

Automated detection of obstructive sleep apnoea at different time scales using the electrocardiogram

Philip de Chazal^{1,3}, Thomas Penzel² and Conor Heneghan^{1,3}

¹ BiancaMed Ltd., NovaUCD, UCD, Belfield, Dublin 4, Ireland

² Division of Pulmonary Diseases, Philipps University Marburg Medical Center, Baldingerstraße, 35043 Marburg, Germany

³ Department of Electronic and Electrical Engineering, University College Dublin, Belfield, Dublin 4, Ireland

E-mail: philip.dechazal@biancamed.com, conor.heneghan@ucd.ie and penzel@med.uni-marburg.de

Received 14 February 2004, accepted for publication 18 June 2004

Published 22 July 2004

Online at stacks.iop.org/PM/25/967

doi:10.1088/0967-3334/25/4/015

Abstract

An automated classification algorithm is presented which processes short-duration epochs of surface electrocardiogram data derived from polysomnography studies, and determines whether an epoch is from a period of sleep disordered respiration (SDR) or normal respiration (NR). The epoch lengths considered were 15, 30, 45, 60, 75, and 90 s. Epochs were labeled as 'NR' or 'SDR' by a human expert, based on standard polysomnography interpretation rules. The automated classification algorithm was trained and tested on a database of 70 overnight ECG recordings from subjects with and without obstructive sleep apnoea syndrome (35 used for training, 35 for independent validation). Depending on the epoch length, the classifier correctly labeled between 87% (15 s epochs) and 91% (60 s epochs) of the epochs in the test set. Accuracy was lowest for the shortest (15 s) and longest (90 s) epoch lengths, but the analysis was relatively insensitive to choice of epoch length. The classifications from these epochs were combined to form an overall summary measure of minutes-of-SDR, allowing per-subject classification.

Keywords: Electrocardiogram, sleep apnoea, linear discriminant analysis, classification

1. Introduction

Sleep apnoea (AASM Task Force 1999) is a cardiorespiratory disorder characterized by brief interruptions of breathing during sleep. Typical sleep patterns of a sufferer involve heavy

snoring interspersed with obstruction of the upper airway, leading to waking and gasping for breath. Often the sufferer has no recollection of the sleep interruptions that can occur hundreds of times in a night. The primary health implications (Bradley and Flora 2000) of sleep apnoea are its impact on the heart (increased levels of hypertension, coronary arterial disease, arrhythmias, increased accident levels due to sleepiness, and quality of life issues). Sleep apnoea is typically divided into two classes; central sleep apnoea (CSA) in which respiratory drive is absent or inhibited, and obstructive sleep apnoea (OSA) in which upper airway collapse is responsible for disrupted respiration. While central events are often seen in subjects with OSA, pure CSA is relatively rare. Obstructive sleep apnoea, however, is not a rare condition. It occurs in 2% to 4% of middle-aged adults (Young *et al* 1993) and in 1% to 3% of preschool children (Gislason and Benediktsdottir 1995). Overall it is estimated that there are 10 to 20 million sufferers in the U.S. alone. However, despite the fact that apnoea has such health and quality of life implications, there is a surprisingly low public and medical awareness of the illness. Of the 10 to 20 million sufferers in the U.S., it is estimated that only 10 to 15% have been diagnosed (Young *et al* 1997). Given appropriate diagnosis, the outlook for sufferers is encouraging, since an effective therapy exists. This therapy is called continuous positive airway pressure (CPAP). It consists of a small pump supplying air via a facemask during sleep, at sufficient pressure to keep the soft tissues of the upper airway open, and it is well agreed that when used properly CPAP effectively eliminates periods of disordered breathing in OSA subjects.

Overnight polysomnography (PSG), a laboratory sleep study, is regarded as the gold standard of sleep apnoea diagnosis (Whitney *et al* 1998). A PSG monitors several body functions during sleep and requires that the patient be hooked up to multiple probes (usually 16+) while staying overnight in a laboratory setting. It is widely agreed that PSG is a thorough and reliable test. However, it also receives its share of criticism. Firstly, PSG is inconvenient since it requires the patient to stay in hospital for one night. Secondly, it is an expensive process. This high cost is due to the need for the study to take place in a hospital setting, the requirement to have a sleep technician in attendance overnight, and the need to manually 'score' the resultant measurements. Thirdly, many sleep centers worldwide are currently operating at full capacity and PSG usually suffers from a low availability reflected in up to 6 month-long waiting lists for testing. Hence, techniques which provide a reliable diagnosis of sleep apnoea with fewer and simpler measurements, and without the need for a specialized sleep laboratory may be of benefit.

One physiological signal that has considerable promise for providing a simple low-cost technique for assessment of OSA is the surface electrocardiogram (ECG). As far back as 1984, Guilleminault proposed that examination of the RR-intervals derived from the ECG could provide a useful screening technique for OSA (Guilleminault *et al* 1984). His suggestion was based upon the observation that apnoea events are often associated with characteristic bradycardia/tachycardia variations in heart rate. Automated detection and quantification of these so-called cyclical variations in heart rate (CVHRs) was therefore suggested as a convenient screening methodology. However, to our knowledge this idea has not yet been incorporated into common clinical practice, though a variety of recent publications have revisited this original idea (Roche *et al* 1999, Stein *et al* 2003). Interest in this field was particularly renewed by the *Computers in Cardiology Challenge of 2000* (Moody *et al* 2000), which sought the development of algorithms for automatic detection of periods of OSA solely through analysis of the ECG over one-min time-scales. Two of the current authors (PDC and CH) successfully developed a technique to perform this task, and this paper extends the results reported by de Chazal *et al* (de Chazal *et al* 2003) to consider the reliability of detecting episodes of OSA over a range of time-scales from 15 to 90 s. Demonstrating reliability over

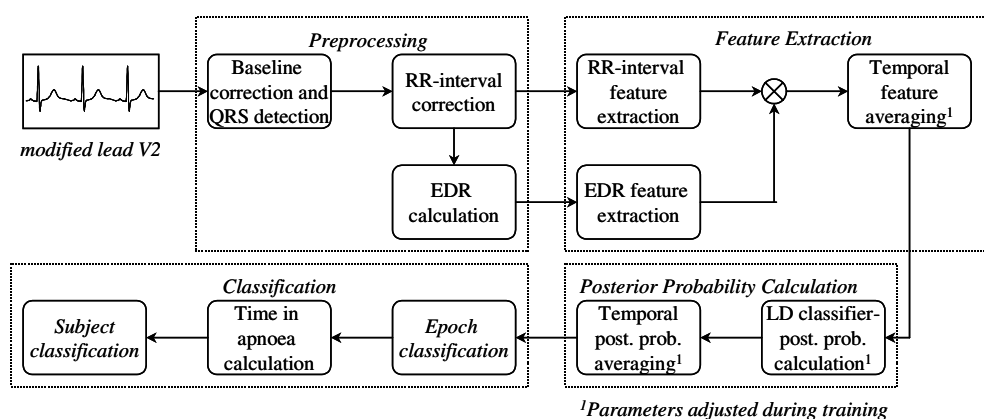


Figure 1. Schematic representation of an automated system for the detection of obstructive sleep apnoea using single lead ECG measurements. The processing steps include detection of QRS complexes leading to an RR-interval sequence, and the calculation of an ECG-derived respiratory (EDR) signal. Using these RR-intervals and the EDR signal, features were constructed for each epoch of recording. Using a linear discriminant (LD) classifier, each epoch can then be assigned a probability of containing sleep disordered respiration. The probability can be thresholded at 0.5 to produce an *epoch classification*. Combining these epoch classifications over the recording results in a *subject classification*, which can be used by a physician as a screening measure. The training data was used to determine the optimal values for the adjustable parameters for temporal feature averaging, LD classifier and temporal posterior probability averaging steps of the system.

