J. Sleep Res. (2010) 19, 612–619 Actigraphy

doi: 10.1111/j.1365-2869.2010.00835.x

Wrist actigraphic scoring for sleep laboratory patients: algorithm development

DANIEL F. KRIPKE, ELIZABETH K. HAHN, ALEXANDRA P. GRIZAS, KEP H. WADIAK, RICHARD T. LOVING, J. STEVEN POCETA, FARHAD F. SHADAN, JOHN W. CRONIN and LAWRENCE E. KLINE

Scripps Clinic Sleep Center, La Jolla, CA, USA

Accepted in revised form 16 December 2009; received 10 November 2009

SUMMARY Wrist actigraphy is employed increasingly in sleep research and clinical sleep medicine. Critical evaluation of the performance of new actigraphs and software is needed. Actigraphic sleep–wake estimation was compared with polysomnographic (PSG) scoring as the standard in a clinical sleep laboratory. A convenience sample of 116 patients undergoing clinical sleep recordings volunteered to participate. Actiwatch-L recordings were obtained from 98 participants, along with 18 recordings using the newer Spectrum model (Philips Electronics), but some of the actigraphic recordings could not be adequately aligned with the simultaneous PSGs. Of satisfactory alignments, 40 Actiwatch recordings were used as a training set to empirically develop a new Scripps Clinic algorithm for sleep–wake scoring. The Scripps Clinic algorithm was then prospectively evaluated in 39 Actiwatch recordings and 16 Spectrum recordings, producing epoch-by-epoch sleep–wake agreements of 85–87% and kappa statistics averaging 0.52 (indicating moderate agreement). Wake was underestimated by the scoring algorithm. The correlations of PSG versus actigraphic wake percentage estimates were $r = 0.6690$ for the Actiwatch and $r = 0.2197$ for the Spectrum. In general, using a different weighting of activity counts from previous and subsequent epochs, the Scripps Clinic algorithm discriminated sleep–wake more successfully than the manufacturer's Actiware algorithms. Neither algorithm had fully satisfactory agreement with PSG. Further evaluations of algorithms for these actigraphs are needed, along with controlled comparisons of different actigraphic designs and software.

KEYWORDS actigraph, algorithm, motor activity, polysomnography, sleep, wake

INTRODUCTION

Wrist actigraphic recording has become increasingly important in sleep research and clinical sleep medicine. In a review of 2007 research abstracts, 23% of studies with human sleep recording employed actigraphy alone or in combination with polysomnography (PSG) to measure sleep (Loving *et al.*, 2008). Because actigraphy is often employed for studies in which large numbers of subjects are recorded for several nights (or 24-h intervals), we estimate that the percentage of research sleep recordings that utilize actigraphy now exceeds 25% on a per-night basis.

Actigraphic sleep estimation is often employed without attention to the validity of the sleep-scoring software or its optimization for a particular application. It has been shown that actigraphic algorithms sometimes perform differently on different subject groups (Cole *et al.*, 1992; Jean-Louis *et al.*, 2001a). Moreover, because the sensitivity and response characteristics of commercial actigraphs vary widely, the scoring algorithm is likely to score sleep poorly when applied to the output from a different actigraph than that for which it was optimized. Several examples of this poor performance are

Correspondence: Daniel F. Kripke, MD, Scripps Clinic Sleep Center, W-207, 10666 North Torrey Pines Road, La Jolla, CA 92037, USA. Tel.: +1-858-554-8839; fax: +1-858-554-8492; e-mail: kripke.daniel@scrippshealth.org

available (de Souza *et al*., 2003; Verbeek *et al*., 1994), in which we have seen algorithms from our own academic colleagues applied to different actigraphs from those with which they were derived.

We were interested in the performance of one of the most popular commercial actigraphs, the Actiwatch-L, a product developed by Minimitter (Bend, OR, USA) and taken over by Respironics, now absorbed into Philips Electronics (Philips Respironics, OR, USA). We could locate very little empirical basis for the derivation of its Actiware scoring software. Moreover, several validation studies suggested that the Actiware algorithms lack specificity for sleep; that is, the scoring underestimates wake during nocturnal PSG (Kushida *et al*., 2001; Paquet *et al*., 2007). A simple change in the algorithm improved its performance (Chae *et al*., 2009). In this study, we applied an empirical optimization approach that had served to develop more successful algorithm parameters for earlier actigraphs (Cole *et al*., 1992; Jean-Louis *et al*., 2001a). A new and independent Scripps Clinic scoring algorithm for Actiwatch was derived. Because the same manufacturer has developed a new actigraph, the Spectrum model, which employs the same Actiware sleep-scoring routines, we also evaluated the application of our Scripps Clinic algorithm to the Spectrum actigraph.

MATERIALS AND METHODS

Without systematic selection, patients scheduled for PSG were asked if they would participate in this study (Table 1). Almost all the PSGs were performed partly to screen for sleep apnea, but in some PSGs the evidence for apnea was not sufficient to require treatment. Usually participants were recruited on the evening they had arrived at the Sleep Laboratory to prepare for clinical PSG. The study was explained and written informed consent was obtained. Subjects consented to wear an actigraph (Actiwatch-L or Spectrum) for a brief period before their clinical PSGs, during the PSGs and for a short duration thereafter. They also consented to make their PSG data available for research use.

