Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals

Z Beattie, Y Oyang, A Statan, A Ghoreyshi, A Pantelopoulos, A Russell and C Heneghan

Fitbit Research, 199 Fremont Street, San Francisco, CA 94105, United States of America
E-mail: cheneghan@fitbit.com

Received 21 May 2017, revised 30 August 2017
Accepted for publication 2 October 2017
Published 31 October 2017

Abstract

Objective: This paper aims to report on the accuracy of estimating sleep stages using a wrist-worn device that measures movement using a 3D accelerometer and an optical pulse photoplethysmograph (PPG). Approach: Overnight recordings were obtained from 60 adult participants wearing these devices on their left and right wrist, simultaneously with a Type III home sleep testing device (Embletta MPR) which included EEG channels for sleep staging. The 60 participants were self-reported normal sleepers (36 M: 24 F, age = 34 ± 10, BMI = 28 ± 6). The Embletta recordings were scored for sleep stages using AASM guidelines and were used to develop and validate an automated sleep stage estimation algorithm, which labeled sleep stages as one of Wake, Light (N1 or N2), Deep (N3) and REM (REM). Features were extracted from the accelerometer and PPG sensors, which reflected movement, breathing and heart rate variability. Main results: Based on leave-one-out validation, the overall per-epoch accuracy of the automated algorithm was 69%, with a Cohen’s kappa of 0.52 ± 0.14. There was no observable bias to under- or over-estimate wake, light, or deep sleep durations. REM sleep duration was slightly over-estimated by the system. The most common misclassifications were light/REM and light/wake mislabeling. Significance: The results indicate that a reasonable degree of sleep staging accuracy can be achieved using a wrist-worn device, which may be of utility in longitudinal studies of sleep habits.
Keywords: sleep stages, actigraphy, photoplethysmogram, wearable

(Some figures may appear in colour only in the online journal)

1. Introduction

While polysomnography remains the gold standard for measurement and definition of sleep stages, recent years have seen the development of technologies that allow reasonable estimation of sleep in more natural home environments, and over longer periods of time (multiple days or weeks). The most widespread and accepted technique for measurement of natural sleep is actigraphy, in which sensitive accelerometers are worn on the wrist, and pattern classification algorithms are developed to map the measured signal to sleep/wake patterns. The reported accuracy of such sleep/wake measurement systems varies from 80–95% depending on the population studied and the particular technical implementation (e.g. Paquet et al (2007) and Blackwell et al (2008)). In particular, actigraphy systems may have algorithms in different modes, which will lead to either over-estimation (typically) or under-estimation of sleep time (less commonly) (Blackwell et al 2011). Nevertheless, actigraphy is an established method of assessing sleep/wake parameters under certain conditions (Ancoli-Israel et al 2003).

One limitation however of actigraphy-based solutions however is that they only provide sleep/wake patterns, and do not provide any insight into the sleep architecture. To overcome this limitation, a variety of technical solutions for simplified ambulatory sleep stage measurement have been proposed. Single channel EEG systems have been proposed by a number of groups, and indeed several products (e.g. the Zeo sleep headband (Shambroom et al 2012), and the SleepProfiler (Finan et al 2016)) have been successfully developed as either consumer- or medical-grade products. However, while single channel EEG can provide good levels of accuracy, the experience of wearing a headband is still considered quite intrusive by many consumers or research participants. Alternative approaches to measuring sleep stages have been based on measurement of physiological variables such as heart rate, respiration or arterial tone which are known to be affected by sleep stage (Redmond and Heneghan 2006, Willemen et al 2014, Fonseca et al 2015, Muzet et al 2016, de Zambotti et al 2017).

To be clear, we do not expect a system based on photoplethysmography and movement to provide a ‘gold standard’ estimation of sleep stages, which will need EEG, EOG and EMG. Instead, what we are trying to achieve is an automated method to determine the correlations of true sleep stages to corresponding physiological metrics such as movement, heart rate, heart rate variability and breathing rate, and to develop a machine learning algorithm that will use these correlations to approximate the most likely underlying sleep stages.

In this paper, we report on the development of a wrist-worn device which combines 3D accelerometer data and optically obtained photoplethysmogram signals to estimate sleep stages.

