Automatic Sleep Scoring in Normals and in Individuals With Neurodegenerative Disorders According to New International Sleep Scoring Criteria

Peter S. Jensen, Helge B. D. Sorensen,* Helle L. Leonthin,† and Poul Jennum†*

Abstract: The aim of this study was to develop a fully automatic sleep scoring algorithm on the basis of a reproduction of new international sleep scoring criteria from the American Academy of Sleep Medicine. A biomedical signal processing algorithm was developed, allowing for automatic sleep depth quantification of routine polysomnographic recordings through feature extraction, supervised probabilistic Bayesian classification, and heuristic rule-based smoothing. The performance of the algorithm was tested using 28 manually classified day-night polysomnograms from 18 normal subjects and 10 patients with Parkinson disease or multiple system atrophy. This led to quantification of automatic versus manual epoch-by-epoch agreement rates for both normals and abnormals. Resulting average agreement rates were 87.7% (Cohen's Kappa: 0.79) and 68.2% (Cohen's Kappa: 0.26) in the normal and abnormal group, respectively. Based on an observed reliability of the manual scorer of 92.5% (Cohen's Kappa: 0.87) in the normal group and 85.3% (Cohen's Kappa: 0.73) in the abnormal group, this study concluded that although the developed algorithm was capable of scoring normal sleep with an accuracy around the manual interscorer reliability, it failed in accurately scoring abnormal sleep as encountered for the Parkinson disease/multiple system atrophy patients.

Key Words: Automatic sleep scoring, AASM Manual for the Scoring of Sleep and Associated Events, New international scoring criteria, Severe neurodegenerative disorders.

(*J Clin Neurophysiol* 2010;27: 296–302)

Numerous automatic sleep analysis systems have been developed based on the well-established manual from Rechtschaffen and Kales (R&K) since its publication 4 decades ago (Rechtschaffen and Kales, 1968). The first prominent one was by Martin et al. (1972) who presented the first fully digital R&K-based automatic system. It was based on a pattern-recognition technique that, in five healthy subjects and with an epoch length of 30 seconds, was in 80.8% agreement with the gold standard of manual sleep stage classification. Stanus et al. (1987) presented a system that used a combined technique of wave detection and Bayesian reasoning. Based on 15 healthy subjects and 15 patients with depressive disorders and insomnia, this system showed an epoch-by-epoch agreement of 75% and 70%, respectively, using 20-second long epochs. Schaltenbrand et al. (1993) presented a neural network-based system (a multilayer perceptron model) which by means of 30-second-long epochs showed agreements of 84.5% for a group of 20 healthy subjects and 81.0% for a group of 20 insomnia patients. Hybrid rule- and case-based reasoning was employed in a system developed by Park et al. (2000*a*). Using 30-second epochs, Park's system showed agreement rates of 87.5% and 82.7% when tested on three healthy subjects and three patients with obstructive sleep apnea, respectively. Park et al. (2000*b*) also developed another hybrid system that was presented in the same year. Together with his coworkers, he proposed a hybrid neural network and rule-based expert system which, when tested on two healthy subjects, showed an agreement rate of 85.9% (30-second epochs). More recently, Anderer et al. (2007) introduced an automatic expert system called Somnolyzer 24×7. Without revealing many details of the implemented techniques, Anderer reported that the system demonstrated an agreement rate of 80% in large-scale validation tests based on 286 recordings from both healthy subjects and four different groups of sleep-disordered patients.

All these figures indicate that state-of-the-art automatic reproductions of the old R&K manual produce reliability scores of 75% to 87.5% and 70% to 82.7% in recordings from healthy and sleep-disordered subjects, respectively. In comparison, several reviews have reported similar scores for the healthy subjects but clearly lower scores for the sleep-disordered subjects. In particular, one review (Park et al., 2000*a*) has reported a typical reliability of only 65% to 75% for the sleep-disordered subjects, whereas another review (Norman et al., 2000) has reported an even lower one of 50% to 70% for the same group.

To set a frame of reference for the reported automatic versus manual agreement rates, typical numbers of intrasite manual versus manual agreement rates, i.e., concordances between two (or more) manual scorers from the same sleep center/clinic, should be kept in mind. According to Piñero et al. (2004), two manual scorers typically reach an agreement of 70% to 90% for normal subjects. However, for sleep-disordered subjects, the reported concordance between the scorers is often significantly lower, in some cases even as low as 50%.

As a matter of fact, none of the R&K-based automatic systems have managed to find general acceptance in the clinical practice, because most of them have suffered from widely accepted drawbacks of the traditional manual itself. Some of the R&K manual's most criticized points have been low temporal resolution, ignorance of spatial information, insufficient number of stages, and low correspondence between electrophysiological activity and stages (Hasan et al., 1996; Himanen and Hasan, 2000; Kubicki et al., 1982; Lairy, 1976). In particular, the manual as such has been criticized for being unfit for the use in diseased individuals with multiple electrophysiological abnormalities, because in these individuals, the EEG/electromyogram (EMG) activity pattern is often disturbed to a degree that makes the use of scoring rules fitted for normal sleep inadequate.

