Estimating actigraphy from motion artifacts in ECG and respiratory effort signals

2016 Physiol. Meas. 37 67

(http://iopscience.iop.org/0967-3334/37/1/67)

OPEN ACCESS
IOP Publishing | Institute of Physics and Engineering in Medicine Physiological Measurement
Physiol. Meas. **37** (2016) 67–82 doi:10.1088/0967-3334/37/1/67

Estimating actigraphy from motion artifacts in ECG and respiratory effort signals

Pedro Fonseca[1,2], Ronald M Aarts[1,2], Xi Long[1,2], Jérôme Rolink[3] and Steffen Leonhardt[3]

[1] Philips Research, High Tech Campus 34, 5656 AE Eindhoven, The Netherlands
[2] Department of Electrical Engineering, Eindhoven University of Technology, Postbus 513, 5600MB Eindhoven, The Netherlands
[3] Philips Chair for Medical Information Technology, RWTH Aachen University, Pauwelsstrasse 20, D-52074 Aachen, Germany

E-mail: pedro.fonseca@philips.com

Received 18 August 2015, revised 5 October 2015
Accepted for publication 12 October 2015
Published 7 December 2015

Abstract

Recent work in unobtrusive *sleep/wake* classification has shown that cardiac and respiratory features can help improve classification performance. Nevertheless, actigraphy remains the single most discriminative modality for this task. Unfortunately, it requires the use of dedicated devices in addition to the sensors used to measure electrocardiogram (ECG) or respiratory effort. This paper proposes a method to estimate actigraphy from the body movement artifacts present in the ECG and respiratory inductance plethysmography (RIP) based on the time-frequency analysis of those signals. Using a continuous wavelet transform to analyze RIP, and ECG and RIP combined, it provides a surrogate measure of actigraphy with moderate correlation (for ECG+RIP, $\rho = 0.74$, $p < 0.001$) and agreement (mean bias ratio of 0.94 and 95% agreement ratios of 0.11 and 8.45) with reference actigraphy. More important, it can be used as a replacement of actigraphy in *sleep/wake* classification: after cross-validation with a data set comprising polysomnographic (PSG) recordings of 15 healthy subjects and 25 insomniacs annotated by an external sleep technician, it achieves a statistically non-inferior classification performance when used together with respiratory features (average κ of 0.64 for 15 healthy subjects, and 0.50 for a dataset with 40 healthy and insomniac subjects), and when used together with respiratory and cardiac features (average κ of 0.66 for 15 healthy subjects, and 0.56 for 40 healthy and insomniac subjects). Since this method eliminates the need for a dedicated actigraphy device, it reduces the number

of sensors needed for *sleep*/*wake* classification to a single sensor when using respiratory features, and to two sensors when using respiratory and cardiac features without any loss in performance. It offers a major benefit in terms of comfort for long-term home monitoring and is immediately applicable for legacy ECG and RIP monitoring devices already used in clinical practice and which do not have an accelerometer built-in.

Keywords: sleep staging, actigraphy, ECG, respiratory effort

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

1. Introduction

Although recent years have seen significant advances on unobtrusive sleep measurements, overnight polysomnographic recordings (PSG) by expert technicians in dedicated laboratories remain the gold standard for sleep medicine (Iber *et al* 2007). However, PSG still suffers from severe drawbacks. Laboratory facilities, dedicated equipment and qualified personnel are very expensive; in addition, PSG is uncomfortable and can have a negative impact on normal sleep, being basically impossible to perform on long-term beyond one or two consecutive nights. All these factors motivated research in the area of unobtrusive sleep monitoring which can be used in a home setting. Actigraphy, in particular, has been quite popular in the assessment of sleep-wake disturbances (Sadeh and Acebo 2002). It consists of the measurement of gross body movements and is usually performed with accelerometers mounted on the wrist, or on other limbs. The amount of body movements is quantified into so-called 'activity counts' in epochs of fixed size (typically 30 s), which are then used to estimate whether the subject was awake or asleep during that period (Cole *et al* 1992). It has been indicated by the American Academy of Sleep Medicine (AASM) as a valid auxiliary method to evaluate patients with circadian disorders and sleep-wake disturbances, and also to evaluate their response to treatments of insomnia and circadian disorders (Morgenthaler *et al* 2007).

Despite its growing popularity, the performance of actigraphy-based *sleep*/*wake* classifiers has been shown to be limited, and inferior to PSG (Paquet *et al* 2007, Fonseca *et al* 2013). In an effort to improve this, additional modalities such as cardiac and respiratory activity have been also analyzed in the context of sleep and shown to improve actigraphy-based *sleep*/*wake* classification (Long *et al* 2014). However, they require the inclusion of sensors which can measure those modalities, such as electrodes for electrocardiogram (ECG) or respiratory inductance plethysmography (RIP) belts for respiratory effort. The addition of these sensors to the traditional actigraphy means an increase in setup and data analysis complexity since the signals all need to be synchronized. Additionally, it also means a decrease in patient comfort since more sensors need to be worn.

Given the value of cardiac and respiratory modalities in ambulatory monitoring, and the fact that these characteristics cannot be captured with simple actigraphy sensors, the goal of this work is to eliminate the need for dedicated actigraphy sensors when cardiac and/or respiratory sensors are already used for sleep monitoring at home.

1.1. Body movement artifacts on ECG signals

The effect of body movements on ambulatory ECG signals has been thoroughly explored, especially in the context of signal improvement prior to clinical diagnosis of cardiac disorders.

In particular, the last few years saw an increasing number of studies on methods to compensate (Romero 2010), eliminate (Liu 2010), and otherwise detect (Lee *et al* 2012) body movement artifacts (BMAs) on ECG signals (Clifford and Moody 2012). However, very few of those studies have focused on the topic of quantifying these artifacts. The exception was the work of Pawar *et al* (2007) where the type of movements was classified based on the artifacts measured with ambulatory ECG.

One of the most commonly observed characteristics in literature is that portions of ECG affected by BMAs have different spectral properties than unaffected portions of the signal. Figure 1(a) shows the normalized power spectrum density (PSD) plots of the ECG signal during periods with high (activity counts above 10) and with low body movement activity (activity counts equaling 0) for the synchronized ECG and actigraphy (with a Philips Actiwatch) recordings of 9 healthy subjects during the night. Although there is a significant overlap in the spectral content of the two signals, it is clear that periods with higher activity correspond to a higher power below 5 Hz, and also have a richer spectral content in the bands between 30 and 60 Hz.