Sometimes the actigraph was placed on the patient's wrist by the Research Assistant 1–3 h before the recording commenced, allowing up to several hours of waking recording. The laterality of wrist placement was not systematically controlled. At other times, the actigraph was placed on the subject by the PSG technician shortly before the lights were turned out at the start of PSG recording. The PSG technicians usually removed the actigraphs from a few minutes to an hour after awakening the subjects. The actigraph was later connected to its interface to recover the digital file of the wrist activity and light recording.

Table 1 Characteristics of study participants

	Mean	*SD*	*Range*
Actiwatch subjects n = 98			
Age (years)	51	13	23–82
AHI	26	28	0–124
PLMI	13	23	0–112
Spectrum subjects n = 18			
Age (years)	54	29	28–80
AHI	24	27	0–94
PLMI	15	28	0–108

AHI, apneas + hypopneas per hour index; PLMI, periodic leg movement index.

The PSGs and simultaneous infrared video were recorded digitally with Compumedics Profusion PSG 2.10.0.131 software on Compumedics E Series platforms. Efforts were made in advance to synchronize the time clocks of the computers used to record the PSGs with those used to initialize the actigraphs. At a later time, the chief technician used Compumedics software (Charlotte, NC, USA) to assist in scoring the PSGs and to output Excel files listing the sleep scoring in 30-s epochs, the timing of each scored periodic limb movement, and the timing of each scored arousal. When the night PSG technician had noted the lights-out time and lights-on time on the Compumedics PSG, these times were transferred to the Excel record, but these notations were not always available. About half way through the study, we started scoring the lights-out and lights-on moments visualized from the digital video accompanying the PSG. Inadvertently, some of the PSG and video records had been terminated at least a few seconds before the lights were turned on and the patient awakened.

After the actigraphic files and Excel data files were assembled, an investigator (D.F.K) estimated the lights-off time and the lights-on time from each Actiwatch light and activity record. Likewise, the lights-out and lights-on times were estimated from the PSGs. A subset of some of the earlier recordings was previously described by Chae *et al*. (2009). The 30-s PSG sleep stage scores were reformatted to scores of 1 for wake and 0 for sleep stages 1–4 or rapid eye movement. Actigraphic activity counts (whether originally from 15 or 60-s epochs) were adjusted to 30-s epochs. A macro routine then plotted the interpolated activity counts and the PSG scoring of sleep versus wake on the computer screen, using the lights-out and lights-on estimates for the initial alignment. The investigator then iteratively varied the alignment and actigraph compression parameters to achieve the best possible correspondence. The best correlation between the activity counts and the PSG wake–sleep states for each 30-s PSG epoch was the criterion used to indicate the best alignment, but correlations of 0.8 or 0.9 could not be expected, as the scores of 1 and 0 for PSG wake and sleep could not be precisely proportional to the widely varying actigraphic activity counts. Actigraphic sleep–wake scores were not utilized for optimizing alignment, so that actigraphic scoring would not bias the alignments.

After ranking the 98 Actiwatch recordings by alignment correlation, each two recordings of adjacent rank were paired, so that one of the pair could be randomly assigned to a training set of 49 recordings and the other member of the pair could be reserved for a prospective validation set of 49 records. This assured that training and validation sets had comparable alignments.

Using another Excel Visual Basic macro, a highly iterative program was applied to each Actiwatch recording to search for the optimal parameters for a sleep–wake scoring algorithm to score each Actiwatch epoch. The algorithm was initiated in the form D = Scaler * $\Sigma(b_{-10}x_{-10} + b_{-9}x_{-9}.... \quad b_0x_0.... \quad b_{10}x_{10})$, where D was the scaled polynomial sum of activity scores for 21 epochs of 30 s; Scaler was an overall scaling parameter; x_{-10} represented the activity count for the 10th epoch preceding the 30-s epoch being scored; x_{-9} represented activity for the ninth epoch preceding the epoch being scored; x_0 represents the activity count of the epoch being scored; x_{10} represented the 10th epoch after that being scored, etc.; and b_{-10} etc. represented a scaling parameter for each corresponding value of x. When $D \geq 1$, the epoch was scored wake – otherwise sleep, according to this algorithm. The optimal parameters were determined for each of the 49 training recordings individually. Then, from examination of the distributions of each coefficient b_n, especially the median b_n for the 49 recordings, an optimal set of parameters was selected (Table 2), along with a Scaler value of 0.204. It was estimated that the optimal parameter was 0.0 for the 3rd–10th epochs after the epoch being scored, so those epochs were omitted from further iterations of the scoring algorithm.

A secondary rescoring algorithm was derived from previous work (Jean-Louis *et al*., 2001a; Webster *et al*., 1982). This secondary algorithm scanned the primary algorithm scoring of 30-s epochs until 10 epochs in a row (5 min) of continuous sleep were noted. All of the initial epochs before the first 5 min scored continuously as sleep were rescored as wake. Unlike the Actiware software (Minimitter) (and the study of Chae *et al*., 2009), no secondary rescoring algorithm was applied at the end of the record, where scores of alternating wake and sleep were retained, based on the belief that there is no evidence that the last minutes of sleep should be rescored when occasional epochs of wake are observed just prior to the last awakening.

