Sleep Medicine 21 (2016) 47–56

Contents lists available at ScienceDirect

Sleep Medicine

journal homepage: www.elsevier.com/locate/sleep

Original Article

Assessing sleep architecture and continuity measures through the analysis of heart rate and wrist movement recordings in healthy subjects: comparison with results based on polysomnography

Alain Muzet [a], Sandra Werner [a,*], Gil Fuchs [a], Thomas Roth [b], Jay B. Saoud [a], Antoine U. Viola [a], Jean-Yves Schaffhauser [a], Rémy Luthringer [c]

[a] PPRS, Paris, France
[b] Sleep Disorders Center, Henry Ford Hospital, Detroit, MI, USA
[c] Index Ventures, Geneva, Switzerland

ARTICLE INFO

Article history:
Received 7 October 2015
Received in revised form 18 January 2016
Accepted 25 January 2016
Available online 17 February 2016

Keywords:
Sleep scoring
Automatic analysis
Sleep descriptors
Inter-scorer reliability

ABSTRACT

Objective: The objective of the study was to evaluate the reliability of a new methodology for assessing sleep architecture descriptors based on heart rate and body movement recordings.
Methods: Twelve healthy male and female subjects between 18 and 40 years of age, without sleep disorders and not taking any drug or medication that could affect sleep, were recorded continuously during five consecutive nights. Together with the standard polysomnography, heart rate was recorded with a Holter and wrist movements by actimetry.

Of the 60 recorded nights, 48 artifact-free nights were analyzed by two independent and well-trained visual scorers according to the rules of the American Academy of Sleep Medicine. Sleep stages were assigned to every 30-s epoch. In parallel, the same nights were analyzed by the new methodology using only heart rate and actimetry data, allowing a 1-s epoch sleep stage classification. Sleep architecture was measured for 48 nights, independently for the two manual scorings and the automatic analysis.
Results: Over 42 nights, the intra-class correlation coefficient, used to assess the consistency or reproducibility of quantitative measurements made by different observers, was classified as excellent when all 12 descriptors were combined. Analyses of the individual descriptors showed excellent interclass correlation for eight and good for four of the 12.
Conclusion: The automatic analysis of heart rate and body movement during sleep allows for the evaluation of sleep architecture and continuity that is equivalent to those obtained by manual scoring of polysomnography. The technique used here is simple and robust to allow for home sleep monitoring.

1. Introduction

In considering the future of sleep medicine, experts' advice is to incorporate telemedicine and remote monitoring in future healthcare delivery. Thus, there is a growing interest in portable monitoring devices that are capable of assessing sleep characteristics reliably in real-world settings such as one's home. In the past few years, there have been several systems or devices offered to monitor either sleep and wake or sleep stages in more person-friendly, efficient, and economical ways than the standard polysomnography (PSG). For this purpose, wrist actimetry has been used for several decades in sleep–wake rhythm research.

There are some benefits to utilizing actimetry in sleep research: it is convenient to use, inexpensive, and can be used for extended periods of time. There have been numerous studies to validate the use of actimetry compared to PSG recordings [1–5]. Most of these studies concluded that actimetry provides poor sleep onset measure, generally by underestimating latency, as well as poor detection of short awakenings, generally overestimated when compared to PSG. In addition, actimetry cannot provide measures of sleep stages, sleep cycles, and REM sleep rhythmicity. Most of the time, actimetry is limited to patients suffering from circadian rhythm disorders and for the evaluation of total sleep time [1,4].

Some devices are more oriented toward the use of autonomic variables which are often associated with motor activity [6–10]. Thus, over a 24-h period, heart rate and heart rate variability are moderately influenced by the circadian clock but they appear to be sleep-stage dependent [11]. Other devices use autonomic variables and/or motor activity together with some limited electroencephalographic (EEG) or electro-oculographic (EOG) recordings [12,13]. The use of the latter is still subject to the same limitations of PSG recordings in being unpractical and limited to night-time periods.

* Corresponding author. PPRS, 12 rue de la Paix, 75002 Paris, France.
E-mail address: swe@pprs-research.com (S. Werner).

http://dx.doi.org/10.1016/j.sleep.2016.01.015

In a recent review, Kelly et al. discussed the pros and cons of these new devices based on their technical features [14]. With respect to sleep monitoring, they underlined two major factors limiting the scope of practical use: comfort and cost. They also stressed the difficulty of having a single algorithm applicable to various populations and the necessity to store large amounts of data in a limited space. In addition, most of these systems are yet to prove having any clinical value without published validation reports in comparison with PSG data. The authors remarked that gold standard laboratory PSG is in itself imprecise as the inter-scorer reliability is approximately 85% on average. According to the authors, "this value sets an upper limit on what can be expected of an automated algorithm."

However, longitudinal home monitoring is certainly one of the major advantages of portable monitoring devices. Such an approach not only allows the evaluation of the intrinsic sleep variability from night to night but also the correlation of sleep with the subject's daytime activities or environmental exposure. In addition, it avoids the obvious limitations of PSG such as limited number of successive recordings performed in an unfamiliar sleep environment.

We have recently developed a methodology (Somno-Art methodology, referred below as HMSS for 'Heart rate and Movement Sleep Stager') which uses both 1-Hz heart rate data and wrist movement values to score sleep and sleep stages. These basic data have been recorded through an off-the-shelf recording system (Holter ECG and wrist actimeter) during several 24-h periods. However, as a first step, and because PSG recordings allowing comparisons were performed only during the night sleep, results obtained during the daytime period are not presented or discussed here.

2. Materials and methods

2.1. Participants

Twelve subjects (six males and six females) 18–40 years old were recruited to participate in this study. All subjects were in good physical and mental health. None was on any drugs or medications that could affect sleep. All subjects signed a written informed consent that informed them about the nature and risks of the study. All subjects received financial compensation for their participation.

2.2. Study procedure

The study protocol was subjected to critical review and it was consistent with current knowledge of risks and benefits of the investigation, as well as with moral, ethical, and scientific principles governing clinical research as set out in the Declaration of Helsinki and the guidelines on Good Clinical Practice. The research project was approved by the ethics commission Ärztekammer Berlin, Friedrischstrasse 16, 10969 Berlin, on 26 March 2012.

Subjects stayed in the clinical research unit for five consecutive days and nights and were released after a follow-up examination the morning after the last night. On the day of admission, the subjects were wired to Holter-ECG equipment and a wrist actimeter was attached. Holter-ECG and actimeter data were recorded continuously during the whole experiment (with the exception of the time required for service and shower) while PSGs were performed every night.

2.3. Measurements

2.3.1. PSG recordings

PSG was recorded using Compumedics GRAEL amplifiers using Compumedics ProFusion PSG 3. Sampling rate was 512 Hz (optimal) and 256 Hz (minimal) and only a high pass filter of 0.3 Hz was applied (recommended by Compumedics to suit Grael DC recording). The amplitude range (displayed, converted into EDF) for electroencephalogram (EEG) was 2 mV (±80 μV, ±80 μV); 2 mV (±250 μV, 80 μV) for electro-oculogram (EOG); 2 mV (±125 μV, 80 μV) for electromyogram (EMG), and 2 mV (±1 mV, ±1 mV) for electrocardiogram (ECG).

