Med Biol Eng Comput (2006) 44: 105–110
DOI 10.1007/s11517-005-0002-4

ORIGINAL ARTICLE

Jennifer Caffarel · G. John Gibson · J. Phil Harrison
Clive J. Griffiths · Michael J. Drinnan

Comparison of manual sleep staging with automated neural network-based analysis in clinical practice

Received: 27 July 2005 / Accepted: 18 October 2005 / Published online: 26 January 2006

Abstract We have compared sleep staging by an automated neural network (ANN) system, BioSleep™ (Oxford BioSignals) and a human scorer using the Rechtschaffen and Kales scoring system. Sleep study recordings from 114 patients with suspected obstructed sleep apnoea syndrome (OSA) were analysed by ANN and by a blinded human scorer. We also examined human scorer reliability by calculating the agreement between the index scorer and a second independent blinded scorer for 28 of the 114 studies. For each study, we built contingency tables on an epoch-by-epoch (30 s epochs) comparison basis. From these, we derived kappa (κ) coefficients for different combinations of sleep stages. The overall agreement of automatic and manual scoring for the 114 studies for the classification {wake | light-sleep | deep-sleep | REM} was poor (median $\kappa = 0.305$) and only a little better ($\kappa = 0.449$) for the crude {wake | sleep} distinction. For the subgroup of 28 randomly selected studies, the overall agreement of automatic and manual scoring was again relatively low ($\kappa = 0.331$ for {wake | light-sleep | deep-sleep | REM} and $\kappa = 0.505$ for {wake | sleep}), whereas inter-scorer reliability was higher ($\kappa = 0.641$ for {wake | light-sleep | deep-sleep | REM} and $\kappa = 0.737$ for {wake | sleep}). We conclude that such an ANN-based analysis system is not sufficiently accurate for sleep study analyses using the R&K classification system.

J. Caffarel · J. P. Harrison · C. J. Griffiths · M. J. Drinnan (✉)
Medical Physics, Freeman Hospital,
NE7 7DN Newcastle upon Tyne, England
E-mail: Michael.Drinnan@nuth.northy.nhs.uk
Tel.: +44-191-2336161
Fax: +44-191-2330290

G. J. Gibson
Respiratory Medicine, Freeman Hospital,
Newcastle upon Tyne, England

Keywords Sleep stages · Classification · Polysomnography · Neural networks

1 Introduction

Manual scoring of sleep studies is frequently performed in patients with sleep disorders but it is time-consuming. Since in most patients sleep disorders are of a respiratory nature, it is important to document respiratory disturbances such as apnoeas [3, 9], but also the effect of these disturbances on the sleep architecture. Over the years, research has been undertaken to find methods of automating sleep staging in order to reduce analysis time and increase the reliability of the results. Many automated systems now exist, the most promising of which are based on neural networks using frequency analysis of EEG recordings [12, 14].

Validation studies on various automated sleep analysis systems have reached contradictory conclusions: some support the efficiency of automated systems, while other authors are cautious and recommend expert supervision by an experienced scorer [1, 5, 16].

One current state-of-the-art system is BioSleep™ (Oxford BioSignals), which uses automated neural network (ANN) techniques for sleep staging. A potential advantage of such systems, in addition to ease of use and speed of analysis, is that they can assess sleep state on a second-by-second basis to detect important events, such as micro-arousals, occurring on a much finer timescale than that achieved by traditional scoring methods.

Automated neural network systems such as BioSleep that are not tied to the R&K criteria potentially offer substantial benefits over manual assessment and over earlier automated systems. However, most clinical sleep studies are still scored using the Rechtschaffen and Kales staging system based on 30 s epochs. Although it is possible for an ANN to produce a "pseudo-R&K" hypnogram based on 30 s epochs, it is not clear how well this corresponds to a conventionally staged R&K

hypnogram. For BioSleep, the limited validation reported used "visual inspection" to ascertain agreement between the pseudo-R&K hypnogram and manual scorers http://www.oxford-biosignals.com/downloads/BioSleep in use tmp.pdf. There is little published information on reliability and the available studies have used few subjects (6 "healthy adult volunteers" [15] and 20 "consenting subjects" [13]). Furthermore, these studies used data collected from healthy subjects, and hence information on use in individuals undergoing clinical investigation is lacking.

