J. Sleep Res. (2011) Regular Research Paper

doi: 10.1111/j.1365-2869.2011.00944.x

Validation of an automated wireless system to monitor sleep in healthy adults

JOHN R. SHAMBROOM[1], STEPHAN E. FÁBREGAS[1] and JACK JOHNSTONE[2]

[1]Zeo Sleep Research Center, Zeo, Inc., Newton, MA, USA and [2]Valley Sleep Center, Burbank, CA, USA

Accepted in revised form 3 July 2011; received 9 March 2011

SUMMARY The availability of a reliable system to record sleep stage measures easily and automatically in ambulatory settings could be of utility for research and clinical work. The aim of this study was to evaluate a novel wireless system (WS) that does not require skilled preparation for the automatic collection and scoring of human sleep. Twenty-nine healthy adults underwent concurrent sleep measurement via the WS, polysomnography (PSG) and an actigraph (ACT) in a sleep laboratory for one assessment night preceded by an acclimation night. The PSG recordings were scored by two experienced trained technicians from separate laboratories. Each recording was scored by both technicians to Rechtschaffen and Kales (R&K) criteria. The WS and ACT were compared with each of the PSG scores and a consensus PSG score, and the PSG scores were compared with each other. Inter-rater agreement was assessed for each pair over all pooled epochs by percentage agreement, Cohen's kappa and intraclass correlation coefficient. The WS agreement with each of the two PSG scores for sleep stages was 75.8 and 74.7%, respectively. WS agreement with each of the two PSG scores for sleep/wakefulness was 92.6 and 91.1%, ACT agreement with PSG was 86.3 and 85.7%. The PSG scorers' agreement with each other for sleep stages was 83.2%, and for sleep/wakefulness was 95.8%. The findings from the current study indicate that the WS may provide an easy to use and accurate complement to other established technologies for measuring sleep in healthy adults.

KEYWORDS actigraphy, ambulatory, automated, inter-rater agreement, sleep scoring, wireless

INTRODUCTION

The ability to investigate the sleep of individuals is of scientific and clinical interest. Sleep patterns have most commonly been examined with polysomnography (PSG), wrist actigraphy and subjective self-report instruments. Each method of measuring sleep has strengths and limitations. PSG is the gold standard assessment methodology for sleep and it provides detailed information on sleep staging, latencies to sleep and specific sleep stages, as well as the number of arousals from sleep.

Actigraphy is an alternative to PSG as an objective indirect measurement of sleep and wakefulness (Ancoli-Israel *et al.*, 2003; Paquet *et al.*, 2007). Actigraphy has the advantages that it is less costly and less intrusive than PSG, is relatively easy to use in ambulatory settings, and utilizes automated scoring algorithms that reduce the need for manual interpretation of the recordings. These advantages enable its use for multiple nights in longitudinal studies of sleep/wakefulness patterns. However, there is no set standard for the collection or scoring of sleep using actigraphy, making interpretation of data from different systems difficult (Acebo and LeBourgeois, 2006). In addition, because actigraphy uses movement as a surrogate for wakefulness, it is limited almost exclusively to the detection of sleep and wakefulness, and its correlation with PSG is moderate. Moreover, actigraphy tends to overestimate total

Correspondence: John R. Shambroom, Shambroom Associates, LLC, 384 Edmands Road, Framingham, MA 01701, USA. Tel.: 781-929-5807; fax: 508-877-6102; e-mail: jshambroom@verizon.net

sleep time (TST) in healthy and sleep-disordered subjects because actigraphy is prone to misinterpreting quiet wakefulness (e.g. lying in bed) as sleep (Kushida *et al*., 2001; Pollak *et al*., 2001).

Subjective sleep assessments are a common method for obtaining information about the sleep of subjects in research and clinical populations. Although sleep diaries and questionnaires offer a way to monitor sleep and habits easily and affordably over long periods of time (such as days or weeks), such instruments suffer limitations. The data collected by such means are subjective, and may not always be accurate reports of sleep in both healthy and sleep-disordered populations (Feige *et al*., 2008; Lichstein *et al*., 2006; Means *et al*., 2003). In addition, sleep diaries are limited to sleep and wakefulness with no ability to monitor sleep stage information. Subjective measures of sleep also, by nature, require subjects to schedule the time and effort to input data manually onto a form, making compliance challenging. The instruments used for subjective sleep assessments also vary widely within the literature, making comparisons between reports difficult.

The development and validation of low-cost, easy-to-use, portable sleep recording devices, with automated algorithms to distinguish among sleep stages and wakefulness, has important implications for sleep medicine and research. We evaluated a novel wireless system (WS), developed for monitoring sleep in the home and other environments (Zeo, Inc., Newton, MA, USA). The system utilizes a dry fabric headband for collecting a single-channel signal from the forehead, which is transmitted wirelessly to a base station where sleep stages are scored in real time by an automated algorithm. The resulting sleep stage information is available in summarized form at the conclusion of the night. The aim of the present study was to validate the WS prospectively by comparing sleep stage and sleep/wakefulness assessments to those measured by PSG and an actigraph (ACT) in the laboratory.

METHODS

Subjects

Twenty-nine subjects were enrolled and completed the protocol. However, three consecutive sleep recordings were lost due to a technician error. Of the 26 remaining subjects, 50% were female and the average age [± standard deviation (SD)] was 38 (± 13) years, ranging from 19 to 60 years. Inclusion criteria were as follows: healthy adults (18–65 years), with no physical or mental health complaints, with a body mass index between 18.6 and 30.0, and a willingness to abstain from alcohol, nicotine and illicit drugs for the 24 h prior to each recording. Reported presence of any sleep disorder, excessive daytime sleepiness, recent shift work or travel outside of the time zone within the prior month and pregnancy were exclusionary. The study protocol was approved by the Western Institutional Review Board of Olympia, Washington. Subjects provided written informed consent and were compensated for their participation in the study.

Protocol

Polysonographic, WS and ACT data were collected simultaneously for each subject for two nights occurring within a 1-week period. The first night was an acclimation night for subjects to adjust to sleeping with the equipment in a laboratory setting.