For comparison, the same Actiwatch records were rescored for wake–sleep with the manufacturer's program Actiware 5.01.0007. The Actiware algorithm sleep scores were aligned with the 30-s PSG epochs, using the same offset and compression parameters that had been previously optimized for raw actigraphic counts. When part of a 30-s epoch had been scored sleep and part wake by Actiware, the sleep–wake score interpolated to the larger part of the 30-s record was used, but when there was an exact tie, the sleep–wake score of the prior Actiware epoch was selected.

Table 2 Optimal scoring parameters for relative epochs

Epoch	*Parameter*
x_{-10}	0.0064
x_{-9}	0.0074
x_{-8}	0.0112
x_{-7}	0.0112
x_{-6}	0.0118
x_{-5}	0.0118
x_{-4}	0.0128
x_{-3}	0.0188
x_{-2}	0.0280
x_{-1}	0.0664
x_0	0.0300
x_1	0.0112
x_2	0.0100
x_3	0.0000

All parameters for 3–10 epochs after the current epoch were 0.0.

RESULTS

Actiwatch-L subjects recorded were 64 men and 34 women. Some characteristics are shown in Table 1.

The mean PSG recording was 7 h and 44 min (SD 39 min). The mean wrist activity recording was 9 h and 2 min (SD 56 min). The maximal alignment indicated that two recorded 15-s Actiwatch epochs or 0.5 1-min epoch averaged 0.9998 (SD = 0.0029) of the duration of a 30.0-s PSG epoch. The correlations of sleep–wake (scored as 0 or 1) and aligned interpolated 30-s actigraphic counts ranged from $r = 0.2285$ to $r = 0.6394$. The mean alignment correlation for the 49 training recordings was $r = 0.3952$ and the mean for the 49 prospective recordings was $r = 0.3950$.

Polysomnographic-scored arousals and leg movements were also explored as potential correlates of actigraphic counts, which might be used for alignment. In general, the correlations of wrist activity with arousals and with leg movements (periodic limb movements of sleep; PLMS) recorded by PSG were very low. Indeed, these correlations were often negative, presumably because arousals and leg movements were not scored during PSG wake epochs, when wrist activity was highest. The average correlation for activity versus arousals was $r = 0.008 \pm$ SD 0.074, and most of the correlations were negative. The average correlation for leg movements (in 30 records with PLMS sufficiently frequent to analyze) was $r = -0.039 \pm 0.107$, and most correlations were negative. Therefore, arousals and leg movements could not be employed for optimizing the alignment of PSGs and actigraphs.

From the optimized scores for the 49 training records, a plot was constructed of the alignment correlation versus the percent agreement between the PSG and actigraphic recordings for epoch-by-epoch sleep–wake scores. Those recordings with particularly poor alignment correlations had notably poorer sleep–wake scoring agreement percentages. Thus, those recordings with alignment correlations of $r < 0.32$ were excluded, leaving 40 recordings remaining in the training set and 40 in the prospective validation set. It appeared that the alignment of actigraphic and PSG records for most of the 80 records retained were accurate within 15 s or less.

The training set of 40 recordings was then scored with the fixed values of the optimal b_n coefficients that had been selected (Table 2) and a common value for the Scaler. This experiment was repeated with several Scaler values both before and after applying the secondary rescoring algorithm (Table 3). The first two rows of Table 3 show results for the training set with two different Scaler values, before applying secondary rescoring.

Table 3 Results of varying scoring parameters for Scripps Clinic and Actiware algorithms

Scripps Clinic scoring algorithms	*Agreement*	κ	*PSG Wake*	*Acti Wake*	$r_{Wake\%}$ *(PSG versus Acti)*
Training set, Scaler = 0.204	0.8884	0.4903	0.1644	0.1058	0.6160
Training set, Scaler = 0.280	0.8848	0.5179	0.1644	0.1364	0.5854
Training set, Scaler = 0.204, 2A	0.8915	0.5143	0.1644	0.1109	0.5712
Training set, Scaler = 0.240, 2A	0.8902	0.5282	0.1644	0.1253	0.5738
Training set, Scaler = 0.280, 2A	0.8887	0.5416	0.1644	0.1416	0.5808
Training set, Scaler = 0.300, 2A	0.8865	0.5428	0.1644	0.1491	0.5785
Training set, Scaler = 0.320, 2A	0.8824	0.5416	0.1644	0.1585	0.5360
Training set, Scaler = 0.360, 2A	0.8730	0.5377	0.1644	0.1777	0.4245
Prospective set, Scaler = 0.300, 2A	0.8595	0.5054	0.1725	0.1689	0.4329
Prospective set, Scaler = 0.300, 2A, one outlier excluded	**0.8700**	**0.5178**	**0.1755**	**0.1586**	**0.6690**
Actiware scoring algorithms	*Agreement*	κ	*PSG Wake*	*Acti Wake*	$r_{Wake\%}$ *(PSG versus Acti)*
Threshold low, 2A	0.8192	0.4345	0.1724	0.1857	0.0834
Threshold medium, 2A	0.8741	0.4426	0.1724	0.1248	0.2232
Threshold high, 2A	0.8693	0.3936	0.1724	0.0774	0.6611
Threshold high, one outlier excluded, 2A	**0.8696**	**0.3991**	**0.1755**	**0.0760**	**0.7354**
Auto-threshold, 2A	0.8466	0.3106	0.1724	0.0770	0.1294