Our aim was to investigate the validity of an ANN system (the BioSleep system) for sleep stage scoring in relation to the Rechtschaffen and Kales classification. We have used overnight sleep studies from a database of patients being investigated for suspected obstructive sleep apnoea syndrome (OSA), over a large age range and of both genders.

2 Method

2.1 Subjects

Six-hour polysomnographic studies were recorded in 114 snoring patients (91 male) being screened for OSA as part of a prospective study of laser palatoplasty between 1998 and 2001. Signals recorded continuously during sleep were: central EEG (CzA1 and OzA1), EOG, chin EMG, ribcage and abdomen movements and airflow (for assessment of apnoea and hypopnoea). The subjects are summarised in Table 1. The sleep stages were defined as follows: wake, light-sleep (stages I and II), deep-sleep (stages III and IV or slow wave sleep) and rapid-eye movement (REM) sleep.

2.2 Manual analysis

The 114 studies were analysed using the Rechtschaffen and Kales system [10] based on 30 s epochs. All signals recorded during the study were used by the human scorer for the sleep stage analysis.

A quarter of the studies (28 subjects) were randomly selected for analysis by a second blinded human scorer, to assess scorer reliability. Details of this subgroup and the total population are summarised in Table 1.

2.3 Automated analysis

We analysed the sleep EEG recordings from all 114 subjects using the BioSleep ANN system. Initially we analysed the CzA1 channel and subsequently we analysed the OzA1 channel to investigate whether changing the channel used for the analysis influenced the outcome. All settings were set to their default values (central EEG derivation with EOG).

As with other ANNs, BioSleep generates hypnogram derived statistics (such as sleep onset, sleep offset, sleep latency, sleep efficiency...), micro-arousal statistics (total number of arousals, micro-arousal index) and "pseudo-R&K" derived statistics (R&K sleep stage classification). The pseudo R&K hypnograms were used for the statistical analysis.

2.4 Statistical assessment

The pseudo R&K hypnograms of all 114 subjects were compared with the results obtained by the first scorer. Comparisons were made using contingency tables constructed on an epoch-by-epoch basis for each subject (720 30-s epochs per study) and calculating the proportion of chance agreement (i.e. the proportion of times the automated and human scoring should agree by chance alone). Overall corrected agreement (taking into account the proportion of chance agreement) was assessed using Cohen's kappa coefficient: κ = (observed agreement − expected agreement)/(1 − expected agreement).

This was calculated for the two combinations of sleep stages: (a) {wake | light-sleep | deep-sleep | REM} and (b) {wake | sleep}. Cohen's kappa coefficient is a standard statistic for such comparisons [6]. The statistical comparisons were made for CzA1 and again for OzA1.

Similarly, we assessed inter-scorer reliability for the 28 subjects analysed by both human scorers. To check that the 28 studies were representative of the studies as a whole, we calculated automatic/manual agreement for the 28 cases, using the CzA1 configuration.

The above agreement tests were also carried out in a subset of patients (47 records) from our database with a low apnoea/hypopnoea index (AHI $< 10\ h^{-1}$), in order to verify that the disturbed sleep pattern in patients with OSA did not affect agreement.

Finally, since total sleep time is the denominator for AHI, the most commonly used clinical index of severity in this population, we measured overall agreement for total sleep time using the method of Bland and Altman.

3 Results

The kappa coefficients for each subject are shown in Fig. 1 ({wake | light-sleep | deep-sleep | REM} sleep stages) and Fig. 2 ({wake | sleep}). The agreement between the automated and manual analyses was generally poor as shown by a wide spread of κ values, even for the crude {wake | sleep} comparison (114 studies, Fig. 2) from −0.456 to 0.898 (-1 representing much worse than chance, 0 chance and 1 total agreement). Most subjects had κ values between 0 and 0.5 with a median of 0.305 (Fig. 1). Changing the EEG configuration of the automated analysis system (using OzA1 instead of CzA1) did not improve the level of agreement (see second column on Figs. 1, 2).

