EDITORIAL

Robust detection of heart beats in multimodal data

IOP Publishing | Institute of Physics and Engineering in Medicine Physiological Measurement

Physiol. Meas. **36** (2015) 1629–1644 doi:10.1088/0967-3334/36/8/1629

Editorial

Robust detection of heart beats in multimodal data

Ikaro Silva[1], Benjamin Moody[1], Joachim Behar[2,3], Alistair Johnson[1,2], Julien Oster[2], Gari D Clifford[4,5] and George B Moody[1]

[1] Institute for Medical Engineering and Science, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA
[2] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK
[3] Department of Biomedical Engineering, Technion, Israel Institute of Technology, Haifa, Israel
[4] Department of Biomedical Informatics, Emory University, 201 Dowman Dr, Atlanta, GA 30322, USA
[5] Department of Biomedical Engineering, Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332, USA

E-mail: ikaro@mit.edu

Abstract

This editorial reviews the background issues, the design, the key achievements, and the follow-up research generated as a result of the PhysioNet/Computing in Cardiology (CinC) Challenge 2014, published in the concurrent focus issue of *Physiological Measurement*. Our major focus was to accelerate the development and facilitate the comparison of robust methods for locating heart beats in long-term multi-channel recordings. A public (training) database consisting of 151 032 annotated beats was compiled from records that contained ECGs as well as pulsatile signals that directly reflect cardiac activity, and other signals that may have few or no observable markers of heart beats. A separate hidden test data set (consisting of 152 478 beats) is permanently stored at PhysioNet, and a public framework has been developed to provide researchers with the ability to continue to automatically score and compare the performance of their algorithms. A scoring criteria based on the averaging of gross sensitivity, gross positive predictivity, average sensitivity, and average positive predictivity is proposed. The top three scores (as of March 2015) on the hidden test data set were 93.64%, 91.50%, and 90.70%.

Keywords: ECG, blood pressure, multimodal, beat detection, PhysioNet Challenge, heart rate, data fusion

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

1. Introduction

The ubiquitous presence of digital bedside monitors in the delivery of health care has provided an unprecedented opportunity to develop software that can robustly estimate a patient's condition. Over the past years, several large databases have been developed with concurrent recordings of multiple physiological signals, including electrocardiogram (ECG), blood pressure (BP), electroencephalogram (EEG), respiration (RESP), photoplethysmogram (PPG), and others (Welch *et al* 1991, Moody and Mark 1996, Terzano *et al* 2001, Saeed *et al* 2011). Because some of these signals carry information pertaining to the cardiovascular system (figures 1 and 2), the PhysioNet/Computing in Cardiology Challenge 2014 (the Challenge) sought to discover the optimal methods for reliably detecting heart beats by combining information from simultaneously recorded physiological waveforms (Moody *et al* 2014).

The development of software for automatic detection of heart beats (or heart rate) using either single channels of ECGs or pulsatile waveforms has a long history of accomplishments (see, for instance, Chang *et al* 2009, Chen *et al* 2009, Hamilton and Tompkins 1986, Kohler *et al* 2002, Liu *et al* 2010, Li and Clifford 2012, Mendelson 1992, Moody and Mark 1982, Okada 1979, Pahlm and Sörnmo 1984, Pan and Tompkins 1985, Portet *et al* 2005, Starmer *et al* 1973, Zong *et al* 2003a, 2003b). In addition, there have also been studies proposing methods for heart beat or rate estimation from records containing multiple ECG leads and/or extra pulsatile channels (for a review, see Pahlm and Sörnmo 1984). Gritzali *et al* (1989) used the length transform as a way to project the squared amplitude of multiple ECG channels into a single axis for improved peak detection.

Yu *et al* (2006) used a cohort of trauma patients to develop a method for reliable heart rate estimation by combining ECG and PPG heart rate estimates on the basis of their waveform quality. However, the study was limited to only 158 randomly selected 7 s data samples of trauma patients collected during helicopter transport, and compared only heart rate. Although one could expect to learn the relationships between signal quality measures and physiological changes, an enormous database of scenarios would be needed (e.g. see Behar *et al* 2013a).

An alternative approach to these essentially static methods is to incorporate temporal dynamics into the learning method to leverage the vast lengths of data. Feldman *et al* (1997) used Kalman filtering frameworks to robustly calculate heart rate from the ECG and the PPG pulsatile waveforms collected from 85 records with maximum duration of four hours (12 from an operating room, 60 from an adult ICU and 13 from a pediatric ICU). Unfortunately, no generally optimal method for combining estimates was proposed. Subsequently, Tarassenko and Townsend (2005) extended the approach to weight the fusion step by the inverse of the Kalman filter's covariance. However, this did not account for large changes due to artifacts occurring on multiple channels. Li *et al* (2008) solved this issue by including non-linearly weighted signal quality indices. The authors used a database of 6000 h of simultaneously acquired waveform from 437 ICU patients and developed a Kalman filter and signal quality-based approach to fusing both temporal and signal quality information to accurately identify changes in heart rate, and in subsequent works, blood pressure and respiration rate (Li *et al* 2008, Nemati *et al* 2010). Importantly, the authors included a significant number of pathological events, although not an exhaustive selection.

Although most physiological signals carry information that helps us differentiate cardiac events or cycles from other physiology (such as rapid breathing) or noise (e.g. movement), there have been few other attempts to combine many of the observables into a single estimator of heart rate or heart beat timing. Moreover, the inaccuracies are rarely reported and many publications simply use a convenient detector, chosen often for simplicity, as a pre-processing step. The potential errors this introduces and its lack of reproducibility or consistency is