2A: secondary scoring adjustment applied. If not noted, this adjustment was not applied.
Agreement: portion of the record for which PSG scoring and actigraphic scoring agreed, for example, 1.0000 would represent 100% agreement.
Kappa: Cohen's kappa is a measure of scoring concordance corrected for scoring biases.
PSG Wake: portion of the record scored Wake by PSG.
Acti Wake: portion of the actigraphic record scored Wake by the algorithm.
r_{Wake}(PSG versus Acti): correlation between PSG and actigraphic scoring of Wake% for $n = 40$ or $n = 39$ records.
All Actiware scoring results were for the prospective validation set ($n = 40$ or $n = 39$).
PSG, polysomnogram.
Bold values emphasize the appropriate contrast between the best *Scripps clinic algorithms* and *Actiware algorithm* results.

On an epoch-by-epoch basis, the Scaler value of 0.204 gave slightly better agreement of actigraphic sleep–wake scoring versus PSG as the standard, and a slightly better correlation between PSG wake percentage and actigraphic wake percentage for the 40 training records. On the contrary, with a Scaler of 0.204, the actigraphic scoring estimated that only 10.584% of the recordings were wake, markedly below the PSG estimate of 16.44%. With the Scaler at 0.280, which increased the *D*-scores, actigraphic scoring yielded a higher wake estimate of 13.65% and, thus, less bias, as well as a higher kappa. Cohen's kappa statistic for concordance indicates the percentage of the scoring agreement between actigraphic scoring and PSG not attributable to chance, providing a correction for effects of varying biases upon scoring agreement. When the secondary rescoring algorithm (which scores more wake at the start of the sleep record) was applied, the percent wake was increased for each value of the Scaler, as might be expected. A Scaler value of 0.300 with the secondary algorithm was selected as the best compromise, yielding little bias in estimated Wake percentage as referenced to the PSGs, very good epoch-by-epoch agreement, a good correlation of actigraphy and PSG wake percentages from night to night, and the best kappa. A Scaler of 0.300 with secondary rescoring was therefore considered the optimized scoring algorithm for the training set.

Table 3 also provides the scoring of the prospective data set by the best Scripps algorithm ($n = 40$). It was noted that in the prospective validation set, there was one outlier with very poor scoring agreement, resulting from a long interval of high actigraphic activity scored by PSG as sleep (a pattern sometimes observed when the actigraph is situated such that it responds to respiratory or cardiac motion). Tabulation of the prospective results is also shown in Table 3 excluding this outlier, leaving 39 records for prospective scoring. As compared with the training set, in the prospective validation set with outlier excluded, the algorithm yielded slightly lower scoring agreement and kappa, almost the same bias, and a slightly better correlation between recordings of estimated actigraphic and PSG wake percentages.

For comparison, results from the manufacturer's Actiware scoring program are shown at the bottom of Table 3, using the same 40 or 39 prospective validation recordings aligned identically, and incorporating Actiware's secondary scoring algorithm using the 5-min immobility criterion recommended by Chae *et al*. (2009). Actiware results were compiled using its low, middle and high thresholds of 20, 40, and 80 counts, and its automatic threshold, which averaged 194 counts for the prospective validation set after removing the outlier. As might be expected, the higher the threshold for activity to be scored as wake, the lower the percent wake (Wake%) scored by Actiware (Table 3). The low-threshold Actiware scoring had little bias in Wake%, but its epoch-by-epoch agreement and correlation across records were very poor by comparison to the other algorithms. The high threshold produced an excellent correlation of the Wake% between Actigraph and PSG records, and fairly good epoch-by-epoch agreement, but the bias was severe and kappa was quite low. With the high

Table 4 Activity counts/epoch of four actigraphs during manual shaking

	Spectrum1	*Spectrum2*	*Actiwatch1*	*Actiwatch2*
Gentle	277	266	268	127
Hard	1657	1542	2226	1096
Ratio hard/gentle	6.0	5.8	8.3	8.6

threshold excluding the outlier, Actiware scored only 7.60% Wake, when the PSG had scored 17.55%. In comparison with the Actiware high-threshold scoring, the Scripps Clinic algorithm for Actiwatch with Scaler of 0.300 and secondary scoring produced markedly less bias, superior epoch-by-epoch scoring kappa, and almost as high a correlation between actigraph and PSG for estimated Wake% (Table 3, both highlighted in bold type). The kappa value for the Scripps Clinic algorithm prospective scoring [$\kappa = 0.5178$ (0.4787–0.5569, 95% CI)] indicated substantially better concordance of scoring than the Actiware algorithm with the best kappa ($P < 0.01$).