The wireless system

The wireless system uses proprietary dry silver-coated fabric sensors in a headband (Fig. 1) to collect electrophysiological signals from the forehead with a single bi-polar channel located at approximately Fp1–Fp2. The electrophysiological signal includes contributions from the electroencephalogram (EEG), eye movements and the frontalis muscle. The headband is fully adjustable to accommodate use by different individuals and is worn such that it is tight enough to be secure, but loose enough to minimize discomfort.

The headband contains a 12-bit analog to digital converter and preprocessing unit which amplifies and filters the electrophysiological signal. The signal is captured at 128 samples per second and filtered within a second-order bandpass frequency of 2–47 Hz. The use of the dry sensor necessitates the use of a low frequency cut-off that is higher than the recommended 0.3 Hz, as the spectrum below 2 Hz is contaminated by excessive noise. The resulting signal is transmitted to a base station using an ultra-low-power propriety wireless protocol at 2.4 GHz.

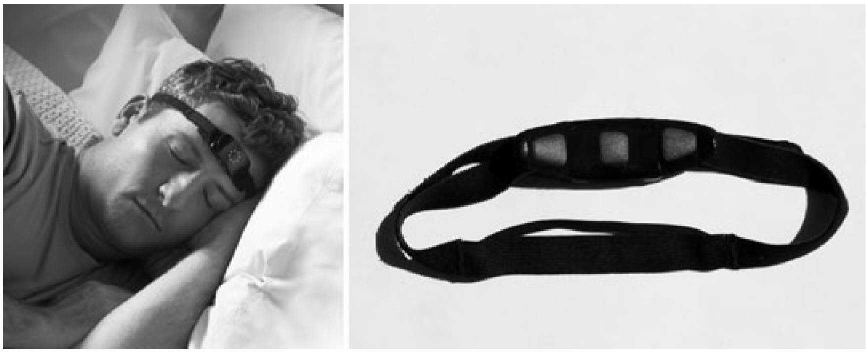

Figure 1. The headband, about the size of a tennis headband, is comprised of three dry silvercoated fabric sensors and a sealed plastic enclosure containing electronics for signal sampling and transmission to a base station for processing.

A microprocessor within the base station calculates the sleep stage from the signal in real time utilizing artificial neural network technology. The neural network uses a combination of time and frequency dependent features derived from the signal to create a best estimate of sleep stage corresponding to those described by Rechtschaffen and Kales (R&K). A reduced set of sleep stages are reported that include: wakefulness, rapid eye movement (REM) sleep, light sleep (combined Stages 1 and 2), and deep sleep (combined Stages 3 and 4) (Rechtschaffen and Kales, 1968). Prior to training, the neural network does not have a priori information about the correlation of any time or frequency feature with sleep stage. Optimization of the algorithm through training compensates for the lack of information below 2 Hz by the inclusion of available features in the determination of sleep stage. A sleep stage is assigned to each 2-s interval of recording; these are then smoothed using a 2-min moving average window, and a result is reported once every 30 s.

A previously collected independent data set of nocturnal sleep recordings from 18 healthy adults, aged 21–60 years (33% female), was used for the training and optimization of the algorithm by iteratively improving the correlation between the algorithm output and the determination made by a human sleep scorer using R&K guidelines. The present analysis was conducted on a new data set using new sleep scorers. None of the subjects in the training set were included in the present analysis, and neither of the PSG scorers was involved in algorithm training.

Procedure

Polysomnography (Cadwell Easy III; Cadwell Laboratories Inc., Kennewick, WA, USA) was recorded using the standard 10–20 system electrode placement with a referential system montage (with the reference electrode placed at Cz). The sleep scoring montage included six EEG channels that were derived for offline scoring (F3-A2, F4-A1, C3-A2, C4-A1, O1-A2, O2-A1), left electro-oculogram (LOC-A2), right electro-oculogram (ROC-A1), a bi-polar submental electromyogram (EMG1–EMG2) and a respiration belt. The EEG and EMG channels were recorded at 200 samples s^{-1} and filtered within a bandpass frequency of 0.3–35 Hz. An ACT (Actiwatch 64; Mini Mitter Philips/Respironics, Bend, OR, USA) was placed on the subject's non-dominant wrist and the wireless system headband was undocked from the bedside display and placed on the subject's head. A sleep technician monitored each recording session by closed-circuit television with audio and streaming data from the PSG. Subjects were required to stay in bed for a minimum of 6 h and either awoke spontaneously or were awakened in the morning at their requested rise time. Lights out and lights on times were noted.

The three recording systems (PSG, WS and ACT) were time-synchronized before each recording. Polysomnography records were scored manually in 30-s epochs according to standard R&K criteria by two sleep scorers (PSG1 and PSG2). The two scorers were trained in separate laboratories: one is a registered polysomnographic technologist trained for work in both clinical and research settings (PSG1), the other was trained exclusively for work in a research setting (PSG2). ACT data were collected in 30-s epochs and scored for sleep and wakefulness at medium wake threshold sensitivity (wakefulness threshold value of 40 activity counts), according to Actiware 5.0 software (Mini Mitter Philips/Respironics, Bend, OR, USA). Scored data for the WS were extracted from the flash memory card.

Data analysis

The WS was compared with the PSG data scored according to R&K criteria. For sleep stage comparisons, PSG data scored as Stage 1 and Stage 2 were combined into a single category for comparison with the WS 'light sleep'. Similarly, PSG data scored as Stage 3 and Stage 4 were combined into a single category for comparison with the WS 'deep sleep'. For sleep/wakefulness comparisons, PSG, ACT and WS records were analyzed with binary scores (0 = wakefulness, 1 = sleep). Epochs that were scored by all three systems were included in analyses.

Pooled epoch-by-epoch agreement was established for sleep stages between the two PSG scorers and WS by calculating percent agreement and Cohen's kappa for each pair (WS versus PSG1, and WS versus PSG2, PSG1 versus PSG2). Cohen's kappa measures the agreement between two systems beyond what would be expected from chance alone (Cohen, 1960). A kappa value of 0–0.2 is considered essentially no agreement, 0.2–0.4 low agreement, 0.4–0.6 moderate agreement, 0.6–0.8 high agreement and 0.8–1.0 nearly perfect agreement (Landis and Koch, 1977). Contingency tables are provided for descriptive purposes showing the frequency of agreement/disagreement of epochs each system scored as wakefulness, REM sleep, light sleep or deep sleep. Percentage agreement and positive predictive values (PPV) for sleep stages were computed from the tables. Chi-squared analysis was used to assess differences in agreement between scorer pairs.