Table 1 Age, AHI and proportion of time spent in each sleep stage for the whole population and subgroup analysed by two scorers (based on the readings of the first human scorer)

	114 Studies (91 male)		28 Studies (26 male)	
	Mean ± SD	Range	Mean ± SD	Range
Age (years)	43.8 ± 8.7	26–68	42.3 ± 8.5	28–61
AHI (h^{-1})	16.7 ± 16.8	0.0–97.9	15.6 ± 15.8	1.2–77.6
Wake (%)	22.0 ± 14.3	1.4–69.2	20.4 ± 14.7	2.4–64.4
Light-sleep (%)	58.1 ± 13.3	23.2–98.6	59.9 ± 12.9	25.0–82.2
Deep-sleep (%)	9.2 ± 6.4	0.0–33.9	8.1 ± 6.7	0.1–22.4
REM (%)	10.7 ± 6.1	0.0–28.8	11.6 ± 6.9	2.1–28.8

By contrast, the inter-observer agreement in the subgroup of 28 subjects gave appreciably higher κ values concentrated above $\kappa = 0.5$ with median values of 0.641 and 0.737 for the two classifications used (Figs. 1, 2). To ensure that the 28 studies used were representative and were not those with better agreement, the agreement between ANN and the human scorer for these 28 cases is also given (fourth column). As can be seen, the kappa coefficients are similar to those in the overall group.

The contingency tables (Table. 2, 3) represent an overall epoch-by-epoch comparison of the 114 studies (automatic vs. manual), given as a proportion of the total number of epochs (82,080 epochs). The epochs classified identically by the scorer and automated system lie in the bold-highlighted diagonal. From Table 2, we can see that the greatest misclassification is due to the ANN staging epochs as REM sleep when the manual scorer staged these as light sleep (13.9% of all epochs). As can be seen from Table 3, even with the crude wake/sleep staging, there was agreement in only 82.2% of epochs, which, when corrected with the proportion of chance agreement (0.666), resulted in a κ value of 0.467.

The agreement tests carried out on the subset of patients with low AHI gave very similar results to those obtained from the total database and thus are not presented here.

Figure 3 shows a Bland and Altman plot of the agreement between the automated sleep staging system and the human scorer for total sleep time. The overall agreement for the total sleep time using this method is −6.9 ± 50.6 min (mean of differences ± SD of differences).

4 Discussion

Agreement between two observers or techniques is generally considered good if κ values are significantly greater than 0.5 (where $\kappa = 0$ represents chance agreement). Only a minority of results for the automated–manual comparison exceeded this value; even for the crude wake/sleep comparison (median $\kappa = 0.449$). For the more detailed {wake | light-sleep | deep-sleep | REM} comparison the level of agreement was inevitably worse (median $\kappa = 0.305$). By contrast, inter-rater agreement was relatively good, as illustrated by κ values much greater than chance agreement. Thus the poor agreement in the automatic–manual comparison for the 114 studies cannot be attributed to poor reliability of the human scorer.

BioSleep was not developed as an R&K staging system, its prime use being a "*second-by-second quantifica-*