PSG recordings were performed for five consecutive nights over 8 h (within a typical time period between 23:00 and 07:00).

These recordings included five EEG leads used for the visual scoring, according to the international 10–20 system: F4-M1, C3-M2, C4-M1, O1-M2, and O2-M1; two EOG leads: E1-M1 and E2-M1; two Chin EMGS: EMG1-EM and EMG2-EM; two ECG leads: ECG1 and ECG2; one nose thermistor and one ambient noise marker.

2.3.2. HMSS recordings

2.3.2.1. Actimetry. Non-dominant wrist activity was recorded using an Actisleep+ (Actigraph LLC, 49 East Chase Street, Pensacola, FL 32502, USA) activity monitor. The acceleration data were sampled with a 12-bit analog to digital converter at 30 Hz and stored in a raw, non-filtered/accumulated format in the units of gravity (g). The raw data were stored directly on the non-volatile flash memory of the device and then subsequently downloaded using the Actilife 6 data analysis software (Actigraph LLC). Raw data were then filtered and cumulated in 1-s comma-separated values files that were used for sleep analysis using the HMSS methodology.

The wrist actimetry was measured through the vector magnitude of accelerations obtained every second in the three dimensions of the space and its value is given in counts per second.

2.3.2.2. Holter ECG. A 12-lead Holter ECG (CardioMem CM3000 recorders, Getemed, Teltow, Germany) was used for monitoring over 24 h, including a break of approximately 1 h for service and subjects' personal care (eg, for taking a shower).

Digital signal acquisition was performed at 1024 Hz/12 Bit, and then the times of the ECG R waves were extracted in the unit of samples. Successive inter beat intervals (R–R intervals) were expressed in seconds.

The following equation was used to retrieve heart rate data from R–R intervals:

$$HR = 60/RR(in\ s).$$

2.4. Data processing

Precise synchronization between actimeter, Holter ECG, and PSG recorder was required. Prior to each sleep recording, the internal clocks of the three recording systems were synchronized. Afterward, as the PSG also included ECG recordings, a visual inspection was performed to detect any temporal gaps between the ECG trace and the Holter ECG. All PSG recorded data were converted into European Data Format (EDF) in order to be processed on a computer screen for visual analysis [15]. They were scored by two independent and experienced scorers, and sleep stages were assigned to every 30-s epoch according to the American Academy of Sleep Medicine (AASM) rules [16].

The primary data used by the HMSS software were 1 Hz heart rate, derived from successive R–R intervals, and 1 Hz wrist actimetry.

Some occurring events such as cardiac arousals (sudden increase in heart rate followed by a return to initial values) often associated with occurring body movements (see Fig. 1) or changes in steady state of average heart rate to higher or lower values are detected and quantified.

At the same time, a Support Vector Machine (SVM) is used to distinguish between REM and any other sleep stage, when required by the rule-based system. A cross-validation was made as

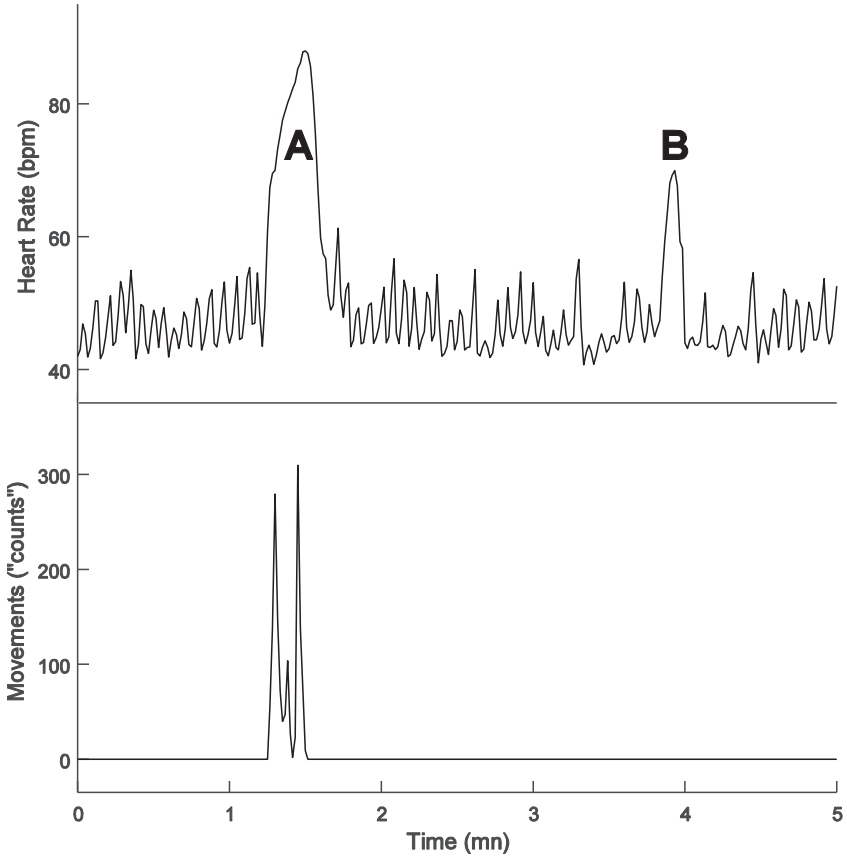

Fig. 1. Examples of cardiac arousals associated (A) or not associated (B) with body movements.

follows: the SVM has been trained on six different nights (no more than one night from any one subject) that were randomly selected and tested on six other nights (no more than one night from any one subject) extracted from our database. The Gaussian kernel was used (bandwidth parameter = 2.2; tradeoff parameter = 10). The SVM was trained using the F-SVC algorithm [17]. The inputs fed to the SVM are:

- The low-frequency power of the centered heart rate (frequency < 6 mHz)
- The low-frequency power of the centered HR in the range [2,20] mHz
- The maximum amplitude of the HR spectrum in the range [0.20–0.33] Hz
- The ratio of the maximum amplitude of the HR spectrum in the range [0.20–0.33] Hz to the maximum amplitude of the HR spectrum in the range [0.15–0.20] Hz

The two last features are estimated using a 10th-order autoregressive model of the HR on a 48-s moving window whose parameters are estimated by simultaneous minimization of the forward and backward prediction error.

During the training phase, the input vectors corresponding to sleep stages are only used, while wake was excluded. The target is the combination of reference hypnograms (codebook: REM in PSG A or PSG B = +1; other sleep stages = −1).

The output of the SVM is low pass filtered to reduce the output variability. An example of the results obtained is shown in Fig. 2. After applying a threshold to this curve, we obtain the binary curve (decision function) that is used to evaluate the performance of the REM/non-REM classifier (in this example, the probability of good classification was 0.8906). In this particular case, the first high REM probability seen at the beginning of the night was not considered as being in contradiction with the rules. As can be seen, this REM/non-REM classifier can be used all night long to detect REM episodes.