Sleep/wakefulness agreement was assessed by calculating percentage agreement and PPV for wakefulness and sleep for each pair (WS versus PSG1, WS versus PSG2, ACT versus PSG1, ACT versus PSG2). The percentage agreement and average PPV were also calculated to assess concordance between scorers (PSG1 versus PSG2).

A consensus score (PSGC) was derived from PSG1 and PSG2 wherein only epochs for which there was agreement between the two sleep scorers were included. This yielded the following additional pairs for analysis of sleep stages (WS versus PSGC) and sleep/wakefulness (WS versus PSGC, ACT versus PSGC).

Entire night agreement was established for sleep stages between the two PSG scorers and the WS by calculating percentage agreement, PPV for sleep stages and Cohen's kappa for each of the 26 subjects.

The following averaged nightly summary sleep measures were also calculated for each system: TST, sleep onset latency to the first epoch of sleep (SOL), latency to persistent sleep of

10 continuous min (LPS), wakefulness after sleep onset (WASO), sleep efficiency (SE), the number of awakenings lasting at least 2 min (NA) and wakefulness time during sleep (WTDS) between the first and last sleep epochs in the recordings. In addition, the following summary statistics were calculated for PSG1, PSG2 and WS: time in REM sleep (T_{REM}), time in light sleep (T_{light}), time in deep sleep (T_{deep}) and latency from the onset of the first epoch of sleep to the onset of the first epoch of REM sleep (REML). Normally distributed parameters (T_{REM}, T_{deep}) were tested by one-way repeated-measures analyses of variance (ANOVAS). If the model was significant ($P < 0.05$), individual pairs were compared by Tukey's honestly significant difference (HSD). Non-normally distributed parameters (determined by Shapiro–Wilk W tests, except T_{REM} and T_{deep}) were tested by Kruskal–Wallis tests for variance. If the model was significant ($P < 0.05$), individual pairs were compared by least significant difference between mean rank tests. Intraclass correlation coefficients (ICC) were calculated for each pair of scores for each summary measure to assess the agreements between systems. This method was chosen as the most appropriate method for assessing interobserver agreement, as it is sensitive to differences in the means of the observations (Fisher, 1954). Scatterplots were created comparing the WS and ACT to PSG. Pearson's correlation coefficient was calculated from the linear regressions.

RESULTS

An example of a typical night of data collection from all recording systems (from a 41-year-old female) is presented in Fig. 2, and provides a heuristic view of the abilities of the WS and ACT to score sleep and wakefulness. Horizontal bars depict sleep over time from each of the four scores. Representative samples of the raw WS headband signal together with the PSG for each sleep stage is available online (Fig. S1).

Epoch-by-epoch sleep stage

The WS reported a sleep stage for 97.4% of the 20 098 epochs between lights out and lights on, reporting unknown for the remainder. Both PSG scorers reported on 99.9% of epochs. Thus all systems yield was 97.3%, and the corresponding 19 556 epochs were included in the analysis.

Percentage agreement and Cohen's kappa for sleep stages are shown in Table 1. Agreement between the WS and PSG1 was greater than agreement between WS and PSG2 by 1.1% ($P = 0.01$).

Consensus between the PSG scorers was achieved on a total of 16 262 epochs for sleep stage scoring and 18 733 epochs for sleep/wakefulness scoring. Agreement between the WS and PSGC was higher than between the WS and the individual scorers by at least 5% (Table 1).

Fig. 3 presents contingency tables showing the observed frequencies of the number of epochs each system scored as wakefulness, REM, light or deep sleep. Agreement between the WS and PSG1 and PSG2 was 62 and 56%, respectively, for wakefulness, 85 and 79% for REM sleep, 82 and 80% for light sleep, and 60 and 67% for deep sleep (Fig. 3a,b). The frequency of epochs scored as light sleep by the WS when either PSG scorer scored wakefulness was 18 and 23%, respectively, and the frequency scored as REM by the WS when either PSG scorer scored wakefulness was 16 and 18%. The frequency of epochs scored as light by the WS when either PSG scorer scored deep was 39 and 33%, respectively.

The WS system agreement with the consensus score PSGC for each stage was: wakefulness 64%, REM 86%, light 86%, deep 71% (Fig. 3c). The frequency of epochs scored as light sleep by the WS when the PSGC score was wakefulness was 15%, and the frequency scored as REM sleep by the WS when

Table 1 Pooled epoch-by-epoch sleep stage agreement between systems

	% Agreement	*Cohen's kappa*
WS		
versus PSG1	75.8	0.62
versus PSG2	74.7	0.60
versus PSGC	81.1	0.70
PSG1		
versus PSG2	83.2	0.74

Epoch-by-epoch percentage agreement and Cohen's kappa between pairs of systems for sleep staging.
WS, wireless system; polysomnography (PSG)1, PSG scorer 1; PSG2, PSG scorer 2; PSGC, PSG consensus score. PSG scored according to Rechtschaffen and Kales guidelines.

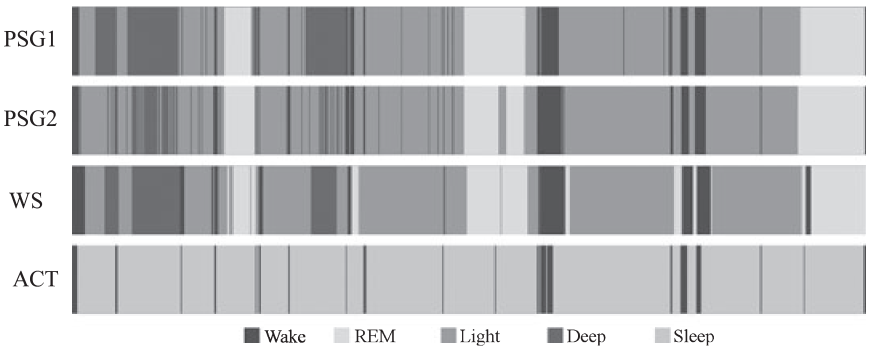

Figure 2. Representative data from one subject (41 years, female). Total recording time = 6 hours 43 minutes. PSG1–PSG scorer 1, PSG2 – PSG scorer 2, WS – wireless system, ACT – actigraph.