the shorter time scales is of particular interest since apnoeas and hypopnoeas typically have a duration of 10–20 s, so reliable detection over these time scales may allow mapping to individual disordered breathing events.

2. Aim

The aim of this study was to assess the performance of a classifier for processing short-duration epochs of surface ECG data acquired from overnight studies. The classifier determines whether an epoch is from a period of sleep disordered respiration (SDR) or normal respiration (NR). The epoch classifications are then combined to determine the minutes-of-SDR per subject. The epoch lengths considered were 15, 30, 45, 60, 75, and 90 s. The same system (see figure 1) used by de Chazal *et al* (de Chazal *et al* 2003) was used in this study. For each epoch length the performance of the system was optimized using training data and independent data used to assess the classification performance.

This paper is organized as follows. Section 3 describes the database of subjects and measurements. Section 4 details the analysis methodology of the paper including a description of the algorithm steps and the optimization of the algorithm for different epoch lengths. In section 5, we detail the results of our automated system. In section 6, a discussion of our results is presented, and we summarize our conclusions regarding the potential usefulness of this system in section 7.

3. Subjects and measurements

The database of ECG signals used in the *2000 Computers in Cardiology Conference Challenge* was used in this study (Penzel 2000). It contains 70 night-time recordings of a single continuous

ECG signal of approximately 8 h duration. The ECG recordings were extracted from a larger database of simultaneously recorded polysomnogram measurements provided by Philipps-University, Marburg, Germany. The ECG signal was sampled at 100 Hz, with 16-bit resolution, with one sample bit representing $5 \mu\text{V}$. The standard sleep laboratory ECG electrode positions were used (modified lead V2).

The database does not contain episodes of pure central apnoea or of Cheyne–Stokes respiration; all apnoeas and hypopnoeas in these recordings are either obstructive or mixed. The subjects of these recordings were men and women between 27 and 63 years of age (mean: 43.8 ± 10.8 years) with weights between 53 and 135 kg (mean: 86.3 ± 22.2 kg). The sleep recordings originated from 32 subjects (25 men, 7 female) that were recruited for previous studies on healthy volunteers and patients with obstructive sleep apnoea. Four subjects contributed a single recording each, twenty-two subjects contributed two recordings each, two subjects contributed three recordings each, and four subjects contributed four recordings each. The duration of the recordings varied between 401 and 578 min (mean: 492 ± 32 min). The Apnoea–Hypopnoea index (AHI) ranges from 0 to 93.5 in these recordings. The initial scoring of apnoeas and hypopnoeas in the recordings was done according to standard criteria based on oronasal airflow, respiratory movement, and pulse oximetry (American Academy of Sleep Medicine (AASM) Task Force 1999). Hypopnoeas were defined as intermittent drops in the volume of air entering the lungs on each breath below 50% of normal, accompanied by drops in oxygen saturation of at least 4%, and followed by compensating hyperventilation.

For the purposes of this study one expert (author—T Penzel) rescored the database as follows: each recording was visually examined and periods of normal respiration (NR) and sleep disordered respiration (SDR) identified. Periods of SDR contained apnoea or hypopnoea and associated recovery breaths. No differentiation between apnoea and hypopnoea events was made when events of disordered breathing were scored. If two or more disordered breathing events were not separated by any normal breathing the entire event was scored as one continuous disordered breathing event.

The standard AHI criteria recommended by the American Academy of Sleep Medicine (AASM Task Force 1999) for classing each *recording* were adapted so that the duration of SDR could be used to discriminate between ‘normal’, ‘borderline’ and ‘apnoea’ subjects. The assessment based on the duration of the disordered breathing was as follows:

- Recordings classed as apnoea contained at least 100 min of disordered respiration. There were 40 recordings in this class. All recordings in this class had an AHI above 15 events per h.
- Recordings classed as borderline contained between five and 99 min with disordered breathing. There were ten recordings in this class. The recordings revealed either mild apnoea, up to an AHI of 15 events per h, or obstructive snoring in otherwise healthy subjects.
- Recordings classed as normal (or control) contained fewer than five min with disordered breathing. There were 20 recordings in this class. All recordings had an AHI of less than 5 in this class.

The database was divided into two sets each containing 35 recordings. The first set (Dataset 1) was used to optimize the classification algorithm at each epoch length and the second set (Dataset 2) was used to provide an independent performance assessment. The recordings were split so each set had 20 apnoea, 5 borderline and 10 normal recordings. Seventeen subjects contributed recordings to both the released and withheld-set. Eight subjects’ recordings were only in the withheld-set, while the recordings of the remaining seven subjects were in the released-set only.

4. Analysis methods

A schematic flowchart of the system used in this study is shown in figure 1. It provides two outputs; the first output is a sequence of epoch classifications as 'NR' or 'SDR', and the second output provides an overall summary of the presence of clinically significant apnoea and is derived on the basis of the epoch annotation sequence. The system has a number of adjustable parameters that were set so that the per-epoch classification performance of the algorithm for the different epoch lengths was maximized. The parameters that could be adjusted included:

- (1) the number of epochs averaged in the feature averaging step;
- (2) the parameter values of the linear discriminant classifier used to calculate the posterior probabilities; and
- (3) the number of epochs averaged in the posterior probability averaging step.

The system in figure 1 is an epoch-based system that processes features based on the timing of QRS complexes and features calculated from an ECG-estimated respiratory signal. The motivation for the design of this system was that a number of previous studies that have shown that features based on the timing of QRS complexes (Guilleminault *et al* 1984, Penzel *et al* 1990, Hilton *et al* 1999, Roche *et al* 1999) are useful for apnoea identification. In addition, a variety of studies have shown that respiratory information can be extracted from the amplitude of the ECG (Moody *et al* 1986, Travaglini *et al* 1998) although these studies were not carried out solely in the context of apnoea detection. The study of Penzel *et al* (Penzel *et al* 2002a) showed that features based on heart rate variability and respiratory information were useful for apnoea identification.