1.2. Body movement artifacts on RIP signals

In contrast with the ECG, the effect of BMAs on RIP signals has been much less studied, the exceptions being the work of Motto *et al* (2004), Keenan and Wilhelm (2005), and Aoude *et al* (2006). What seems clear is that RIP-based measures of respiratory effort are particularly sensitive to body movements during periods of physical activity and even during sleep (Keenan and Wilhelm 2005). Although this makes respiratory monitoring more challenging, this modality seems like a good candidate for the estimation of actigraphy during the night. Figure 1(b) illustrates the normalized PSD plots of the RIP signal during periods with high and low body movement activity, in the same conditions as indicated before for ECG. The spectral content of the signal in these two conditions is clearly different: the characteristic peak between 0 and 1 Hz under low activity conditions (which corresponds to the average breathing rate) is less clear in the presence of high activity; in addition, the spectral content of the signal for frequencies above 1 Hz is much higher in the presence of high activity, making it easier to distinguish from artifact-free segments of the signal.

1.3. Objective

The objective of this paper is to propose and evaluate a technique for quantifying BMAs in ECG and RIP signals during sleep. This will be achieved by estimating, based on time-frequency analysis with a Continuous Wavelet Transform (CWT), the local power of each signal at different frequencies. A selected set of CWT coefficients are used to estimate the intensity of BMAs in the signal recordings. The potential of using this estimation as a method to detect body movements and as a replacement of wrist-actigraphy for *sleep/wake* classification is evaluated.

2. Materials and methods

2.1. Data sets

The data set used in this study comprised full single-night polysomnographic (PSG) recordings of 40 subjects (21 females). 15 subjects (10 females) had a Pittsburg Sleep Quality Index (Buysse *et al* 1989) of less than 6 and had no regular sleep complaints nor earlier diagnosis of

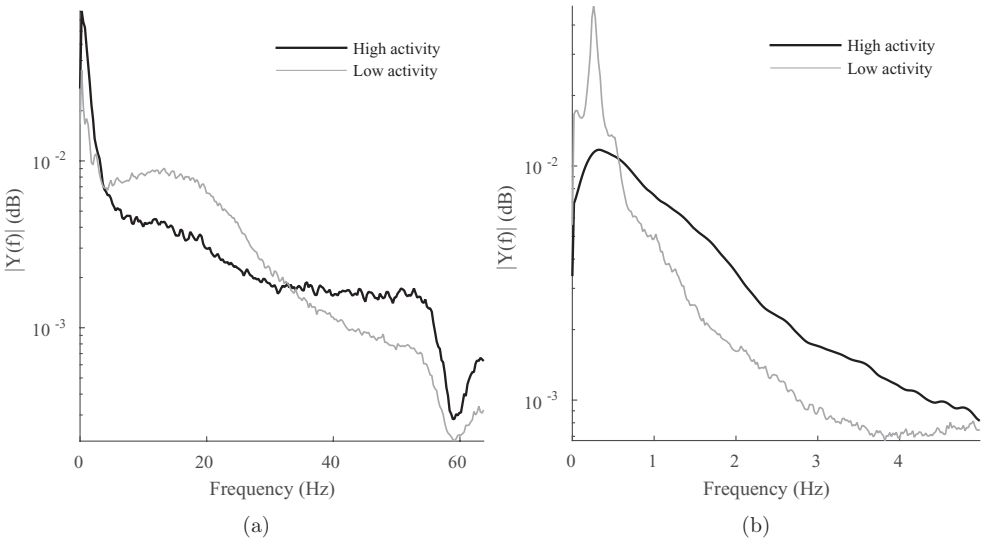

Figure 1. Normalized signal spectrum on periods with high and low body movement intensities for (a) ECG and (b) RIP signals. The low power of the ECG signal around 60 Hz is the result of a notch filter used during acquisition to filter out the mains hum from the power line.

sleep disorders (age 31.0 ± 10.4 years, sleep efficiency (SE) $87.8 \pm 9.6\%$ and total sleep time (TST) 6.67 ± 1.16 hour). The remaining 25 (11 female) subjects (age 45.6 ± 14.9 years, SE $73.6 \pm 17.2\%$ and TST 4.88 ± 1.41 h) had a self-reported Insomnia Severity Index score above 14 and were classified as insomniacs (Bastien *et al* 2001). The PSG data of six subjects was recorded at the Philips Experience Lab of the High Tech Campus, Eindhoven, The Netherlands, during 2010 (Vitaport 3 PSG, TEMEC). The PSG data of the remaining 34 subjects was recorded at the Sleep Health Center, Boston, USA, during 2009 (Alice 5 PSG, Philips Electronics). All PSG montages had at least the recommended AASM set (including redundancy) for sleep scoring, namely EEG F4-M1, F3-M2, C4-M1, C3-M2, O2-M1 and O1-M2, left and right EOG, and submental chin EMG (Iber *et al* 2007). Both studies were approved by the local ethics committees and all participants signed an informed consent form. Sleep stages were scored by trained sleep technicians blind to the participants' condition (healthy or insomniac), in four classes (*wake*, *REM*, *N1*, *N2*, *N3*) according to the AASM guidelines (Iber *et al* 2007). In the scope of this study, *REM*, *N1*, *N2*, and *N3* were all merged into a single *sleep* class. Each PSG recording comprised, besides the standard signals required for sleep scoring, lead II ECG, chest RIP and synchronized actigraphy (Actiwatch Spectrum, Philips Electronics) worn on the wrist of the non-dominant arm and configured to log activity counts in epochs of 30 s. QRS complexes were detected and localized from ECG signals using a combination of a Hamilton-Tompkins detector (Hamilton and Tompkins 1986, Hamilton 2002) and a post-processing localization algorithm (Fonseca *et al* 2014). Prior to actigraphy estimation with CWT, all ECG signals were downsampled to a common sampling rate of 128 Hz. Before feature extraction and actigraphy estimation, RIP signals were downsampled to a common sampling rate of 10 Hz.

2.2. Continuous wavelet transform

The presence of movement artifacts changes the frequency content of ECG and RIP signals in different frequencies than those which normally contain information about actual cardiac and

respiratory processes. Furthermore, the intensity and duration of the movements will cause the artifacts to have different characteristics: more intense, longer movements should lead to artifacts with a higher (overall) energy than less intense or shorter movements. Time-frequency analysis methods can be used to measure the energy of the signals over different frequency bands and over time, which in turn can be used to estimate these characteristics and provide an estimate of gross body movements.

One of the most popular methods for time-frequency analysis, the discrete short-time Fourier transform (STFT), uses the discrete Fourier transform (DFT) together with a window of arbitrary but usually fixed length to calculate the frequency content of segments of a signal. The main disadvantage of STFT is that it relies on a window of constant size and although its length is arbitrary, there is a trade-off in its choice. A longer window will lead to an increase in frequency resolution, but the poorer time resolution will make it less adequate to capture the characteristics of shorter artifacts. A shorter window length improves time localization but will lead to a lower frequency resolution. Typical physiological signals include a combination of events such as spikes (e.g. ECG QRS complexes), transients (e.g. caused by folding and stretching of skin under ECG electrodes during body movements) and oscillations (e.g. respiratory waveform during stable breathing). Although the STFT is very well suited for the analysis of the latter, the limited time resolution, unavoidable when a higher frequency resolution is desirable, makes it less suited for the analysis of events with short duration. When both types of signals are present in the data, and both are of interest, the Wavelet Transform offers a more flexible compromise between time and frequency resolution.