Sensitivity analyses

Eight of the training set and nine of the prospective set were recorded with 60-s actigraphic epochs. There were no significant differences between 15- and 60-s actigraphic epochs in the performance of the Scripps Clinic scoring algorithm for Actiwatch: that is, the scoring of the interpolated 30-s epochs agreed equally with PSG scoring. However, with the Actiware scoring, the middle- and high-threshold scores from interpolated 60-s epochs were not as accurate as those derived from 15-s scores. There were 35 recordings split between undisturbed and CPAP titration portions, of which 20 were in the prospective set. For the Scripps Clinic algorithm, no significant differences in scoring bias or kappa was found between the group of recordings with split CPAP titration recordings and those recordings that were not split. The correlation between actigraphic kappa and scoring bias was $r = -0.21$ (NS). There were no significant differences in kappa or bias between male and female participants. Likewise, for the Scripps Clinic algorithm, there were no significant correlations of kappa or scoring bias with a participant's age, number of obstructive apneas, AHI (apneas + hypopneas per hour index) or periodic leg movement index (PLMI). The scoring kappas of the Actiware algorithms were significantly correlated with age: $r = 0.36$, $r = 0.41$, $r = 0.38$ and $r = 0.40$ for the low, middle, high and automatic thresholds ($P \sim 0.01$), that is, scoring was more accurate for older participants. However, the kappas of the Actiware algorithms were not significantly correlated with gender, AHI or PLMI.

To compare the electromechanical sensitivity of the Actiwatch and Spectrum actigraphs, the four actigraphs used in this research were tightly taped together and then subjected to several minutes of gentle shaking, followed by several minutes of hard shaking. As shown in Table 4, the responses of the two new Spectrums were very similar, but the responses of the older Actiwatches differed considerably both from the Spectrum actigraphs and from each other.

Algorithm applied to Spectrum actigraphs

After the Actiwatch recordings had been completed, 18 Spectrum recordings were obtained in the same way from 12 men and 6 women (see Table 1). The Spectrum recordings were aligned to the PSG in the same way, yielding a mean alignment of $r = 0.3747$. Two records with alignment < 0.30 were excluded, leaving 16 recordings to be tested prospectively with the same optimal Scripps Clinic algorithm developed for the Actiwatch. The results of the Scripps Clinic algorithm prospectively applied to Spectrum records are shown at the top of Table 5. Although epoch-by-epoch agreement was slightly less, kappa for the Spectrum records was a fraction better than with Actiwatch-L. The Scripps Clinic algorithm produced

Table 5 Results of prospective Scripps Clinic algorithm and Actiware scoring for Spectrum records

Scripps Clinic scoring algorithm	*Agreement*	κ	*PSG Wake*	*Acti Wake*	$r_{Wake\%}$ *(PSG versus Acti)*
Prospective set, Scaler = 0.300, 2A	**0.8504**	**0.5209**	**0.2109**	**0.1580**	**0.2197**
Actiware scoring algorithms	*Agreement*	κ	*PSG Wake*	*Acti Wake*	$r_{Wake\%}$ *(PSG versus Acti)*
Threshold low, 2A	0.8408	0.4729	0.2062	0.1614	0.3074
Threshold medium, 2A	0.8405	0.4323	0.2062	0.1190	0.2531
Threshold high, 2A	0.8321	0.3511	0.2062	0.0805	0.2072
Auto-threshold, 2A	0.8288	0.3179	0.2062	0.0677	0.1959

2A: secondary scoring adjustment applied.
Agreement: portion of the record for which PSG scoring and actigraphic scoring agreed, for example, 1.0000 would represent 100% agreement.
Kappa: Cohen's kappa is a measure of scoring concordance corrected for scoring biases.
PSG Wake: portion of the record scored Wake by PSG.
Acti Wake: portion of the actigraphic record scored Wake by an actigraph algorithm.
r_{Wake}[PSG versus Acti]: correlation between PSG and actigraphic scoring of Wake% for $n = 16$.
All scoring results were for the Spectrum prospective validation set with two outliers excluded ($n = 16$).
PSG, polysomnogram.
Bold values emphasize the appropriate contrast between the best *Scripps clinic* and *Actiware algorithms* results.

considerable bias in underestimating the Wake% in the Spectrum records, and the correlation of Wake% estimates between PSG and Spectrum records was unsatisfactory.

The aligned Spectrum activity counts were exported in a form readable by Actiware 5.01.0007, as we thought it more important to test scoring performance of identical Actiware software for both actigraphs than to evaluate the newer version of Actiware software designed for the Spectrum instruments, which scores with the same algorithm. The outcomes are shown at the bottom of Table 5. The low-threshold recording produced epoch-by-epoch agreement and kappa almost as good as the Scripps Clinic algorithm, and there was slightly less bias in the underestimated Wake%. Because Actiware does not score the first four and last four epochs of each aligned record (which tended to be waking time), the corresponding PSG Wake% was slightly lower for Actiware evaluations than that used for the Scripps Clinic algorithm. The correlation of wakefulness estimates by Spectrum and PSG was slightly better for the Actiware low threshold than for the Scripps Clinic algorithm, but none of these correlations was statistically significant or significantly different from each other. The higher Actiware thresholds produced progressively poorer results with the Spectrum records. The mean automatic threshold was 108.