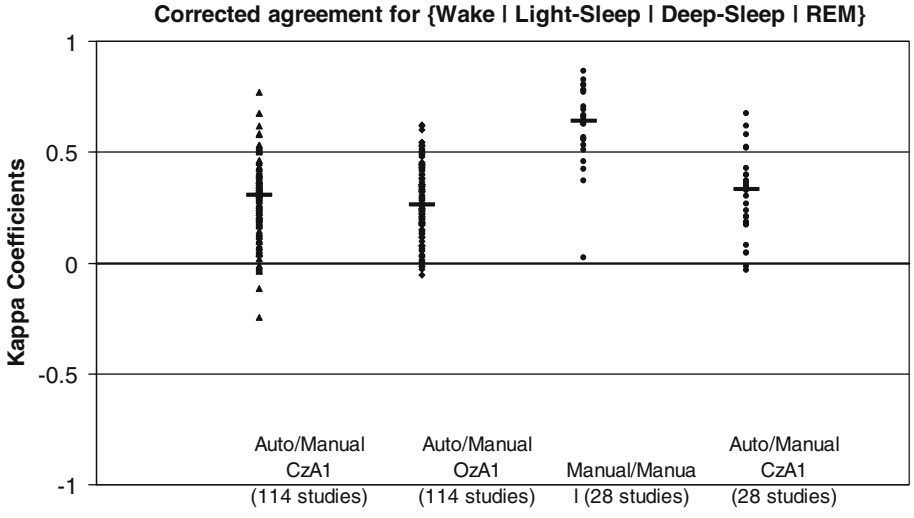

Fig. 1 Comparison of kappa coefficients for {wake | light-sleep | deep-sleep | REM} sleep stage combination (with median values)

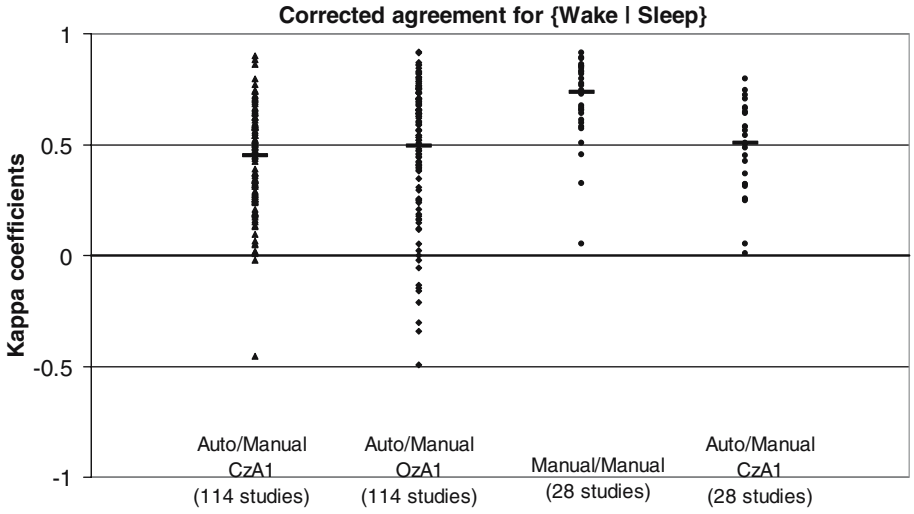

Fig. 2 Comparison of kappa coefficients for {wake | sleep} (with median values)

tion of the sleep/wake continuum with a resolution that far exceeds that of rule-based staging'' [8] and hence it may be more appropriate for the detection of micro-arousals than for conventional staging. Although the originators rightly state[1] that ''*Rechtschaffen and Kales rules for sleep staging suffer from a number of limitations*'', a hypnogram is incorporated in the analysis as clinicians are more used to this method of analysing sleep recordings. As its name indicates, the output is not a true R&K hypnogram and is denoted as a pseudo-R&K hypnogram [7].

One possible cause of misclassification is the recognition of REM sleep. In its earlier version, BioSleep did not attempt to detect this stage correctly as it only used EEG and EMG as input signals but this was later modified to accept EOG recordings (''*with the EOG and EMG being used to assist in the scoring of REM sleep*'' [8]). Furthermore, it should be noted that the output does not include an ''unknown'' state and hence the system will always force an epoch into one of the classes it identifies (wakefulness, stage 1, stage 2, stage 3, stage 4, REM) [8], even if the input signal is subject to noise and/or is non-existent. However, if one of the electrodes were to become permanently disconnected, this would not necessarily be identifiable in the results.

Although the mean total sleep time derived from automated analysis was similar to that for automated analysis, the very wide limits of agreement (Fig. 3) imply considerable inaccuracy in the individual.

The performance of this ANN in sleep staging has previously been evaluated only in small numbers of presumably healthy subjects. In one study [13] overnight home recordings from 20 subjects were used. The neural network was trained with overnight sleep recordings obtained from only ''*9 healthy female*'' subjects [8] and only ''nonartefactual data'' were used [11]. Another group collected data from 22 patients with obstructive sleep apnoea but calculated agreement from the summary statistics rather than using an epoch-by-epoch comparison basis [2].