The continuous wavelet transform (CWT) measures the similarity between a signal and an analyzing wavelet function $\psi(t)$ for different dilations of the wavelet, at different locations in the signal. It is defined as

$$T(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\psi^*\left(\frac{t-b}{a}\right)\mathrm{d}t, \tag{1}$$

where ψ^* is the complex conjugate of the wavelet function ψ, a is the scaling (dilation) parameter and b is the location parameter. The wavelet power spectrum at scale a and location b is given by

$$P(a,b) = \left|T(a,b)\right|^2. \tag{2}$$

In contrast with the STFT, which is limited to sinusoidal functions with fixed scales, the CWT allows a wide range of wavelet functions to be used, and to be computed at arbitrary scales and locations. When applied to discrete signals, the CWT is computed over a discretized time-frequency grid. The integral is approximated by a summation over a number of samples for each time step location b.

A large number of wavelet functions has been proposed over the last decades, but one family of wavelets, the Gaussian Wavelet (Daubechies 1990), is widely used for biomedical signal analysis. It is given by

$$\psi^{(N)}(t) = \frac{\mathrm{d}^N}{\mathrm{d}t^N}\left[\pi^{-1/4}\mathrm{e}^{-t^2/2}\right]. \tag{3}$$

The Mexican Hat (Daubechies 1990) is a scaled version of the second derivative, up to a sign, of the Gaussian wavelet, and is given by

$$\psi(t) = \frac{2}{\sqrt{3}}(1-t^2)\mathrm{e}^{-t^2/2}. \tag{4}$$

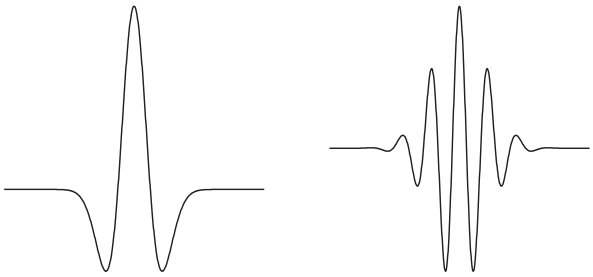

Figure 2. Wavelet functions used for CWT: (left) Mexican Hat, (right) Gaussian order 16.

After experimentally comparing the performance of different wavelet function families and a different number and range of scales, the ECG signal was analyzed with a Mexican Hat wavelet function (figure 2, left) with 320 scales chosen such that the length of the wavelets spanned between 1 and 60 s. The RIP signal was analyzed with a Gaussian wavelet of order 16 (figure 2, right) with 150 scales chosen such that the length of the wavelets spanned between 1 and 30 s. Figures 3(a) and (b) illustrate examples of the CWT computed on short ECG and RIP segments with BMAs, respectively.

Since the wavelet function is shifted sample by sample over the entire signal, P in (2) effectively corresponds to a single value for each sample and for each scale. Since the goal is to estimate a value of actigraphy for each 30 s epoch, a feature vector is built by computing a number of sample statistics over the values of each epoch,

$$F[n] = \{P_{M(a_1)}[n], ..., P_{M(a_A)}[n], P_{m(a_1)}[n], ..., P_{m(a_A)}[n], \bar{P}_{(a_1)}[n], ..., \bar{P}_{(a_A)}[n]\}, \tag{5}$$

where $P_{M(a_i)}[n]$, $P_{m(a_i)}[n]$, and $\bar{P}_{(a_i)}[n]$ are the maximum, minimum and average wavelet power for scale a_i for the epoch starting at sample n, respectively,

$$P_{M(a_i)}[n] = \max(\{P[n, a_i], ..., P[n+N-1, a_i]\}) \tag{6}$$

$$P_{m(a_i)}[n] = \min(\{P[n, a_i], ..., P[n+N-1, a_i]\}) \tag{7}$$

$$\bar{P}_{(a_i)}[n] = \frac{1}{N} \sum_{b=n}^{n+N-1} P[b, a_i]. \tag{8}$$

This yields a total of 960 features for the ECG signal, and 450 features for the RIP signal.

The purpose of these sample statistics is to capture the power of different types of artifacts in the signal. BMAs which last for a large portion of the entire epoch should yield a high average wavelet power for some scales. BMAs which last only a few seconds, or even a fraction of a second, should yield a high maximum wavelet power. When the minimum wavelet power is lower, on those same scales, it suggests that the epoch also contains portions of the signal which are not affected by BMAs. Note that this will allow the feature vector to represent artifacts with short and long durations, in contrast with the STFT, for which the average spectral density would only able to discriminate longer artifacts lasting a significant portion of the entire epoch.

The feature vectors are computed for the ECG signal, F_{ECG}, for the RIP signal, F_{RIP}, and for the combination of the two signals by concatenating the two previous vectors, $F_{\text{ECG+RIP}}$.

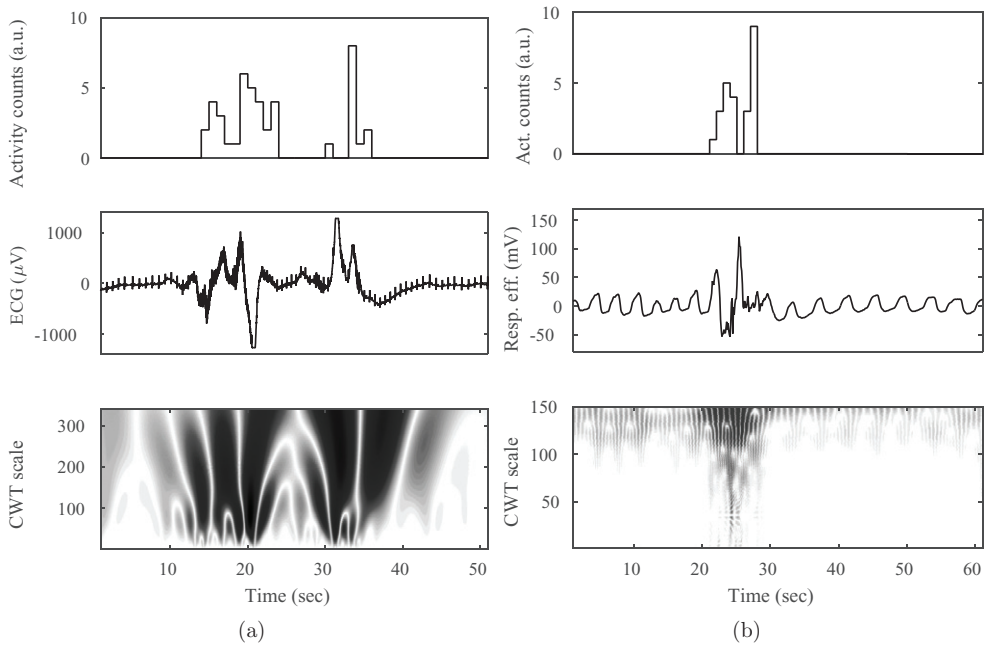

Figure 3. Example of the effect of a body movement artifact on (a) ECG and (b) RIP signals and corresponding CWT with a Mexican Hat and a Gaussian order 16 wavelet, respectively. The top plots of each figure indicate the corresponding activity counts measured with actigraphy, for reference.