The expert annotations provided with the database are event-based i.e. the start and finish of an annotation correspond to the start and finish of the respiration event. As our system is epoch-based the first step was to map the expert annotations to epoch-based annotations. To achieve this the annotation time sequence was divided into epochs and the annotation of each epoch was assigned to 'SDR' if for more than half the epoch the corresponding annotation time sequence was disordered breathing, otherwise the epoch was assigned to 'NR'. Figure 2 shows this process for 30 s epochs. Using this method six sets of epoch-based annotations were produced for the different epoch lengths between 15 and 90 s.

The system shown in figure 1 has four major steps (1) preprocessing, (2) feature extraction, (3) posterior probability calculation and (4) classification. These steps are described below.

4.1. Preprocessing

The purpose of the preprocessing step was to determine the timing of QRS complexes and estimate the respiratory signal using the ECG.

4.1.1. Detection of QRS complexes and RR-intervals. All the features used in this study required QRS detection times. A 'QRS detection time' is loosely defined as the time of occurrence of the QRS complex in an ECG signal. Different algorithms have been optimized to reliably capture different components of the QRS complex (e.g., R-peak, point of steepest up-slope, onset point, etc.). In this study, QRS detection times were generated automatically for all recordings using a previously described algorithm (Engelse and Zeelenberg 1979). This algorithm provides detection times that occur close to the onset of the QRS complex. On this database, this QRS detection method was found to be 98.6% accurate in identifying true QRS complexes (de Chazal *et al* 2003).

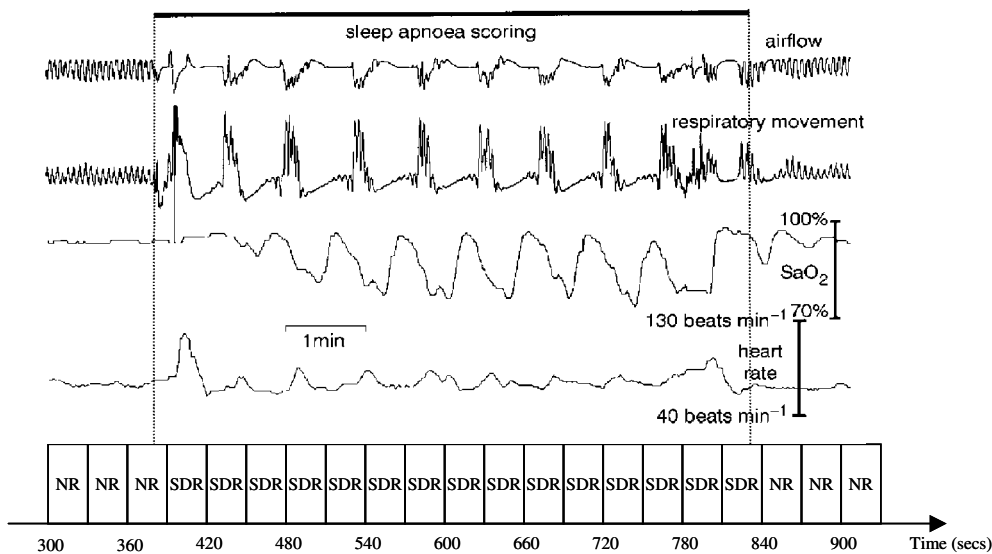


Figure 2. Indicative example of mapping polysomnogram interpretation to epoch labels. In this example, the epoch length is set at 30 s. At time $t = 380$ s, a sequence of repeating apnoeas/recovery breaths occurs and ends at $t = 822$ s. The labeling of the epochs is determined by whether more than 50% of the interval is disordered breathing.

RR-intervals were defined as the interval between successive QRS detection points. Due to limited signal quality which leads to errors in the automatically generated QRS detections, the RR-interval sequences generated from a set of QRS detection times contained physiologically unreasonable times. A first preprocessing step prior to calculating the ECG features was to calculate a corrected RR-interval sequence where all intervals were physiologically reasonable. The following automatic algorithm was developed for this purpose.

Suspect RR-intervals were found by applying a median filter of width five to the sequence of RR-intervals. This provided a robust estimate of the expected value for each RR-interval. Significant variations from this expected value led to it being flagged as a suspect RR-interval. Suspect RR-intervals could be due to either spurious QRS detections, or missed QRS complexes.

Spurious QRS detections were found by comparing the sum of adjacent RR-intervals with the robust RR-interval estimate. If this sum was numerically closer to the robust estimate than either of the individual RR-intervals, then a spurious detection was deemed to be present. The two RR-intervals were merged to form a single RR-interval.

Conversely, we determined heuristically that if an RR-interval was a factor of 1.8 times or greater than the robust estimate then it was probable that one or more QRS complexes were missed. This scenario can occur due to significant background noise levels, or contact electrode noise. To estimate (interpolate) the times of the missing QRS complexes the RR-interval was divided by the sequence of integers 2, 3, 4, ..., until it best matched the robust estimate of the RR-interval. The single RR-interval was then subdivided by the appropriate integer to form a series of new detections.

4.1.2. ECG-derived respiratory signal. During the breathing cycle, the body-surface ECG is influenced by electrode motion relative to the heart and by changes in thoracic electrical impedance as the lungs fill and empty with air. The effect is most obviously seen as a slow

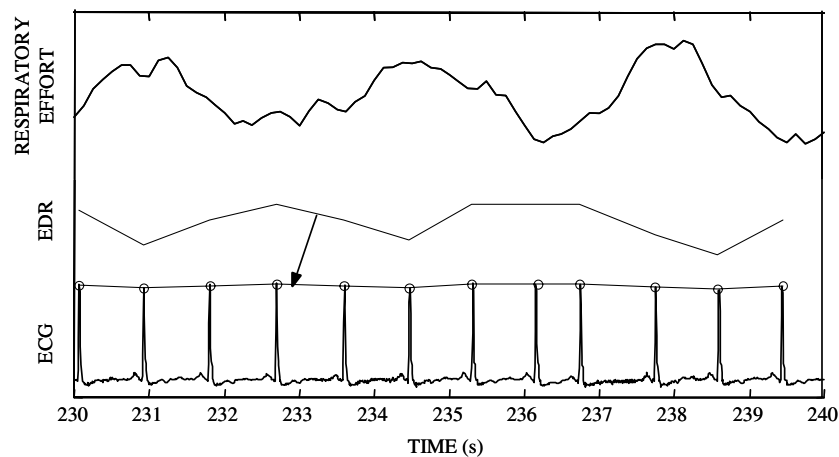


Figure 3. An illustration of the ECG derived respiratory (EDR) signal. The top trace shows recorded ribcage motion using inductance plethysmography over a 10 s period. The bottom trace shows the recorded ECG, and the middle trace shows a scaled EDR signal formed by following the modulation of the ECG signal.

modulation of the ECG amplitude at the same frequency as the breathing cycle (Moody *et al* 1986, Travaglini *et al* 1998). To quantify this modulation effect, the original unprocessed ECG signal was filtered with two median filters to remove the baseline wander. The original ECG signal was processed with a median filter of 200 ms width to remove QRS complexes and P waves. The resulting signal was then processed with a median filter of 600 ms width to remove T waves. The signal resulting from the second filter operation contained the baseline of the ECG signal, which was then subtracted from the original signal to produce a baseline corrected ECG signal.