In sensitivity analyses of the Scripps Clinic algorithm scoring of Spectrum records, females had higher kappas ($P < 0.05$), and the leg movement index was correlated with lower kappa ($r = -0.56$, $P < 0.05$) and greater bias ($r = 0.54$, $P < 0.05$), but these associations with nominal $P < 0.05$ would not be significant after correction for multiple testing.

DISCUSSION

In this study, we developed a new Scripps Clinic sleep–wake scoring algorithm for use with the Actiwatch-L wrist actigraph, using a sample of Sleep Center patients undergoing clinical PSGs. The optimal algorithm was:

$$\begin{aligned} D = 0.30 \times [&0.0064 \times x_{-10} + 0.0074_{-9} + 0.0112 \times x_{-8} \\ &+ 0.0112 \times x_{-7} + 0.0118 \times x_{-6} + 0.0118 \times x_{-5} \\ &+ 0.0128 \times x_{-4} + 0.0188 \times x_{-3} + 0.0280 \times x_{-2} \\ &+ 0.0664 \times x_{-1} \\ &+ 0.0300 \times x_0 + 0.0112 \times x_1 + 0.100 \times x_2] \end{aligned}$$

where epoch x is scored wake if $D \geq 1.00$, but otherwise it is scored sleep, and x_{-10} to x_2 represent the Actiwatch interpolated activity counts for 30-s epochs from 10 epochs before the epoch being scored (i.e. x_{-10}) to two epochs after epoch x being scored (i.e. x_2). A postscoring algorithm is then applied, rescoring all 30-s epochs scored sleep to wake until the first 10 consecutive epochs have $D < 1.00$.

This Scripps Clinic algorithm was validated with a prospective set of actigraphic records and clinical PSGs. However, it has been validated mainly for use with the Actiwatch and only then for in-bed records with a mostly middle-aged sample of sleep clinic patients, most with some degree of sleep apnea and periodic limb movements (Table 1). It has not yet been tested with 24-h recordings because of the difficulties of obtaining 24-h ambulatory PSG data for comparison.

The Spectrum actigraph is claimed by the manufacturer to have the same sensitivity as the Actiwatch and to utilize the same sleep-scoring algorithm. However, the Spectrum has a newer accelerometric transducer. Table 4 would suggest that there may be some differences in response between the Actiwatch-L and Spectrum designs or, alternatively, variability between instruments may develop after they have been in use for a time. These tentative conclusions require further evaluation because only two actigraphs of each type were tested and because the testing method was so crude. Like most clinical laboratories, we lack satisfactory instrumentation for regularly recalibrating wrist actigraphs, although previous experience suggests that actigraph responses drift after the hard use that actigraphs receive. Difficulty maintaining instruments in calibration is one of the problems of current clinical actigraphy. Nevertheless, with close examination of the performance of the instruments, we cannot confidently attribute discrepancies between Actiwatch-L and Spectrum performance to differences in sensitivity. There may have been subtle chance differences between the samples tested, and the Spectrum alignments produced slightly lower alignment correlations but, in general, the Spectrums did not appear to perform quite as well with the Scripps Clinic algorithm. Sleep–wake discrimination with the Spectrum model was somewhat disappointing both with the manufacturer's software and with the Scripps Clinic algorithm, but possibly better performance could be obtained if the algorithm parameters were optimized from a training set of Spectrum recordings.

In general, the Scripps Clinic algorithm performed better in scoring sleep–wake against a PSG standard than did any of the manufacturer's Actiware 5 algorithms. The superiority of the Scripps Clinic algorithm was less notable with the outlier retained, and one might argue that without PSG, actigraph users would have little means of excluding such outliers. A successful algorithm should estimate the amount of sleep in the recording rather accurately and should accurately reflect the contrasting variation in the amount of sleep between subjects (as indicated by the correlation of Wake% estimated by actigraph and by PSG across subjects). The correlations of Wake% for the Scripps Clinic algorithm were quite modest compared with what has been achieved in studies using other actigraphs (Ancoli-Israel *et al.*, 1997, 2003; Blackwell *et al.*, 2008; Edinger *et al.*, 2004; Jean-Louis *et al.*, 2001a; b; Verbeek *et al.*, 1994).

It is also desirable, although perhaps less important, that the scoring of individual epochs be accurate. Because an algorithm estimating that all epochs in the records were sleep could perform with close to 80% epoch-by-epoch agreement, the kappa statistic is a better measure of scoring performance than the simple percentage of epoch-by-epoch scoring agreement. Although substantially higher than that obtained by the manufacturer's algorithms, the mean kappa of 0.52 for the Scripps Clinic algorithm with Actiwatch and Spectrum

reflected only moderate epoch-by-epoch accuracy. Similar to our result with the manufacturer's algorithm, another study of kappa with Actiware scoring of Actiwatch data coincidentally found an identical kappa of 0.399 (Gale *et al*., 2005). Like ours, another study found that scoring accuracies in subjects with and without apnea were similar (Wang *et al*., 2008).