Besides the CWT method, an analogous analysis was performed after computing the STFT for 256 frequencies using Hanning windows of 60 s centered on each 30 s epoch of the ECG and the RIP signals. Since the results were inferior to the CWT and failed to yield a statistically non-inferior *sleep*/*wake* classification performance when compared with reference actigraphy, the results obtained with the STFT will not be described further and only the results obtained with the CWT will be presented and discussed.

2.3. Feature selection and regression

After the feature vectors are calculated with the CWT method for the ECG, RIP and combined ECG and RIP signals, an actigraphy value expressing the intensity of gross body movements is finally estimated.

This is achieved by finding the linear combination of the feature values calculated for each epoch which best quantifies those artifacts. The parameters of this linear combination can be obtained by multiple linear regression to a reference, in this case, simultaneously measured actigraphy.

This method generates a large number of coefficients (960 features for the ECG signal, 450 for RIP and 1410 for the combination of both), but not all are needed for the estimation procedure. In fact, there is a strong correlation between several pairs of these coefficients so it is actually advantageous to use only a subset for estimation. This allows the model to be strongly simplified, which in turn can help prevent phenomena such as overfitting.

The subset of features used in the linear model and the regression parameters for the selected features were automatically computed by means of stepwise regression (Draper and

Smith 1998) with the reference actigraphy. This procedure was performed separately for the features obtained for each modality (ECG, RIP, and for the combination of ECG and RIP). The stepwise regression search was performed with cross-validation, i.e. the subjects were first divided in two folds and the model was estimated on each fold while the fit of each coefficient to the data was evaluated on the other. The estimated model and the estimation error, per iteration, were obtained by averaging the results obtained for each validation fold. It has been shown that classical stepwise regression, where regression and evaluation are calculated on the same data set, can bias not only the parameter estimation, but also the tests used to determine whether variables should be included or excluded in the model. Cross-validation alleviates these issues, and increases the robustness of the selection and regression procedure (Fox 1991).

The result is, for each modality $m = \{\text{ECG}, \text{RIP}, \text{ECG} + \text{RIP}\}$, a list of N_m features included in the linear model, $S_m = \{s_1, ..., s_{N_m}\}$ and the corresponding regression coefficients, $C_m = \{\beta_1, ..., \beta_{N_m}\}$. These can be used to compute an estimate of surrogate actigraphy, X_m for each epoch n,

$$X_m[n] = \sum_{i=1}^{N_m} \beta_i F_m[n, s_i], \qquad s_i \in S_m, \beta_i \in C_m, \tag{9}$$

where $F_m[n, s_i]$ is the i-th selected coefficient in the feature vector of epoch n for modality m.

In order to evaluate this method, the estimated actigraphy was compared, for each modality, with the reference actigraphy measured with the Actiwatch using Pearson's correlation coefficient. For each subject, the correlation was computed between the regression estimate obtained with cross-validation on the final iteration of the stepwise regression procedure (the iteration yielding the final list of features) and the reference. The resulting values were then averaged over all subjects. In addition, the estimated and reference actigraphy signals were pooled across all subjects, and the corresponding correlation coefficients were computed. In the pooled case, the correlation coefficient obtained for each modality was tested against the null-hypothesis of no correlation. A Bland–Altman analysis was performed to evaluate the agreement between surrogate and reference actigraphy.

2.4. Sleep/wake classification

Surrogate actigraphy was also evaluated in terms of its capacity to distinguish the presence or absence of movements. Note that this is particularly relevant in a sleep classification scenario, where it is often more important to be able to identify whether the subject has moved rather than to quantify the amplitude of her movements. To compare the agreement in the detection of movements with reference and surrogate actigraphy, we used a simple criteria: if the actigraphy value is above a given threshold, the subject was considered to move. Traditional metrics of agreement such as sensitivity, positive predictive value (PPV) and Cohen's kappa coefficient of agreement were computed for different actigraphy thresholds.

In order to evaluate whether surrogate actigraphy is at least as discriminative of *sleep* and *wake* states as reference actigraphy, the surrogate estimates were used together with a set of features shown in earlier work (Long *et al* 2014) to provide the best classification performance in this task. These feature sets comprise, besides actigraphy (`ac`), five respiratory features including the standard deviation of respiratory frequency over nine epochs (`sdf`), high frequency components (`hfc`) (Redmond and Heneghan 2006), sample entropy (`se`) (Costa *et al* 2002), Dynamic Time Warping (`dtw`) and Dynamic Frequency Warping (`dfw`) (Long *et al*

Table 1. Feature sets used to train the different classifiers and compare performance with and without actigraphy, and with surrogate actigraphy.

Feature set	Features (#)	Modality[a]	Feature set in Long *et al* (2014)
F_{R}	`dtw`, `dfw` (2)	R	F_{DW}
F_{RA}	`ac`, `dtw`, `dfw` (3)	A, R	$F_{\mathrm{AC-DW}}$
F_{RX}	X_{RIP}, `dtw`, `dfw` (3)	A[b], R	n.a.
F_{RC}	`sdf`, `hfc`, `se`, `dtw`, `dfw`, `mhr` (6)	R, C	n.a.
F_{RCA}	`ac`, `sdf`, `hfc`, `se`, `dtw`, `dfw`, `mhr` (7)	A, R, C	F_{ARC2}
F_{RCX}	$X_{\mathrm{ECG+RIP}}$, `sdf`, `hfc`, `se`, `dtw`, `dfw`, `mhr` (7)	A[c], R, C	n.a.

[a] A: actigraphy, R: respiratory effort, C: cardiac activity.
[b] Surrogate actigraphy, estimated from RIP.
[c] Surrogate actigraphy, estimated from ECG and RIP.

2014) and one cardiac feature, mean heart rate (`mhr`). All features were extracted from ECG and RIP signals on non-overlapping epochs of 30 seconds.

The combination of features in the different feature sets is indicated in table 1. The first three sets were used to test surrogate actigraphy computed from the RIP signal alone. F_{R}, and F_{RA} correspond to the sets found by Long *et al* (2014) to give the best performance when respiratory effort and respiratory effort plus actigraphy were used, respectively. These served as benchmark for the performance obtained when surrogate actigraphy computed from RIP was used together with the respiratory features (F_{RX}) instead of reference actigraphy.