From the *DTU Department of Electrical Engineering, Technical University of Denmark, Copenhagen; †Danish Centre for Sleep Medicine, Department of Clinical Neurophysiology, Glostrup University Hospital, Glostrup, Denmark.
Address correspondence and reprint requests to Peter Steen Jensen, Technical University of Denmark, Oesterbrogade 88 A, 2 th., 2100 Copenhagen Oe, Denmark; e-mail: pedesjensen@ofir.dk or peje@nru.dk.

ISSN: 0736-0258/10/2704-0296

In many neurodegenerative diseases, the process of sleep is disturbed. Disease-related pains constitute the main reason for the sleep disturbances; however, other factors also play an important role. The disturbances can for instance also be a direct result of lesions in some of the brain areas that control sleep, paralysis, and side effects from treatments given to control the symptoms from the neurodegenerative disease itself.

Parkinson disease (PD), paralysis agitans, is a degrading brain illness that causes loss of nerve cells in vital parts of the brain. It occurs in approximately 2 of 1000 people in the general population, and according to the International Classification of Sleep Disorders, the prevalence may be as high as 20% of persons aged older than 60 years. Slightly more men than women are affected by the disease. PD is caused by a lack of dopamine in the basal ganglia, but the exact reason why the dopamine-creating nerve cells die is still to be revealed.

Because of the lacking dopamine, people with PD may not only experience a wide range of characteristic movement-related symptoms, including tremor, rigidity, slowness of movement, gait problems, and problems with balance and coordination, but also typically experience memory problems, speech problems, depression, and severe sleep problems. The latter include problems with both impairment of night time sleep and excessive daytime sleepiness. In fact, recent estimates from International Classification of Sleep Disorders reveal that between 60% and 90% of all medically treated PD patients experience sleep problems. Insomnia, sleep fragmentation, circadian rhythm disturbances, periodic limb movements with and without restless leg syndrome, rapid eye movements (REM) behavior disorder, and sleep apnea are all examples of sleep disorders that are reported to occur more often in PD patients than in normal controls (Tse et al., 2005; Yanagisawa, 2006). In particular, REM behavior disorder is reported to occur in about one-third of all PD patients whereas occurring only very rarely in normal controls.

Multiple system atrophy (MSA) is a progressive neurodegenerative disease in which multiple areas of the nervous system experience degeneration. Most affected areas are the basal ganglia, cerebellum, and brain stem. MSA is characterized by a combination of autonomic dysfunction, parkinsonism, and a failure of muscular coordination (Gilman et al., 1999). The disorder develops gradually and is most often diagnosed in men older than 60 years. It is estimated to occur in 2 to 15 individuals per 100,000; however, because these estimations rely on the fact that many individuals with MSA are not correctly diagnosed, the actual incidence number may in practice be much higher. At the moment, little is known about the mechanisms at work in MSA. Consequently, the causes of the disease's characteristic nerve cell degeneration are still unclear.

MSA causes symptoms similar to PD; hence, complaints about sleep disorders and excessive daytime sleepiness are also very common among MSA patients. Some of the most commonly reported sleep disorders in MSA are REM behavior disorder, sleep fragmentation, daytime somnolence, and sleep-related breathing disturbances including obstructive sleep apnea, vocalization, and nocturnal laryngeal stridor (Plazzi et al., 1997). Sleep problems in MSA seem to be associated with more severe motor symptoms, longer disease duration, depression, and longer duration of medical treatment, whereas daytime somnolence seems to be significantly associated with disease severity in MSA (Ghorayeb et al., 2002).

Motivated by the recognized problems with the traditional R&K manual, in 2007, the American Academy of Sleep Medicine (AASM) released a new international sleep scoring manual: "AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications" (Iber et al., 2007). Despite that its publication has certainly received attention and raised debate across sleep centers/clinics around the world (Parrino et al., 2009), no studies have yet reported anything about the new manual's usefulness as a foundation for the development of an automated sleep scoring system nor have any studies yet presented evaluations of how well such an AASM-based automated system performs compared with the most prominent of the earlier developed R&K-based systems, regarding automatic sleep scoring in healthy and neurologically diseased subjects.

With the purpose of presenting an automatic sleep scoring system that is founded directly on the new AASM criteria, we have developed an algorithm based on an as close as possible reproduction of the new manual's recommendations, definitions, and scoring rules. We have evaluated our algorithm by performing a direct comparison of manual and automatic scorings within a group of healthy (normal) subjects as well as within a group of neurologically diseased (abnormal) patients with either PD or MSA, thereby allowing us to address whether the new manual overcomes the typical problems encountered when scoring severely disturbed sleep.

METHODS

Subjects and Polysomnographic Recordings

Twenty-eight day-night polysomnographic (PSG) recordings were selected from a database at the Danish Centre for Sleep Medicine, Glostrup University Hospital, Copenhagen. Eighteen of the recordings were from normal (healthy) subjects who had no diagnosed severe neurodegenerative disorder, and the remaining 10 were from patients with either of the two severe neurodegenerative disorders, PD or MSA.

The normal group consisted of nine males and nine females aged 40.3 ± 8.9 years (range: 30–62 years), and the patient group consisted of six males and four females aged 61.6 ± 8.3 years (range: 47–72 years). The age and sample size differences between the two groups were secondary, because we did not intend to perform intergroup comparisons.