Polysomnographic studies provide a perspective. In one study, comparing different scorers of the same PSG, the epoch-by-epoch agreement for scoring wake was only 82% (Norman *et al*., 2000), actually less than what was achieved with our best actigraphic scoring algorithm (Table 3). However, in a recent study, largely using second-night recordings from control subjects, multiple PSG scorers of the same digital record achieved an epoch-by-epoch kappa for sleep–wake of 0.81 (Danker-Hopfe *et al*., 2009), considerably better than what the Scripps Clinic algorithm achieved. When two scorers hand-score the same PSG record, there can be no alignment problems in epoch-by-epoch scoring, but when we compare scoring of actigraphic and PSG epochs, errors of time alignment of epochs as well as errors in the PSG scoring are compounded with the errors of the actigraphic algorithms. Our problems with aligning actigraphic and PSG epochs were a serious limitation of this study, which may have led to underestimates of algorithm accuracy.

An interesting finding of this study was how poorly the 30-s wrist actigraphic counts correlated with periodic limb movements and with brief arousals. In this sample, the actigraphic record would have been of no value in recognizing periodic limb movements in sleep, nor would it provide a useful proxy for the arousal index. Actigraphs placed directly on the lower limbs, for example situated over the tibialis anterior tendon, might perform much better in recognizing periodic limb movements (Kazenwadel *et al*., 1995; Sforza *et al*., 1999).

As shown in Fig. 1, the Scripps Clinic algorithm for Actiwatch has a similar form and rather similar epoch weights as the preceding algorithm developed by Jean-Louis *et al*. for sleep scoring with the Actillume instrument (Ambulatory Monitoring, Ardsley, New York, NY, USA; Jean-Louis *et al*., 2001a). Both algorithms derive a scoring parameter from a combination of activity counts of the epoch being scored with several epochs preceding and subsequent. The preceding epochs are weighted more highly than subsequent epochs. An interesting feature of the Scripps Clinic algorithm is that the activity count for the 30-s preceding the epoch being scored was weighted more highly than the activity count for the actual epoch being scored. Both of these algorithms, in turn, resemble in form the scoring algorithm developed by Cole and Kripke for an early-model wrist actigraph of quite different mechanical design (Cole *et al*., 1992). It appears that the form of the optimized algorithms was similar among these three developments, despite the different epoch lengths and different mechanical and electronic properties of the activity sensors. On the contrary, the scaling factors necessarily differed for transducers of quite different sensitivity, epochs of different durations and scaling to, for example, eight-bit or 12-bit epoch count storage allocations. We would anticipate that scaling

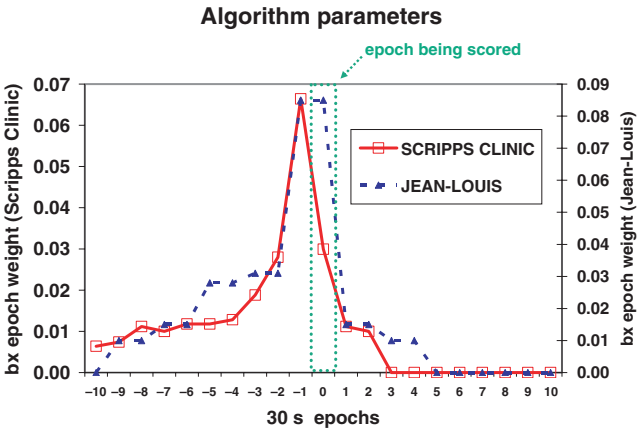

Figure 1. The parameters of the Scripps Clinic scoring algorithm for Actiwatch are compared with the Jean-Louis *et al*. algorithm parameters for scoring Actillume recordings (Jean-Louis *et al*., 2001a). Because the Jean-Louis *et al*. algorithm was derived for 60-s epochs, to depict it for 30-s epochs the parameters were rescaled and plotted by two points with the same *y*-value for each 60 s. The box of dotted lines represents the 30-s epoch being scored (epoch 0 on the *x*-axis). Epochs −10 to −1 represent 30-s epochs preceding the epoch being scored, whereas epochs 1–10 represent epochs subsequent to the epoch being scored. The left-hand *y*-axis shows the *b* parameters for the Scripps Clinic algorithm, and the right-hand *y*-axis gives the comparable parameters for the Jean-Louis algorithm.

factors would have to be optimized for each different instrument and for variations in recording parameters. It is notable that the algorithms supplied in the Actiware 5 program weighted equally the activity 2 min preceding and 2 min subsequent to the epoch being scored for sleep–wake, a weighting pattern that did not seem optimal in the empirically derived weights of Cole and Kripke, Jean-Louis *et al*. or the Scripps Clinic algorithm.