The last three feature sets were used to test surrogate actigraphy computed from combined ECG and RIP. F_{RCA} corresponds to the feature set found by Long *et al* (2014) to give the best performance when respiratory, cardiac and actigraphy features were used. F_{RC} corresponds to the same feature set but without actigraphy, and was used to establish the baseline performance improvement due to actigraphy. F_{RCX} comprises, besides the same respiratory and cardiac features, surrogate actigraphy computed from ECG and RIP.

A Bayesian linear discriminant (LD) classifier with time-varying prior probabilities similar to the one used by Long *et al* (2014) was also used in this work to classify *sleep* and *wake*. Despite its simplicity, this classifier has been shown to perform very well in this task (Redmond *et al* 2007, Long *et al* 2014). For each epoch, the selected class (*sleep* or *wake*) is the class ω_i that maximizes the posterior probability given a feature vector F (Duda *et al* 2000),

$$\omega_i(F) = \arg\max_i \left[\mathrm{g_i}(F)\right] \tag{10}$$

where the discriminant function $\mathrm{g_i}$ for each class is given by

$$\mathrm{g_i}(F) = -\frac{1}{2}(F-\mu_i)^t\ \Sigma^{-1}(F-\mu_i) + \ln \mathrm{P}(\omega_i, t) \tag{11}$$

and where μ_i is the average feature vector for class i, Σ is the pooled covariance matrix for both classes, and $\mathrm{P}(\omega_i, t)$ is the prior probability for class i at time (since lights off) t. The time-varying prior probabilities are estimated during training by computing, for each epoch, the relative frequency of each class (according to the ground-truth) on all recordings on the training set (Redmond *et al* 2007).

Separate classifiers were trained and evaluated with each of the feature sets indicated in table 1 using a leave-one-subject-out cross-validation procedure. The classification performance obtained with each feature set was evaluated using Cohen's κ coefficient of agreement (Cohen 1960), a measure of agreement between the result of *sleep*/*wake* classification and the reference annotations by the sleep technicians. This metric is particularly useful to evaluate performance in the presence of imbalanced classes, as is the case with *sleep* and *wake*. Since earlier work (Long *et al* 2014) only reports the performance in a set of healthy subjects, our results were analyzed separately for healthy subjects ($N = 15$) and for the entire data set ($N = 40$), allowing for a fair comparison with the results obtained in that work.

In order to evaluate whether the classifier with surrogate actigraphy provides an equivalent performance to the classifier with reference actigraphy, an equivalence test with a confidence interval (CI) of 90% was used (Walker and Nowacki 2011). Since the performance values (κ) do not follow a normal distribution, the CI was determined using a non-parametric method based on Wilcoxon's signed rank test (Hollander *et al* 2013). Equivalence was statistically established if the 90% CI for difference in performance using surrogate actigraphy instead of reference actigraphy was fully contained in the interval established by a predefined margin δ, i.e. $(-\delta, \delta)$. The margin δ was chosen according to the recommendations of Walker and Nowacki (2011), i.e. as the the lower bound of the 90% CI for difference in performance using actigraphy instead of only cardiac and/or respiratory features. This is the same as using a two one-sided test procedure, where equivalence is established if the null hypotheses

$$H_{01} : \text{med}(\kappa_a) - \text{med}(\kappa_b) < -\delta, \tag{12}$$

and

$$H_{02} : \text{med}(\kappa_a) - \text{med}(\kappa_b) > \delta \tag{13}$$

can be both rejected with $p < 0.05$. med(κ) stands for the median κ obtained with a given condition for all subjets in the data set. In case only H_{01} (12) can be rejected, then non-inferiority (but not equivalence) is statistically established.

In addition, traditional performance metrics of precision, sensitivity, specificity, accuracy, and the area under the precision-recall curve were computed for the classification results obtained with each feature set.

3. Results

Table 2 indicates the correlation values obtained for each modality. The lowest pooled correlation is obtained with ECG, 0.62, followed by RIP, 0.65. The algorithm performs best when combining both modalities, achieving a pooled correlation coefficient of 0.74 and an average correlation coefficient of 0.67. All correlation coefficients were found to be significant with $p < 0.001$.

A visual inspection of the Bland–Altman plot between surrogate (EST) and reference (REF) actigraphy pooled across all subjects revelead heteroscedasticity, confirmed by a Spearman's correlation of 0.81 ($p < 0.001$) between the absolute value of the differences and the average actigraphy values. Comparing log transformed data, as suggested by Bland and Altman (1999), we obtained as mean difference ($\log_{10}$EST $-$ $\log_{10}$REF) a value of -0.03 and 95% limits of agreement (LoA) [-0.98, 0.93] (figure 4(a)). Back-transforming (with antilog) the results we obtained values relating to the ratios between the two measurements. The pooled geometric mean bias ratio between EST and REF was 0.94 (average 1.06 ± 0.62 across all subjects) with 95% agreement ratio between 0.11 and 8.45 (0.15 ± 0.07 and 9.23 ± 10.87).

Table 2. Correlation between estimated and reference actigraphy, and number of coefficients used from each signal.

				Nr. ECG coeff.				Nr. RIP coeff.			
Signal(s)	ρ	95% CI	$\bar{\rho}$	P_m	P_M	$\bar{P}$	*tot*	P_m	P_M	$\bar{P}$	*tot*
ECG	0.62[a]	(0.61, 0.62)	0.56	27	38	3	68				
RIP	0.65[a]	(0.65, 0.66)	0.64					60	77	12	149
ECG, RIP	0.74[a]	(0.73, 0.74)	0.67	15	28	1	44	57	73	5	135

[a] $p < 0.001$.
Note: P_m, P_M and $\bar{P}$ indicate the number of coefficients selected, for each modality, from the wavelet power statistics (minimum, maximum and average, respectively) in (5). *tot* indicates the total number of coefficients selected for each modality.

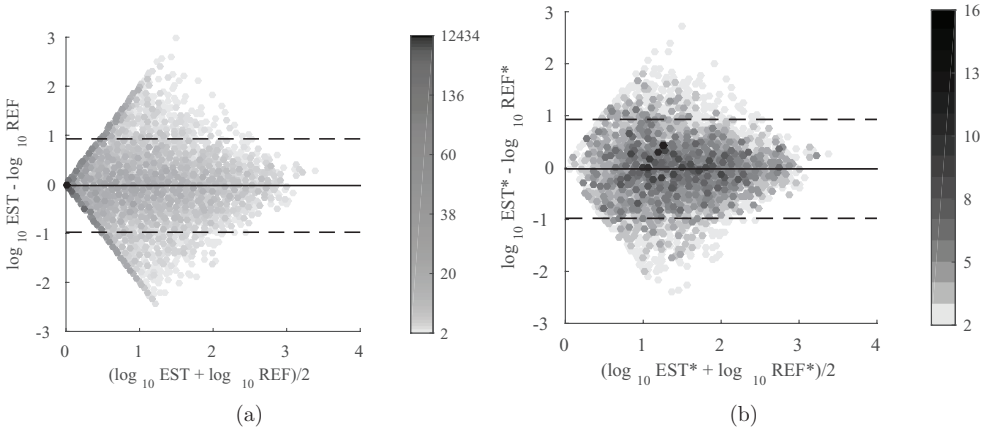

Figure 4. Bland–Altman plots (a) after log transformation, with mean bias ratio and 95% LoA indicated by the solid and dashed lines respectively, and (b) after restricting the plot to points where both methods agree regarding the presence of body movements, i.e. where both reference and surrogate actigraphy have a value greater or equal than 1.