A sample point of an ECG-derived respiratory signal (EDR) was then obtained by calculating the area enclosed by the baseline corrected ECG in the region 100 ms beyond the QRS detection point. This EDR signal is an unevenly sampled sequence, with samples corresponding to the QRS detection times (see figure 3). In a separate study we showed that this method correlates well with plethysmogram signals (correlation coefficient of 0.78) and is insensitive to body position (O'Brien *et al* 2004).

4.2. Feature extraction

The processing steps outlined above resulted in discrete index sequences of the RR-intervals and the EDR signal. Based on these, the following features generated for each epoch that represented the pertinent information for classification were calculated:

- mean epoch and recording RR-interval
- standard deviation of the epoch and recording RR-interval
- the first five serial correlation coefficients of the RR-intervals (Ambardar 1995)
- the NN50 measure (variant 1), defined as the number of pairs of adjacent RR-intervals where the first RR-interval exceeds the second RR-interval by more than 50 ms
- the NN50 measure (variant 2), defined as the number of pairs of adjacent RR-intervals where the second RR-interval exceeds the first RR-interval by more than 50 ms
- two pNN50 measures, defined as each NN50 measure divided by the total number of RR-intervals

- the SDSD measures, defined as the standard deviation of the differences between adjacent RR-intervals
- the RMSSD measure defined as the square root of the mean of the sum of the squares of differences between adjacent RR-intervals
- the Allan factor $A(T)$ evaluated time scales T of 5, 10, 15, 20 and 25 s (depending on epoch length) where the Allan factor is defined as $A(T) = \frac{E\{[N_{i+1}(T) - N_i(T)]^2\}}{2E\{N_{i+1}(T)\}}$, $N_i(T)$ is the number of QRS detection points occurring in a window of length T stretching from iT to $(i + 1)T$, and E is the expectation operator
- interval-based power spectral density of the RR-intervals (number of features dependent on epoch length)
- mean epoch and recording EDR amplitude
- standard deviation of the epoch and recording EDR amplitude
- the power spectral density of the EDR signal (number of features dependent on epoch length)

The RR-interval measurements are routinely used in heart rate variability analysis (Task Force of ESC and NASPE 1996, Teich *et al* 2000). It is worth noting that none of the measures listed above consider changes in the morphology of the ECG due to altered conduction mechanisms in the heart. It is implicitly assumed that the processes leading to apnoea occur at a location external to the heart and thus do not directly affect the generated cardiac potentials.

The interval-based RR-interval PSD was calculated in the following way (DeBoer *et al* 1984). A sequence of RR-intervals was associated with each epoch. The index for this sequence was beat number, not time. The mean RR-interval for that epoch was removed from each value, to yield a zero-mean sequence. The sequence was zero-padded to length $N = 256 \times \text{epoch length}/60$ where the epoch length is expressed in s, and the fast Fourier transform (FFT) was taken of the entire sequence. The magnitudes of the FFT coefficients were squared to yield a periodogram estimate of the PSD, which had high variance. Averaging of four adjacent frequency bins yielded an $N/4$ -point PSD estimate of which only the first $N/8$ points were used as features (due to the symmetry of the upper and lower PSD point estimates). The x -axis has units of cycles/interval. The EDR power spectral density was calculated in an identical fashion, with the spectral variable also defined as cycles/interval.

4.2.1. Feature averaging. Before the features were used as input to the classifier a smoothing operation over epochs was performed. The feature value output of the smoothing operation for an epoch was the average of the feature values for that epoch with the surrounding epochs. Different widths of filter were trialed and an optimal filter length chosen.

4.3. Per-epoch posterior probability calculation

In this study, posterior probability calculation was carried out using linear discriminant analysis (Ripley 1996). A discriminant value was derived for each sample feature vector and then mapped to a posterior probability of apnoea. Optimization of the classifier parameters was achieved by the method of maximum likelihood using the training data using 'plug-in' values (Ripley 1996). Equations for this can be found in de Chazal *et al* (de Chazal *et al* 2003).

The classifier model considered in this study implicitly assumes that the feature data have a class-dependent Gaussian distribution. Classifier performance will be degraded when the actual feature statistics differ significantly from this assumption. Therefore, where appropriate, a transformation was applied to the features so that the histogram of the transformed feature more closely approximated a Gaussian distribution. For our choice of features, all PSD,

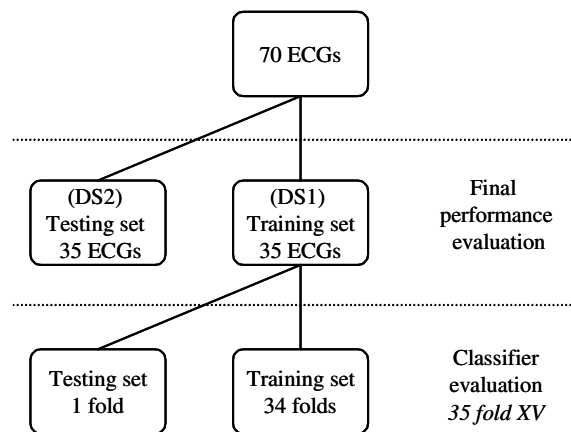


Figure 4. The division of the data into Dataset 1 (DS1) and Dataset 2 (DS2) used in the selection of a configuration at each epoch length. Dataset 1 was used for evaluating different classifier configurations and utilized the cross-validation (XV) scheme. Final performance evaluation of the chosen configuration was performed by retraining the configuration on all of DS1 and evaluating on an independent set (DS2).

RMSSD, and SDDSD features required a logarithmic transform. The Allan variance features required a square-root transform.

4.3.1. Posterior probability averaging. Before the posterior probabilities were thresholded to determine the final classification for each epoch, the probability values were averaged over adjacent epochs. Different widths of filter were trialed and an optimal filter length chosen.

4.4. Per-epoch and per-subject classification

A final classification for each epoch was obtained by thresholding the posterior probability estimate at a value of 0.5. If the posterior probability was equal to or below 0.5 then the epoch was assigned to NR otherwise it was assigned to SDR.

The screening for clinically significant OSA for each recording was achieved by calculating the average time spent in SDR from the epoch classifications. To classify on a per-subject basis the same thresholds applied to expert annotations were used i.e. *normal*: minutes-of-SDR ≤ 5 ; *borderline*: minutes-of-SDR $> 5, < 100$; *apnoea*: minutes-of-SDR ≥ 100 .