All PSG recordings included, as a minimum, two EEG channels, a left and a right electrooculography (EOG) channel, and a submental EMG channel according to standard R&K criteria.

Manual reference hypnograms were obtained by allowing two experienced physicians from Danish Centre for Sleep Medicine independently (blinded from each other) perform visual scoring of each of the recordings using AASM rules with 30-second epochs. This lead to determination of the intrasite manual versus manual agreement rate, which provided a valid frame of reference for the algorithm.

Algorithm Details

The developed signal processing algorithm comprises six automatically functioning modules performing (1) loading and preparation of data, (2) preanalysis filtering, (3) feature extraction, (4) feature conditioning, (5) sleep stage classification, and (6) hypnogram smoothing, as illustrated in Fig. 1.

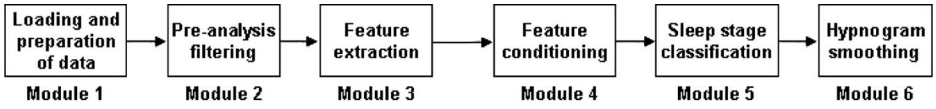

FIGURE 1. Architecture of developed sleep scoring algorithm.

Loading and Preparation of Data

The first module of the algorithm loads and prepares PSG recording data presented in EDF format. Five different traces—two central EEGs (C_3-A_2 and C_4-A_1), left and right EOG, and one submental EMG—are extracted from the recording to yield the signals on which automatic sleep scoring will be based. Subsequent preparation of the five signals (including ear lobe referencing of EEGs and EOGs as well as simple unit conversion) allows them to be in as close accordance with the technical AASM specifications as possible.

Preanalysis Filtering

The second module applies AASM-recommended preanalysis filtering on each of the five signals to emphasize inherent valuable sleep scoring-related information. Specifically, the EEGs and EOGs are bandpass filtered between 0.3 and 35 Hz and the EMG between 10 and 100 Hz, using two separate bandpass filters implemented as Parks-McClellan optimal equiripple finite impulse response filters (McClellan and Parks, 1973). Each implemented filter has a relatively high order. Both filters possess acceptably narrow transitions bands and provide strong and satisfactory attenuation of >35 dB in the stopbands.

In addition, the second module applies a simple 50-Hz notch filter on all five signals to suppress interfering noise from the mains. The filter attenuates the 50-Hz components with 135 dB.

Feature Extraction

In the third module, each 30-second epoch of the signals gets parameterized by a 16-dimensional feature vector. Eleven of the 16 features are extracted from the 2 EEGs, 3 are extracted from the 2 EOGs, and 2 are extracted from the single EMG, as illustrated in Fig. 2.

Ten of the 11 EEG features represent each of the two EEGs' relative spectral power in the five frequency bands 0.5 to 2, 2 to 6, 4 to 7, 8 to 13, and 12 to 14 Hz, which come directly from the definitions in the AASM manual. Power spectrum estimation uses averaging of eighteenth order AutoRegressive models fitted to half-overlapping 5-second segments. The last EEG feature represents epochal information about K complex occurrence. This feature is extracted through a combined peak-tracking and peak-excluding procedure. Simple peak tracking is used first to identify all deflections in the EEGs being distinct, conjugate, and in-phase, and then a series of exclusion criteria is applied on the identified deflections to differentiate between true and artifactual K complex activity.

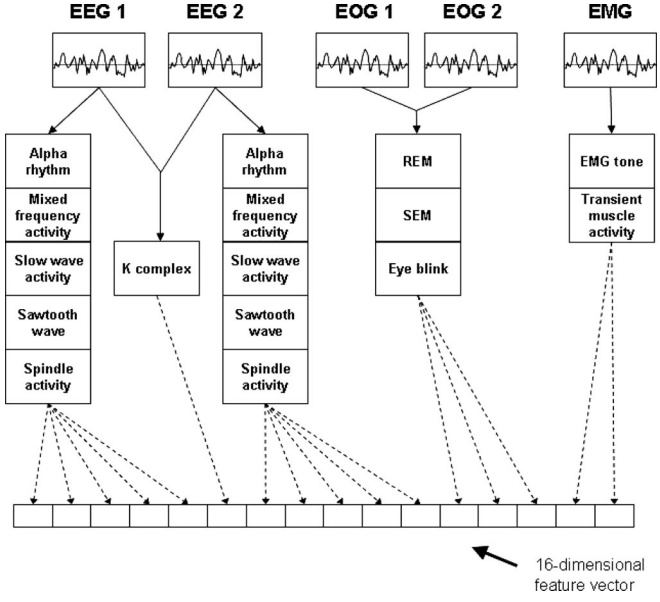

FIGURE 2. Schematic overview of the algorithm's feature extraction procedure. Each of the signals in the top represents a 30-second epoch.