2. Materials and methods

2.1. Participants

Sixty participants between the ages of 18 and 60 took part in this study. The participants were judged to be healthy sleepers by self-assessment (i.e. no previous diagnosis of sleep disorders such as sleep apnea, insomnia etc). Participants were not on drugs or medication known to affect sleep. Table 1 summarizes the demographic information on the participants. Each participant underwent the protocol for a single night study only.
2.2. Study procedures

The sleep study was conducted using an Embletta MPR with ST+ Proxy (Natus Medical, Pleasanton, CA). This is a home polysomnogram device capable of measuring multiple channels of EEG, ECG, EOG, airflow, etc. Since this study was concerned with sleep stage measurement, the measured channels were EEG, ECG, and EOG; airflow and respiratory effort were not recorded. The equipment was set up by one of two registered polysomnogram technologists, typically between 8 PM and 9 PM, and the equipment was set to auto-record from 10 PM onwards. Study participants had the option of sleeping in their home, or at a hotel room, dependent on their personal preference for convenience. The study participants also wore two commercially available personal trackers (Fitbit Surge, Fitbit, San Francisco, CA), configured in a data logging mode to record raw accelerometer and photoplethysmogram signals. These wrist-worn devices record an optical photoplethysmogram signal using a green LED and receiver, and also record a 3D accelerometer trace. Participants wore devices on both the left and right wrist, in case there was any systematic difference between the sleep stage classification that can be obtained from dominant and non-dominant wrists.

2.3. Measurement details and polysomnogram scoring

Scoring was done according to current AASM rules (Berry et al 2015). Scoring was done on a 30 s epoch level. The scoring was done blindly with respect to the automated algorithm described below. There were two independent scorers. To ensure consistency of scoring, prior to the analysis both scorers jointly scored selected PSG records and verified that their inter-rater reliability was consistent with AASM guidelines (Rosenberg and Van Hout 2013). The measured kappa between the two scorers was 0.72.

2.4. Data pre-processing

An important aspect of the data processing was synchronization between the PSG and the personal tracker data streams. This was achieved by (a) initially setting the clock on both devices to the same time to achieve approximate time alignment, and (b) aligning the extracted heart rate traces from the PSG ECG and the two devices, to get beat-by-beat time accuracy.

The Fitbit device records an optical photoplethysmogram (PPG) using a green LED and photodiode detector. This data is stored on the memory of the device and downloaded for offline analysis. A peak detector algorithm has been developed to find the peaks in the PPG signal. The time between PPG peaks (PP-interval) is taken as a surrogate for the RR intervals obtained from an ECG. In general PPG signals are more prone to motion artefact than ECG,

Table 1. Demographic information on study participants.

| Description          | Value   |
|----------------------|---------|
| Number of participants | 60 (36 M, 24 F) |
| Age (years)          | 34 ± 10 |
| BMI (kg m⁻²)         | 28 ± 6  |
and in the case of excess motion, the peak detection algorithm does not return any estimated peaks.

2.5. Automated sleep stage classification

Actigraphy (motion-based sleep tracking) is a well-accepted methodology for tracking basic sleep/wake patterns, and is widely used in the sleep research and clinical community. The basic premise of actigraphy is that the patterns of movement associated with sleep and wake states can be learned by automated classification algorithms. These algorithms are trained on a known set of annotated sleep logs across a population of interest, and can then be prospectively applied to new sleep records. As it is known that sleep stage can affect physiological variables such as heart rate and respiration, we adopted a similar methodology where we obtained annotated sleep records from a representative population, and used machine learning techniques to develop an automated sleep staging algorithm. Specifically, we developed features on a 30 s time scale (to match the time scale on which the sleep study is scored). The generated features were based on motion, heart rate variability, and estimated breathing rate parameters. In total, an initial set of 180 features was calculated for each 30 s epoch and fed into an automated classifier together with the gold standard labels. This classifier outputs a label of ‘Wake’, ‘Light’, ‘Deep’, or ‘REM’ for each 30 s epoch, and the gold standard labels are used to choose the optimal set of classifier parameters. As discussed later on, in order to assess the performance of the system, it is typical to train the classifier using 59 of the 60 records, and then assess its performance on the 60th withheld data set. Further improvements in classifier accuracy beyond the raw output of the classifier were achieved by implementing a set of simple post-processing rules which penalized unlikely physiological patterns (e.g. seeing a single isolated wake epoch during a period of sustained deep sleep epochs). In a similar fashion, an isolated deep sleep epoch is converted to the label of the surrounding epochs.