Our study was not intended as an overall evaluation of the automated sleep analysis system. In particular, the detection of micro-arousals [17] was not studied and only the agreement with a human scorer for R&K classification has been investigated. Although the outcome of this study was not very encouraging due to the low level of agreement, this does not necessarily imply that this approach is inappropriate for sleep analysis. On the contrary, there is a need for further studies of the relative merits of classification obtained by clustering using neural networks (as done by BioSleep) and the traditional R&K method of scoring. In addition, the detection of respiratory-related disturbances was not studied here.

Table 2 Contingency table for {wake | light-sleep | deep-sleep | REM} representing proportion of total number of epochs for all 114 studies (automatic analysis vs. manual scoring)

Auto	Manual			
	Wake	Light	Deep	REM
Wake	**0.122**	0.058	0.007	0.014
Light-sleep	0.054	**0.264**	0.016	0.032
Deep-sleep	0.002	0.119	**0.064**	0.002
REM	0.042	0.139	0.005	**0.058**

Table 3 Contingency table for {wake | sleep} representing proportion of total number of epochs for all 114 studies (automatic analysis vs. manual scoring)

Auto	Manual	
	Wake	Sleep
Wake	**0.122**	0.079
Sleep	0.098	**0.699**

[1]Tarassenko L, Braithwaite E. BioSleep analysis technique for the evaluation of sleep EEG, http://www.oxford-biosignals.com/admin/files/AASM_response.pdf

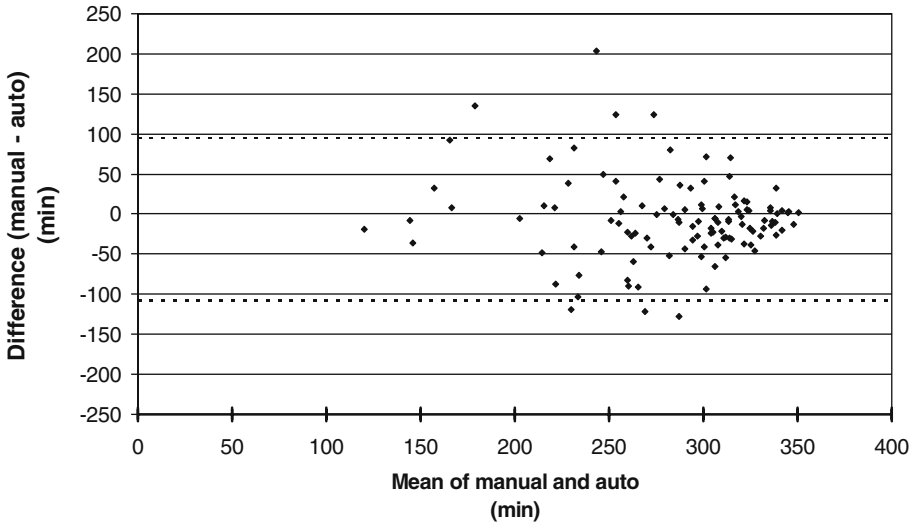

Fig. 3 Bland and Altman plot showing agreement in total sleep time, for 114 studies. The *dotted lines* represent the limits of agreement at mean ± 2SD, i.e. -6.9 ± 101.2 min

While this is an important part of the sleep assessment, automated systems devised to undertake this time-consuming task have been validated elsewhere [3, 9].

It should be pointed out that BioSleep was originally designed to detect micro-arousals based on 1 s epochs and not to classify 30 s epochs using the R&K method of classification. Its ability to detect micro-arousals has already been shown [4]. For a human scorer, the staging of arousals is time-consuming and the level of precision is far less than that obtained with automated systems. Thus automated sleep staging systems could be complementary to the traditional R&K classification method by providing additional information.