The three EOG features express the epochal amount of REM, slow eye movements, and eye blinks. The REM feature is extracted by subjecting the two EOGs to a peak-tracking and peak-excluding procedure, which identifies EOG deflections that are distinct, conjugate, and out-of-phase. The same procedure is used in the extraction of the slow eye movements feature, this time combined with an initial lowpass filtration of the two EOGs. The employed lowpass filter is a fourth-order Butterworth filter with cut-off frequency at 0.5 Hz and implemented as a zero-phase forward and reverse filter (bidirectional filtering) to ensure linear phase response (Powell and Chau, 1991). The eye blink feature is extracted from the averaged EOGs as the relative spectral power in the 0.5- to 2-Hz band. Power spectrum estimation is again based on averaging of eighteenth order AutoRegressive models fitted to half-overlapping 5-second segments. Only segments with peak-to-peak amplitudes >60 μV contribute.

The two EMG features quantify each EMG epoch's content of general background activity (EMG tone) and transient burst activity. Using a simple statistical analysis of the EMG trace, the former EMG feature is extracted as the standard deviation of each EMG epoch. The latter EMG feature is extracted via an extended method that uses max/min analysis of short (0.25-second) epochal segments as well as a combined amplitude and duration thresholding procedure.

Feature Conditioning

All the extracted 16-dimensional feature vectors are conditioned in the algorithm's fourth module to prepare them for sleep stage classification. The influence from large artifactual feature values is minimized through application of an optimized division/truncation procedure. Specifically, each feature vector element is divided by a fixed feature-specific percentile value and then truncated if >1. As a result of the procedure, all the individual elements are mapped to the interval 0 to 1.

Sleep Stage Classification

Sleep stage classification is completed in the fifth module of the algorithm. Subjected to a probabilistic Bayesian classifier which has been trained using a supervised learning strategy, each of the conditioned feature vectors is categorized as belonging to one of the five sleep stages—stage W, stage R, stage N1, stage N2, or stage N3. This leads to a determination of a provisional automatic AASM-based hypnogram.

Hypnogram Smoothing

Finally, in the algorithm's sixth module, the provisional hypnogram is smoothed through a heuristic rule-based smoothing technique. The provisional hypnogram may contain excessively many (nonphysiological) stage transitions because the Bayesian classifier treats all epochs individually, without taking information from surrounding epochs into consideration when classifying a given epoch. Hence, the classifier has an inherent tendency to be hypersensitive to differences between neighboring epochs, leading to unreasonably many oscillations in the provisionally scored sleep stages.

The hypnogram smoothing is completed using a series of heuristic contextual rules that aim at reducing the number of tran-

sitions. Based on the scoring rules in the AASM manual, the following 9 rules were defined in collaboration with an experienced physician from Danish Centre for Sleep Medicine.

Rule 1 (Continued Stage W Scoring). Whenever a continuous series of N1-epochs is surrounded by two W-epochs, then change the scoring of all the N1s to Ws if, and only if, the continuous N1-series has a maximum total duration of 5 minutes.

Rule 2 (Continued Stage W Scoring). Whenever two or more consecutive W-epochs are followed by an epoch with a transition, then the transition-epoch should be rescored as W if at least three of the six epochs that follow the transition-epoch also have been scored as W's.

Rule 3 (Continued Stage R Scoring). Whenever two or more consecutive R-epochs are followed by an epoch containing a transition, then the transition-epoch should be rescored as R if at least 4 of the following 20 epochs also have been scored as R's.

Rule 4 (Continued Stage R Scoring). Whenever a continuous series of N1-epochs is preceded by at least two consecutive R-epochs and followed by at least one R-epoch, then all the N1's should be rescored as R's if, and only if, the continuous N1-series has a maximum total duration of 10 minutes.

Rule 5 (Continued Stage N2 Scoring). Whenever two or more consecutive N2-epochs are followed by an epoch with a transition to N3, then the N3 should be rescored as N2 if at least two of the six epochs that follow the N3-epoch also have been scored as N2's.

Rule 6 (Continued Stage N2 Scoring). Whenever two or more consecutive N2-epochs are followed by a transition to N1, then the N1 should be rescored as N2 if at least three of the four epochs that follow the N1-epoch also have been scored as N2's.

Rule 7 (Continued Stage N2 Scoring). Whenever a continuous series of W-epochs is preceded by at least one N2/N3-epoch and followed by three or more consecutive N2-epochs, then all of the W's should be rescored as N2's if, and only if, the continuous W-series has a maximum total duration of 15 minutes.

Rule 8 (Continued Stage N3 Scoring). Whenever a single N2-epoch is surrounded by groups of two or more consecutive N3-epochs, then the N2 should be rescored as N3.

Rule 9 (Continued Stage N3 Scoring). Whenever a continuous series of W-epochs is preceded by three or more consecutive N3-epochs and followed by at least one N3-epoch, then all of the W's should be rescored as N3's if, and only if, the continuous W-series has a maximum total duration of 5 minutes.

RESULTS

To assess the performance of the developed algorithm, the 28 manually classified day-night PSGs were used as train/test material for the algorithm's classifier. The analyzed PSGs were applied to the algorithm using the leave-one-out method. This method was employed to provide the maximum number of tests from the relatively small sample size, resulting in 28 individual tests. In each test, 27 manually classified PSGs served as training material for the algorithm's classifier and 1 PSG as test material. Each PSG was used only once as test material.