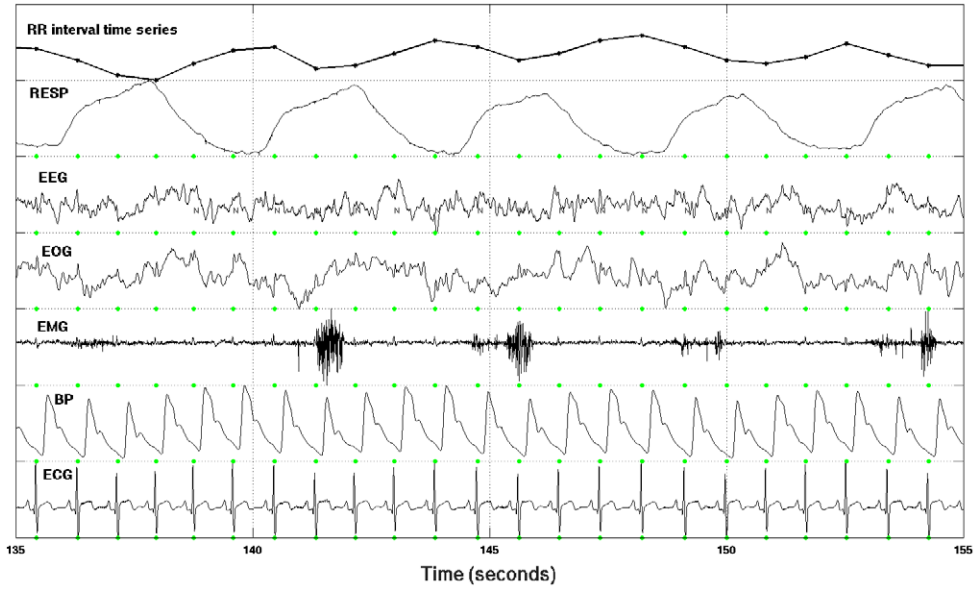

Figure 1. Example waveforms used in the Challenge. The beat annotations are marked in green. The RR interval time series derived from beat annotations is displayed for comparison with the RESP signal. Note that the EMG signal contains observable cardiac artifact. See table 1 for definition of signal labels.

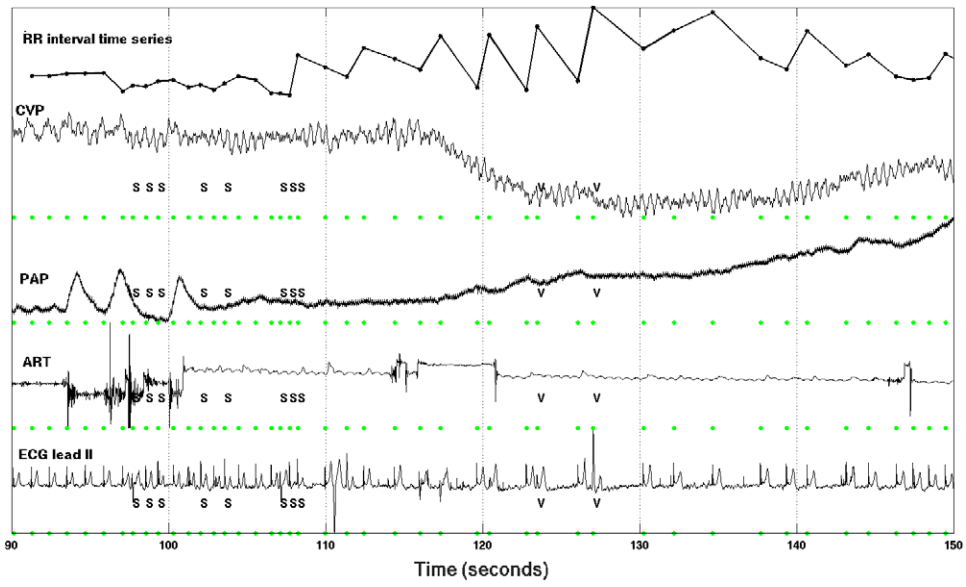

Figure 2. Example waveforms used by the Challenge containing abnormal beats. The beat annotations are marked in green. The RR interval time series derived from beat annotations is displayed for comparison with the CVP signal. The abnormal beats are labelled: S (Supraventricular premature or ectopic beat), and V (premature ventricular contraction). In addition, the normal beat also have tracings of pacemaker activity. See table 1 for definition of signal labels.

widely ignored. Furthermore, QRS detectors are often designed and tested on clean and rather limited databases, which are not representative of the application domain, such as the ICU, where noise and recording types can be very different to that of the databases. It should also be noted that detectors are rarely evaluated as a function of their final application (such as estimation of ST elevation, QT interval, respiration rate, etc). Performing fair comparisons of multi-channel algorithms developed in different data sets and without clearly defined metrics can be difficult because of the variability in the selection of these criteria. The 2014 Challenge, follows some of the themes of the 2013 Challenge (Silva *et al* 2013, Clifford *et al* 2014), exploring issues related to accurate heart rate detection. In particular, the two major aims of the 2014 Challenge and of this focus issue were to facilitate the development and comparison of robust methods for locating heart beats in long-term multi-channel recordings. A public database was compiled from records that contained ECGs, pulsatile signals that directly reflect cardiac activity and other signals that may have few or no observable markers of heart beats. A permanent hidden test data set is kept at PhysioNet and a public framework has been developed to provide researchers with the ability to automatically score and compare the performance of their algorithms. All algorithms that were successfully scored remain privately archived at PhysioNet. This allows us to efficiently re-calculate and publish performance statistics in case of changes or improvements in either the data sets or scoring criteria. Open source algorithms from the Challenge and from this focus issue are available from PhysioNet (http://physionet.org/challenge/2014/sources/).

2. Overview of the Challenge

2.1. Data description

The Challenge data set used for this focus issue has been modified with respect to the original Challenge data set (Moody *et al* 2014) in order to account for feedback received at the end of the competition. More specifically, the training set was augmented with 100 records from the original hidden test set, in an attempt to generate a more realistic and difficult training group. Thus, the public training set consists of 200 records, while the hidden test set consists of 210 records. The data sets contained signals with a maximum duration of 10 min, but several records were shorter than 10 min. The minimum, mean, and standard deviation of the record lengths, in seconds, for the hidden test set (training set) was: 13.9 (19.9), 521.7 (563.1), 160.6 (118.2).

The cohort consisted of human adults, including both patients with a wide range of cardiac irregularities and healthy volunteers. A subset of patients had implanted cardiac pacemakers. Each record contained one ECG signal and at least three additional signals (table 1). Several records included multiple pulsatile signals. The waveforms were sampled at a rate between 250 and 360 Hz; though in any given record all signals were sampled at the same fixed frequency. The signal types included arterial blood pressure (ART), general blood pressure (BP), carbon dioxide (CO_2), central venous pressure (CVP), ECG, electroencephalogram (EEG), electromyography (EMG), EOG, pulmonary arterial pressure (PAP), general pressure (pressure), nasal or abdominal respiration (RESP), oxygen level (SO_2), and stroke volume (SV).