4.5. Optimization and assessment of the system at each epoch length

Dataset 1 was used to optimize the per-epoch classification performance of the system at each epoch length. A number of different classifier configurations were constructed by varying the length of the feature averaging and posterior probability averaging filters. For each configuration the classifier was trained and the classification performance determined. The classifier performance measure chosen to compare the configurations was overall accuracy (Willems *et al* 1990) and it was estimated using the cross-validation scheme (Ripley 1996). The feature data in Dataset 1 was divided into 35 folds with each fold containing the feature data from one recording (see figure 4). The feature data from 34 records was used to determine the classifier parameters and then the overall accuracy of the system was determined on the

Table 1. The number of normal respiration (NR) and sleep disordered respiration (SDR) epochs from the epoch-based expert annotations at different epoch lengths.

Epoch length (s)	Dataset 1			Dataset 2		
	NR	SDR	Total	NR	SDR	Total
15	42 135	26 026	68 161	4 2776	26 211	68 987
30	21 068	13 005	34 073	21 383	13 102	34 485
45	14 042	8 670	22 712	14 245	8 740	22 985
60	10 534	6 497	17 031	10 700	6 533	17 233
75	8 428	5 190	13 618	8 556	5 225	13 781
90	7 023	4 323	11 346	7 139	4 345	11 484

remaining fold. This process was then repeated 35 times with a different fold used each time to measure the system performance. The overall accuracy of the system was the average of the 35 individual fold results. Once the overall accuracy was determined for each configuration the best performing configuration for each epoch length was selected for a second performance estimation.

While the above process facilitated choosing between different configurations the overall accuracy figures obtained were biased estimates of per-epoch performance as Dataset 1 was used to develop the system. To obtain an unbiased assessment of performance the selected configuration for each epoch length was retrained using all feature data in Dataset 1 and then the system assessed by processing Dataset 2. This assessment is unbiased as Dataset 2 was not used at any point in the development of the system. For this assessment the overall accuracy, sensitivity, specificity, positive and negative predictivity (Willems *et al* 1990) and Kappa coefficients (Cohen 1960) were determined.

5. Results

5.1. Annotation mapping and feature generation

Table 1 shows the number of NR and SDR epochs after mapping the expert event-based annotations to the epoch-based annotations using epoch lengths of 15, 30, 45, 60, 75 and 90 s. Results are shown for Dataset 1 and 2. For all epoch lengths there were approximately 62% NR and 38% SDR epochs. Accordingly the prior probabilities of the linear discriminant model were set to 0.62 for the NR class and 0.38 for the SDR class.

Table 2 shows the number of features at each epoch length. As the epoch length increases, more QRS detection points are available which results in more features (e.g., more spectral features). The number of features for an epoch varied between 36 for the 15 s epoch to 120 for the 90 s epoch. It is worth noting that there were two types of mean and standard deviation features. The epoch-based features were calculated using values from all the recording and hence had the same value for all epochs in a recording.

5.2. Optimizing the per-epoch classification performance at each epoch length using training data

Table 3(a–f) shows the classification accuracy at different levels of feature and posterior averaging for the six epoch lengths. Note that for epoch lengths 15 and 30 s, some intermediate results have not been shown for the sake of brevity. All of these results have been estimated

Table 2. The number of features used by the algorithm at different epoch lengths.

Epoch length (s)	RR-interval based features									EDR-based features			
	PSD	mean	std	NN50	pNN50	SDSD	RMSSD	A(T)	ser. corr.	PSD	mean	std	Total
15	8	2	2	2	2	1	1	1	5	8	2	2	36
30	16	2	2	2	2	1	1	3	5	16	2	2	54
45	24	2	2	2	2	1	1	4	5	24	2	2	71
60	32	2	2	2	2	1	1	5	5	32	2	2	88
75	40	2	2	2	2	1	1	5	5	40	2	2	104
90	48	2	2	2	2	1	1	5	5	48	2	2	120

Table 3. The per-epoch classification accuracy (%) determined using the cross-validation process on Dataset 1 for different levels of temporal input feature and posterior probability averaging for (a) 15 s, (b) 30 s, (c) 45 s, (d) 60 s, (e) 75 s and (f) 90 s epochs.

	Feature averaging (epochs)																	
	(a)	1	9	17	25	33	(b)	1	5	9	13	17	(c)	1	3	5	7	9
	1	75.6	87.6	88.8	89.0	88.9	1	81.2	88.7	89.8	90.0	89.9	1	83.3	87.9	89.1	89.5	89.8
9	81.0	88.4	89.1	89.1	89.0	5	85.7	89.6	90.1	90.1	90.0	3	86.9	89.0	89.7	89.8	90.0	
17	81.1	88.9	89.2	89.2	89.0	9	86.2	89.9	90.2	90.1	90.0	5	87.5	89.6	89.9	90.0	90.1	
25	81.0	88.9	89.1	89.1	89.0	13	86.0	90.1	90.3	90.2	90.0	7	87.6	89.6	90.0	90.1	90.1	
33	80.8	88.8	89.0	89.0	88.9	17	85.7	90.0	90.2	90.1	89.9	9	87.5	89.8	90.1	90.1	90.0	
	(d)	1	3	5	7	9	(e)	1	3	5	7	9	(f)	1	3	5	7	9
1	84.3	88.5	89.4	89.7	89.6	1	84.6	88.6	89.2	89.3	89.1	1	85.1	88.6	88.9	88.8	88.7	
3	87.5	89.6	89.9	89.8	89.6	3	87.5	89.2	89.5	89.4	89.2	3	87.4	89.2	89.2	89.1	88.7	
5	88.2	89.9	90.1	89.9	89.7	5	88.0	89.5	89.6	89.6	89.4	5	87.7	89.1	89.1	89.1	88.8	
7	88.0	90.0	90.0	89.9	89.8	7	87.5	89.5	89.6	89.7	89.5	7	87.3	88.9	89.1	89.0	88.7	
9	87.9	89.9	89.9	89.9	89.7	9	87.2	89.3	89.6	89.6	89.5	9	87.2	89.0	89.0	88.8	88.6	

using the cross-validation process on Dataset 1. These accuracies are reported on a per-epoch basis. It can be observed that both temporal and posterior probability averaging improve the classifier accuracy significantly.