In our algorithm, the SVM output is first computed on the whole night. The SVM output (REM/non-REM probability) is used only if required by a rule-based system.

Then, taking into account the dynamics of the occurring events, the amplitude of the heart rate changes calculated over different period lengths, and the REM/non-REM probability, a set of more than 40 rules is applied to generate the sleep stage classification (see Fig. 3).

In using this process, HMSS final scoring is performed every consecutive 1-s epochs during the night. The basis of the methodology used by the software has resulted in an international patent filed in 2012 (PCT/EP2012/059074). However, several of the rules used here are not disclosed as they are proprietary.

The usual statistical descriptors of sleep architecture and continuity are sleep and REM latencies, amounts of various sleep stages in minutes, and the number of awakenings. From these measures, complementary variables such as sleep efficiency and number of sleep cycles are derived. Therefore, the architecture of sleep was evaluated by using the following measures and in accordance with the AASM modified scoring rules:

- *Sleep Onset Latency (SOL)*: elapsed time between light out time and the first occurrence of any sleep stage other than stage W.
- *REM sleep Latency (REML)*: elapsed time between sleep onset and the first occurrence of REM sleep.

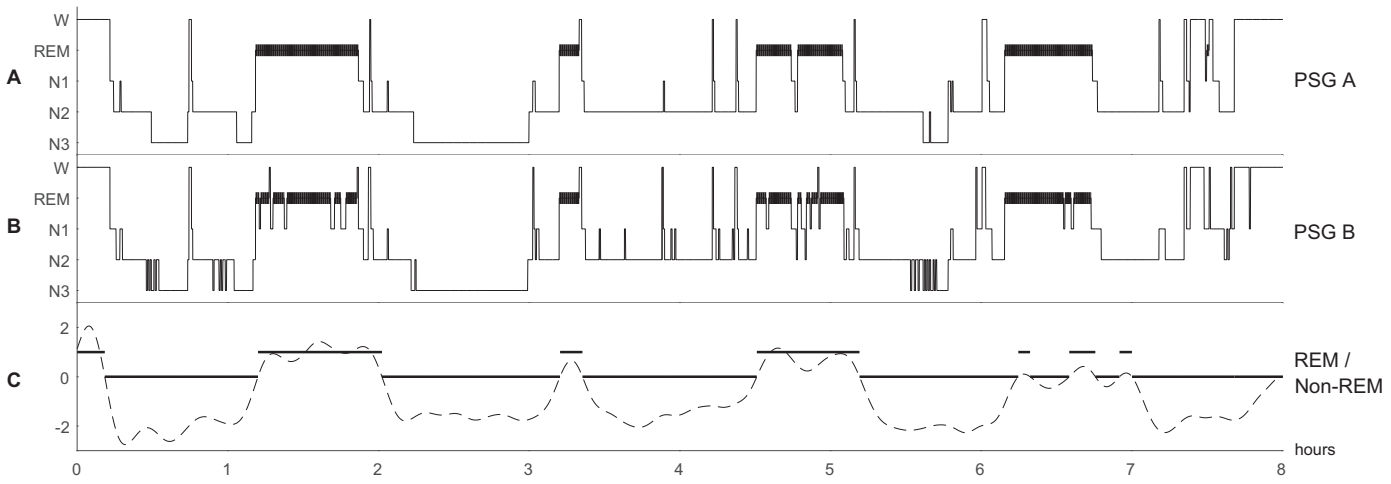

Fig. 2. Representation of the filtered continuous SVM output. The two upper curves (A and B) represent the PSG hypnograms of a night of the test set. The dashed-dotted line is the filtered output of the SVM. After applying a threshold to this curve, we obtain the binary curve (decision function) that is used to evaluate the performance of our REM/non-REM classifier.

- *Wake After Sleep Onset (WASO)*: cumulative time of wake episodes occurring between sleep onset and lights-on time.
- *Time spent in each sleep stage:* cumulative time spent in each sleep stages (N1, N2, N1 + N2, N3, or REM sleep) between sleep onset and lights-on time.
- *Total Sleep Time (TST)*: cumulative time spent in N1, N2, N3, and REM sleep from sleep onset to lights-on time.
- *Sleep Efficiency Index (SEI)*: ratio of total sleep time to time in bed.
- *Number of awakenings (NAW)*: number of wake episodes exceeding 15 s, between sleep onset and lights-on time.
- *Number of Sleep Cycles (NSC)*: a sleep cycle was defined as the elapsed time between sleep onset and the end of the first REM sleep phase (first cycle) or the end of one REM sleep phase to the end of the following REM sleep phase (second and following cycles). A REM sleep phase can be constituted by several REM sleep episodes. Two successive REM sleep episodes were considered as parts of the same REM sleep phase when they were separated by less than 20 min of any other sleep stage [18].

2.5. Statistical analyses

To demonstrate consistency among the two manual scoring methods and the HMSS methodology, a two-way mixed interclass correlation coefficient (ICC) model, where subject effects are random and rater effects are fixed, to assess the inter-rater reliability (IRR) was utilized [19].

ICC, one of the most commonly used statistics for assessing IRR for ordinal, interval, and ratio variables, is a suitable statistic for assessing consistency or reproducibility of quantitative measurements between two or more raters or rating methodologies of the same quantity. The ICC incorporates the magnitude of the disagreement to compute IRR estimates, with larger-magnitude disagreements resulting in lower ICCs than smaller-magnitude disagreements. An ICC estimate of 1 indicates perfect agreement (ie, changing in the same direction for the values being evaluated or directional) and 0 indicates only random agreement (values increase by one method and decrease by another method, non-directional). Cicchetti [20] provides commonly cited cutoffs for qualitative ratings of agreement based on ICC values, with IRR being poor for ICC values less than 0.40, fair for values between 0.40 and 0.59, good for values between 0.60 and 0.74, and excellent for values >0.74.

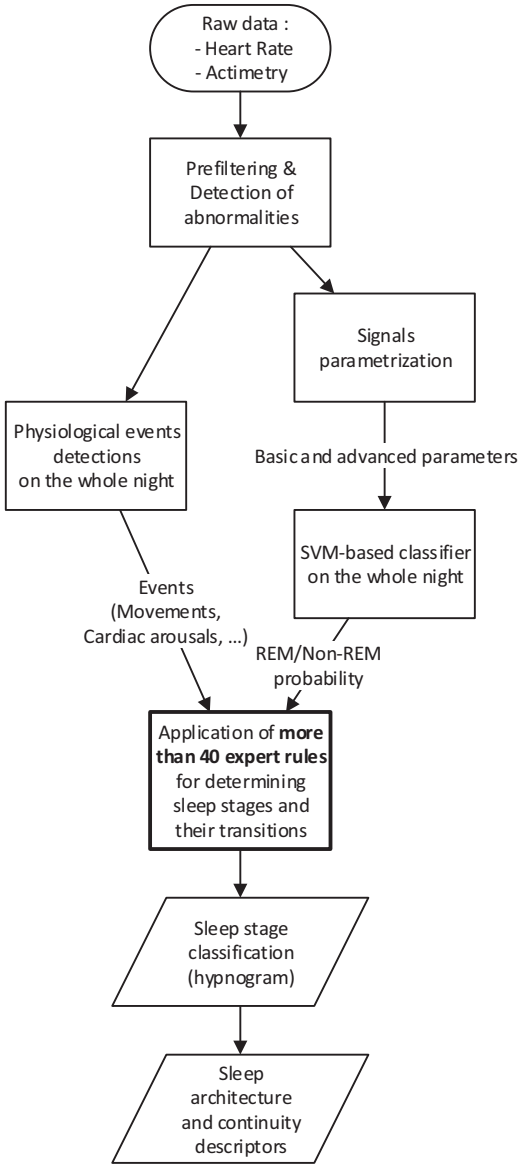

Fig. 3. Description of the steps of the Somno-Art software.