A total of 303 510 beats were annotated (152 478 in the test set and 151 032 in the training set). All beats were manually verified by at least two humans, but errors in beat locations are likely to still exist (particularly on annotations derived from pulsatile signals and with no visible QRS in the ECG waveform to validate the fiducial point). The beats were annotated under a wide range of unusual conditions, including pacemaker activity, supraventricular tachycardia, cardiac massage, electrocautery interference, premature ectopic beats, defribillation,

Table 1. Number of signal waveforms per data set.

Signal name	Acronym	Test	Training
Arterial blood pressure	ART	135	61
General blood pressure	BP	25	116
Carbon dioxide level	CO_2	79	39
Central venous pressure	CVP	123	57
Electrocardiogram	ECG	210	200
Electroencephalogram	EEG	25	110
Electromyogram	EMG	8	44
Electrooculogram	EOG	8	44
Pulmonary arterial pressure	PAP	122	6
General pressure	Pressure	149	83
Respiration	RESP	119	213
Oxygen level	SO_2	1	23
Stroke volume	SV	1	23

fusion of paced and normal beats, flutter, and ventricular fibrillation. Roughly 95% of the beats were normal beats. For the revised data sets, no specific beat labels were provided to competitors (all beats purposely labelled as Normal by default).

2.2. Scoring criteria

Competitors submitted software that was run on the hidden test set in order to generate the competitor's beat annotations (see the section below for more details on the scoring environment). The participant's annotations on the hidden test data set were then compared to the reference annotations using the beat-by-beat algorithm defined by the ANSI/AAMI EC38 and EC57 standards, as implemented by the 'bxb' and 'sumstats' tools from the WFDB software package (Goldberger *et al* 2000). A tolerance window of 300 ms centered at the reference fiducial point was used in order to define a correctly detected beat. Each entry's output was evaluated on four performance statistics:

$$Se_{\text{gross}} = 100 \cdot \frac{TP}{TP + FN} \tag{1}$$

$$PPV_{\text{gross}} = 100 \cdot \frac{TP}{TP + FP} \tag{2}$$

$$Se_{\text{average}} = \frac{100}{n} \cdot \sum_{i=1}^{n} \frac{TP_i}{TP_i + FN_i} \tag{3}$$

$$PPV_{\text{average}} = \frac{100}{n} \cdot \sum_{i=1}^{n} \frac{TP_i}{TP_i + FP_i} \tag{4}$$

where TP, FP, and FN denote true positives (correctly detected beats), false positives (erroneously identified beats outside of the tolerance window or additional estimated beats within a tolerance window), and false negatives (undetected reference beats) respectively, and TP_i, FP_i, and FN_i denote the statistics for an individual record. The overall score for each entry was the

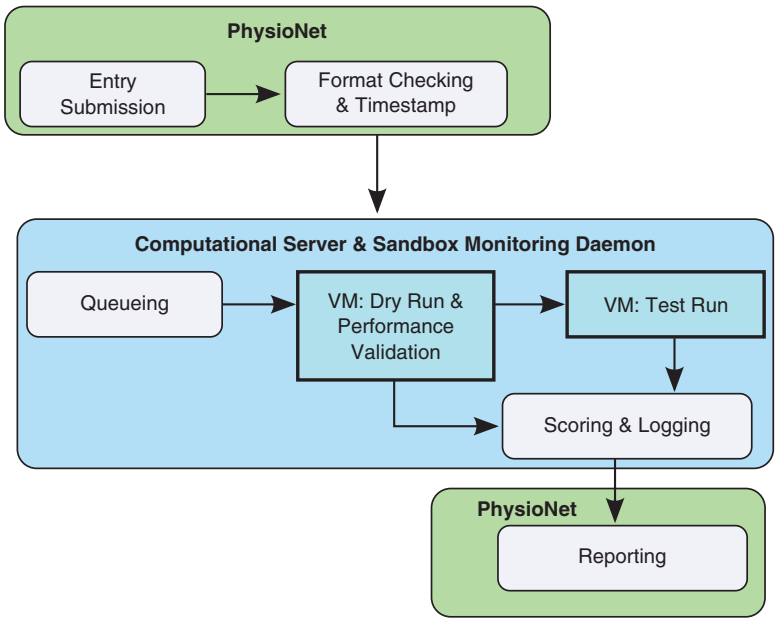

Figure 3. Diagram describing the process for automatic evaluation of Challenge entries.

average of these four statistics, equations (1)–(4). No distinction was made regarding beat types (normal and abnormal beats were treated equally).

2.3. Scoring environment

An automated scoring framework was developed on PhysioNet (Goldberger *et al* 2000) in order to grade the entries on the hidden test data set (figure 3). Competitors submitted their entries in the form of a 'zip' or 'tar' archive that included everything needed to compile and run their software on a GNU/Linux system, together with the complete set of annotations that they expected their program to produce for the records in the training set. This format allowed us to validate and score entries completely automatically, notifying competitors as soon as their entries were scored. The median response time, from the moment the user submitted an entry to PhysioNet, to the moment their scores were reported back to PhysioNet, was 64 min (including the processing of 200 training records for code validation and processing 200 hidden test records for scoring).

The competitor's algorithm was limited to 6×10^{10} CPU instructions per record. In the original Challenge, entries were allowed to run for at most 40 s per record, but we found that the exact running time was impossible to control with any precision. Feedback statistics on the number of CPU instructions used by the entry were provided via PhysioNet's web interface. If the program reached its CPU instruction limit, it was stopped at that point and scored based on the annotations it had already written.

Each time an entry was uploaded to the PhysioNet web server, it was first checked for proper formatting and then transferred to a virtual 'sandbox' system. A cloned copy of the sandbox was created for each entry. The scoring system would then unpack the archive and run the entry's setup script (compiling any code if necessary). After the initial setup, the entry code was executed individually on each record of the training set. If the program could not be

compiled, or did not produce the same annotations that the submitter obtained when running the code on the training set on their own machines, the evaluation stopped, and an error message were sent back to the submitter.

Once an entry was verified to be producing the same output as expected by the entrant on the training set, the scoring system then proceeded to compute the annotations on the hidden test set. The annotation files were collected, scored by 'bxb' and 'sumstats' as described above, and the final scores sent back to the submitter. Any errors which occurred during this portion of the evaluation were ignored, and we did not allow the program to report back any information about the test set apart from the final aggregate scores.