The manual classifications of the analyzed PSGs enabled the performance of the algorithm to be evaluated in terms of concordance level between the scorings from the algorithm and the human scorer. Using a simple epoch-by-epoch comparison of the automatic and manual hypnograms, this led to quantification of a series of automatic versus manual agreement rates and corresponding Cohen's kappa coefficients, as illustrated in Fig. 3.

Figure 3 illustrates that when trained/tested on all 28 PSGs, the algorithm obtained an overall automatic versus manual reliability score of 87.7% (range: 75.2%–93.6%) for the normal PSGs and 68.2% (range: 49.6%–85.0%) for the abnormal PSGs. Furthermore, the algorithm obtained an overall Cohen's kappa of 0.79 (range: 0.63–0.88) and 0.26 (range: 0.00–0.61) for the normal and abnormal PSGs, respectively. These figures indicate that although the algorithm generally succeeded in classifying the normal PSGs in fine agreement with the human scorer, it provided classifications of the abnormal PSGs that were not concordant with the corresponding manual scorings.

Four specific recordings were selected to highlight the worst and best performance of the algorithm in scoring the normal and abnormal PSGs. The resulting four automatic hypnograms are presented together with their corresponding manual hypnograms in Figs. 4 to 7.

From a visual inspection of the hypnograms in Figs. 4 to 7, it is seen that the overall automatic versus manual agreement is substantial for the abnormal and normal PSGs with the highest concordances (Figs. 5 and 7, respectively). Conversely, the overall agreement is seen to be only acceptable for the normal PSG with the lowest concordance (Fig. 6) and unacceptably low for the abnormal PSG with the lowest concordance (Fig. 4). Although these findings confirm that the algorithm's performance was highest among the normal PSGs, they do not directly confirm that the algorithm generally failed in classifying the abnormal PSGs. Rather, they suggest that the algorithm was capable of classifying an abnormal PSG accurately when the PSG contained only few short periods of sleep interspersed between long and dominating periods of wakefulness.

Tables 1 and 2 summarize the overall percentage of concordance between automatic and manual scorings of normal and abnormal PSGs, respectively. Concordance was expressed on a stage-by-stage basis with all epochs across all subjects within the given PSG group pooled. Manual scoring results were listed vertically and automated scoring results horizontally at each stage. Stage-specific sensitivity was highlighted along the main diagonal. Stage-specific specificity was calculated at each stage using the expression TN/(FP + TN), with TN meaning the number of true negatives, FP the number of false positives, and TN the number of true negatives. Finally, the number of epochs within each stage was listed in the last column to the right.

Among the normal PSGs, the sensitivity of stage W was the highest (96.6%) and that of stage N1 the poorest (59.3%). The specificity was highest for stage R (98.5%) and poorest for stage W (94.7%). All in all, satisfactory results were found for all the stages.

Among the abnormal PSGs, the highest and poorest sensitivities were of stage W (95.4%) and stage N3 (6.59%), and the highest and poorest specificities were of stage N3 (100%) and stage W (39.6%). Clearly, despite a few high figures, in this case, the overall picture was unfortunately unsatisfactory.

The general high sensitivity of the algorithm's classification of stage W reveal that the algorithm succeeded in classifying most of the manually classified stage W epochs, both within the normal and abnormal PSGs. However, the general low specificity of the same stage also reveal that the algorithm had a general tendency to score stage W too often, a tendency that was more pronounced within the group of abnormal PSGs.

A very likely explanation for the high sensitivity of stage W in both groups is that this stage was indeed the one with the highest relative occurrence in both the manually classified normal (59%)

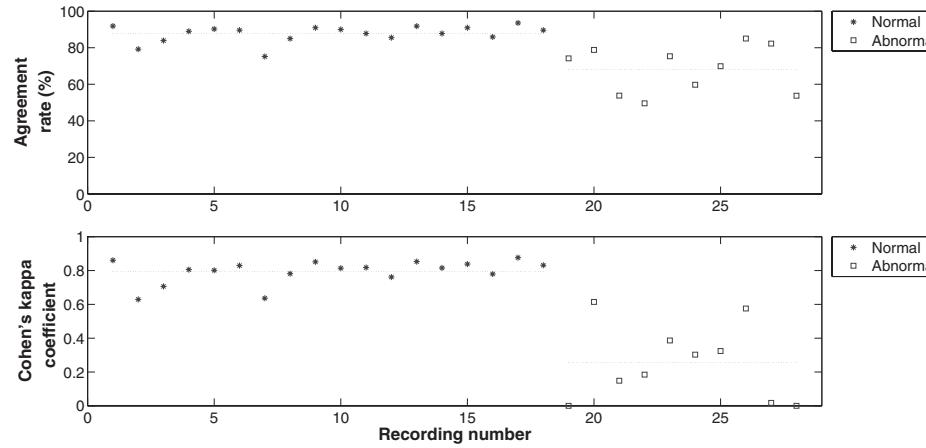

FIGURE 3. Test results from the 28 PSGs. Horizontal lines indicate mean values in each group.