To be more specific on the types of features used, they can be considered in three groups: motion-based features, heart rate-based features, and breathing-based features. The motion based features include

- Activity count over 30 s (e.g. integrated area under the accelerometer signal)
- Rotation magnitude (using the 3D accelerometer to combine the maximum–minimum of each axis)
- Time since last significant movement
- Time till next significant movement

The heart rate features utilize the fact that when a person is relatively stationary, the optical PPG sensor can detect the time of each pulse, so that interbeat intervals (RR values) can be obtained reliably (we have confirmed a high degree of accuracy separately by simultaneously measuring ECG signals in reference data sets). By having a set of RR intervals, we can then calculate standard heart rate variability (HRV) metrics (Task Force 1996) over the epoch of interest, such as

- HF power (0.15–0.4 Hz)
- LF power (0.04–0.15 Hz)
- VLF power (0.015–0.04 Hz)
- RMSSD
- pNN50
- Delta RR
- Mean heart rate
Finally, an instantaneous breathing rate estimate on a 1 s time scale can be formed by taking the spectrum of the RR intervals, and choosing the frequency of maximum power within a plausible band of breathing frequencies. Given this breathing rate estimate on a 1 s basis, this can then be used to form spectral features of the breathing rate such as

- HF power (0.15–0.4 Hz)
- LF power (0.04–0.15 Hz)
- VLF power (0.015–0.04 Hz)

Variants of the features can then be obtained by calculating these parameters on different time scales (e.g. calculating HRV parameters solely on the 30 s epoch, or using 2, 5 or 10 min windows centered on the epoch).

This leads to a full set of 180 features in which there is significant redundancy between features. While optimal performance on a cross-fold validation can be obtained by using the full set of features, we implemented a standard recursive feature elimination in Python’s SCIKIT library to end up with a set of 54 features which provided good performance, and these are what is reported on in the Results section.

We also used the SCIKIT library to explore different types of classifiers: specifically we looked at linear discriminant classifiers, quadratic discriminant classifiers, random forests and support vector machine approaches. On this data set, the linear discriminant mode worked marginally better than the others, so it was chosen as the final model. LDA models also tend to be easy to implement and robust to variations in data sets.

2.6. Statistical analysis

In a multi-class problem, where you wish to compare the labeling from two different observers, a useful analytical tool is the confusion matrix, and the associated metric entitled Cohen’s kappa $\kappa$ (Cohen 1960). It is defined as

$$\kappa = \frac{\% \text{Observed Agreements} - \% \text{Agreements by chance}}{1 - \% \text{Agreements by chance}}.$$

This is a measure of inter-rater agreement, where the two raters in our case are the expert sleep technician (who scored the polysomnograph recordings) and the automated sleep staging system. $\kappa$ is a chance-adjusted measure of agreement which varies from 1 for perfect agreement to 0 for a performance no better than chance. Cohen proposed that $\kappa$ be interpreted as follows: values $\leq 0$ as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement.

A weighted kappa could also be useful, as in practice not all misclassifications are of equal importance (e.g. mislabeling deep sleep as wake is a ‘worse’ error than labeling a light sleep epoch as REM). However other researchers in this field have tended to provide unweighted kappa scores, so for ease of comparison that is what we report.

Results are obtained using a standard leave-one-out validation process to assess the performance of our system.

As well as kappa, we also report on overall accuracy (performance) which is the percentage of epochs correctly labeled relative to the gold standard, as well as class-specific sensitivity and specificity, and level of agreement for ease of comparison with other work. The sleep sensitivity is defined as the percentage of true sleep epochs correctly labeled as sleep (combining
Light, Deep and REM into one stage); the wake specificity is defined as the percentage of true wake epochs correctly labeled as wake; the PSG-tracker agreement for reach of Light, Deep and REM is defined as the percentage of true epochs in those classes correctly labeled by the tracker (de Zambotti 2017).

As well as per-epoch accuracy, overall durations are of importance in assessing the quality of sleep, so we also provide results on the estimated durations of sleep in the various labeled stages.