For this focus issue, a maximum of 20 submissions were allowed per author (not counting entries that were not scored). The submitter could choose to designate an entry as a 'dry run' by including a file named 'DRYRUN' in the archive; in this case, the entry would be tested on the training set, but not on the test set, and would not count against the user's limit of 20 entries.

The test environment consisted of a virtual 64 bit CPU running Debian GNU/Linux 7. The virtual system provided a single CPU core, 2 GB of memory, and 1 GB of virtual disk space for the program to use. In addition to the standard Debian packages, the test environment included a variety of open-source compilers, libraries, and utilities, including the WFDB software package (version 10.5.22), GNU Octave (version 3.6.2) (Eaton *et al* 2009), and OpenJDK (version 7u55). This system was hosted using KVM on a computational server with an 8-core, 2.6 GHz Opteron CPU and 32 GB of RAM; we allowed the server to run up to six virtual machines to evaluate up to three entries in parallel. Users were provided with the system information described above, and encouraged to develop their entries on their own replica of this open source environment.

3. Review of key algorithms in the Challenge

In general, each algorithm consisted of several (or all) of the following seven stages, as we now describe.

3.1. Signal phenotyping and selection

Before analysing a signal it is important to first verify that the signal contains the information you expect (i.e. an ECG signal is actually an ECG). Sometimes the signals are labelled incorrectly, or do not contain the information that one would expect from the label. An example of this is when the ECG channel has a very low amplitude and the baseline wander is strong and synchronous with respiration. More worrying, some archiving agents used in data collection (which are not designed for clinical use) report the wrong label for the signal. Finally, as an artefact related to the heart beat can manifest on atypical signals (such as the EEG), there is potential to automatically detect the presence of this artefact and incorporate the additional source of information only when appropriate (e.g. only utilise the EEG if it contains information relating to the heart contraction). Vollmer (2014) selected the signal type by applying the methods they had developed for ABP and ECG simulatenously: they classified the signal as ABP if the resulting RR series was more regular than the RR series produced by the other method (and if not, then the signal was classified as an ECG). Note that while this is designed to classify a signal as 'ECG' or 'ABP', it also incorporated other signals so long as the RR interval was sufficiently regular. De Cooman *et al* (2014) assumed that the ECG signal was labelled correctly and subsequently ran a peak detection algorithm on the ECG. The authors then estimated the power spectral density (PSD) of the resultant RR time series and identified

a frequency band centred on the dominant peak (i.e. the heart rate). The remaining signals were similarly processed (peak detection followed by PSD estimation), and they were used only if there was a high correlation present in the selected frequency band.

3.2. Signal quality assessment

A strongly related area to that of signal phenotyping and selection is that of signal quality. Several entrants used signal quality indices (SQIs) to identify trustworthy segments of data. In particular, Johnson *et al* (2014) and (2015) used a suite of SQIs developed in earlier works to do this, achieving the highest score in the Challenge and the second highest score in this focus issue. Pimentel *et al* (2014) used an estimate of signal quality as a 'confidence' measure in the input of a hidden semi-Markov model, down weighting the impact of peaks detected on the ECG or ABP if the signal quality was low. Vollmer (2014) used the difference between a smoothed windowed maximum and a smoothed windowed minimum: if this difference was too low then the signal was considered bad quality, equivalently considered as a check on the amplitude of pulses on the waveform. Johannesen *et al* (2014) used physiologic constraints to filter waveforms: there should be at least 10 beats per 60 s of recording. Some entrants, including De Cooman *et al* (2014) and Vollmer (2014), used the regularity of the resultant RR series as a surrogate for signal quality. In normal sinus rhythm, subsequent RR intervals tend to be of similar duration as previous RR intervals. Consequently, a high standard deviation in the first difference of the RR series indicates abrubtly changing RR interval durations, and this was frequently used to determine quality of the underlying signal. It is worth noting, however, that many arrhythmias also cause highly irregular RR series, which will be discussed further later.

3.3. Preprocessing

Prior to the application of a peak detection algorithm, it was highly beneficial for competitors to perform some level of preprocessing. The aim of preprocessing was to increase the presence of the heart beat pulse while reducing the presence of noise, i.e. to improve the signal to noise ratio (SNR). It was very common for participants to low pass filter the data, as most cardiac information is contained below 40 Hz. Interestingly, Pimentel *et al* (2014) used a low pass filter with a 3 dB cut off of 16 Hz, which undoubtedly corrupted the morphology of the ECG waveform. However, as the only feature of interest is the location of the QRS peak, this corruption is irrelevant, and it has been previously shown for feotal ECG waveforms that quite liberal cut off frequencies provide better resolution of the peak locations (Behar *et al* 2014). Looking beyond frequency domain filtering, Johnson *et al* (2014) used a Mexican hat filter which better resolved peaks than more commonly applied rectangular window filters. Vollmer (2014) drifted from this approach and used a nonlinear trimmed average, followed by a smoothed maximum/minimum step, to create a square like waveform which better resolved QRS complexes.

3.4. Peak detection

Peak detection is a well studied field for both the ECG (Pahlm and Sörnmo 1984, Kohler *et al* 2002) and ABP (Li *et al* 2009, Li and Clifford 2012) signals. This is the core of any ECG processing algorithm, as correct determination of the location of heart beats is key for any subsequent analysis. Open source peak detectors have been available for the ECG for decades (Pan and Tompkins 1985), and similarly for the ABP (Zong *et al* 2003a). The typical approach

(and that of Pan and Tompkins 1985) is the sequential application of a difference filter (to amplify steep waveforms i.e. the QR and RS slopes), a squaring operation (to amplify peaks and act as a full wave rectifier), and finally a windowed average (to reduce noise). This method was used for *gqrs*, which was the sample entry in the Challenge. Pimentel *et al* (2014) did not directly detect peaks, but rather treated the heart beat as one of two states in a Markov model and treated the states as peak detections. Hoog Antink *et al* (2014) treated peak detection as a blind deconvolution problem, aiming to extract the peaks (assumed to be a Dirac delta train) from the measured signals by estimating the transfer functions between the Dirac delta function and each corresponding signal. Amplitude thresholding is applied to the extracted source signal to determine the final peak locations. Gierałtowski *et al* (2014) used a slope detector and achieved good performance. For the ABP signal, and pulsatile waveforms in general, one of the more common approaches uses the slope sum function (Zong *et al* 2003a, Li *et al* 2009). This involves calculating a cumulative sum across a window of the first difference of a signal, and thresholding on this new signal to estimate peak locations. For pulses with high initial slopes (e.g. ABP, PPG) this technique has been reasonably effective, and was used by many entrants including Johnson *et al* (2014) and Pimentel *et al* (2014). De Cooman *et al* (2014) treated the maximum in consecutive 300ms windows as peaks, and this simple algorithm was surprisingly effective, though undoubtedly sensitive to noise. Finally, Pangerc and Jager (2014) used a similar approach to Pan and Tompkins (1985), with an addition of morphological smoothing to improve robustness against noise.