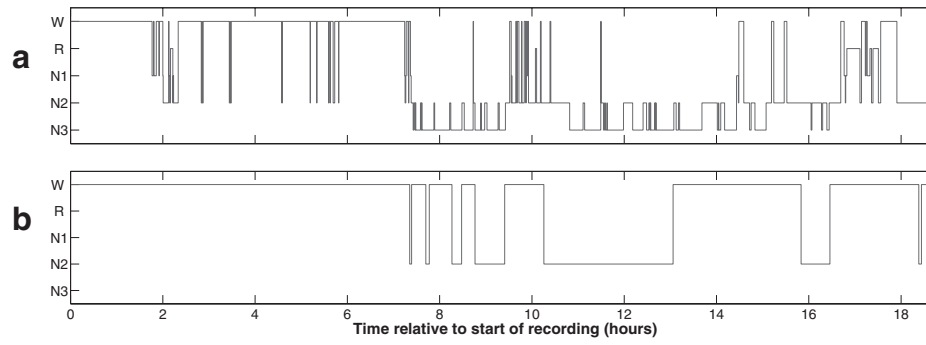

FIGURE 4. Full day-night hypnograms for the abnormal PSG (recording 22 in Fig. 3) with the lowest automatic versus manual agreement. **a**, Manual hypnogram. **b**, Automatic hypnogram.

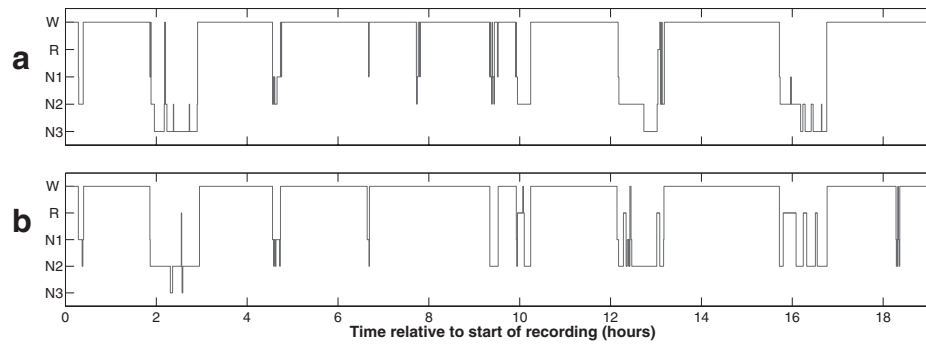

FIGURE 5. Full day-night hypnograms for the abnormal PSG (recording 26 in Fig. 3) with the highest automatic versus manual agreement. **a**, Manual hypnogram. **b**, Automatic hypnogram.

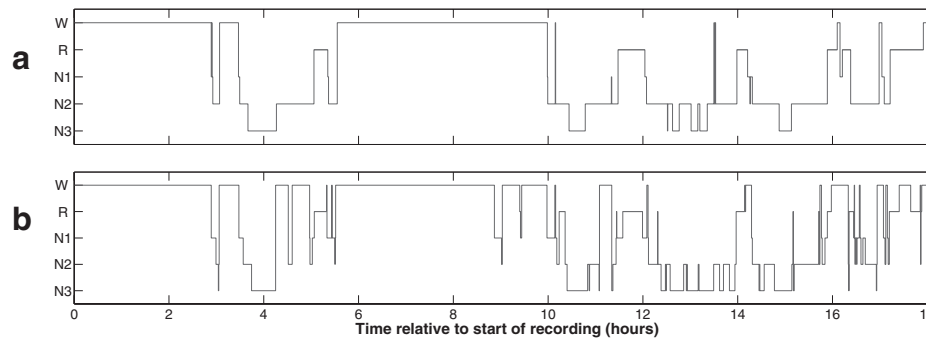

FIGURE 6. Full day-night hypnograms for the normal PSG (recording 7 in Fig. 3) with the lowest automatic versus manual agreement. **a**, Manual hypnogram. **b**, Automatic hypnogram.

and abnormal (64%) PSGs, which obviously made this stage the most probable.

The high relative occurrence of stage W in the training material is likely also to be the explanation for the low specificity of stage W in both groups, because this might have led to a suboptimal training of the algorithm's probabilistic classifier. Nevertheless, because the relative occurrence of stage W was almost the same in both groups, it cannot explain why the relative number of false-

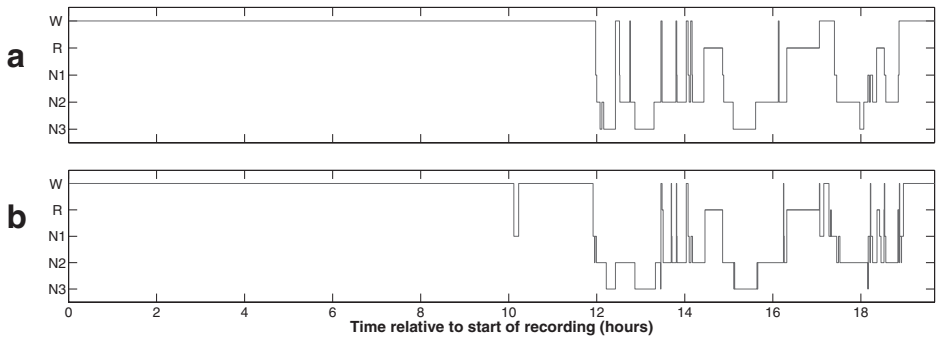

FIGURE 7. Full day-night hypnograms for the normal PSG (recording 17 in Fig. 3) with the highest automatic versus manual agreement. **a**, Manual hypnogram. **b**, Automatic hypnogram.