Expressing the agreement ratios as function of the average actigraphy values (Euser *et al* 2008) we obtain

$$-1.62\ \mathrm{AVG} < \mathrm{EST} - \mathrm{REF} < 1.58\ \mathrm{AVG}, \tag{14}$$

with $\mathrm{AVG} = (\mathrm{EST} + \mathrm{REF})/2$.

Correlation analysis was used to test the effects of age, BMI, sleep efficiency and total sleep time on the correlation between surrogate and reference actigraphy, mean bias ratio and on the 95% limits of ratio agreement obtained for each subject, but no significant effects were found. Also, no significant differences between males and females were found after a t-test comparison of these performance metrics.

Figure 5(a) illustrates the histograms of (reference) actigraphy values for the classes *sleep* and *wake* for all subjects. Optimal separation (assuming equal prior probabilities) is achieved with a threshold of 30 (indicated with a dashed line in the figure). Using this threshold as average actigraphy in equation (14) we obtain as LoA [−48.6, 47.4].

Figure 5(b) illustrates the agreement on presence of body movements (as defined by actigraphy above a given threshold) detected with surrogate and reference actigraphy for different metrics. For a threshold of 30, we obtain a kappa of 0.58, a sensitivity of 0.61 and a PPV of

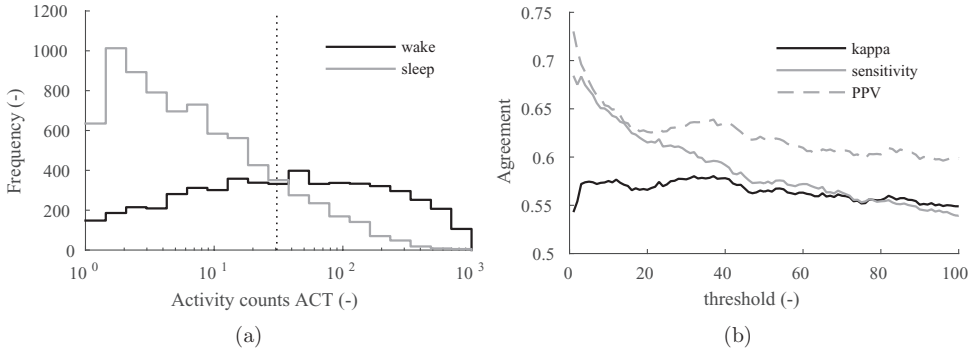

Figure 5. (a) histogram of actigraphy values for *sleep* and *wake* classes, with optimal separation indicated with dotted line, and (b) agreement on presence of body movements detected with surrogate and reference actigraphy for thresholds below 100.

0.63. Overall, agreement is higher for lower thresholds, with sensitivity and PPV achieving a maximum of 0.68 and 0.73 respectively, for a threshold of 1.

Tables 3 and 4 indicate the performance of *sleep*/*wake* classification for each feature set for the healthy subjects and for all subjects, respectively. Note that the results obtained with the same feature sets used by Long *et al* (2014) for healthy subjects are slightly different than those reported in that paper. The small differences are explained by the different handling of epochs where features could not be computed (e.g. no QRS complexes detected due to motion artifacts): Long *et al* excluded those expochs while we linearly interpolated missing values in order to classify all epochs in each recording.

In case only respiratory features and actigraphy are used, surrogate actigraphy (F_{RX}) provides a comparable average performance across all metrics to the reference feature set with standard actigraphy (F_{RA}) both for healthy as well as for all subjects. The kappa obtained was found to be significantly higher than when respiratory features are used alone, without actigraphy (F_{R}). Furthermore, as illustrated in figures 6(a) and (b), the performance of the feature set with surrogate actigraphy (F_{RX}) is statistically non-inferior to the performance with reference actigraphy ($\kappa_{\mathrm{RX}} - \kappa_{\mathrm{RA}}$).

In case respiratory and cardiac features are used together with actigraphy, surrogate actigraphy (F_{RCX}) once again provides a comparable average performance to the reference feature set with standard actigraphy (F_{RCA}) both for healthy as well as for all subjects. Furthermore, surrogate actigraphy provides a statistically significantly higher κ when compared to the feature set without actigraphy (F_{RC}) and as illustrated in figure 6(c) and 6(d), the performance of that feature set ($\kappa_{\mathrm{RCX}} - \kappa_{\mathrm{RCA}}$) is statistically non-inferior to the performance with reference actigraphy.

4. Discussion

As indicated in table 2, the CWT performs overall better when combining both modalities, achieving a correlation coefficient of 0.74 for ECG+RIP (significant with $p < 0.001$). The correlations obtained for ECG and RIP were similar, suggesting that this method is equally able to quantify body movements in both signals. The correlation increase obtained by combining ECG and RIP suggests the BMA information in each signal is somewhat complementary, and the best estimate is obtained by combining both.

Table 3. Mean and standard deviation of classification performance for healthy subjects ($N = 15$), per feature set.

Feature set	Precision	Sensitivity	Specificity	Accuracy	AUC_{PR}	κ
F_R	0.61 ± 0.17	0.67 ± 0.10	0.96 ± 0.01	0.92 ± 0.04	0.66 ± 0.13	0.58 ± 0.10
F_{RA}	0.69 ± 0.16	0.69 ± 0.13	0.97 ± 0.01	0.94 ± 0.02	0.73 ± 0.12	0.64 ± 0.11[b]
F_{RX}	0.69 ± 0.15	0.70 ± 0.11	0.97 ± 0.01	0.93 ± 0.04	0.74 ± 0.13	0.64 ± 0.09[b]
F_{RC}	0.72 ± 0.19	0.67 ± 0.16	0.96 ± 0.04	0.93 ± 0.04	0.73 ± 0.12	0.63 ± 0.14
F_{RCA}	0.72 ± 0.18	0.72 ± 0.15	0.96 ± 0.04	0.94 ± 0.04	0.77 ± 0.12	0.67 ± 0.14[b]
F_{RCX}	0.69 ± 0.20	0.76 ± 0.13	0.95 ± 0.05	0.93 ± 0.05	0.76 ± 0.13	0.66 ± 0.16[a]

[a] $p < 0.05$.
[b] $p < 0.001$, after a two-tailed Wilcoxon-signed rank test of whether the performance (κ) obtained after adding (surrogate) actigraphy to F_R and to F_{RC} was significantly different than without actigraphy.