(a)

		PSG1			
		W	R	L	D
WS	W	1 667	63	374	9
	R	439	2 812	947	11
	L	480	427	8 365	1299
	D	92	20	574	1 977

Total = 19 556

(b)

		PSG2			
		W	R	L	D
WS	W	1 700	83	330	0
	R	537	2 967	698	7
	L	686	652	8 566	667
	D	114	32	1148	1 369

Total = 19 556

(c)

		PSGC			
		W	R	L	D
WS	W	1 562	61	218	0
	R	376	2 625	521	6
	L	423	332	7 680	537
	D	85	18	491	1 327

Total = 16 262

(d)

		PSG1			
		W	R	L	D
PSG2	W	2 446	51	540	0
	R	49	3 036	640	9
	L	180	235	8 910	1417
	D	3	0	170	1 870

Total = 19 556

Figure 3. Contingency tables for sleep stage recording systems. Contingency tables for sleep stage scoring pairs. Each row and column represents a state scored by a recording system for each 30-s epoch. Elements along the main diagonal of the table (upper left to lower right) indicate those epochs for which both systems in the pair were in agreement. Conversely, elements off the main diagonal indicate those epochs for which the two systems were not in agreement. W: wakefulness; R: rapid eye movement sleep; L: light sleep; D: deep sleep; WS: wireless system:, polysomnography (PSG)1: PSG scorer 1; PSG2: PSG scorer 2; PSGC: PSG consensus score. PSG scored according Rechtschaffen and Kales guidelines.

the PSGC score was wakefulness was 17%. The frequency of epochs scored as light sleep by the WS when the PSGC score was deep sleep was 29%.

On an epoch-by-epoch basis, the average agreement between the scorers PSG1 and PSG2 (using the method proposed by Norman *et al.*, 2000) for each stage was: wakefulness 86%, REM 86%, light 85%, deep 70% (Fig. 3d). One of the scorers scored deep more frequently than the other (17 versus 10% of epochs).

The PPV for each sleep stage represents the probability that the PSG was in agreement with the determination of the WS for a given epoch. For wakefulness the PPV was (WS versus PSG1, WS versus PSG2, WS versus PSGC) 78.9, 80.5, 84.8%, REM sleep 66.8, 70.5, 74.4%, light sleep 79.1, 81.0, 85.6%, and deep sleep 74.2, 51.4, 69.1%.

Entire-night agreement

Box-plots of night-by-night distributions of sleep stage agreement and kappa are shown for each pair of scorers in Fig. 4. The average (± SD) entire-night sleep stage agreement for the 26 subjects was 75.9% (7.0%) for the WS versus PSG1, 74.7% (8.5%) for the WS versus PSG2, and 81.2% (7.4%) for the WS versus PSGC. The average agreement between PSG scorers was 83.1% (8.1%). Pearson's correlation coefficients by night with sleep stage were (WS versus PSG1, WS versus PSG2): REM sleep (0.69, 0.69), light sleep (0.64, 0.75), deep sleep (0.71, 0.63). Pearson's correlation coefficients for PSG1 versus PSG2 were 0.84 for REM sleep, 0.84 for light sleep, and 0.74 for deep sleep. Scatter-plots are available online (Fig. S2).

Epoch-by-epoch sleep versus wakefulness

Percentage agreements and PPV for sleep and wakefulness are shown in Table 2. The WS comparisons against PSGC showed higher values for percentage agreement, PPV for sleep, and PPV for wakefulness versus the comparisons made against PSG1 or PSG2 alone.

Comparisons of ACT against the human scorers showed lower percentage agreements and PPV for sleep compared to the WS and lower PPV for wakefulness. ACT measures were not higher when comparisons were made against PSGC than for comparisons to either PSG1 or PSG2 alone.

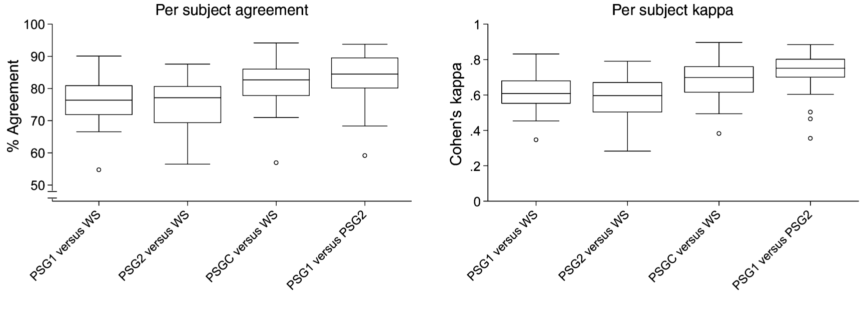

Figure 4. Distributions of per subject entire night percentage agreement and Cohen's kappa for each pair of sleep stage scorers. Each box shows the median and inter-quartile range. The whiskers show the upper and lower adjacent values.

Table 2 Epoch-by-epoch sleep/wakefulness agreement between systems

	% Agreement	*PPV Sleep (%)*	*PPV Wakefulness (%)*
WS			
versus PSG1	92.6	94.2	78.9
versus PSG2	91.1	92.3	80.5
versus PSGC	93.6	94.8	83.5
ACT			
versus PSG1	86.3	90.5	50.1
versus PSG2	85.7	89.2	55.8
versus PSGC	87.6	91.2	53.4
PSG1			
versus PSG2	95.8	97.5	85.6

Epoch-by-epoch percentage agreement, PPV sleep, and PPV wakefulness between pairs of systems for sleep/wakefulness determinations.
PPV, positive predictive value; WS, wireless system; ACT, actigraph; polysomnography (PSG)1, PSG scorer 1; PSG2, PSG scorer 2; PSGC–PSG consensus score. PSG scored according to Rechtschaffen and Kales guidelines.