The assessment of the ICC was based on the assumption that ratings from multiple raters for a set of parameters collected from multiple subjects are composed of a true score component and measurement error component. The model can be written using the Mixed Factorial Design in the form:

$$X_{ij} = \mu + r_i + c_j + rc_{ij} + e_{ij}$$

where Xij is the rating provided to subject i (random) by rater j (fixed); μ is the mean of the true value for variable X; r_i is the deviation of the true value from the mean for subject I; c_j represents the degree that rater j systematically deviates from the mean; rc_{ij} represents the interaction between subject deviation and rater deviation; and e_{ij} is the measurement error.

The ICC was calculated based on all three rating methods included in mode l. When the lower bound of the 95% confidence interval of the ICC resulted in a fair IRR based on the classification above, ICC values were computed for all pairwise comparisons to identify the possible causes for the low IRR.

The ICC was also calculated separately for each of the five nightly recordings but as the sample size was 8–11 subjects per night based on valid artifact-free nightly recordings, only the lowest and highest ICC values (range) are presented in Table 1.

All statistical analyses were performed with SPSS version 22 (IBM Corporation).

3. Results

Table 1 presents the results obtained on a subgroup of 42 nights. These nights correspond to 48 artifact-free nights of the total 60 recorded nights (with a range of 8–11 valid cases per night) minus the six nights used for the training of the cross-validation of the SVM. The artifacts were primarily due to missing heart rate values for extended and repeated periods during the night. For these 48 nights, PSGs were manually scored by two independent scorers (A and B) while the heart rate and wrist movements were automatically analyzed by the HMSS methodology.

Following each scoring process, sleep architecture was calculated and a comparison was performed in order to evaluate the three sleep evaluation results (Table 1). Table 1 presents the averaged values of each sleep architecture and continuity descriptors obtained for the three scorings calculated from the two manual scorings (PSG A and PSG B) and the automatic HMSS scoring.

The overall ICC for all descriptors combined was excellent in the comparisons of PSG A vs PSG B (0.996), HMSS vs PSG A (0.991), and HMSS vs PSG B (0.988). Among the 12 descriptors considered in these analyses, eight showed an 'excellent' IRR (ICC > 0.74). Excellent IRR based on observed ICC values were noted for sleep latency, REM sleep latency, Wake After Sleep Onset, amount of stage N3, amount of REM sleep, Total Sleep Time, Sleep Efficiency, and Number

Table 1
Intraclass correlation coefficient (ICC) between the three analyses of the sleep descriptors ($N = 42$).

	PSG A Mean (SD)	HMSS Mean (SD)	PSG B Mean (SD)	ICC (all three scorings) (lower bound) ICC range[a]	ICC (two scorings only)
Sleep latency (min)	10.3 (8.4)	8.7 (7.2)	9.1 (7.7)	**0.952***** (0.920) 0.876–0.990	
REM latency (min)	69.4 (36.3)	62.1 (26.8)	74.7 (41.3)	**0.764***** (0.605) 0.600–0.940	
WASO (min)	31.1 (32.9)	38.9 (36.1)	27.2 (29.9)	**0.958***** (0.930) 0.762–0.993	
Stage N1 (min)	25.6 (8.0)	28.5 (9.3)	49.5 (24.7)	**0.652**** (0.419) 0.433–0.748	PSGA vs PSGB: **0.576*** PSGA vs HMSS: **0.699**** PSGB vs HMSS: **0.499***
Stage N2 (min)	240.2 (27.7)	235.0 (37.8)	229.2 (26.3)	**0.699**** (0.498) 0.273–0.917	PSGA vs PSGB: **0.861***** PSGA vs HMSS: **0.568*** PSGB vs HMSS: **0.411***
Stage N1 + N2 (min)	265.8 (30.8)	263.4 (39.6)	278.7 (38.2)	**0.731**** (0.494) 0.528–0.902	PSGA vs PSGB: **0.895***** PSGA vs HMSS: **0.545*** PSGB vs HMSS: **0.452***
Stage N3 (min)	66.1 (27.3)	71.8 (36.3)	72.8 (32.5)	**0.787***** (0.645) 0.267–0.867	
REM sleep (min)	106.3 (22.7)	96.5 (24.7)	91.8 (23.3)	**0.877***** (0.794) 0.647–0.952	
TST (min)	438.2 (33.2)	431.8 (34.9)	443.3 (29.7)	**0.953***** (0.922) 0.826–0.990	
Sleep efficiency (%)	91.4 (6.8)	90.1 (7.3)	92.4 (6.1)	**0.952***** (0.920) 0.820–0.990	
Number of awakenings	14.1 (6.3)	14.7 (4.0)	20.6 (9.9)	**0.768***** (0.612) 0.569–0.871	
Number of sleep cycles	5.2 (0.8)	5.5 (0.9)	5.2 (0.9)	**0.662**** (0.436) 0.323–0.930	PSGA vs PSGB: **0.931***** PSGA vs HMSS: **0.222** PSGB vs HMSS: **0.335**

[a] Lowest to highest ICC values when analysis performed separately for each of the five nightly recordings.
Commonly-cited cutoffs for qualitative ratings of agreement based on ICC values [20]: poor for ICC values < 0.40; fair for 0.40 to 0.59 (*); good for 0.60 to 0.74 (**); excellent ≥ 0.75 (***).
Boldface numbers correspond to main ICC values obtained from the test.

of awakenings. The four remaining descriptors showed a 'good' IRR (ICC values between 0.60 and 0.74). These were noted for the amount of stages N1, N2, N1 + N2, and number of sleep cycles.

Sleep Onset Latency was very similar between the three methods with very high ICC (ICC = 0.952). The mean REM sleep latency was shorter with HMSS by 7 min than with PSG A and by 12 min than with PSG B. However, the IRR was 'excellent' (ICC = 0.764) between the three methods as these differences were mainly due to a few nights, and the scoring methods were consistent and directional over such nights. The mean time spent in WASO was similar between the three methods with an ICC value of 0.958.

The mean time spent in N1 was similar between PSG A (25.6 min) and HMSS (28.5 min) while it was quite different with PSG B (49.5 min). Pairwise comparison revealed 'good' IRR for PSG A vs PSG B (ICC = 0.604) and PSG A vs HMSS (ICC = 0.699), while IRR was 'fair' for the PSG B vs HMSS (ICC = 0.499).