Table 4. Mean and standard deviation of classification performance for all subjects ($N = 40$), per feature set.

Feature set	Precision	Sensitivity	Specificity	Accuracy	AUC_{PR}	κ
F_R	0.60 ± 0.19	0.58 ± 0.18	0.93 ± 0.04	0.83 ± 0.11	0.65 ± 0.15	0.45 ± 0.16
F_{RA}	0.65 ± 0.19	0.60 ± 0.19	0.94 ± 0.04	0.85 ± 0.11	0.69 ± 0.16	0.50 ± 0.17[b]
F_{RX}	0.63 ± 0.20	0.62 ± 0.21	0.93 ± 0.07	0.85 ± 0.11	0.69 ± 0.17	0.50 ± 0.18[a]
F_{RC}	0.67 ± 0.20	0.65 ± 0.21	0.92 ± 0.10	0.85 ± 0.11	0.71 ± 0.15	0.53 ± 0.17
F_{RCA}	0.71 ± 0.19	0.65 ± 0.21	0.94 ± 0.08	0.87 ± 0.10	0.74 ± 0.15	0.56 ± 0.17[b]
F_{RCX}	0.70 ± 0.21	0.67 ± 0.22	0.93 ± 0.09	0.87 ± 0.10	0.73 ± 0.16	0.56 ± 0.19[a]

[a] $p < 0.01$.
[b] $p < 0.001$, after a two-tailed Wilcoxon-signed rank test of whether the performance (κ) obtained after adding (surrogate) actigraphy to F_R and to F_{RC} was significantly different than without actigraphy.

Regarding the Bland–Altman analysis (figure 4), the heteroscedasticity found in the data suggests that this method is less suited for precisely quantifying moderate to large movements. This can be partly explained by the different location of the sensors: reference actigraphy was measured on the wrist, and surrogate actigraphy was calculated from sensors placed on the chest. Larger movements of one of the parts of the body may simply not have an equal correspondence in terms of amplitude on the other.

Furthermore, it is clear that a number of points in the Bland–Altman plot correspond to a value of zero from either surrogate or from reference actigraphy figure 4(a). This can also be attributed to the different location of the sensors. For small movements of either the wrist or the chest, there isn't always a corresponding movement in the other body location. Overall, however, the method achieves a reasonable agreement in the detection of small body movements. For a threshold of 1, it yields a sensitivity and PPV of 0.68 and 0.73 respectively, meaning that close to 70% of all wrist movements were detected with surrogate actigraphy on the chest, and vice versa. This sensitivity to movements of low amplitude is confirmed by the 95% LoA calculated as function of average actigraphy with (14) and which, for a value of 30 (which gives an optimal separation between *sleep* and *wake*) are moderate ($[-48.6, 47.4]$). Restricting the Bland–Altman plot to points where both methods agree regarding the presence of body movements (figure 4(b)), the agreement between (log) actigraphy estimations is more clearly visible. From an analysis of the plot structure we note that for very large values of actigraphy (above 10^2) agreement is better than for intermediate values. A possible explanation for this is that very large, isolated movements, which likely correspond to movements of the entire body, affect both sensor readings in the two locations in a comparable way.

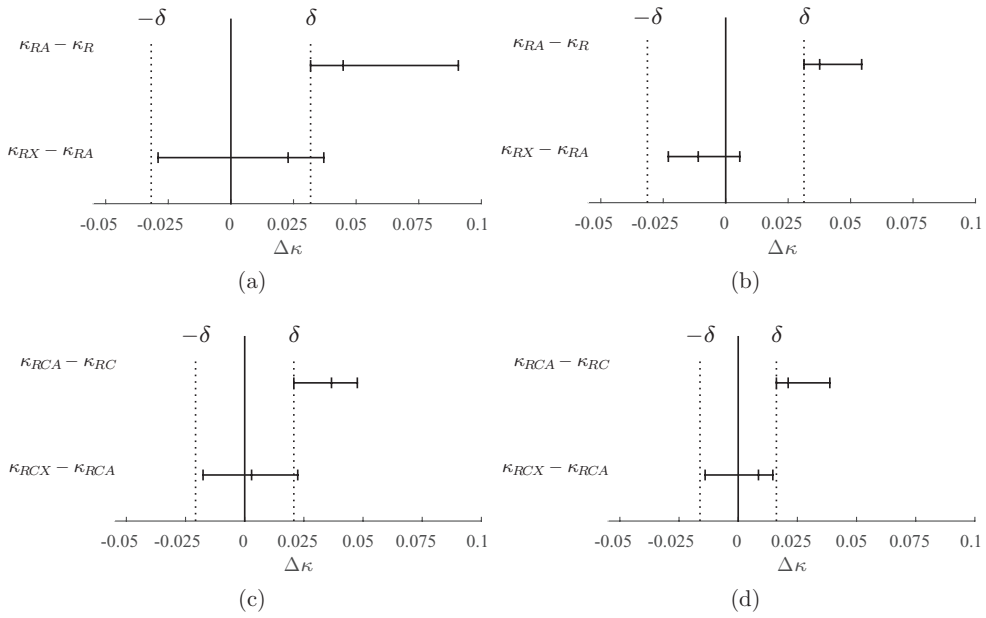

Figure 6. Confidence intervals (90%) for difference in κ obtained with different feature sets with (surrogate) actigraphy using respiratory features for (a) healthy and (b) all subjects, and using respiratory and cardiac features for (c) healthy and (d) all subjects. The lower bound of the CI for $\kappa_{\mathrm{RA}} - \kappa_{\mathrm{R}}$ and $\kappa_{\mathrm{RCA}} - \kappa_{\mathrm{RC}}$ is used to establish the equivalence margin δ for the comparisons $\kappa_{\mathrm{RX}} - \kappa_{\mathrm{RA}}$ and $\kappa_{\mathrm{RCX}} - \kappa_{\mathrm{RCA}}$ respectively. Equivalence for a given comparison is established if the corresponding CI is fully comprised between $(-\delta, \delta)$ (indicated with the vertical dashed lines). Non-inferiority is established if the lower bound of the CI is larger than $-\delta$. In this case, non-inferiority with surrogate actigraphy can be established for healthy subjects and equivalence can be established for all subjects.

These results suggest that surrogate actigraphy is adequate for quantifying small body movements and, overall, to detect the presence of movements of any intensity and henceforth, to classify *sleep* and *wake*. This is confirmed by the results of *sleep*/*wake* classification, where surrogate actigraphy achieves statistically non-inferior performance to reference actigraphy when combined with RIP and/or ECG signals, both for healthy as well as for all subjects.