A further analysis of interest is to determine whether there is a systematic bias in either under-estimating or over-estimating the durations in the various sleep stages. We used a paired two-tailed $t$-test under the null hypothesis that there is no difference between the mean estimated duration values. Box plots of the estimated durations and the error between the reference and estimated values are useful visualizations of the system outputs. Finally, a modified Bland and Altman plot of the errors is included.

### 3. Results

Table 2 shows the sleep characteristics of the study population. As expected from the recruiting criteria, this population shows a fairly typical profile of sleep durations and sleep architecture. Table 3 shows the confusion matrix for the left-hand wrist-worn device, calculated using a leave-one-out validation approach, where out of $n$ records, $n - 1$ records are used to train the classifier, and then the remaining record is tested; we then add up the resulting confusion matrices for all $n$ records tested in this way. This shows that the most common misclassification errors are light classed as REM, REM classed as light, deep classed as light, and wake classed as light. Similar results were obtained for the right-hand worn devices.

The overall Cohen’s kappa for the leave-one-out cross validation on the left-hand recordings was $0.52 \pm 0.14$. This corresponds to an overall per-epoch accuracy of 69%.

As well as showing an epoch-level classification performance, it is meaningful to consider the results on a per-subject basis also. Figure 1 shows a scatter plot of the true duration in each stage versus the estimated duration for all participants in a leave-one-out validation. To provide context on the subjective nature of the hypnogram that can be produced by the automated sleep stage classifier system, figure 2 shows hypnograms for three of the participants in the database which have the worst, average, and best kappa scores. It can be seen that the kappa scores of 0.5 (corresponding to an ‘average’ case) provide a quite meaningful view of the overall sleep architecture.
An important aspect of the classifier and post-processing system is to minimize any potential bias towards a particular sleep stage. For example, if we would like to use this system to estimate population level means of the distribution of times in the various stages, it is desirable that the classifier has minimal bias. Figure 3(a) shows a boxplot for the duration in each sleep state across the 60 participants. (The boxplot shows the the median, interquartile ranges, and maximum/minimum of the data which falls within a distance less than 1.5 times the interquartile range from the 25th and 75th percent values.) For ease of interpretation, figure 3(b) shows a similar boxplot for the difference between the true value and the estimated value. Table 4 shows the data in numerical form and uses a paired \( t \)-test to determine if there is a likely bias in the classifier to over- or under-estimate the duration in any particular stage. Based on this data, we cannot detect a statistically significant difference between the estimate durations in wake, light to deep sleep, but there is evidence of slight overestimation of the duration of REM sleep.

Figure 4 provides modified Bland and Altman plots of the errors in duration estimation versus the reference value for each stage.

Table 5 shows the sensitivity of the classifier to determine each particular stage of sleep, and also its overall sensitivity to sleep and wake.

We compared the kappa values obtained from the left and right hands of the subject to see if there was any bias towards either the dominant or non-dominant hands providing higher accuracy. The kappa distribution was 0.52 ± 0.14 for the left hand, and 0.53 ± 0.14 for the right hand, with no statistically significant difference between the two hands.

### Table 3. The overall confusion matrix for all participants in a leave-one-out cross validation. The values in parentheses show the percent of that stage labeled. This includes wake periods prior to sleep onset and after wake up.

| Predicted stage | Wake | REM | Light | Deep |
|-----------------|------|-----|-------|------|
| Actual (%) stage | Wake (15.7%) | 6116 | 640 | 2035 | 33 |
|                 | REM (19.1%)  | 424  | 7653 | 2566 | 47 |
|                 | Light (50.8%) | 1995 | 3598 | 19681 | 3179 |
|                 | Deep (14.4%) | 145  | 314  | 2583 | 5056 |

Figure 1. Predicted versus actual time spent in each sleep stage—results from leave-one-out validation evaluation.
4. Discussion and conclusions

This study shows that it is feasible to estimate sleep stages with a reasonable degree of accuracy from optical pulse signals and accelerometers. In particular, the addition of heart rate (and the corresponding ability to estimate breathing rate metrics) gives a significant advantage over earlier accelerometer-only actigraphy systems. This is consistent with the physiological observations of previous researchers that there are different patterns of heart rate variability in the various sleep stages (Penzel et al. 2003). In the development of this overall classifier, we experimented with using only movement-derived features or only heart beat derived features. In both cases, we saw kappa values in the range of 0.25–0.3, which is significantly worse than when combined, and also qualitatively indicates that there is approximately equal importance to the movement and heart beat features in the overall four-stage classification.