The mean time spent in N2 was similar between the three analyses (PSG A = 240.2 min; HMSS = 235.0 min; PSG B = 229.2 min). However, ICC was excellent for the comparison PSG A vs PSG B (ICC = 0.861), 'fair' for the comparison PSG A vs HMSS (ICC = 0.568), and 'poor' for the comparison PSG B vs HMSS (ICC = 0.411).

Such findings suggest non-directional differences between PSG B and HMSS where assessed values by the two scoring methods were randomly higher in some cases and lower in others. The average values show that the increase in time spent in N1 was partly compensated for by a decrease in time spent in N2 for the PSG B analysis.

Therefore, we calculated the ICC for N1 + N2. The cumulated time spent in these two stages was higher for PSG B (278.7 min) than for PSG A (265.8 min) and HMSS (263.4 min). The ICC value for the three scorings was 'good' (0.731) while it was 'fair' for the comparisons PSG A vs HMSS (0.545) and PSG B vs HMSS (0.452), and 'excellent' for the comparison PSG A vs PSG B (0.895).

The mean time spent in N3 was similar between the three analyses (PSG A = 66.1 min; HMSS = 71.8 min; PSG B = 72.8 min) and the IRR was 'excellent' (ICC = 0.787).

The mean time spent in REM sleep was similar between the three analyses (PSG A = 106.3 min; HMSS = 96.5 min; PSG B = 91.8 min) and here again the IRR was 'excellent' (ICC = 0.877).

Similarly, TST and sleep efficiency were similar between the three methods and the IRR was in both cases 'excellent' (TST: ICC = 0.953; sleep efficiency: ICC = 0.952).

The mean number of awakenings was higher with PSG B (20.6/night) than with PSG A (14.1/night) and with HMSS (14.7/night). However, the overall IRR was 'excellent' for this descriptor (ICC = 0.768).

For number of sleep cycles, the mean was slightly higher with the HMSS analysis (5.5) than with the two manual methods (PSG A = 5.2; PSG B = 5.2) and the ICC was 'good' (0.662). The detailed comparisons showed that the ICC was 'poor' for both PSG A vs HMSS (0.222) and PSG B vs HMSS (0.335), and 'excellent' between PSG A and PSG B (0.931).

In addition, as the manual scoring was made every 30-s epoch of the PSG recordings and HMSS scoring was made every 1-s epoch, the latter was converted into transformed HMSS 30-s epochs in order to compare the two scoring methods. In doing so, a subsequent variability was introduced into this comparison. The global 'page by page' agreement between HMSS and the manual scorer A was of 73.3%, while it was 74.9% between the manual scorer B and HMSS and 88% between the two manual scorers (see Table 2).

An example of corresponding hypnograms for one single night is presented in Fig. 4.

Table 2
Matrices of confusion between the two manual and the HMSS scorings (cumulated values over the 42 nights). (Global scores between PSG A and HMSS = 73.3% and between PSG B and HMSS = 74.9%.)

		nPages	PSG A pages				
			W	N1	N2	N3	REM
			3656	1816	20388	5474	8934
HMSS pages	**W**	4000	**76.7%**	19.5%	2.9%	0.3%	2.5%
	N1	2392	9.4%	**51.9%**	3.9%	0.3%	3.2%
	N2	19736	9.6%	17.0%	**76.3%**	26.1%	23.3%
	N3	6033	0.2%	0.5%	9.8%	**72.2%**	0.7%
	REM	8107	4.1%	11.1%	7.0%	1.1%	**70.2%**
		nPages	PSG B pages				
			W	N1	N2	N3	REM
			3376	3017	19973	5829	8073
HMSS pages	**W**	4000	**83.0%**	17.6%	2.6%	0.2%	1.8%
	N1	2392	6.3%	**42.8%**	3.5%	0.5%	1.9%
	N2	19736	6.8%	20.0%	**78.9%**	23.4%	22.0%
	N3	6033	0.4%	0.7%	7.9%	**74.8%**	0.7%
	REM	8107	3.5%	19.0%	7.0%	1.1%	**73.6%**

Boldface numbers correspond to full agreement scores.

4. Discussion

The use of EEG to score different stages of sleep started in the first half of the last century [21]. The polysomnographic approach was then developed and sleep scoring rules were adopted by the sleep community [16,22]. Although origin and functions of the different stages of sleep are still debated, it is obvious that brain waves collected on the skull are just local outputs reflecting the changes in brain states. These brain states clearly affect other systems such as motor activity [23] and sympathetic and parasympathetic cardiac regulation [24].

As previously mentioned, actimetry underestimates sleep onset latency, has poor detection of short awakenings, which are often overestimated when compared to PSG measures, and it cannot provide measures of sleep stages, sleep cycles, or REM sleep rhythmicity. However, detecting a body movement can be useful in the determination of the probability of observing a transition from N2 to N3 or from N2 to REM [25]. We also know that small movements can be observed in REM sleep, while large movements are generally followed by transitions from a deeper to a lighter sleep stage or to wake [26].

Similarly, heart rate alone does not provide details of the sleep structure, although several attempts have been made to test it. The spectral analysis of heart rate variability allows distinguishing between low and high frequencies (LF and HF, respectively) and they are used in sleep studies to quantify the modulation of sympathetic and parasympathetic actions of the autonomic nervous system [27,28]. A few studies have been conducted to investigate the interrelations between heart rate variability and the sleep EEG [29,30]. Among different measures extracted from the heart rate variability, the LF/HF ratio provides an indirect measure of sleep fragmentation [31].

However, combining the measure of motor activity together with heart rate could be a possible way of exploring the dynamics of sleep. These two variables are often related as major movements are accompanied by heart rate accelerations which cease when movements stop. But heart rate modifications can occur independently of movement occurrences. They can be changes in average values, variability amplitude, or rhythm regularity. All these modifications can be measured by specific calculated parameters. Therefore, we made the hypothesis that changes of sleeping brain stages could be explored as well by looking simultaneously at the heart rate variations and the occurring body movements.