3.5. Delay correction

As the signals in the Challenge were acquired from a variety of locations in the body, it was important to correct for the delay of these signals. In terms of the ABP and PPG, this delay is often called the pulse transit time (PTT). As the ECG is treated as the true time of the heart beat for annotation purposes, most algorithms focused on shifting peaks detected on other signals backward in order to match the ECG. Vollmer (2014) shifted detections on the ABP signal by 260 ms by default, or if possible by the median delay between peaks on the ECG and ABP signals (calculated over 20 s). Johnson *et al* (2014) shifted ABP peaks by 200 ms by default, or by the average delay between ECG and ABP detections over 60 s. Interestingly, Vollmer (2014) had a dynamic delay across the signal, updated every 20 s, while Johnson *et al* (2014) had a static shift for each 10 minute segment. Pimentel *et al* (2014) shifted the ABP signal by a fixed 40 ms, but estimated peaks jointly from the ABP and the ECG signals making exact record-wise alignment less of a necessity. Pangerc and Jager (2014) estimated the relationship between pulse rate and the PTT using a univariate regression, and utilized the PTT which best matched the current pulse rate for each beat. Gierałtowski *et al* (2014) used a default delay of 280 ms or averaged the delay for all ECG and ABP detections if they were available. Finally, Hoog Antink *et al* (2014) used a default delay of 200 ms or the maximum lag in a cross-correlation between the reference signal (usually the ECG) and the examined signal (usually the ABP).

3.6. Fusion

Fusion refers to the combination of peaks across various signals, all of which correspond to the same QRS complex. Fusing data across channels is a surprisingly non-trivial task, particularly when the source data come from different transducers (see Li *et al* 2008 and Nemati *et al* 2010). Identifying when one should ignore a signal segment, or weight together parameters assessed on it with those from other channels can be problematic, and in essence has to be learned from a large data set, and optimised for a given application. In Johnson *et al* (2014)

and (2015), the authors found that rather than weighting segments by quality, a higher accuracy was found by simply switching between segments with higher signal quality. Johannesen *et al* (2014) used a voting scheme, where a time series of beat detections, convolved with a tapered cosine, were averaged and peaks from this average waveform determined the final beat location. De Cooman *et al* (2014) had a similar voting system using rectangular windows and required agreement of $\lfloor \frac{D}{2} \rfloor + 1$ signals. Vollmer (2014) used an SQI to determine safe beats which were averaged to produce the final annotation set.

3.7. Search back

A final post processing step is sometimes applied which involves reviewing the current peak detections and deciding if any are false positives or if there are potential false negatives. One common procedure in peak detection algorithms is the process of searching backwards through the data with different thresholds when a beat is not detected. This can significantly improve performance when the amplitudes or noise levels in the data change frequently. A simple approach is to decrement or increment any threshold by a given percent every few seconds as in Clifford (2002). De Cooman *et al* (2014) used the ratio of subsequent RR intervals to determine if a beat had been missed, and guessed the location of missed beats using the last observed RR interval. Vollmer (2014) also used sudden increases or decreases in the RR interval to determine whether a beat had been missed. The popular open source algorithm *epltd* (Hamilton 2002) has a detailed search back procedure to ensure no beats are missed.

4. Review of articles in the focus issue

The top scores for entries graded on the revised hidden test data set is displayed in figure 4 and table 2. The C sample entry consisted of a single lead QRS detector only ('gqrs' function from the WFDB toolbox). The M-code sample entry consisted of a QRS detector and a BP detector from the WFDB toolbox for MATLAB/Octave (Silva and Moody 2014). As of March 2015, 12 teams submitted a total of 83 entries that were scored in the new environment.

Pangerc and Jager (2015) obtained the highest score reported in this focus issue, table 3 (improving on their sixth place from the Challenge). They used the MIT-BIH Arrhythmia database, the long-term ST database, the MIT-BIH polysomnographic database and the MGH database (Goldberger *et al* 2000) together with the Challenge training set in order to train their algorithm. They made use of the ECG and BP signals and performed peak detection using their custom ECG and BP pulse detectors, signal quality estimation to exclude bad ECG segments and pulse transit time estimation in order to map the ECG pulses to the corresponding BP pulses. Their QRS detector (*repdet*) provided a much improved performance over *gqrs* when evaluated on the Challenge training set. This is most likely due to the inclusion of a step in the detector to identify ECG records with paced beats (and the associated QRS detection correction)—there were 12 such records in the training set according to the authors. Successfully identifying these records significantly improved the authors performance over their official Challenge entry (Pangerc and Jager 2014).

Johnson *et al* (2015) achieved the highest score in the Challenge and second highest in this focus issue. Their algorithm made use of previously published signal quality indices for ECG and ABP in order to decide whether the physiological information extracted from these biosignals were reliable. The authors attempted to add biosignals other than ECG and ABP to their algorithm, but due to the limited number of operations allowed by the Sandbox (see figure 3) they were not able to test if it added any value on the test set.