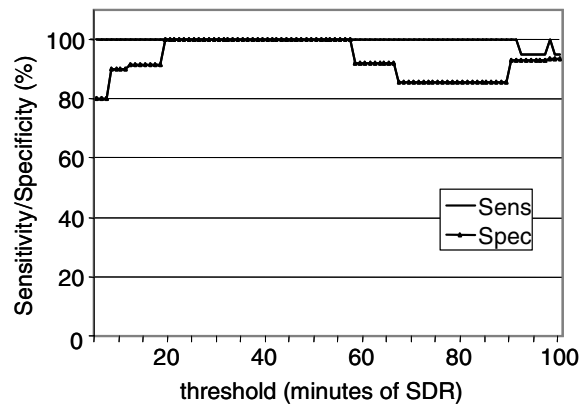
5.3. Classifier performance on Dataset 2: per-epoch and per-subject results

The best performing configurations for each epoch length (as assessed using cross-validation on Dataset 1) were retrained using all recordings of Dataset 1 and then tested using all recordings of Dataset 2. The performance of these classifiers at the various time scales is shown in table 4. The average classification accuracy observed on a per-epoch basis is 89.5% over the chosen epoch lengths, and this is relatively insensitive to chosen epoch length, though as expected performance starts to drop off at both excessively long and short epoch lengths.

In clinical practice, performance on a per-epoch basis is not as significant as on a per-subject basis. Specifically, it would be desirable to use the number of epochs of disordered breathing as a parameter for measuring the severity of the sleep apnoea. Accordingly, we converted the number of epochs to minutes-of-SDR for all records in Dataset 2 for the different epoch lengths. Table 5(a) shows the results of these calculations for all 35 records in the tests set, arranged in descending order of severity. We applied the same thresholds

Table 4. Independent per-epoch performance assessment on Dataset 2 of the best performing classifier (as determined on Dataset 1) at each epoch length.

Epoch length (s)	Feat. avg. (epochs)	PP. avg. (epochs)	TP	FN	FP	TN	Acc. (%)	Spec. (%)	Sens. (%)	+P (%)	-P (%)	Kappa stat.
15	17	17	22 919	3292	5771	37 005	86.9	86.5	87.4	79.9	91.8	0.73
30	9	13	11 698	1404	2230	19 153	89.5	89.6	89.3	84.0	93.2	0.78
45	7	7	7 839	901	1362	12 883	90.2	90.4	89.7	85.2	93.5	0.79
60	5	5	5 823	710	936	9 764	90.4	91.3	89.1	86.2	93.2	0.80
75	7	7	4 594	631	719	7 837	90.2	91.6	87.9	86.5	92.5	0.79
90	3	3	3 826	519	663	6 476	89.7	90.7	88.1	85.2	92.6	0.78

**Figure 5.** Per-subject sensitivity and specificity versus threshold level for the 60 s classifier.

used for categorizing each recording as ‘normal’, ‘borderline’ or ‘apnoea’ (see section 3) to the predicted minutes-of-SDR for the 60 s classifier to form the predicted classifications of table 5(b). Figure 5 shows a plot of the sensitivity and specificity when the system is configured to make a two-way per-subject decision (i.e. the outcome of the system is a positive or negative test for clinically significant apnoea). The plot shows the variation in sensitivity and specificity as the threshold is varied between 5 and 100 minutes-of-SDR. For example, at a threshold of 50 minutes-of-SDB then all recordings with an expert minutes-of-SDB of 50 or less were classed as ‘not clinically significant apnoea’ and all those above as ‘clinically significant apnoea’. The same threshold was applied to the predicted minutes-of-SDB and a comparison then made of the expert and predicted annotations to calculate the sensitivity and specificity. Depending on the threshold setting, the sensitivity varies between 80 and 100% and the specificity between 95 and 100%. Any threshold between 19 and 57 minutes-of-SDR resulted in complete agreement between the expert and predicted annotations.

6. Discussion

The results in table 4 show that an automated classification system can recognize epochs of sleep disordered respiration with a high degree of accuracy (approximately 90%, Kappa value 0.80) as compared to an expert human observer. In terms of per-subject performance the 60 s classifier correctly classified 31 subjects and misclassified 4 subjects (see table 5(b)). This corresponds to a per-subject accuracy of 88.6% with a Kappa value of 0.89. It categorized one apnoea subject as borderline, one borderline subject as apnoea, and two controls as

Table 5. Minutes of sleep disordered respiration determined by (a) the expert, and (b) predicted by the classifiers at each epoch length. (c) The expert and predicted per-subject class using the 60 s epoch classifier: A denotes apnoea, B borderline and N normal. Results shown for all recordings of Dataset 2.

Subject		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
(a) Expert		517	489	439	433	426	409	377	341	325	324	314	293	291	268	240	207	200	168
(b) Predicted	Epoch length = 90 s	522	476	392	387	392	447	279	383	294	342	419	215	329	287	288	236	123	225
	Epoch length = 75 s	526	483	414	416	396	445	278	398	290	325	424	243	304	283	276	240	133	233
	Epoch length = 60 s	519	487	412	416	399	440	314	397	302	341	423	218	286	275	270	243	116	231
	Epoch length = 45 s	526	493	421	441	406	440	303	383	311	339	419	236	291	284	274	248	110	257
	Epoch length = 30 s	522	492	439	437	408	446	297	412	299	342	413	255	244	282	277	252	117	287
	Epoch length = 15 s	517	498	440	479	398	449	269	460	284	335	408	242	148	267	299	250	120	347
(c) Expert		A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Predicted		A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A

Subject		19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
(a) Expert		121	120	98	66	58	12	12	3	3	2	2	1	1	0	0	0	0
(b) Predicted	Epoch length = 90 s	105	207	63	138	98	14	50	0	0	5	0	0	14	0	0	11	0
	Epoch length = 75 s	68	194	53	120	48	13	6	0	1	0	1	4	14	0	3	15	0
	Epoch length = 60 s	99	202	92	127	90	12	19	0	0	5	2	0	8	0	2	12	0
	Epoch length = 45 s	90	232	112	123	86	18	36	1	0	5	0	2	5	0	0	13	0
	Epoch length = 30 s	77	261	123	129	86	11	30	5	0	6	0	1	5	2	1	13	0
	Epoch length = 15 s	116	358	139	167	83	5	54	3	0	5	4	4	6	6	4	17	0
(c) Expert		A	A	B	B	B	B	B	N	N	N	N	N	N	N	N	N	N
Predicted		B	A	B	A	B	B	B	N	N	N	N	N	B	N	N	B	N

borderline. It is worth noting that no apnoea subjects were classed as normals and vice versa (i.e. the system was only incorrect by at most one category). This is a very encouraging result as the study by Collop (Collop 2002) showed a high degree of variability between scorers of respiratory events with Kappa values of 0.31 for scorers counting apnoea and hyponoea events. Thus the error rate between our system and the human expert used in this study is much less than the expected error between experts.

A novel aspect of this work was to consider the impact of carrying our per-epoch classifications at different time-scales. While there is some dependence on the time-scale over which classification is performed, it appears to be significantly robust across time scales from 15 to 90 s. However, we begin to see a reduction in accuracy at both short- and long-time scales as expected. At short time-scales, there is an insufficient number of QRS complexes in order to form statistically meaningful estimates of the autonomic activity or breathing changes associated with apnoea. We note also that 15 s is close to the minimum time-scale over which apnoeas are clinically defined. At longer time-scales, the effect of temporal over-averaging is seen as sections of disordered and normal respiration are mixed together, blurring the distinction between a normal respiration and sleep disorder respiration epoch. From the point of view of optimizing the per-epoch accuracy, the optimum time scale appears to be close to 60 s. A reasonable physiological interpretation of this is that 60 s is probably quite close to the natural time scale associated with an apnoea/hypopnoea, recovery breaths, and the corresponding autonomic disturbances.