In our analysis, specific rules were used to determine the chances of being in a specific sleep stage as well as the time and type of transition from one stage to another. These rules, based on the knowledge

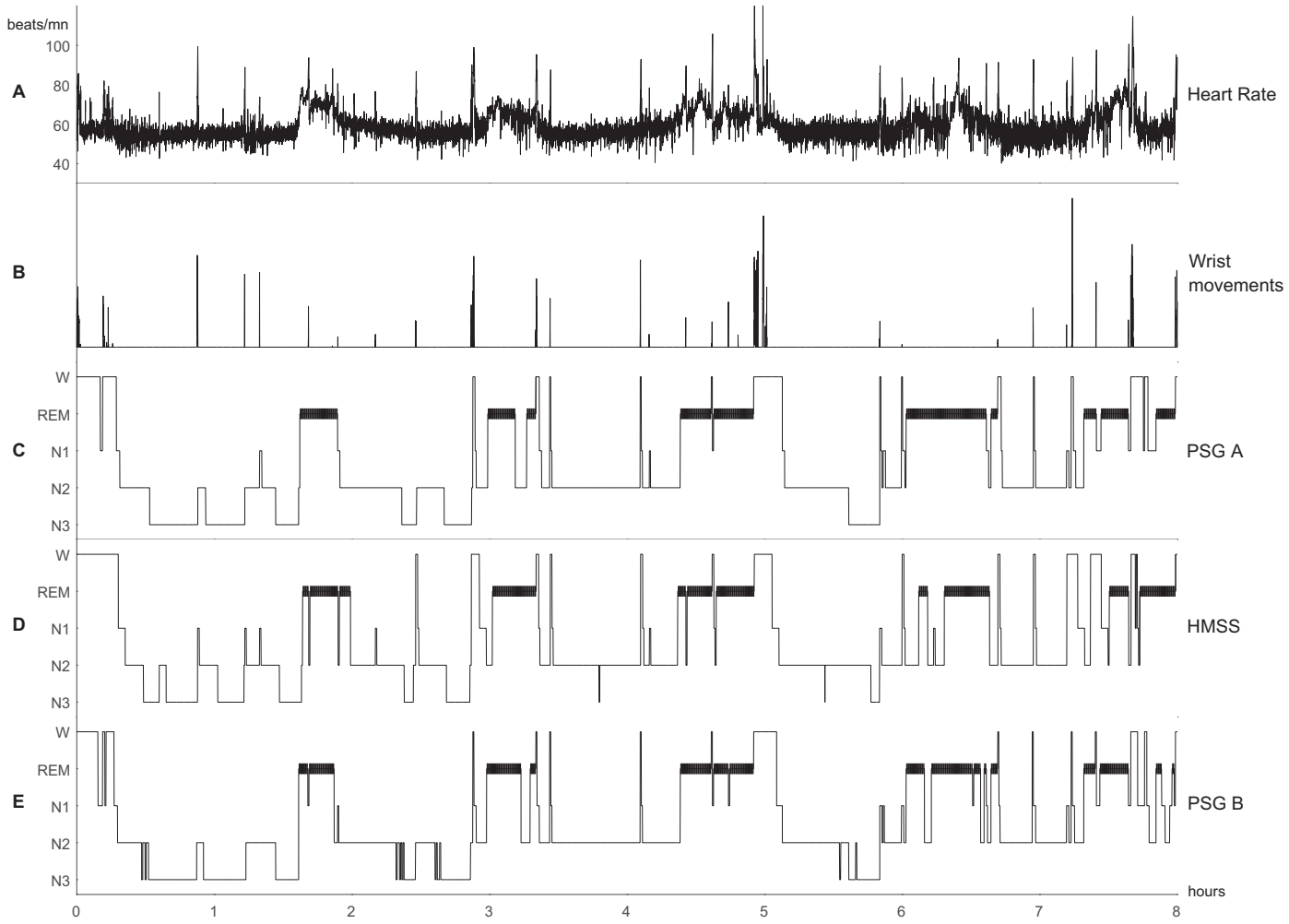

Fig. 4. Comparative hypnograms between HMSS and visual scorers. The graph shows the time evolution of instantaneous heart rate (A), wrist movements (B), and the corresponding hypnograms obtained from the PSG A (C), HMSS scoring (D), and PSG B (E).

accumulated over more than 50 years of polysomnography practice by the first author of this paper, used the REM/non-REM sleep stage probability and sudden occurring events such as body movements or cardiac arousals as well as changes in heart rate averages or variability ranges. As an example of one rule, a body movement associated with a cardiac arousal occurring in N3 leads to a transition to a lighter stage of sleep or to wake, depending on the duration of the movement and the amplitude of the cardiac arousal. As another example, any occurring body movement delays the transition from N2 to N3 by several minutes (depending on the sleep cycle) while it does not block the transition from N2 to REM after a delay of 1 min [25]. Similarly, sleep onset occurs when the average heart rate shows a sustained drop without any simultaneous body movement occurring during the last few minutes. At the end of this rule application step, the sleep stage classification is obtained and the sleep architecture and continuity descriptors are calculated from the final hypnogram (see Fig. 3).

The results presented here show that scoring sleep stages with algorithms and pre-established rules applied to heart rate dynamics and wrist movements produces sleep architecture and continuity descriptors that are consistent with those obtained with manually scored PSG. Among the 12 sleep descriptors which were compared, eight showed 'excellent' IRR between the three scoring methods, while the four remaining descriptors (time spent in N1, N2, and N1 + N2, and number of sleep cycles) showed 'good' IRR only. It is important to note that the decreased ICC value for times spent in N1 and N2 were mainly due to differences between HMSS and only one of the PSG analyses (PSG B). Further exploration of this finding revealed that for PSG B analysis, the increase in time spent in N1 was in part compensated for by a decrease in time spent in N2 and also in REM sleep. The consistency between the sleep architecture and continuity descriptors associated with manual scorings of PSGs and the automatic HMSS system can be seen in Fig. 4, showing the hypnograms scored from one single night by the three methods.

Manual sleep stage scoring is a time-consuming subjective process with inter-scorer and even intra-scorer variability. The inter-scorer variability depends on the experience and the rule adherence of the manual scorers [32,33]. An average value of approximately 85% is generally given [16] but its variation is large and ranges from 76% to 88% [34,35]. In the present study, the average value of inter-scorer agreement (88%) was globally within the same range. The intra-scorer agreement is also quite revealing; it is generally thought to be above 95% but lower percentages such as 83% have also been reported [34]. Such intra-scorer variability is obviated with an automated system like HMSS where repeated analyses give identical results.

Differences in sleep scoring are generally due to periods with unclear EEG and EOG signals, and in most cases these are periods which confound between stages N1 and N2 or between stage N1 and REM sleep. In the latter case, background EEG is quite similar

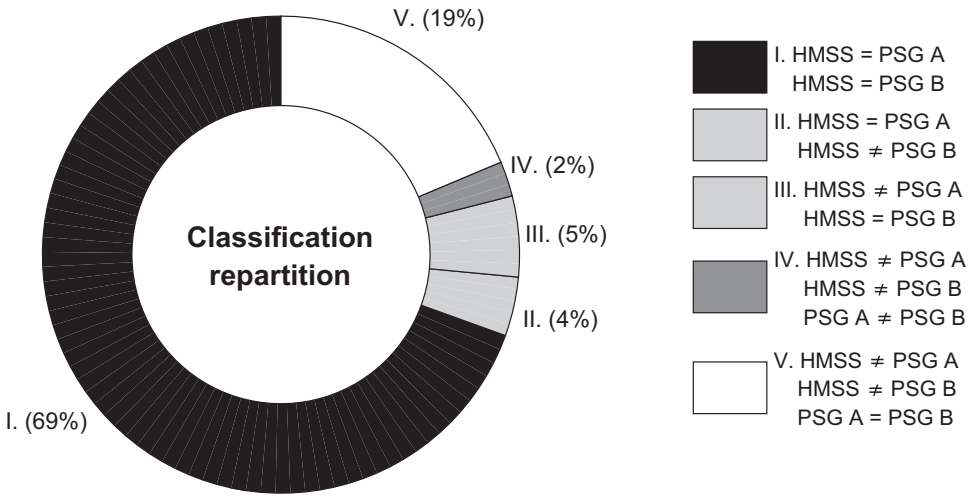

Fig. 5. Repartition of correct and incorrect sleep stage classification of HMSS compared to PSG A and PSG B.

and what distinguishes them most is the presence of either slow eye movements (stage N1) or rapid eye movements (REM sleep). However, the latter are not continuously present in REM sleep as they appear in bursts, and therefore, there may be several consecutive 30-s epochs during REM sleep with no rapid eye movement occurring [36]. Precise distinction between these sleep stages requires careful monitoring of both EEG and EOG traces, which is not always possible outside of the sleep laboratory.