While no single feature is simply predictive of sleep stages, by visually examining the time course of individual features versus sleep stage, we can provide some physiological insights. Our interpretation of the most significant features is as follows. Movement features correlate with previously known insights from actigraphy. For example, an activity count feature which includes the magnitude and duration of movement during the 30s epoch is easily interpreted as being correlated with wake. Periods of near-constant heart rate and low movement are associated with deep sleep. Periods with a high degree of short-term heart rate variability (e.g. as seen in the LF and HF spectral features) and relatively little movement are associated with REM. The advantage of machine learning algorithms such as the linear discriminant classifier is that it can pull together these different aspects of the data which would be hard to do by inspection.
The results are comparable with previous attempts at estimating sleep stage information from simplified signals such as heart rate, movement and respiration (respiratory inductance plethysmography). Fonseca et al. (2015) provided a useful table of previous attempts to classify sleep stages on the basis of movement, cardiac and respiration signals. For example,
they achieved a Cohen’s kappa coefficient of 0.49 and an accuracy of 69%, using ECG and respiration signals obtained using a respiratory inductance plethysmogram. The best performing combination in the literature is reported by Willemin et al, using a combination of ECG, respiration inductance plethysmography, and actigraphy (Willemen et al 2014). They obtained a Cohen’s kappa of 0.56 from 85 night’s recordings (note however that they had a very tightly clustered demographic with an average age of 22.1 ± 3.2 years).

More recently, de Zambotti et al reported on the accuracy of sleep staging using optical plethysmography and movement using a wearable device on the finger (de Zambotti et al 2017). They reported overall sleep sensitivity of 95.5% and wake specificity of 48.1%, which is similar to the figures we report in table 5 (94.6% and 69.3%). The figures for per-stage agreement are also similar (the ring configuration reports per-stage agreement of Light: 64.6%, Deep 60.9%, REM 61.9%—the wrist-worn device described here gives figures of Light 69.2%, Deep 62.4%, REM 71.6%). The slightly higher figures reported by the wrist-worn device may be due to a better quality optical signal plethysmogram available at the wrist versus the finger, or potentially algorithmic optimizations. However, the fact that a similar signal set of movement and optical plethysmogram gives broadly comparable results in an independent study is encouraging for the overall validity of this approach.

Figure 4. Modified Bland and Altman plots for the four defined classes. The x-axis shows the true stage duration from the study, and the y-axis shows the true duration minus the estimated duration.

Table 5. Overall sensitivity, specificity and accuracy figures on an epoch-per-epoch basis.

| Parameter                                      | Percentage |
|------------------------------------------------|------------|
| Overall sensitivity (in detecting sleep)      | 94.6       |
| Specificity (in detecting wake)               | 69.3       |
| PSG-tracker agreement (light sleep)           | 69.2       |
| PSG-tracker agreement (deep sleep)            | 62.4       |
| PSG-tracker agreement (REM sleep)             | 71.6       |
To place these results in context, we can compare the Cohen’s kappa with an estimate of what the Cohen’s kappa would be for two expert humans scoring the same full PSG montage. Based on table 3 of Rosenborg’s paper on the AASM inter-rater discussion (Rosenberg and Van Hout 2013), we estimate that the average Cohen’s kappa between two expert humans is about 0.78, indicating the overall difficulty of sleep staging even in the presence of a full set of signals and human expertise. We believe that kappa values in the range of 0.4 and above can provide useful and meaningful metrics for sleep researchers and clinicians in the future. In particular, the ability to easily monitor sleep architecture over longer periods of time opens up many interesting possibilities for examining the impact of behavioral and pharmaceutical effects on sleep. A wrist-worn device is also more convenient and comfortable for participants than an EEG head-worn device, ECG chest strap or a respiratory inductance plethysmograph.

The current sleep stage classification system is designed for a retrospective analysis at the end of each night when all data is available for processing. It is possible that a real-time version of this approach could be implemented in which the classifier operates on only the data available up till that point of the night, but it is anticipated that it would have lower performance. This was not investigated during this study.