A motivation for analysis of the ECG is to allow Holter recordings to be used in identifying patients likely to have obstructive sleep apnoea. A prime advantage of the proposed technique is that it comes at no additional cost for subjects undergoing Holter monitoring as part of a normal cardiology check up. Moreover, the ECG analysis technique reported here provides comparable or improved performance relative to other proposed ECG-based screening techniques. As is shown in figure 5, depending on the criterion used for defining clinically significant apnoea up to 100% sensitivity and specificity on per-subject basis may be obtained. For comparison, Roche *et al* have presented several reports on automated recognition of subjects with obstructive sleep apnoea syndrome (OSAS) using analysis of heart rate variability (Roche *et al* 1999), inter-beat interval times (Roche *et al* 2002), and wavelet-based analysis of the RR-interval series (Roche *et al* 2003). Using HRV parameters, they obtained a sensitivity of 83% and specificity of 96% on an independent test set of 52 subjects, using a criterion of AHI > 10 as defining OSAS. Using inter-beat intervals, they obtained sensitivity of 87% and specificity of 52% on 124 subjects. Finally, their wavelet-based analysis applied to 147 subjects yielded a sensitivity and specificity of 92% and 90%. Their techniques do not provide any temporal information about the occurrence of the apnoeic events. Thus on a per-subject basis, the performance of our system compares very well with other published systems.

Stein *et al* (Stein *et al* 2003) also reported a technique to identify subjects with OSAS by visual inspection of RR tachograms. A human scorer was trained to recognize characteristic cyclical variations in heart rate (CVHR) associated with obstructive events. The magnitude and frequency of occurrence of these CVHRs were then used to classify 11 control subjects and 46 clinical subjects in terms of OSAS. Of the 46 clinical subjects, 33 had significant OSAS (AHI > 15). The positive predictive accuracy was 86%, and negative predictive accuracy was 94% (which correspond to a sensitivity of 97% and specificity of 77%) in distinguishing subjects with AHI > 15 from those with lower AHIs. However, the need for human scoring may reduce the potential clinical utility of their system.

The per-subject performance of the proposed system also compares favourably with other low-cost screening methodologies such as oximetry, or reduced-parameter cardio-respiratory studies. Oximetry-based screening has been widely suggested for both adult and pediatric

populations. However, a confounding factor is that obstructive events do not always lead to significant oxyhemoglobin desaturation. In addition, different averaging times and movement artifacts may lead to false desaturation detection. Nevertheless, oximetry-based systems have provided reasonable performance in some studies. For example, Brouillette *et al* obtained a positive predictive value (PPV) of greater than 97% (over 349 subjects) in the case of children suspected of having OSA, using overnight pulse oximetry alone (Brouillette *et al* 2000). However, in order to achieve such a high PPV, their test produced significant numbers of inconclusive results (effectively lowering the sensitivity, since an inconclusive result does not rule out the presence of OSA). However, the low cost of oximetry-based systems means that they may play a useful screening role in clinical practice.

Low-cost at-home cardio-respiratory studies have also shown potential clinical utility. Such studies typically include two or more signals such as chest wall movement, pulse oximetry, heart rate, body movement or airflow. Examples of such systems include the NovaSom QSG, which captures oronasal airflow, snoring sound detector, finger pulse oximetry, and thoracic effort (Reichert *et al* 2003), and the Embletta PDS, which also captures thoracic and abdominal effort, pulse rate, oronasal airflow, oximetry, body position, and activity level (Dingli *et al* 2003). These systems have shown promising results in terms of direct comparison with polysomnography, but there are still a significant number of issues related to the number of studies which fail due to data acquisition issues.

Finally, a range of new techniques has been recently proposed for low-cost unattended screening. Peripheral arterial tonometry (PAT) is a useful technique for assessment of peripheral arterial tone, which allows a window into vasoconstriction dynamics, and hence autonomic arousal. A number of studies has shown that PAT measurements are well correlated with the number of observed apnoeic events (Penzel *et al* 2002b). Home-based pulse transit time (PTT) measurements have been shown to provide sufficient information to accurately decide whether a subject requires nasal CPAP therapy (Pitson and Stradling 1998).

Therefore, we would see Holter-based screening for sleep apnoea as providing an additional low-cost screening tool, whose particular advantages relate to ease of use, reliable performance, and familiarity to the cardiology community. Moreover, unlike oximetry or cardio-respiratory studies that are aimed at providing information about sleep disordered breathing only, a Holter-based screening will retain all the clinical benefits of arrhythmia analysis. Given the high prevalence of cardiac co-morbidities in the sleep disordered breathing population, this is potentially a significant advantage in its favour. It is also perfectly reasonable to expect that Holter-based screening may be combined with other physiological measurements such as oximetry or respiratory plethysmography to enhance clinical utility.

The current study has some limitations, which should be noted. An ECG-based system provides no information about sleep stage; in this analysis, we had knowledge of sleep start and end times. In clinical practice, we might rely upon patient event triggering or sleep logs to access such information. This study was restricted to subjects with purely obstructive or mixed apnoeas and hypopnoeas. The performance on subjects with central sleep apnoeas or Cheyne–Stokes respiration is unknown. We also note that we did not specifically control for either cardiac history or pharmacological agents such as beta-blockers, which may be expected to alter cardiac autonomic response.

7. Conclusion

Analysis of a single channel of surface ECG can be used to identify sleep disorder respiration events associated with obstructive apnoea, and subjects with OSAS with a high degree of reliability. There is a slight dependence on epoch length, but provided appropriate processing

is carried out, epoch lengths from 15 to 90 s can all be reliably used in forming an assessment. Epoch lengths close to 60 s appear to be most robust.

Acknowledgments

The authors acknowledge initial scoring of the polysomnogram data by Ludger Grote, MD, and the efforts of the Computers in Cardiology Competition Committee in collecting and disseminating the ECG-polysomnogram database. The authors are also grateful to Walter McNicholas, Philip Nolan and Patricia Boyle for their helpful insights on polysomnogram scoring and autonomic function during sleep.