A major reason for scoring difference is the artificial cutting into successive 30-s epochs. The occurrence of a remarkable sign (first K-complex or spindle, first burst of rapid eye movements, etc.) has to be located within the epoch. However, if this change occurs close to the middle of the epoch, one can qualify this epoch as a continuation of the stage given to the previous epoch or qualify it as a new sleep stage. Therefore, some approximation in the clinical evaluation of the sleep architecture and continuity descriptors gets introduced.

For this reason, HMSS scoring was developed to use 1-s consecutive epochs. This was facilitated by the fact that based on instantaneous heart beat values and movement recorded every second, HMSS scoring method is not dependent on ratios between amounts of brain wave activities, compared to PSG scoring, to determine a sleep stage such as in the case of the transition from N2 to N3. Therefore, HMSS scoring is more precise and due to its rapidity, the processing time is not much affected.

By converting 30 independent 1-s epochs into a single 30-s epoch (simple majority rule), an artificial variability was introduced, and a precise time-synchronization between the new 30-s epochs of HMSS and the corresponding 30-s manual scoring epochs was not always guaranteed. However, as the page to page agreement between independent scorers constitutes a traditional way to compare those methods, we performed it too. As an example, the confusion matrices between the HMSS scoring and the two PSG scorings are presented in Table 2. These matrices were calculated over a total of 40,268 30-s epochs included in the 42 nights. They reveal that the main differences are related to the confusion between REM sleep and stages N1 and N2. The mean overall agreement between HMSS sleep stage scoring and the manual scorings was lower than the inter-scorer one (73.3% between PSG A and HMSS, 74.9% between PSG B and HMSS against 88% between the two manual scorings). Fig. 5 shows the repartition of correct and incorrect HMSS classification compared to the PSG ones. From this figure it appears that globally in 78% of the cases the HMSS classification is in agreement with both or at least one of the two manual scorings. It has to be underlined that in 2% of the cases, there is no agreement between any of the three scorings.

As such, HMSS overall agreement was higher than the value of 65.4% obtained in normal subjects by Hedner et al. [8] using the analysis of autonomic signals from an ambulatory recording system. In their case, stages N1 and N2 were grouped in a single sleep stage called 'light sleep'. Therefore, it could be hypothesized that their overall agreement value would have been somewhat lower if they would have distinguished N1 from N2. HMSS overall agreement can also be compared to the epoch-by-epoch inter-site agreement between large samples of visual scorers, ranging from 71% to 76% [37,38].

It has been shown that manual sleep staging is less reliable when sleep is highly fragmented [37,38]. This is one of the criticisms of the usual classification in time-locked epochs. It has been proposed to use adaptive segmentation of the epoch duration in order to give a better representation of this phenomenon [39]. Thus, the rigid constraint of 30-s epochs could be avoided and sleep stage would begin or end at more precise times. This would give a more accurate representation of the sleep structure. However, scoring by 30-s pages instead of much shorter ones reduces the number of sleep stage changes and saves a lot of scoring time. Depending on the experience of the scorer, scoring time can vary considerably from less than 30 min with experienced scorers and good quality recordings to 2 h. A mean value of 1 h 15 min to 1 h 20 min has been reported for experienced scorers [34].

Considering the above, it is important to stress that there are no intra-scorer reliability concerns when using HMSS software as the algorithms used are not subject to change over the scoring duration. In addition, a full-night sleep scoring and its subsequent calculation of the sleep architecture is performed in less than 1 min. These are valuable advantages given the fact that the simplicity and ease of the recording system make sleep evaluation possible anywhere and for any number of successive nights.

Although the heart rate and body movement data were recorded continuously, including daytime, only the night-time period has been analyzed. In fact, the data collected during the daytime were limited to the sole heart rate and wrist movements. During this period, subjects were not maintained in their bed, they were not allowed to nap, they were moving around in the experimental facilities, and therefore, it was not possible to record a PSG simultaneously. Thus, it was not possible to compare PSG sleep scoring data to HMSS scoring during daytime. This is why, in this preliminary study, we have not considered the data collected during

daytime, as our first objective was to score sleep stages and compare the sleep architecture obtained by this method with the one obtained from the PSG. In a further step, we shall be considering the possible occurrence of napping or extended low vigilance periods by applying the same approach.

4.1. Current study limitations

The accuracy of the classification has been tested in normal adult subjects' recordings only. We are fully aware that this sleep scoring method will need to be tested in sleep-disordered patients. Therefore, our next step will be testing HMSS in OSAS patients as well as in patients suffering from insomnia or patients exhibiting periodic leg movements. We also make the assumption that it will not be possible to explore patients with autonomic dysfunction or heart rate abnormalities by the same method.

5. Conclusion

The results presented in this study show that scoring sleep stages from heart rate dynamics and wrist movements produces sleep architecture and continuity descriptors that are consistent with those obtained with manually scored PSG.

If we consider that results obtained by the two systems are equivalent, the main advantages of using HMSS system in healthy subjects are quite obvious. Recording heart rate and body motility is much easier than recording polysomnography and it can be performed without the intervention of any specialized staff, in any environment and repeatedly for several nights. In addition, the processing of sleep stage scoring and the calculation of the sleep measures is shorter than 1 min. Therefore, the HMSS system can be seen as a complementary way of exploring normal sleep to the classical PSGs, especially in the home and for repeated nights.

Conflict of interest

Gil Fuchs, Antoine U. Viola, Jean-Yves Schaffhauser and Alain Muzet, Jay B. Saoud and Sandra Werner declare that they have conflicts of interest, as PPRS is shareholder of V-Watch. Alain Muzet, Rémy Luthringer and Jean-Yves Schaffhauser are shareholders of the V-WATCH company. Rémy Luthringer is consultant for Medicxi Ventures (formerly Index Ventures Life Sciences) and Chief Executive Officer of Minerva Neurosciences, Inc. Thomas Roth has received consultancy fees and grants from V-WATCH company. He also received consultancy fees from Merck, Purdue, Glaxo SmithKline, Jazz, Johnson & Johnson, Proctor & Gamble, Cypress, Pfizer. He obtained grants from Cephalon, Merck, Transcept, and was paid for lectures by Purdue.

