vittello M, Silber MH, Kushida CA, Chesson AL: Digital analysis and technical specifications. *J Clin Sleep Med* 2007;3:109–120.
- 14 Barbanoj MJ, Clos S, Romero S, Morte A, Giménez S, Lorenzo JL, Luque A, Dal-Ré R: Sleep laboratory study on single and repeated dose effects of paroxetine, alprazolam and their combination in healthy young volunteers. *Neuropsychobiology* 2005;51:134–147.
- 15 Saletu B, Prause W, Anderer P, Mandl M, Freidl M, Weiss M, Aigner M, Mikova O, Saletu-Zyhlarz GM: Insomnia in somatoform pain disorder: sleep laboratory studies on differences to controls and acute effects of trazodone, evaluated by the Somnolyzer 24 × 7 and the Siesta database. *Neuropsychobiology* 2005;51:148–163.
- 16 Svetnik V, Ma J, Soper K, Doran S, Renger JJ, Deacon S, Koblan KS: Evaluation of automated and semi-automated scoring of polysomnographic recordings from a clinical trial using zolpidem in the treatment of insomnia. *Sleep* 2007;20:1562–1574.
- 17 Iber C, Ancoli-Israel S, Chesson A, Quan SF, for the American Academy of Sleep Medicine: The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, ed 1. Westchester, American Academy of Sleep Medicine, 2007.
- 18 Silber MH, Ancoli-Israel S, Bonnet MH, Chokroverty S, Grigg-Damberger MM, Hirshkowitz M, Kapen S, Keenan SA, Kryger MH, Penzel T, Pressman MR, Iber C: The visual scoring of sleep. *J Clin Sleep Med* 2007; 3:121–131.
- 19 Danker-Hopfe H, Anderer P, Zeitlhofer J, Boeck M, Dorn H, Gruber G, Heller E, Loretz E, Moser D, Parapatics S, Saletu B, Schmidt A, Dorffner G: Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *J Sleep Res* 2009;18:74–84.
- 20 Danker-Hopfe H, Kunz D, Gruber G, Kloesch G, Lorenzo JL, Himanen SL, Kemp B, Penzel T, Roschke J, Dorn H, Schloegl A, Trenker E, Dorffner G: Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. *J Sleep Res* 2004;13:63–69.
- 21 Moser D, Anderer P, Gruber G, Parapatics S, Loretz E, Boeck M, Kloesch G, Heller E, Schmidt A, Danker-Hopfe H, Saletu B, Zeitlhofer J, Dorffner G: Sleep classification according to AASM and Rechtschaffen & Kales: effects on sleep scoring parameters. *Sleep* 2009;32:139–149.
- 22 Klösch G, Kemp B, Penzel T, Schlögl A, Rappelsberger P, Trenker E, Gruber G, Zeitlhofer J, Saletu B, Herrmann WM, Himanen SL, Kunz D, Barbanoj MJ, Röschke J, Värri A, Dorffner G: The SIESTA project polygraphic and clinical database. *IEEE Eng Med Biol Mag* 2001;20:51–57.
- 23 Kemp B, Värri A, Rosa AC, Nielsen KD, Gade J: A simple format for exchange of digitized polygraphic recordings. *Electroencephalogr Clin Neurophysiol* 1992;82:391–393.
- 24 Danker-Hopfe H, Schäfer M, Dorn H, Anderer P, Saletu B, Gruber G, Zeitlhofer J, Kunz D, Barbanoj MJ, Himanen SL, Kemp B, Penzel T, Röschke J, Dorffner G: Percentile reference charts for selected sleep parameters for 20- to 80-year-old healthy subjects from the SIESTA database. *Somnologie* 2005;9:3–14.
- 25 Bonnet M, Carley D, Carskadon M, Easton P, Guilleminault C, Harper R, Hayes B, Hirshkowitz M, Ktonas P, Keenan S, Pressman M, Roehrs T, Smith J, Walsh J, Weber S, Westbrook P: ASDA report. EEG arousals: scoring rules and examples. *Sleep* 1992;15:173–184.
- 26 Schulz H: Rethinking sleep analysis: comment on the AASM manual for the scoring of sleep and associated events. *J Clin Sleep Med* 2008;4:99–103.
- 27 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.
- 28 Zammit GK: Insufficient evidence for the use of automated and semi-automated scoring of polysomnographic recordings. *Sleep* 2008;31:449–450.
- 29 Svetnik V, Ma J, Soper K, Doran S, Renger JJ, Deacon S, Koblan KS: Automated sleep scoring with human supervision adds value compared with human scoring alone: a reply to Zammit GK: Insufficient evidence for the use of automated and semi-automated scoring of polysomnographic recordings. *Sleep* 2008;31:449–450, 451.