TABLE 1. Automatic Versus Manual Scoring Agreement Matrix (%) for All Normal Subjects

	W	R	N1	N2	N3	Number of Epochs
W	**96.6**	0.63	2.51	0.31	0	23104
R	7.59	**71.1**	12.1	9.12	0.15	3915
N1	19.8	9.60	**59.3**	10.9	0.32	927
N2	4.65	3.94	7.24	**73.3**	10.9	7415
N3	0.65	0	0	13.5	**85.8**	3828
% Specificity	94.7	98.5	95.8	96.7	97.7	

TABLE 2. Automatic Versus Manual Scoring Agreement Matrix (%) for All Abnormal Subjects

	W	R	N1	N2	N3	Number of Epochs
W	**95.4**	0.81	1.16	2.59	0	13186
R	68.0	**11.1**	6.66	14.3	0	1021
N1	79.4	1.02	**11.8**	7.74	0	788
N2	59.8	4.39	2.87	**32.9**	0	3169
N3	51.4	0.47	0	41.5	**6.59**	2321
% Specificity	39.6	98.6	98.4	91.3	100	

positive stage W classifications was more pronounced within the group of abnormal PSGs. After having carefully examined all the abnormal PSGs, we suggest that the algorithm simply showed a significantly lower stage W specificity when classifying these, because of the fact that it was indeed difficult to identify periods of true sleep in the abnormal PSGs because of multiple EEG abnormalities as well as absence of REM sleep atonia.

Collectively, Tables 2 and 3 have further indicated that the algorithm's overall performance was substantially higher in classifying the normal PSGs than in classifying the abnormal PSGs.

To be able to evaluate whether the algorithm had performed better or worse than could be expected, the reliability of our human scorer was tested by letting a second human scorer (blinded from the results from the first scorer) manually reclassifying all 28 PSGs. This led to determination of a representative value for the manual versus manual concordance level, both within the group of normal and abnormal PSGs. Using again a direct epoch-by-epoch comparison, it was found that the two scorers reached an agreement rate of 92.5% (range: 87.0%–98.4%) for the normal PSGs and 85.3% (range: 75.5%–95.8%) for the abnormal PSGs. In terms of Cohen's kappa coefficient, the agreement between the two scorers was 0.87 (range: 0.77–0.97) and 0.73 (range: 0.51–0.91) for the normal and abnormal PSGs, respectively. Based on these figures, we conclude that although the algorithm has been capable of scoring normal sleep with an accuracy on the level of the manual interscorer reliability, it has failed in accurately scoring abnormal sleep as encountered for the PD/MSA patients.

DISCUSSION

According to the committee behind the new AASM manual, the manual is intended to address visual and automatic classification of both normal and abnormal sleep. Our algorithm has been based directly on the criteria stated in the manual, and therefore the algorithm should provide a means for automatic classification of sleep in both normal and diseased subjects.

The main result of this study was that the general performance of the developed algorithm was high when applied on PSGs from normal subjects but unsatisfactorily low when applied on PSGs from PD/MSA patients. In details, the algorithm showed automatic versus manual agreement rates of 87.7% and 68.2% for the group of normal and abnormal PSGs, respectively, as opposed to observed corresponding manual interscorer reliabilities (agreement rates between two human scorers) of 92.5% and 85.3%.

Comparing the obtained performance results with those from automatic R&K-based systems that represent state-of-the-art within the area of automatic sleep analysis, it can be seen that in scoring of normal sleep, the performance of this algorithm actually exceeded the best of these systems, which provide automatic versus manual reliabilities of 75% to 87.5% for healthy subjects. Conversely, in scoring of abnormal sleep, the algorithm is seen to be inferior to the best of the compared systems because these achieve reliabilities of 70% to 82.7% in recordings from sleep-disordered subjects.

In our opinion, lower classification accuracies between automatic and manual scorings of severely abnormal PSGs, observable in PD/MSA patients, as opposed to scorings of normal PSGs are expectable. This is because the abnormal PSGs typically experience multiple EEG abnormalities as well as absence of REM sleep atonia, and because these phenomena make the abnormal sleep pattern look very different from the normal healthy sleep pattern. Referring to Piñero et al. (2004), we must admit that we are actually a bit surprised to see that our two human scorers reached a level of concordance as high as 85.3% when classifying the 10 abnormal PSGs. However, we think that much of the explanation for this high interscorer reliability is that both scorers apparently classified nearly two-thirds of all epochs as stage W, making a fairly high agreement rate probable simply by chance. We also think that some of the explanation is that the new AASM manual is simply easier to use in visual analysis compared with the old R&K manual, because of the lower number of sleep stages in the new manual.