The ICMJE Uniform Disclosure Form for Potential Conflicts of Interest associated with this article can be viewed by clicking on the following link: http://dx.doi.org/10.1016/j.sleep.10.1016/j.sleep.2016.01.015.

Acknowledgements

This work was supported by Index Ventures and the V-Watch Company. We would like to thank Matthias Mohse and the PAREXEL team in Berlin who run the experimental study, as well as Georg Ferber, Jalloul Elfehri, Emilien Haug and Régis Lengellé for their help in developing the Somno-Art software algorithms.

Appendix: Supplementary material

Supplementary data to this article can be found online at doi:10.1016/j.sleep.2016.01.015.

References

[1] Ancoli-Israel S, Cole R, Alessi C, et al. The role of actigraphy in the study of sleep and circadian rhythms. Sleep 2003;26:342–92.
[2] Paquet J, Kawinska A, Carrier J. Wake detection capacity of actigraphy during sleep. Sleep 2007;30(10):1362–9.
[3] Blackwell T, Redline S, Ancoli-Israel S, et al. Comparison of sleep parameters from actigraphy and polysomnography in older women: the SOF study. Sleep 2008;31(2):283–91.
[4] Sadeh A. The role and validity of actigraphy in sleep medicine: an update. Sleep Med Rev 2011;15(4):259–67.
[5] Meltzer LJ, Walsh CM, Traylor J, et al. Direct comparison of two new actigraphs and polysomnography in children and adolescents. Sleep 2012;35(1):159–66.
[6] Bresler M, Sheffy K, Pillar G, et al. Differentiating between light and deep sleep stages using an ambulatory device based on peripheral arterial tonometry. Physiol Meas 2008;29(5):571–84.
[7] Mendez MO, Matteucci M, Cerutti S, et al. Sleep staging classification based on HRV: time-variant analysis. Conf Proc IEEE Eng Med Biol Soc 2009;2009:9–12.
[8] Hedner J, White DP, Malhotra A, et al. Sleep staging based on autonomic signals: a multi-center validation study. J Clin Sleep Med 2011;7(3):301–6.
[9] Citi L, Bianchi MT, Klerman EB, et al. Instantaneous monitoring of sleep fragmentation by point process heart rate variability and respiratory dynamics. Conf Proc IEEE Eng Med Biol Soc 2011;2011:7735–8.
[10] Meng X, Hong Y, Jinzhong S, et al. Sleep stages classification based on heart rate variability and random forest. Biomed Signal Process Control 2013;8(6):624–33.
[11] Viola AU, Simon C, Ehrhart J, et al. Sleep processes exert a predominant influence on the 24-h profile of heart rate variability. J Biol Rhythms 2002;17(6):539–47.
[12] Levendowski DJ, Popovic D, Berka C, et al. Retrospective cross-validation of automated sleep staging using electroocular recording in patients with and without sleep disordered breathing. Intern Arch Med 2012;5:21.
[13] Shambroom JR, Fabregas SE, Johnstone J. Validation of an automated wireless system to monitor sleep in healthy adults. J Sleep Res 2012;21:221–30.
[14] Kelly JM, Strecker RE, Bianchi MT. Recent developments in home sleep-monitoring devices. ISRN Neurol 2012;2012:768794.
[15] Kemp B, Värri A, Rosa AC, et al. A simple format for exchange of digitized polygraphic recordings. Electroencephalogr Clin Neurophysiol 1992;82:391–3.
[16] Iber C, Ancoli-Israel S, Chesson AL, et al. The AASM manual for the scoring of sleep and associated events: rules, terminology, and technical specifications. 1st ed. Westchester, Illinois: American Academy of Sleep Medicine; 2007.
[17] Thomé M, Lengellé R. F-SVM: a simple and fast training algorithm for soft margin Support Vector Classification. In: IEEE Workshop on Machine Learning for Signal Processing, 16–19 Oct. 2008, Cancun; 2008, pp. 339–44.
[18] Naitoh P, Johnson LC, Lubin A, et al. Computer extraction of an ultradian cycle in sleep from manually scored sleep stages. Intern J Chronobiol 1973;1:223–34.
[19] McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1:30–46.
[20] Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994;6(4):284–90.
[21] Loomis AL, Harvey EN, Hobart GA. Electrical potentials of the human brain. J Exp Psychol 1936;19:249–79.
[22] Rechtschaffen A, Kales A. A manual of standardized terminology, techniques, and scoring systems for sleep stages of human subjects. Los Angeles, USA: Brain Research Institute, UCLA; 1968.
[23] Dement W, Kleitman N. Cyclic variations in EEG during sleep and their relation to eye movements, body motility, and dreaming. Electroenceph Clin Neurophysiol 1957;9:673–90.
[24] Baust W, Bohnert B. The regulation of heart rate during sleep. Exp Brain Res 1969;7:169–80.
[25] Muzet A, Naitoh P, Townsend RE, et al. Body movements during sleep as a predictor of stage change. Psychon Sci 1972;29:7–10.
[26] Muzet A, Schieber JP. Etude de la covariation de variables électroencéphalographiques, motrices et cardiaque au cours du sommeil nocturne chez l'Homme. Arch Sci Physiol 1971;25:467–516.
[27] Busek P, Vankova J, Opavsky J, et al. Spectral analysis of the heart rate variability in sleep. Physiol Res 2005;54:369–76.
[28] Burr RL. Interpretation of normalized spectral heart rate variability indices in sleep research: a critical review. Sleep 2007;30:913–19.
[29] Jurysta F, van de Borne P, Migeotte PF, et al. A study of the dynamic interactions between sleep EEG and heart rate variability in healthy young men. Clin Neurophysiol 2003;114:2146–55.
[30] Dumont M, Jurysta F, Lanquart JP, et al. Interdependency between heart rate variability and sleep EEG: linear/non-linear? Clin Neurophysiol 2004;115:2031–40.
[31] Sforza E, Pichot V, Cervena K, et al. Cardiac variability and heart rate increment as a maker of sleep fragmentation in patients with a sleep disorder: a preliminary study. Sleep 2007;30:43–51.
[32] Penzel T, Zhang X, Fietze I. Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules. J Clin Sleep Med 2013;9(1):89–91.
[33] Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. J Clin Sleep Med 2013;9(1):81–7.

[34] Ruehland WR, O'Donoghue FJ, Pierce RJ, et al. The 2007 AASM recommendations for EEG electrode placement in polysomnography: impact on sleep and cortical arousal scoring. Sleep 2011;34(1):73–81.
[35] 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.
[36] Weber LD, Muzet A, Schieber JP, et al. Characteristics of rapid eye movement production during REM sleep in man: organization and rhythmicity. Electroencephal Clin Neurophysiol 1983;55:151–5.
[37] Danker-Hopfe H, Anderer P, Zeitlhofer J, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. J Sleep Res 2009;18:74–84.
[38] Norman RG, Pal I, Stewart C, et al. Inter observer agreement among sleep scorers from different centers in a large dataset. Sleep 2000;23:901–8.
[39] Penzel T, Hirshkowitz M, Harsh J, et al. Digital analysis and technical specifications. J Clin Sleep Med 2007;3(2):109–20.