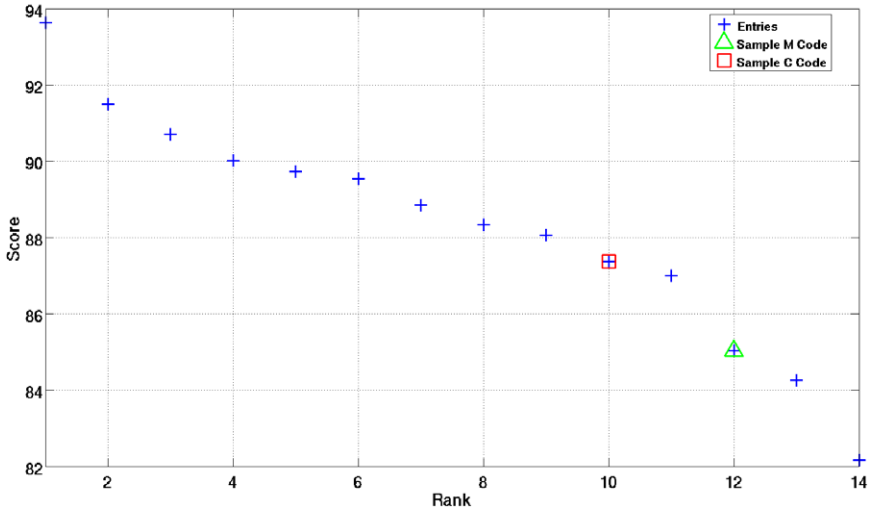

Figure 4. Top scores obtained on the revised data set for the 2014 Challenge. Both the C and M code sample entries are highlighted for comparison. A total of 83 entries from 12 teams were scored through the Sandbox environment on the revised data set as of March 2014 table 2.

Table 2. Results for entries submitted during Phase III of the challenge on a 300 record test set.

Challenge entry	Phase III score (%)
Johnson *et al* (2014)	87.93
Antink *et al* (2014)	87.07
De Cooman *et al* (2014)	86.61
Gieraltowski *et al* (2014)	86.40
Vollmer (2014)	86.22
Pangerc and Jager (2014)	85.13
C-code sample entry	84.49
Johannesen *et al* (2014)	84.42
Pimentel *et al* (2014)	83.47
M-code sample entry	79.28

Note: the sample entries were created by the Challenge organisers and are described in Moody *et al* (2014).

Hoog Antink *et al* (2015) suggested a technique that fuses the peaks detected on the ECG and ABP signals, based on the estimate of the RR intervals. These estimations were performed using a multimodal similarity approach. Three similarity measures were extracted from each of the available ECG and ABP signals, and the final RR estimate was extracted using a Bayesian approach.

De Cooman *et al* (2015) proposed two approaches, where one did not use the signal label. It is indeed possible that some recordings in existing databases have mislabelled signals, and creating an automatic 'signal type labeling' technique might be useful. Their automatic signal labelling approach, unfortunately, performed worse than the approach which used the signal labels provided. The authors argue that this is likely due to the high accuracy of the signal labels in the datasets. They also suggested a majority voting approach for fusing the peaks

Table 3. Results for entries submitted for this focus issue on the revised hidden test set (201 records).

Focus issue entry	Score (%)
Pangerc and Jager (2015)	93.64
Johnson *et al* (2015)	91.50
Antink *et al* (2015)	90.70
DeCooman *et al* (2015)	90.02
Galeotti *et al* (2015)	89.73
*Vollmer M	89.55
Pimentel *et al* (2015)	89.13
Mollakazemi *et al* (2015)	88.85
*Krug J	88.34
Gieraltowski *et al* (2015)	88.07
C-code sample entry	87.38
M-code sample entry	85.04

Note: participants marked with an asterisks (*) do not have a manuscript in this issue but source code is available on PhysioNet. The sample entries were created by the Challenge organisers and are described in Moody *et al* (2014).

from multiple signals, which incorporated the fact that peak localisations on ECG signals are more precise than pulsatile ones. Finally, they also introduced a search back procedure, in case irregular rhythms were detected.

Galeotti *et al* (2015) proposed an algorithm which fused the beats detected on all the available pulsatile signals. However they noted that their Challenge score was not changed over an approach which used the ECG and BP signals only, thus showing that the inclusion of the additional pulsatile signals did not improve the estimation of heart beats on the Challenge test set. The authors also used the MIT-BIH Polysomnographic database for training their algorithm.

Pimentel *et al* (2015) proposed an interesting approach that differs dramatically from the other entries. Whereas other entries detected the peaks on the different signals and fused the localisations of the peaks based on different heuristics, Pimentel *et al* (2015) preprocessed the ECG and ABP signals, and used a machine-learning approach with these pre-processed signals as inputs, to output the final peak locations. They proposed the use of a semi-hidden Markov model, which offers the advantage of incorporating a prior knowledge of the durations in each state (part of the cardiac cycle) of the model.

Mollakazemi *et al* (2015) fused the peaks detected from the ECG and ABP signals. Fusion was performed based on two criteria: (1) number of candidate detection in a defined time window and (2) the regularity of the derived RR time series. The authors did not use any other pulsatile signals than the ECG and ABP.

Gierałtowski *et al* (2015) have proposed an approach where they fuse the peaks detected on multiple signals: ECG, ABP, but also EOG, EMG and EEG. They suggested the use of an in-house QRS detector based on the RS slope, but also used *gqrs*. Their overall approach was as described in the previous subsection.

5. Summary and future directions

A total of 340 Challenge entries were scored the main challenge and 104 for this focus issue, totalling 444 entries from 47 teams. Due to the limited amount of time available on our servers, and to reflect the relatively constrained processing power in wearables and bedside monitors, we chose an upper limit of running at almost 500 times real time (on our servers). This

enforced a trade off between time taken and complexity of the algorithm. Challenge entries favoured simpler, faster algorithms (to fit within the challenge time constraints) versus more complicated potentially more accurate algorithms. Part of this issue is the use of interpretive languages (e.g. MATLAB) over low level languages (e.g. C). Some algorithms, which are too slow in MATLAB, may be perfectly reasonable in C.

One key issue to note is the quality and variety of the underlying data used in the Challenge. In particular, the Challenge data set is not perfectly labelled. A few records in the training set contained incorrect beat annotations. These records were identified by agreement of three independent annotators as being 1033, 1354, 42 511, 2277 with another possible three records: 1195, 1242, 1858 although these were more contentious because the beats were paced and it was difficult to decide whether or not the reference annotations were accurate. An example of erroneous reference annotations for record 2277, which had bigeminy, was due to the fact that the 'normal' beats were not annotated (only the ectopic beats). It is advisable that future work on the Challenge database does not include these records in the training set until this issue is fixed.