5 Conclusions

For over 20 years, research has been undertaken to develop reliable automated EEG analysis systems that could be used in sleep staging to increase precision and reduce time of analysis. One of these is a state-of-the-art automated sleep staging system using ANNs, BioSleep. We evaluated one of this system's outputs, the pseudo Rechtschaffen and Kales hypnogram, against the traditional method as carried out by a human scorer.

The low overall agreement obtained for the 114 studies even for the crude {wake | sleep} distinction ($\kappa = 0.467$) suggests that ANNs cannot yet replace manual Rechtschaffen and Kales scoring in clinical studies. Although the Rechtschaffen and Kales staging classification has existed for 37 years, a replacement has not been found and it seems that state-of-the-art technology has difficulties in mimicking a trained human observer. However, other features of ANNs such as the detection of micro-arousals may be complementary to the Rechtschaffen and Kales system in sleep studies, and this issue requires further study.

Acknowledgements We would like to thank Professor Janet Wilson and Mr. Mohamed Reda for recruiting the patients for this study.

References

1. Andreas S, von Breska B, Magnusson K, Kreuzer H (1993) Validation of automated sleep stage and apnoea analysis in suspected obstructive sleep apnoea. Eur Resp J 6(1):48–52
2. Buchanan F, Wiltshire N, Catterall JR, Kendrick AH (2002) Analysis of sleep stage using a neural network: comparison to manual scoring in patients with obstructive sleep apnoea (OSA). Thorax 57(P50):48–94
3. Chen W, Zhu X, Nemoto T, Kanemitsu Y, Kitamura K, Yamakoshi K (2005) Unconstrained detection of respiration rhythm and pulse rate with one under-pillow sensor during sleep. Med Biol Eng Comput 43(2):306–312
4. Diamantea F, Walker PP, Lowe S, Barr D, Mckown T, Calverley PMA (2002) Comparison of R&K and BioSleep systems for analysis of obstructive sleep apnoea. Thorax 57(P51):48–94
5. Gfullner F, Siemon G (2000) Studies with the fully automated EEG sleep analysis system QUISI. Pneumologie 54(12):580–583
6. Hirshkowitz M, Moore CA (1994) Issues in computerized polysomnography. Sleep 17(2):105–112
7. McGrogan N, Braithwaite E, Tarassenko L (2001) BioSleep: a comprehensive sleep analysis system. In: Proceedings of the 23rd annual international conference of the IEEE Eng Med Bio Soc 2:1608–1611
8. Pardey J, Roberts S, Tarassenko L, Stradling J (1996) A new approach to the analysis of the human sleep/wakefulness continuum. J Sleep Res 5:201–210
9. Penzel T, McNames J, Murray A, de Chazal P, Moody G, Raymond B (2002) Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings. Med Biol Eng Comput 40(4):402–407
10. Rechtschaffen A, Kales A (1968) A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Technical report, UCLA
11. Roberts S, Tarassenko L (1992) New method of automated sleep quantification. Med Biol Eng Comput 30:509–517
12. Roberts S, Tarassenko L (1995) Automated sleep EEG analysis using an RBF network. In: Applications of neural networks. Kluwer, Dordrecht, pp 305–322
13. Royston R, Aldridge M (2002) Clinical evaluation of additional single channel EEG in a home sleep studies programme. Eur Sleep Res Soc JSR 11(Suppl 1):197
14. Schaltenbrand N, Lengelle R, Toussaint M, Luthringer R, Carelli G, Jacqmin A, Lainey E, Muzet A, Macher JP (1996) Sleep stage scoring using the neural network model: comparison between visual and automatic analysis in normal subjects and patients. Sleep 19(1):26–35

15. Stores G, Braithwaite E, Crawford C (2002) Evaluation of a single channel neural network sleep analysis system. Eur Sleep Res Soc JSR 11(Suppl 1):217
16. Villa MP, Piro S, Dotta A, Bonci E, Scola P, Paggi B, Paglietti MG, Midulla F, Ronchetti R (1998) Validation of automated sleep analysis in normal children. Eur Respir J 11:458–461
17. Zamora M, Tarassenko L (1999) The study of micro-arousals using neural network analysis of the EEG. Artificial Neural Networks 7–10 September