Summary sleep measures

Table 3 shows the nightly averaged sleep stage and sleep/wakefulness summary measures for all four recording systems. Kruskal–Wallis tests revealed significant effects of scorer on SOL (df = 3, $\chi^2 = 21.2$, $P < 0.001$), LPS (df = 3, $\chi^2 = 8.4$, $P < 0.05$), WASO (df = 3, $\chi^2 = 8.2$, $P < 0.05$), NA (df = 3, $\chi^2 = 22.2$, $P < 0.001$), WTDS (df = 3, $\chi^2 = 7.9$, $P < 0.05$) and REML (df = 2, $\chi^2 = 26.7$, $P < 0.001$), but no significant effects of scorer on TST [df = 3, $\chi^2 = 4.4$, not significant (NS)], SE (df = 3, $\chi^2 = 7.1$, NS), or T_{light} (df = 2, $\chi^2 = 1.4$, NS). Repeated-measures ANOVAS revealed significant effects of scorer on T_{REM} ($F_{(2,25)} = 9.21$, $P < 0.001$) and T_{deep} ($F_{(2,25)} = 13.83$, $P < 0.001$).

There was a significant difference in T_{REM} between the WS and PSG1 but not the WS and PSG2 (Table 3). T_{deep} measures differed significantly between the WS and each PSG scorer, and T_{deep} also differed significantly between the two PSG scorers. The WS significantly underestimated REML compared to both PSG1 and PSG2. There were no other significant differences in sleep/wakefulness measures between the WS and PSG scorers.

Actigraph significantly underestimated SOL compared to each PSG scorer, and LPS compared to PSG2. Actigraph also significantly overestimated NA compared to PSG1.

Table 4 shows that ICC values between the WS and PSG scorers were greater than or equal to 0.90 for TST and SE. The ICC for SOL between the WS and PSG scorers was between 0.40 and 0.50, as was the ICC between the PSG scorers. Latency to persistent sleep concordance was > 0.80, both between the WS and PSG, and between the two PSG scorers. Intraclass correlation coefficients values between ACT and PSG were at least 0.20 lower than the correlations of the WS to PSG for each measure.

Intraclass correlation coefficients values between the WS and PSG were between 0.50 and 0.75 for all sleep stage

Table 3 Summary sleep measures

	PSG1	*PSG2*	*WS*	*ACT*
TST (min)*	324.6 (11.2)	317.7 (11.4)	335.4 (11.7)	336.9 (8.3)
SE (%)	86.1 (2.6)	84.1 (2.6)	88.9 (2.7)	89.4 (1.1)
SOL (min)*	12.7 (3.1)†	9.7 (2.0)†	7.8 (2.4)	2.4 (0.6)†
LPS (min)*	18.4 (4.2)	22.4 (4.6)†	17.4 (4.0)	9.5 (2.5)†
WASO (min)*	38.8 (6.9)	48.7 (7.7)	32.9 (7.7)	36.8 (3.1)
NA (#)*	2.58 (0.37)†	4.46 (0.66)	3.62 (0.55)¶	6.73 (0.66)†,¶
WTDS (min)*	33.8 (5.8)	43.7 (6.8)	31.2 (7.4)	35.6 (3.1)
T_{REM} (min)*	63.9 (4.5)‡	71.8 (6.3)	80.9 (5.1)‡	–
T_{light} (min)	197.3 (8.3)	206.6 (8.4)	203.3 (9.0)	–
T_{deep} (min)*	63.4 (6.4)‡,§	39.3 (5.7)‡,§	51.2 (5.4)‡	–
REML (min)*	95.3 (10.6)‡	96.0 (10.9)‡	31.9 (8.1)‡	–

Summary sleep measures for each recording system [mean ± standard error of the mean (SEM)]. *Significant main effects by analysis of variance (ANOVA) or Kruskal–Wallis ($P < 0.05$). *Significant main effects by ANOVA or Kruskal-Wallis ($P < 0.05$), †significant pairwise comparison between ACT and PSG1 and/or PSG2 ($P < 0.05$), ‡significant pairwise comparison between WS and PSG1 and/or PSG2 ($P < 0.05$), ¶significant pairwise comparison between ACT and WS ($P < 0.05$), §significant pairwise comparison between PSG1 and PSG2 ($P < 0.05$).
WS, wireless system; ACT, actigraph; polysomnography (PSG)1, PSG scorer 1; PSG2, PSG scorer 2; TST, total sleep time; SE, sleep efficiency; SOL, sleep onset latency; LPS, latency to persistent sleep; WASO, wakefulness after sleep onset; NA, number of awakenings; WTDS, wakefulness time during sleep; T_{REM}, time in rapid eye movement (REM) sleep; T_{light}, time in light sleep; T_{deep}, time in deep sleep; REML, REM latency. PSG scored according to Rechtschaffen and Kales guidelines.

Table 4 Sleep measure correlations

	WS versus PSG1	*WS versus PSG2*	*ACT versus PSG1*	*ACT versus PSG2*	*PSG1 versus PSG2*
TST (min)	0.95	0.92	0.60	0.63	0.98
SE (%)	0.95	0.90	0.36	0.33	0.96
SOL (min)	0.42	0.50	−0.07	0.13	0.48
LPS (min)	0.94	0.81	0.40	0.22	0.89
WASO (min)	0.90	0.85	0.21	0.14	0.92
NA (#)	0.60	0.69	−0.33	0.05	0.39
WTDS (min)	0.85	0.82	0.30	0.23	0.90
T_{REM} (min)	0.51	0.65	–	–	0.77
T_{light} (min)	0.74	0.65	–	–	0.82
T_{deep} (min)	0.65	0.57	–	–	0.52
REML (min)	0.02	−0.06	–	–	0.92

Intraclass correlation coefficients between each PSG scorer, WS and ACT. Coefficients range from −1 to 1, with 1 representing perfect direct correlation, and 0 representing a complete lack of correlation.
WS, wireless system; ACT, actigraph; polysomnography (PSG)1, PSG scorer 1; PSG2, PSG scorer 2; TST, total sleep time; SE, sleep efficiency; SOL, sleep onset latency; LPS, latency to persistent sleep; WASO, wakefulness after sleep onset; NA, number of awakenings; WTDS, wakefulness time during sleep; T_{REM}, time in rapid eye movement (REM) sleep; T_{light}, time in light sleep; T_{deep}, time in deep sleep; REML, REM latency. PSG scored according to Rechtschaffen and Kales guidelines.

measures except for REML, where correlations to the PSG scorers were 0.02 and −0.06. ICC values between PSG scorers were greater than 0.75 for T_{REM} and T_{light}, and 0.52 for T_{deep}.