Moreover, any heart beat detection algorithm should be assessed in the context of the application for which it is intended. These can range from simple heart rate estimation (to identify bradycardia or tachycardia), to subtle estimators of ECG morphology changes (such as heart rate variability studies or late potentials). It is therefore important to consider the composition of the data and the exact metric used to assess accuracy.

The framework presented in the present Challenge could be improved by using an F_1 statistic such as in Behar *et al* (2014) in order to score the performance of the algorithms. Indeed, the F_1 measure is an harmonic mean and it is suited to situation when the average of rates (here *Se* and *PPV*) is desired (Sasaki 2007). In addition, given that the length of some records were shorter than 10 min (with a few as short as a few seconds) it is not advisable to compute gross statistics (for obvious reasons). The statistics should ideally be reported by beat types or condition types if medical annotations are available. This is because the behaviour of some algorithms will likely be different from one rhythm to another (see, for example, Behar *et al* (2013a) where the SQI performance was rhythm dependent). Approximately 95% of the data used in this Challenge were identified to be normal (either by expert labels or algorithms). Although this is probably representative of any clinical recording scenario, evaluating on this data without weighting can lead to statistics which are strongly biased towards normal data. Since it is most important to identify beats during abnormality, it could be argued that the data set should be enriched with more pathological scenarios.

Most algorithms required some thresholds or parameters to be set using the training set data. This is usually performed by trying a couple of sensible values and evaluating how the algorithm performance changes on the training set data, or fixing most parameters and performing an exhaustive search of one or two parameters over a limit range. A better way to identify 'optimal' values for these parameters and their relative importance is using random search as in Behar *et al* (2013b) (code and example available on Physionet[6]).

The purpose of the Challenge was to design algorithms that could locate heart beats in long-term multi-channel recordings. This is particularly interesting in contexts such as: (1) ICU where multiple biosignals are systematically recorded and where the number of false alarm could dramatically be reduced; (2) ambulation—with the increased number of wearable technology where multiple ECG channels can be recorded along other pulsatile signals such as the PPG. The publications in this focus issue have shown an improvement in the range of

[6] www.physionet.org/physiotools/random-search/.

3–4% above the Challenge scores when using multi-channel recordings versus only one channel. This very much highlights the important impact that multi-channel approaches can have in providing a better estimate of the heart rate and the potential application domains such as in false alarm reduction, which is the subject of the 2015 Challenge (Clifford *et al* 2015).

Finally, we note that, despite the limitations of the algorithms and the competition discussed above, the data set created for this Challenge can form the basis of a general testing set. We hope that, as the data set continues to be used in studies, more annotations are contributed by the community to the data set to enable the community to continue to address these limitations.

Acknowledgments

This work was funded in part by NIH/NIGMS grant R01 GM104987.

References

Behar J, Johnson A E, Oster J and Clifford G 2013b An echo state neural network for foetal ECG extraction optimised by random search *Proc. of the Machine Learning for Clinical Data Analysis and Healthcare NIPS Workshop* (*Lake Tahoe, USA*)

Behar J, Oster J and Clifford G D 2014 Combining and benchmarking methods of foetal ECG extraction without maternal or scalp electrode data *Physiol. Meas.* **35** 1569

Behar J, Oster J, Li Q and Clifford G D 2013a ECG signal quality during arrhythmia and its application to false alarm reduction *IEEE Trans. Biomed. Eng.* **60** 1660–6

Chang K M *et al* 2009 Pulse rate derivation and its correlation with heart rate *J. Med. Biol. Eng.* **29** 132–7

Chen L, Reisner A T and Reifman J 2009 Automated beat onset and peak detection algorithm for field-collected photoplethysmograms *Proc. of the Ann. Int. Conf. of the IEEE Engineering in Medicine and Biology Society* pp 5689–92

Clifford G D, Silva I, Behar J and Moody G B 2014 Non-invasive fetal ECG analysis *Physiol. Meas.* **35** 1521

Clifford G D 2002 Signal processing methods for heart rate variability *PhD Thesis* Department of Engineering Science, University of Oxford

Clifford G, Silva I, Moody B, Li Q, Kella D, Shahin A, Kooistra T, Perry D and Mark R 2015 The PhysioNet/computing in cardiology challenge 2015: reducing false arrhythmia alarms in the ICU (http://physionet.org/challenge/2015/)

De Cooman T, Goovaerts G, Varon C, Widjaja D, Willemen T and Huffel S V 2015 Heart beat detection in multimodal data using automatic signal type recognition *Physiol. Meas.* **36** 1691

De Cooman T, Goovaerts G, Varon C, Widjaja D and Van Huffel S 2014 Heart beat detection in multimodal data using signal recognition and beat location estimation *Comput. Cardiol.* **41** 257–60 (www.cinc.org/archives/2014/pdf/0257.pdf)

Eaton J W, Bateman D and Hauberg S 2009 GNU Octave version 301 manual: a high-level interactive language for numerical computations (CreateSpace Independent Publishing Platform)

Feldman J M, Ebrahim M H and Bar-Kana I 1997 Robust sensor fusion improves heart rate estimation: clinical evaluation *J. Clin. Monit.* **13** 379–84

Galeotti L, Scully C G, Vicente J, Johannesen L and Strauss D G 2015 Robust algorithm to locate heart beats from multiple physiological waveforms by individual signal detector voting *Physiol. Meas.* **36** 1705

Gierałtowski J, Ciuchciński K, Grzegorczyk I, Kośna K, Soliński M and Podziemski P 2015 RS slope detection algorithm for extraction of heart rate from noisy, multimodal recordings *Physiol. Meas.* **36** 1743

Gierałtowski J, Ciuchcinski K, Grzegorczyk I, Kosna K, Solinski M and Podziemski P 2014 Algorithm for detection of heart rate from noisy multimodal recordings *Comput. Cardiol.* **41** 253–6 (www.cinc.org/archives/2014/pdf/0253.pdf)

Goldberger A L, Amaral L 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, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals *Circulation* **101** e215–20