References

- Ambardar A 1995 *Analog and Digital Signal Processing* (Boston, MA: PWS)
- American Academy of Sleep Medicine Task Force 1999 Sleep-related breathing disorders in adults: Recommendations for syndrome definition and measurement techniques in clinical research *Sleep* **22** 667–89
- Bradley T D and Flora J S (ed) 2000 Sleep Apnea: implications in cardiovascular and cerebrovascular disease (*Lung Biology in Health and Disease*) vol 146 (New York: Dekker)
- Brouillette R, Morielli A, Leimanis A, Waters K, Luciano R and Ducharme F 2000 Nocturnal pulse oximetry as an abbreviated testing modality for pediatric obstructive sleep apnoea *Pediatrics* **105** 405–12
- Cohen J 1960 A coefficient of agreement for nominal scales *Educ. Psychol. Measurement* **20** 37–46
- Collop N A 2002 Scoring Variability between Polysomnography Technologists in different sleep laboratories *Sleep Med.* **3** 43–7
- DeBoer R W, Karemaker J M and Strackee J 1984 Comparing spectra of a series of point events particularly for heart rate variability data *IEEE Trans. Biomed. Eng.* **31** 384–7
- de Chazal P, Heneghan C, Sheridan E, Reilly R B, Nolan P and O'Malley M 2003 Automated processing of the single lead electrocardiogram for the detection of obstructive sleep apnea *IEEE Trans. Biomed. Eng.* **50** 686–96
- Dingli K, Coleman E L, Vennelle M, Finch S P, Wraith P K, Mackay T W and Douglas N J 2003 Evaluation of a portable device for diagnosing the sleep apnoea/hypopnoea syndrome *Eur. Respir. J.* **21** 253–9
- Engelse W A H and Zeelenberg C 1979 A single scan algorithm for QRS-detection and feature extraction *Comput. Cardiol.* **6** 37–42
- Gislason T and Benediktsdottir B 1995 Snoring, apneic episodes, and nocturnal hypoxemia among children 6 months to 6 years old. An epidemiologic study of lower limit of prevalence *Chest.* **107** 963–6
- Goldberger A L, Amaral A N, Glass L, Hausdorff J M, Ivanov P C, Mark R G, Mietus J E, Moody G B, Peng C K and Stanley H E 2000 Physiobank, Physiokit, and Physionet *Circulation* **101** e215–20
- Guilleminault C, Connolly S J, Winkle R, Melvin K and Tilkian A 1984 Cyclical variation of the heart rate in sleep apnoea syndrome. Mechanisms and usefulness of 24h electrocardiography as a screening technique *Lancet* **321** 126–31
- Hilton M F, Bates R A, Godfrey K R, Chappell M J and Cayton R M 1999 Evaluation of frequency and time-frequency spectral analysis of heart rate variability as a diagnostic marker of the sleep apnoea syndrome *Med. Biol. Eng. Comput.* **37** 760–9
- O'Brien C, Doherty L and Heneghan C 2004 *Comparison of algorithms for derivation of a respiratory signal from surface electrocardiogram measurements* Dept of Electronic and Electrical Engineering, University College Dublin Internal Report
- Moody G B, Mark R G, Zoccola A and Mantero S 1986 Clinical validation of the ECG-derived respiration (EDR) technique *Comput. Cardiol.* **13** 507–10
- Moody G B, Mark R G, Goldberger A L and Penzel T 2000 Stimulating rapid research advances via focused competition: the Computers in Cardiology challenge 2000 *Comput. Cardiol.* **27** 207–10
- Penzel T, Amend G, Meinzer K, Peter J H and von Wichert P 1990 Mesam: a heart rate and snoring recorder for detection of obstructive sleep apnoea *Sleep* **13** 175–82
- Penzel T 2000 The apnoea-ECG database *Comput. Cardiol.* **27** 255–8
- Penzel T, McNames J, de Chazal P, Raymond B, Murray A and Moody G 2002a Systematic comparison of different algorithms for apnoea detection based on ECG recordings *Med. Biol. Eng. Comp.* **40** 402–7
- Penzel T, Fricke R, Jerrentrup A, Peter J and Vogelmeier C N 2002b Peripheral arterial tonometry for the diagnosis of obstructive sleep apnoea *Biomed. Tech. (Berlin)* **47** 315–7
- Pitson D J and Stradling J R 1998 Value of beat-to-beat blood pressure changes, detected by pulse transit time, in the management of the obstructive sleep apnoea/hypopnoea syndrome *Eur. Respir. J.* **12** 685–92

- Reichert J A, Bloch D A, Cundiff E and Votterria B A 2003 Comparison of the NovaSom QSG, a new sleep apnoea home diagnostic system, and polysomnography *Sleep Med.* **4** 213–8
- Ripley B D 1996 *Pattern Recognition and Neural Networks* (Cambridge: Cambridge University Press)
- Roche F, Gaspoz J M, Court-Fortune I, Minini P, Pichot V, Duverney D, Costes F, Lacour J R and Barthelemy J C 1999 Screening of obstructive sleep apnoea syndrome by heart rate variability analysis *Circulation* **100** 1411–5
- Roche F, Duverney D, Court-Fortune I, Pichot V, Costes F, Lacour J R, Antoniadis J A, Gaspoz J M and Barthelemy J C 2002 Cardiac interbeat interval increment for the identification of obstructive sleep apnoea *Pacing Clin. Electrophysiol.* **25** 1192–9
- Roche F, Pichot V, Sforza E, Court-Fortune I, Duverney D, Costes F, Garet M and Barthelemy J C 2003 Predicting sleep apnoea from heart period: a time-frequency wavelet analysis *Eur. Respir. J.* **22** 937–42
- Stein P K, Duntley S P, Domitrovich P P, Nishith P and Carney R M 2003 A simple method to identify sleep apnoea using Holter recordings. *J. Cardiovasc. Electrophysiol.* **14** 467–73
- Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology 1996 Heart rate variability—standards of measurement, physiological interpretation and clinical use *Eur. Heart J.* **17** 354–82
- Teich M C, Lowen S B, Jost B, Vibe-Rheymer K and Heneghan C 2000 Heart rate variability: measures and models *Nonlinear Biomedical Signal Processing* vol 2 ed M Akay (Piscataway, NJ: IEEE Press)
- Travaglini A, Lamberti C, DeBie J and Ferri M 1998 Respiratory signal derived from eight-lead ECG *Comput. Cardiol.* **25** 65–8
- Whitney C W, Gottlieb D J, Redline S, Norman R G, Dodge R R, Shahar E, Surovec S and Nieto F J 1998 Reliability of scoring respiratory disturbance indices and sleep staging *Sleep* **21** 749–57
- Willems J L, Abreu-Lima C, Arnaud P, Brohet C R and Denis B *et al* 1990 Evaluation of ECG interpretation results obtained by computer and cardiologists *Methods Inf. Med.* **29** 308–16
- Young T, Palta M, Dempsey J, Skatrud J, Weber S and Badr S 1993 The occurrence of sleep-disordered breathing among middle-aged adults *N. Engl. J. Med.* **328** 1230–5
- Young T, Evans L, Finn L and Palta M 1997 Estimation of the clinically diagnosed proportion of sleep apnoea syndrome in middle-aged men and women *Sleep* **20** 705–6