Fig. 5 shows scatter-plots of sleep/wakefulness measures per subject per night for TST, SE and WASO. The correlation between the WS and PSG was higher than the correlation between ACT and PSG for these measures.

DISCUSSION

In the present study we observed good overall agreement between the WS and PSG sleep recordings scored according to R&K guidelines by two independent trained technicians. When comparing 'light' (Stages 1 and 2 combined), 'deep' (Stages 3 and 4 combined) and REM sleep, the epoch-by-epoch agreement was above 74% with a kappa of at least 0.60. The per subject night-to-night variability in agreement of the WS to PSG was similar to the variability between the two human scorers, as seen in Fig. 4.

The agreement between the WS and PSG scorers varied somewhat by sleep stage. Specifically, agreement between the WS and PSG scorers on light sleep and REM sleep was reasonably high (between 79 and 85%), and comparable to the PSG scorers' agreements with each other (between 85 and 86%). Agreement between the WS and PSG on deep sleep was somewhat lower (between 60 and 67%), but this was consistent with the scorers' average agreement with each other of 70%. The finding that the PSG scorers were not in high agreement on deep sleep is also consistent with the finding of 69.3% by Norman *et al.* (2000), in their study of five scorers at three centers. The agreement between the WS and PSG scorers for wakefulness (between 56 and 62%) was substantially lower than the agreement between PSG scorers of 86%, and although the difference did not reach significance, WASO was scored less by the WS than PSG (see Table 3). Further, the PPV was lower for REM sleep than the other stages, and the contingency tables (Fig. 3) reveal that the WS scores REM sleep during periods of wakefulness or light sleep as scored by PSG. Refinements to the WS should concentrate on improving these areas of disagreement.

The WS significantly and substantially underestimated REML compared to PSG. There were nine nights for which the WS scored REM within the first 6 min of sleep, possibly indicating a tendency for the technology to score REM in the early lightest stage of sleep. There were six instances for which the WS scored REML within 2 min of either of the PSG scorers, but the WS tended to greatly underestimate REML in the remaining 11 nights. This may be due to the lack of independent EOG channels in the WS. The WS provides a reasonable estimate of aggregate sleep stage measures, although the low correlation of the WS to PSG for REML indicates that the current version of the WS may not be suitable for measuring the latency to REM sleep.

One consideration when evaluating the new automated system for scoring sleep is the state of PSG as the 'gold standard'. The rules for scoring sleep set forth by R&K have several weaknesses in the face of actually describing sleep as a biophysiological process (Himanen and Hasan, 2000). Inter-scorer reliability, even in best-case scenarios, rarely exceeds about 94%, and for scorers trained in different laboratories, as is the case in the present study, rates of about 85% agreement are generally found (Himanen and Hasan, 2000; Norman *et al.*, 2000). Interpretation of results comparing a system to a gold standard that itself is inconsistent presents a challenge. The new sleep scoring standards set forth by the AASM (Iber *et al.*, 2007) were intended, at least in part, to address the limitations of the old standard in terms of interscorer reliability, and show promise of improvement over R&K (Grigg-Damberger, 2009). For example, Danker-Hopfe *et al.* (2009) found that the new standards had significantly higher inter-rater reliability than R&K. Despite this, overall agreement rates in that study were still in the range of 82% and