Gritzali F, Frangakis G and Papakonstantinou G 1989 Detection of the P and T waves in an ECG *Comput. Biomed. Res.* **22** 83–91
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
Hoog Antink C, Brüser C and Leonhardt S 2014 Multimodal sensor fusion of cardiac signals via blind deconvolution: a source-filter approach *Comput. Cardiol.* **41** 805 (www.cinc.org/archives/2014/pdf/0805.pdf)
Hoog Antink C, Brüser C and Leonhardt S 2015 Detection of heart beats in multimodal data: a robust beat-to-beat interval estimation approach *Physiol. Meas.* **36** 1679
Johannesen L, Vicente J, Scully C G, Galeotti L and Strauss D G 2014 Robust algorithm to locate heart beats from multiple physiological waveforms *Comput. Cardiol.* 277–80 (www.cinc.org/archives/2014/pdf/0277.pdf)
Johnson A E, Behar J, Andreotti F, Clifford G D and Oster J 2014 R-peak estimation using multimodal lead switching *Comput. Cardiol.* **41** 281–4 (www.cinc.org/archives/2014/pdf/0281.pdf)
Johnson A E, Behar J, Andreotti F, Clifford G D and Oster J 2015 Multimodal heart beat detection using signal quality indices *Physiol. Meas.* **36** 1665
Kohler B U, Hennig C and Orglmeister R 2002 The principles of software QRS detection *IEEE Eng. Med. Biol. Mag.* **21** 42–57
Liu S H, Chang K M and Fu T H 2010 Heart rate extraction from photoplethysmogram on fuzzy logic discriminator *Eng. Appl. Artif. Intell.* **23** 968–77
Li Q and Clifford G 2012 Dynamic time warping and machine learning for signal quality assessment of pulsatile signals *Physiol. Meas.* **33** 1491
Li Q *et al* 2009 Artificial arterial blood pressure artifact models and an evaluation of a robust blood pressure and heart rate estimator *Biomed. Eng. Online* **8** 13
Li Q, Mark R G and Clifford G D 2008 Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter *Physiol. Meas.* **29** 15
Mendelson Y 1992 Pulse oximetry: theory and applications for noninvasive monitoring *Clin. Chem.* **38** 1601–7 (PMID: 1525987)
Mollakazemi J M, Atyabi S A and Ghaffari A 2015 Heart beat detection using multimodal data coupling method *Physiol. Meas.* **36** 1729
Moody G B and Mark R G 1982 Development and evaluation of a 2-lead ECG analysis program *Comput. Cardiol.* **9** 39–44
Moody G B and Mark R G 1996 A database to support development and evaluation of intelligent intensive care monitoring *Proc. of the IEEE Conf. on Computers in Cardiology* pp 657–60
Moody G B, Moody B and Silva I 2014 Robust detection of heart beats in multimodal data: the PhysioNet/Computing in Cardiology challenge *Comput. Cardiol.* **41** 549–52 (www.cinc.org/archives/2014/pdf/0549.pdf)
Nemati S, Malhotra A and Clifford G D 2010 Data fusion for improved respiration rate estimation *EURASIP J. Adv. Signal Process.* **2010** 10
Okada M 1979 A digital filter for the QRS complex detection *IEEE Trans. Biomed. Eng.* **26** 700–3
Pahlm O and Sörnmo L 1984 Software QRS detection in ambulatory monitoring: a review *Med. Biol. Eng. Comput.* **22** 289–97
Pangerc U and Jager F 2014 Robust detection of heart beats in multimodal data using integer multiplier digital filters and morphological algorithms *Comput. Cardiol.* **41** pp 285–8 (www.cinc.org/archives/2014/pdf0285.pdf)
Pangerc U and Jager F 2015 Robust detection of heart beats in multimodal records using slope- and peak-sensitive band-pass filters *Physiol. Meas.* **36** 1645
Pan J and Tompkins W J 1985 A real-time QRS detection algorithm *IEEE Trans. Biomed. Eng.* **32** 230–6
Pimentel M A, Santos M D, Springer D B and Clifford G D 2014 Hidden semi-markov model-based heart beat detection using multimodal data and signal quality indices *Comput. Cardiol.* **41** 553–6 (www.cinc.org/archives/2014/pdf/0553.pdf)
Pimentel M A, Santos M D, Springer D B and Clifford G D 2015 Heart beat detection in multimodal physiological data using a hidden semi-Markov model and signal quality indices *Physiol. Meas.* **36** 1717
Portet F, Hernández A I and Carrault G 2005 Evaluation of real-time QRS detection algorithms in variable contexts *Med. Biol. Eng. Comput.* **43** 379–85
Saeed M, Villarroel M, Reisner A T, Clifford G, Lehman L W, Moody G, Heldt T, Kyaw T H, Moody B and Mark R G 2011 Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database *Crit. Care Med.* **39** 952

Sasaki Y 2007 The truth of the F-measure *Teach. Tutor. Mater.* 1–5
Silva I, Behar J, Sameni R, Zhu T, Oster J, Clifford G D and Moody G B 2013 Noninvasive fetal ECG: the PhysioNet/computing in cardiology challenge 2013 (www.physionet.org/challenge/2013/)
Silva I and Moody G 2014 An open-source toolbox for analysing and processing PhysioNet databases in MATLAB and Octave *J. Open Res. Softw.* **2** e27
Starmer C F, McHale P A and Greenfield J C 1973 Processing of arterial pressure waves with a digital computer *Comput. Biomed. Res.* **6** 90–6
Tarassenko L and Townsend N 2005 System and method for acquiring data *US Patent* 6839659 (www.google.bj/patents/US6839659)
Terzano M G *et al* 2001 Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep *Sleep Med.* **2** 537–53
Vollmer M 2014 Robust detection of heart beats using dynamic thresholds and moving windows *Comput. Cardiol.* **41** 569–72 (www.cinc.org/archives/2014/pdf/0569.pdf)
Welch J, Ford P, Teplick R and Rubsamen R 1991 The Massachusetts General Hospital-Marquette Foundation hemodynamic and electrocardiographic database—comprehensive collection of critical care waveforms *Clin. Monit.* **7** 96–7
Yu C, Liu Z, McKenna T, Reisner A T and Reifman J 2006 A method for automatic identification of reliable heart rates calculated from ECG and PPG waveforms *J. Am. Med. Inform. Assoc.* **13** 309–20
Zong W, Heldt T, Moody G and Mark R 2003a An open-source algorithm to detect onset of arterial blood pressure pulses *Proc.Computers in Cardiology (21–24 September 2003)* pp 259–62
Zong W, Moody G and Jiang D 2003b A robust open-source algorithm to detect onset and duration of QRS complexes *IEEE Computers in Cardiology* pp 737–40