5. Conclusions

Although recent work in unobtrusive *sleep*/*wake* classification has shown that cardiac and respiratory features are useful, actigraphy remains the single most discriminative modality for this task. Unfortunately it requires an additional dedicated sensor, typically a wrist-worn actigraphy device with an accelerometer. This paper proposes a method to estimate actigraphy from the body movement artifacts present in the ECG and RIP signals based on the time-frequency analysis of those signals. It provides a surrogate measure of actigraphy with moderate correlation with reference actigraphy, and reasonable agreement on detection of body movements of low intensity. More importantly, it can be used as a replacement of actigraphy in *sleep*/*wake* classification without loss in performance. Although the best performance is obtained with ECG and RIP combined, it should be emphasized that the largest gains of adding actigraphy for *sleep*/*wake* classification were found using RIP alone. Importantly, RIP is

one of the most used modalities in home sleep monitoring: together with pulse oximetry and respiratory flow, it is very often used for screening or diagnosis of sleep disordered breathing with home sleep tests.

The consequences are important: in case only respiratory effort and actigraphy are used for *sleep/wake* classification, this method eliminates the need to use an actigraphy device, relying only on the signals measured with a single (RIP) sensor. In case both respiratory effort and ECG are used together with actigraphy, it can also be used to eliminate the need for an actigraphy device, allowing the setup to be restricted to two sensors (ECG and RIP). Although the economic impact of this benefit might seem small given the low price of accelerometers, it offers a major benefit in terms of comfort, in particular for longer-term monitoring. Additionally, it is immediately applicable for legacy ECG and RIP monitoring devices already used in clinical practice and which do not have an accelerometer built-in. Finally, it allows the retrospective analysis of existing sleep data sets for which no actigraphy was collected but where body movements may be of clinical importance.

Disclosures

The authors declare to have no conflicts of interest.

References

Aoude A A, Motto A L, Galiana H L, Brown K A and Kearney R E 2006 Power-based segmentation of respiratory signals using forward-backward bank filtering *Proc. of the Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* vol 1 pp 4631–4

Bastien C H, Vallières A and Morin C M 2001 Validation of the insomnia severity index as an outcome measure for insomnia research *Sleep Med.* **2** 297–307

Bland J M and Altman D G 1999 Measuring agreement in method comparison studies *Stat. Methods Med. Res.* **8** 135–60

Buysse D J, Reynolds C F, Monk T H, Berman S R and Kupfer D J 1989 The pittsburgh sleep quality index: a new instrument for psychiatric practice and research *Psychiatry Res.* **28** 193–213

Clifford G D and Moody G B 2012 Editorial: signal quality in cardiorespiratory monitoring *Physiol. Meas.* **33** E01

Cohen J 1960 A Coefficient of agreement for nominal scales *Educ. Psychological Meas.* **20** 37–46

Cole R J, Kripke D F, Gruen W, Mullaney D J and Gillin J C 1992 Automatic sleep/wake identification from wrist activity *Sleep* **15** 461–9

Costa M, Goldberger A and Peng C K 2002 Multiscale entropy analysis of complex physiologic time series *Phys. Rev. Lett.* **89** 068102

Daubechies I 1990 The wavelet transform, time-frequency localization and signal Analysis *IEEE Trans. Inf. Theory* **36** 961–1005

Draper N R and Smith H 1998 *Applied Regression Analysis* 3rd edn (New York: Wiley)

Duda R O, Hart P E and Stork D G 2000 *Pattern Classification* 2nd edn (New York: Wiley)

Euser A M, Dekker F W and le Cessie S 2008 A practical approach to Bland–Altman plots and variation coefficients for log transformed variables *J. Clin. Epidemiol.* **61** 978–82

Fonseca P, Aarts R M, Foussier J and Long X 2014 A novel low-complexity post-processing algorithm for precise QRS localization *SpringerPlus* **3** 376

Fonseca P, Long X, Foussier J and Aarts R M 2013 On the impact of arousals on the performance of sleep and wake classification using actigraphy *Proc. of the Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* vol 2013 pp 6760–3

Fox J 1991 *Regression Diagnostics: an Introduction* (Newbury Park, CA: Sage)

Hamilton P S and Tompkins W J 1986 Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database *IEEE Trans. Biomed. Eng.* **33** 1157–65

Hamilton P 2002 Open source ECG analysis *IEEE Computers in Cardiology* pp 101–4

Hollander M, Wolfe D A and Chicken E 2013 *Nonparametric Statistical Methods* 3rd edn (New York: Wiley)

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

Keenan D B and Wilhelm F H 2005 Adaptive and wavelet filtering methods for improving accuracy of respiratory measurement *Biomed. Sci. Instrum.* **41** 37–42

Lee J, McManus D D, Merchant S and Chon K H 2012 Automatic motion and noise artifact detection in holter ECG data using empirical mode decomposition and statistical approaches *IEEE Trans. Biomed. Eng.* **59** 1499–506

Liu S H 2010 Motion artifact reduction in electrocardiogram using adaptive filter *J. Med. Biol. Eng.* **31** 67–72

Long X, Fonseca P, Foussier J, Haakma R and Aarts R M 2014 Sleep and wake classification with actigraphy and respiratory effort using dynamic warping *IEEE J. Biomed. Health Inform.* **18** 1272–84

Morgenthaler T *et al* 2007 Practice parameters for the use of actigraphy in the assessment of sleep and sleep disorders: an update for 2007 *Sleep* **30** 519–29

Motto A L, Galiana H L, Brown K A and Kearney R E 2004 Detection of movement artifacts in respiratory inductance plethysmography: performance analysis of a neyman-pearson energy-based detector *Proc. of the Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* vol 1 pp 49–52

Paquet J, Kawinska A and Carrier J 2007 Wake detection capacity of actigraphy during sleep *Sleep* **30** 1362–9 (PMID: 17969470)

Pawar T, Chaudhuri S and Duttagupta S P 2007 Body movement activity recognition for ambulatory cardiac monitoring *IEEE Trans. Biomed. Eng.* **54** 874–82

Redmond S J, de Chazal P, O'Brien C, Ryan S, McNicholas W T and Heneghan C 2007 Sleep staging using cardiorespiratory signals *Somnologie-Schlafforschung Schlafmedizin* **11** 245–56

Redmond S J and Heneghan C 2006 Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea *IEEE Trans. Biomed. Eng.* **53** 485–96

Romero I 2010 PCA-based Noise Reduction in Ambulatory ECGs *IEEE Computing in Cardiology* pp 677–80

Sadeh A and Acebo C 2002 The role of actigraphy in sleep medicine *Sleep Med. Rev.* **6** 113–24

Walker E and Nowacki A S 2011 Understanding equivalence and noninferiority testing *J. Gen. Intern. Med.* **26** 192–6