As mentioned earlier, there exists a sample size difference between the two analyzed groups of normal subjects and PD/MSA patients. Naturally, it would be optimal to have more patients

included in the study so that this intergroup sample size difference could be eliminated but, unfortunately, we have not had the opportunity to do that. Because we have not performed intergroup comparisons of the algorithm's performance but instead compared the performance of the algorithm to the performance of a manual scorer in the two groups separately, and because we are confident that the results from the patient group will not change significantly if we could include eight patients more, we do not think that the intergroup sample size difference affects the validity of the conclusions from this study. Hence, based on our findings, we infer that although the AASM manual is appropriate for use when automating classification of normal sleep, it is as such inappropriate when automating classification of sleep in subjects with severe neurodegenerative diseases. However, we note that there is still a distinct need for further evaluation of the new manual in healthy and diseased PSGs before any such firm conclusions can be made.

Currently, there exists a strong need for a standardized method for sleep scoring in patients with severe neurodegenerative diseases, because the incidence of these diseases is growing (Landrigan et al., 2005). We are confident that a fulfillment of this need will require a very fundamental discussion about the appropriateness of classifying these patients' highly abnormal sleep patterns according to criteria suited for classification of normal sleep.

REFERENCES

Anderer P, Gruber G, Parapatics S, Dorffner G. Automatic sleep classification according to Rechtschaffen and Kales. *Conf Proc IEEE Eng Med Biol Soc*. 2007;3994–3997.

Ghorayeb I, Yekhlef F, Chrysostome V, et al. Sleep disorders and their determinants in multiple system atrophy. *J Neurol Neurosurg Psychiatry*. 2002;72:798–800.

Gilman S, Low PA, Quinn N, et al. Consensus statement on the diagnosis of multiple system atrophy. *J Neurol Sci*. 1999;163:94–98.

Hasan J. Past and future of computer-assisted sleep analysis and drowsiness assessment. *J Clin Neurophysiol*. 1996;13:295–313.

Himanen SL, Hasan J. Limitations of Rechtschaffen and Kales. *Sleep Med*. 2000;4:149–167.

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 its Associated Events: Rules, Terminology and Technical Specifications.* 1st ed. Westchester, IL: American Academy of Sleep Medicine; 2007.

Kubicki S, Herrmann WM, Höller L, Scheuler W. Kritische Bemerkungen zu den Regeln von Rechtschaffen und Kales über die visuelle Auswertung von Schlaf-EEG-Aufzeichnungen. *Z EEG-EMG*. 1982;13:51–60.

Lairy CC. Critical survey of sleep stages; Chairman's summary. In: Koella WP, Levin P. *Sleep*. Basel, Switzerland: Karger; 1976:170–184.

Landrigan PJ, Sonawane B, Butler RN, et al. Early Environmental Origins of Neurodegenerative Disease in Later Life. *Environ Health Perspect*. 2005;113:1230–1233.

Martin WB, Johnson LC, Viglione SS, Naitoh P, Joseph RD, Moses JD. Pattern recognition of EEG-EOG as a technique for all-night sleep stage scoring. *Electroenceph Clin Neurophysiol.* 1972;32:417–427.

McClellan JH, Parks TW. A unified approach to the design of optimum FIR linear phase digital filters. *IEEE Trans Circuit Theory*. 1973;20:697–701.

Norman RG, Pal I, Stewart C, et al. Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep*. 2000;23:901–908.

Park HJ, Oh JS, Jeong DU, Park KS. Automated sleep stage scoring using hybrid rule- and case-based reasoning. *Comput Biomed Res*. 2000*a*;33:330–349.

Park HJ, Park KS, Jeong DU. Hybrid neural-network and rule-based expert system for automatic sleep stage scoring. *Eng Med Biol Soc*. 2000*b*;2:1316–1319.

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.

Piñero P, Garcia P, Arco L, et al. Sleep stage classification using fuzzy sets and machine learning techniques. *Neurocomputing*. 2004;58–60:1137–1143.

Plazzi G, Corsini R, Provini F, et al. REM sleep behavior disorders in multiple system atrophy. *Neurology*. 1997;48:1094–1097.

Powell SR, Chau PM. A technique for realizing linear phase IIR filters. *IEEE Trans Signal Process*.1991;39:2425–2435.

Rechtschaffen A, Kales A. *A Manual of Standardized Technology, Techniques, and Scoring System for Sleep Stages of Human Subjects.* Washington, DC: US Department of Health, Education, and Welfare, Public Health Service—NIH/NIND; 1968.

Schaltenbrand N, Lengelle R, Toussaint M, et al. Sleep stage scoring using the neural network model: comparison between visual and automatic analysis in normal subjects and patients. *Sleep*. 1996;19:26–35.

Stanus E, Lacroix B, Kerkhofs M, Mendlewicz J. Automated sleep scoring: a comparative reliability study of two algorithms. *Electroencephalogr Clin Neurophysiol*. 1987;66:448–456.

Tse W, Liu Y, Barthlen GM, et al. Clinical usefulness of the Parkinson's disease sleep scale. *Parkinsonism Relat Disord*. 2005;11:317–321.

Yanagisawa N. Natural history of Parkinson's disease: from dopamine to multiple system involvement. *Parkinsonism Relat Disord*. 2006;12:40–46.