In summary, this study developed a new Scripps Clinic algorithm for scoring sleep–wake state from wrist activity measured by the Actiwatch instrument. The Scripps Clinic algorithm performed better in our Sleep Center laboratory sample than did the manufacturer's algorithms. Several studies of sleep discrimination with the Actiwatch have obtained less satisfactory specificity than reported in studies of other instruments, but there is no direct evidence whether the Actiwatch design is less capable of discriminating sleep–wake than other actigraphs. Although the general form of the Scripps Clinic algorithm is quite similar to previously derived actigraphic algorithms, it was optimized for the Actiwatch-L instrument and for our Sleep Center patient sample. We hope that other laboratories will be able to evaluate the Scripps Clinic algorithm in a diversity of samples. Further, controlled comparisons between different actigraphs are needed to assess where improvements are needed in electronic design, in sleep-scoring software and perhaps in choice of epoch lengths.

ACKNOWLEDGEMENTS

This research was supported by Scripps Clinic Academic Funds. The manufacturer of the actigraphs (Minimitter)

provided one free copy of Actiware 5.01.007 to update a purchased copy of Actiware-Sleep 3.4, but did not otherwise support the study. The manufacturer had no role in the design, execution, analysis, or preparation of this manuscript. The authors assert no proprietary interest in the algorithm presented. Larry Ley and Donna Jones assisted in scoring polysomnograms. Sleep Clinic technicians and patients generously donated their time and effort to this study. Arthur Dawson, MD, contributed to the planning of the study and reviewed the manuscript.

CONFLICT OF INTEREST

This was not an industry-supported study. In 2005, Minimitter, then the manufacturer of the Actiwatch, provided one free copy of Actiware 5.01.007 to update a purchased copy of Actiware-Sleep 3.4, but did not otherwise support the research. The current manufacturer of Actiwatch and Spectrum, Philips Electronics, had no role in the design, execution, analysis or preparation of this manuscript. D.F.K., E.K.H., A.P.G., K.H.W., R.T.L., F.F.S., J.W.C. and L.E.K. report no conflicts of interest. J.S.P. reports funding from GSK.

REFERENCES

Ancoli-Israel, S., Clopton, P., Klauber, M. R., Fell, R. and Mason, W. Use of wrist activity for monitoring sleep/wake in demented nursing-home patients. *Sleep*, 1997, 20: 24–27.

Ancoli-Israel, S., Cole, R., Alessi, C. A., 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.

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: 283–291.

Chae, K. Y., Kripke, D. F., Poceta, J. S. *et al*. Evaluation of immobility time for sleep latency in actigraphy. *Sleep Med*., 2009, 10: 621–625.

Cole, R. J., Kripke, D. F., Gruen, W., Mullaney, D. J. and Gillin, J. C. Automatic sleep/wake identification from wrist activity. *Sleep*, 1992, 15: 461–469.

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.

Edinger, J. D., Means, M. K., Stechuchak, K. M. and Olsen, M. K. A pilot study of inexpensive sleep-assessment devices. *Behav. Sleep Med*., 2004, 2: 41–49.

Gale, J., Signal, T. L. and Gander, P. H. Statistical artifact in the validation of actigraphy. *Sleep*, 2005, 28: 1017–1018.

Jean-Louis, G., Kripke, D. F., Cole, R. J., Assmus, J. D. and Langer, R. D. Sleep detection with an accelerometer actigraph: comparisons with polysomnography. *Physiol. Behav*., 2001a, 72: 21–28.

Jean-Louis, G., Kripke, D. F., Mason, W. J., Elliott, J. A. and Youngstedt, S. D. Sleep estimation from wrist movement quantified by different actigraphic modalities. *J. Neurosci. Meth*., 2001b, 105: 185–191.

Kazenwadel, J., Pollmacher, T., Trenkwalder, C. *et al*. New actigraphic assessment method for periodic leg movements (PLM). *Sleep*, 1995, 18: 689–697.

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.

Loving, R. T., Joya, F. L., Kripke, D. F., Kline, L. E. and Dawson, A. D. Increased use of actigraphy for sleep research. *Sleep*, 2008, 31[Abstract Supplement]: A345.

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.

Sforza, E., Zamagni, M., Petiav, C. and Krieger, J. Actigraphy and leg movements during sleep: a validation study. *J. Clin. Neurophysiol*., 1999, 16: 154–160.

de Souza, L., Benedito-Silva, A. A., Pires, M. L. N., Poyares, D., Tufik, S. and Calil, H. M. Further validation of actigraphy for sleep studies. *Sleep*, 2003, 26: 81–85.

Verbeek, I., Arends, J., Declerck, G. and Beecher, L. Wrist actigraphy in comparison with polysomnography and subjective evaluation in insomnia. In: A. M. L. Coenen (Ed) *Sleep-Wake Research in the Netherlands*. Dutch Society for Sleep-Wake Research, Leiden, 1994: 163–170.

Wang, D., Wong, K. K., Dungan, G. C., Buchanan, P. R., Yee, B. J. and Grunstein, R. R. The validity of wrist actimetry assessment of sleep with and without sleep apnea. *J. Clin. Sleep Med*., 2008, 4: 450–455.

Webster, J. B., Kripke, D. F., Messin, S., Mullaney, D. J. and Wyborney, G. An activity-based sleep monitor system for ambulatory use. *Sleep*, 1982, 5: 389–399.