A limitation of this study is that it studied people who self-declared as healthy sleepers, with no corresponding clinical evaluation. Specifically, people with high suspicion of sleep apnea or with self-reported insomnia symptoms were not included. The age and BMI range studied, while reasonably representative of the population, does not allow reliable prediction of how such a system would perform in older or higher BMI populations. It also did not include participants under 18 years of age, where actigraphy-based methods have been shown to underestimate sleep (Johnson et al. 2007).

Acknowledgements

This study was funded in full by Fitbit Inc.

ORCID iDs

C Heneghan https://orcid.org/0000-0002-4147-1349

References

Ancoli-Israel S et al 2003 The role of actigraphy in the study of sleep and circadian rhythms Sleep 26 342–92
Berry R B, Brooks R, Gamaldo C E, Harding S M, Lloyd R M, Marcus S L and Vaughn B V 2015 The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, version 2.0 (Darien, IL: American Academy of Sleep Medicine) (https://aasm.org/resources/pdf/scoring-manual-preface.pdf)
Blackwell T et al 2008 Comparison of sleep parameters from actigraphy and polysomnography in older women: the SOF study. Sleep 31 283–91
Blackwell T, Ancoli-Israel S, Redline S and Stone K L 2011 Osteoporotic fractures in men (MrOS) study group. Factors that may influence the classification of sleep-wake by wrist actigraphy: the MrOS sleep study J. Clin. Sleep Med. 7 357–67
Cohen J 1960 A coefficient of agreement for nominal scales Educ. Psychol. Meas. 20 37–46
de Zambotti M, Rosas L, Colrain I M and Baker F C 2017 The sleep of the ring: comparison of the ŌURA sleep tracker against polysomnography Behav. Sleep Med. 21 1–15
Finan P H, Richards J M, Gamaldo C E, Han D, Leeoutsakos J M, Salas R, Irwin M R and Smith M T 2016 Validation of a wireless, self-application, ambulatory electroencephalographic sleep monitoring device in healthy volunteers J. Clin. Sleep Med. 12 1443–50
Fonseca P, Long X, Radha M, Haakma R, Aarts R M and Rolink J 2015 Sleep stage classification with ECG and respiratory effort Physiol. Meas. 36 2027–40
Johnson N L, Kirchner H L, Rosen C L, Storfer-Isser A, Cartar L N, Ancoli-Israel S, Emancipator J L, Kibler A M and Redline S 2007 Sleep estimation using wrist actigraphy in adolescents with and without sleep disordered breathing: a comparison of three data modes Sleep 30 899–905
Muzet A, Werner S, Fuchs G, Roth T, Saoud J B, Viola A U, Schaffhauser J Y and Luthringer R 2016 Assessing sleep architecture and continuity measures through the analysis of heart rate and wrist movement recordings in healthy participants: comparison with results based on polysomnography Sleep Med. 21 47–56
Paquet J, Kawinska A and Carrier J 2007 Wake detection capacity of actigraphy during sleep Sleep 30 1362–9
Penzel T, Kantelhardt J W, Lo C-C, Voigt K H and Vogelmeier C 2003 Dynamics of heart rate and sleep stages in normals and patients with sleep apnea Neuropsychopharmacology 28 S48–53
Redmond S J and Heneghan C 2006 Cardiorespiratory-based sleep staging in participants with obstructive sleep apnea IEEE Trans. Biomed. Eng. 53 485–96
Rosenberg R S and Van Hout S 2013 The American Academy of Sleep Medicine inter-rater reliability program: sleep stage scoring J. Clin. Sleep Med. 9 81–7
SCIKIT 2017 Scikit-learn: Machine learning in Python http://scikit-learn.org/stable/documentation.html
Shambroom J R, Fàbregas S E and Johnstone J 2012 Validation of an automated wireless system to monitor sleep in healthy adults J. Sleep Res. 21 221–30
Task Force 1996 Heart rate variability: standards of measurement, physiological interpretation and clinical use Circulation 93 1043–65
Willemen T, Van Deun D, Verhaert V, Vandekerckhove M, Exadaktylos V, Verbraecken J, Van Huffel S, Haex B and Vander Sloten J 2014 An evaluation of cardiorespiratory and movement features with respect to sleep-stage classification IEEE J. Biomed. Health Inform. 12 661–9