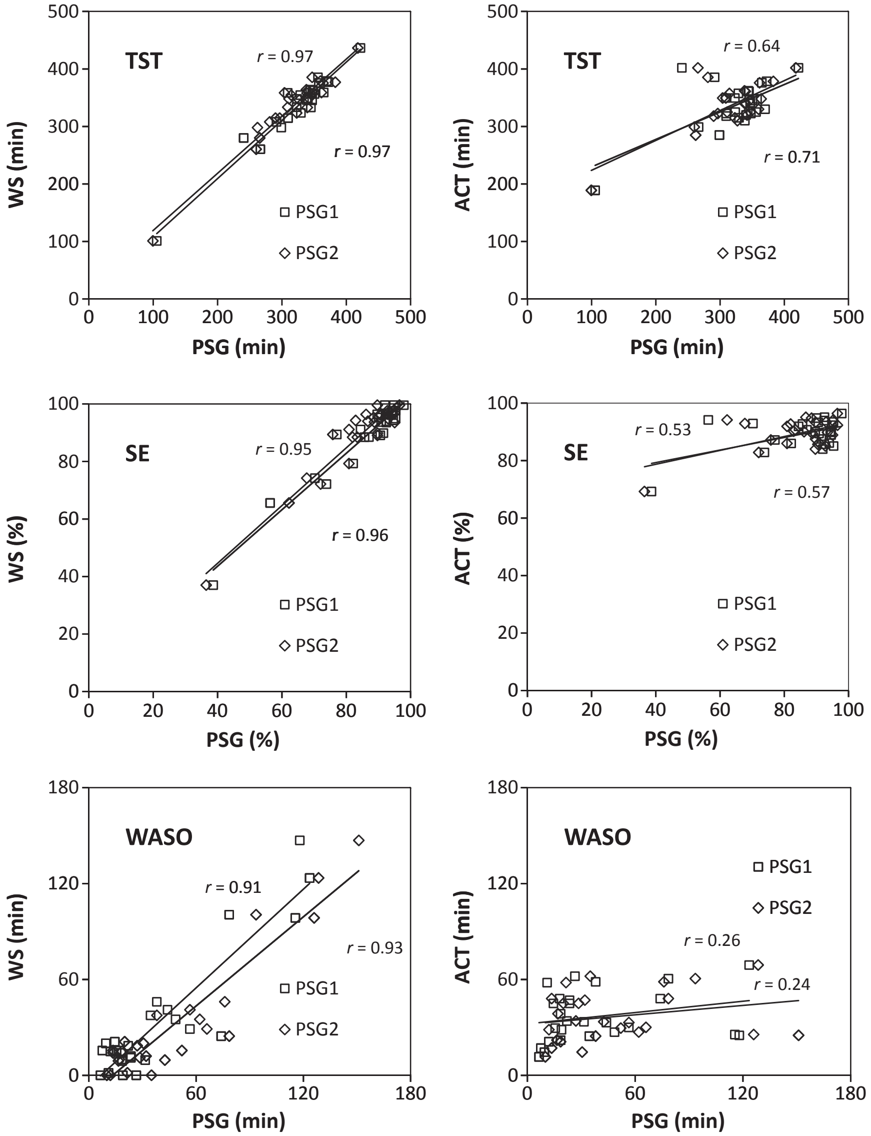

Figure 5. Scatter plots of sleep/wakefulness summary measures. Each point represents the estimate for one subject night by each of two scorers (PSG1 and PSG2). PSG1 – PSG scorer 1, PSG2 – PSG scorer 2, WS – wireless system, ACT – actigraph, TST – total sleep time, SE – sleep efficiency, WASO – wakefulness after sleep onset.

many of the same limitations of these rules as a gold standard still apply. Our results demonstrate that the WS accuracy is similar to but not quite as good as the two human scorers from different sleep centers compared to each other. This reduced accuracy compared to PSG may be a reasonable trade-off for the simplicity and ease of use of the WS in certain circumstances.

A common approach to dealing with inter-rater reliability is to use consensus scoring. This can be a laborious and time-consuming process that requires two technicians to come to agreement on all scored epochs or a third scorer to come to a conclusion upon reviewing the efforts of two independent scorers (Anderer *et al.*, 2005). In this study, we chose to include an analysis where the two human scorers agreed. Compared to true consensus scoring this method has disadvantages, namely the loss of information and risk for stage distribution bias. These risks were minimized in this study, because only 16.8% of epochs were removed from sleep stage analysis and only 4.2% of epochs were removed for sleep/wakefulness analysis, and because the distribution of stages was not affected significantly by the removal of disagreement epochs (see Fig. 3). The results of the epoch-by-epoch consensus comparison, as expected from traditional consensus scoring (Anderer *et al.*, 2005), showed an improved performance of the WS against a truer gold standard. The 81.1% overall agreement and kappa of 0.70 for sleep stage scoring presents what may be considered a more meaningful assessment of the performance of the WS. Regardless, for many of the same reasons that actigraphy is generally accepted as a useful tool for measuring sleep (see, particularly, Tryon, 2004), the WS shows potential as an effective way to measure sleep.

The WS performed better than ACT in differentiating sleep from wakefulness by all measures, and especially PPV for wakefulness. These differences were especially pronounced for nights of low SE. The agreement of ACT with PSG in this study was consistent with findings from previous studies showing agreements ranging from 83 to more than 90% (Ancoli-Israel *et al*., 2003; Kushida *et al*., 2001; Paquet *et al*., 2007; Pollak *et al*., 2001). The correlations between ACT and PSG for summary variables were at the low end of the range described by Tryon (2004), who reported a range of 0.72–0.98 for TST, 0.56–0.91 for SE and 0.49–0.87 for WASO (Fig. 5). A high level of variability in actigraphic equipment and algorithms has been described in the literature (Acebo and LeBourgeois, 2006), and it is important to consider that different wakefulness threshold levels can affect the sensitivity and specificity to sleep and wakefulness (Paquet *et al*., 2007). A medium wake threshold algorithm was used in this study, and an improvement in performance may be expected with a logarithmic regression algorithm (Paquet *et al*., 2007; Pollak *et al*., 2001). This option was not available for the software and equipment used in this study.

Although all subjects in the present study were self-reported good sleepers, and were given a night to acclimate to the laboratory environment, there were still some cases where the sleep quality was low for reportedly healthy sleepers. Lower sleep quality may also have contributed to instances of lower performance of ACT, as it has been reported that concordance of ACT with PSG is lower in nights with lower SE (Ancoli-Israel *et al*., 2003; Kushida *et al*., 2001; Lichstein *et al*., 2006). In contrast, the WS performance was more consistent over the range in sleep quality, and for those sleep measures where the WS had a lower agreement with PSG, the two human scorers exhibited similarly low concordance with each other (for example, with SOL and NA). The differences here, between the WS and ACT, may be due to the fact that actigraphy uses movement or lack of movement as a surrogate to infer a state of wakefulness or sleep, respectively, whereas the WS uses signals coming directly from the brain to differentiate wakefulness and sleep. It is important to note that because of the temporal smoothing in the WS algorithm it is not suitable for scoring single epoch intervals of wakefulness or arousals.

Prior attempts have been made to automate the process of scoring sleep, and with some exceptions (Caffarel *et al*., 2006; Villa *et al*., 1998) results have been generally encouraging (Park *et al*., 2000; Pittman *et al*., 2004; Schaltenbrand *et al*., 1996). Although these systems provide some advantage over manual scoring, they still require substantial time and cost in terms of equipment, set-up and trained expertise because of the need for full PSG. Further advantages could be gained with a system that also requires fewer signals, less expertise and is less obtrusive to the subject. The results of the present study suggest that the WS performs comparably to other systems that have attempted simple recording methods for automated sleep staging, such as Automatic Sleep EEG Analysis (ASEEGA) (Berthomier *et al*., 2007), C STAGE (Prinz *et al*., 1994) and a two electro-oculogram (EOG) montage tested by Virkkala *et al*. (2007). These other systems utilize traditional wired electrodes, requiring preparation and technical expertise. The WS, on the other hand, uses a dry fabric headband that requires no preparation, and has the potential to fill a need for a simple way to record sleep.

There are some limitations of this study which should be considered when evaluating the WS. Only one night's data was analyzed from each subject, all of whom were between 19 and 60 years of age. Internight intrasubject variability was not established, and information about the ability of the WS to score sleep in younger and older populations cannot be inferred from these results. The study subjects were self-reported healthy sleepers, and the WS performance in other populations should not be inferred from these results. Also, these data were collected in a laboratory setting. Additional studies evaluating the WS in the natural in-home environment are needed. The WS yield was slightly lower than the yield for PSG. This may have an effect on the accuracy of certain sleep summary measures on nights with missing epochs. However, given that widespread use of PSG in the home is impractical, this trade-off seems reasonable.

We conclude that the WS shows promise as a relatively accurate system for scoring sleep. The WS incorporates many of the benefits of PSG and actigraphy in one system, and may find utility as an alternative in certain circumstances. With additional validation in broader populations the WS may represent a useful tool for sleep medicine and research.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Paolo DePetrillo and Benjamin Rubin for their work in pioneering the wireless system technology, and would like to thank Richard Harter and Jesse Snider MA for their work in scoring the sleep records. The study was supported by Zeo, Inc., Newton, MA.

DISCLOSURES

This study was supported by Zeo, Inc. John Shambroom has been a salaried employee and consultant for Zeo, Inc. Stephan Fábregas is a salaried employee of Zeo, Inc. Dr Johnstone received financial support for the conduct of this study.

REFERENCES

Acebo, C. and LeBourgeois, M. K. Actigraphy. *Respir. Care Clin. N. Am*.,2006, 12: 23–30.

Ancoli-Israel, S., Cole, R., Alessi, C., Chambers, M., Moorcroft, W. and Pollak, C. P. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*, 2003, 26: 342–392.

Anderer, P., Gruber, G., Parapatics, S. *et al*. An e-health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24 vs. 7 utilizing the Siesta database. *Neuropsychobiology*, 2005, 51: 5–33.

Berthomier, C., Drouot, X., Herman-Stoïca, M. *et al*. Automatic analysis of single-channel sleep EEG: validation in healthy individuals. *Sleep*, 2007, 30: 1587–1595.

Caffarel, J., Gibson, G. J., Harrison, J. P., Griffiths, C. J. and Drinnan, M. J. Comparison of manual sleep staging with automated neural network-based analysis in clinical practice. *Med. Biol. Eng. Comput*., 2006, 44: 105–110.

Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 1960, 20: 37–46.

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.

Feige, B., Al-Shajlawi, A., Nissen, C. *et al.* Does REM sleep contribute to subjective wake time in primary insomnia? A comparison of polysomnographic and subjective sleep in 100 patients. *J. Sleep Res.*, 2008, 17: 180–190.

Fisher, R. A. *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh, 1954.

Grigg-Damberger, M. M. The AASM scoring manual: a critical appraisal. *Curr. Opin. Pulm. Med.*, 2009, 15: 540–549.

Himanen, S.-L. and Hasan, J. Limitations of Rechtschaffen and Kales. *Sleep Med. Rev.*, 2000, 4: 149–167.

Iber, C., Ancoli-Israel, S., Chesson, A. and Quan, S. F. for the American Academy of Sleep Medicine. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specifications*. American Academy of Sleep Medicine, Westchester, IL, 2007.

Kushida, C. A., Chang, A., Gadkary, C., Guilleminault, C., Carrillo, O. and Dement, W. C. Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients. *Sleep Med.*, 2001, 2: 389–396.

Landis, J. R. and Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics*, 1977, 33: 159–174.

Lichstein, K. L., Stone, K. C., Donaldson, J. *et al.* Actigraphy validation with insomnia. *Sleep*, 2006, 29: 232–239.

Means, M. K., Edinger, J. D., Glenn, D. M. and Fins, A. I. Accuracy of sleep perceptions among insomnia sufferers and normal sleepers. *Sleep Med.*, 2003, 4: 285–296.

Norman, R. G., Pal, I., Stewart, C., Walsleben, J. A. and Rapoport, D. M. Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep*, 2000, 23: 901–908.

Paquet, J., Kawinska, A. and Carrier, J. Wake detection capacity of actigraphy during sleep. *Sleep*, 2007, 30: 1362–1369.

Park, H. J., Oh, J. S., Jeong, D. U. and Park, K. S. Automated sleep stage scoring using hybrid rule- and case-based reasoning. *Comput. Biomed. Res.*, 2000, 33: 330–349.

Pittman, S. D., MacDonald, M. M., Fogel, R. B. *et al.* Assessment of automated scoring of polysomnographic recordings in a population with suspected sleep-disordered breathing. *Sleep*, 2004, 27: 1394–1403.

Pollak, C. P., Tryon, W. W., Nagaraja, H. and Dzwonczyk, R. How accurately does wrist actigraphy identify the states of sleep and wakefulness? *Sleep*, 2001, 24: 957–965.

Prinz, P. N., Larsen, L. H., Moe, K. E., Dulberg, E. M. and Vitiello, M. V. C STAGE, automated sleep scoring: development and comparison with human sleep scoring for healthy older men and women. *Sleep*, 1994, 17: 711–717.

Rechtschaffen, A. and Kales, A. *A Manual of Standardized Terminology, Techniques, and Scoring System for Sleep Stages of Human Subjects*. University of California, Brain Information Service/Brain Research Institute, Los Angeles, CA, 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.

Tryon, W. W. Issues of validity in actigraphic sleep assessment. *Sleep*, 2004, 27: 158–165.

Villa, M. P., Piro, S., Dotta, A. *et al.* Validation of automated sleep analysis in normal children. *Eur. Respir. J.*, 1998, 11: 458–461.

Virkkala, J., Hasan, J., Värri, A., Himanen, S.-L. and Müller, K. Automatic sleep stage classification using two-channel electrooculography. *J. Neurosci. Methods*, 2007, 166: 109–115.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1. Representative signals from the wireless system headband and PSG system depicting 30 second epoch examples for wakefulness, REM, light, and deep sleep.

Figure S2. Scatter plots of sleep stage summary measures by night. Each point represents the estimate for one subject night by each of two scorers (paired combinations of WS, PSG1, and PSG2). PSG1 – PSG scorer 1, PSG2 – PSG scorer 2, WS – wireless system.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.