Published in final edited form as:
Physiol Meas. 2015 August ; 36(8): 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]

Ikaro Silva: ikaro@mit.edu

[1]Institute for Medical Engineering & Science, Massachusetts Institute of Technology, USA

[2]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK

[3]Department of Biomedical Engineering, Technion, Israel Institute of Technology, Haifa, Israel

[4]Department of Biomedical Informatics, Emory University, USA

[5]Department of Biomedical Engineering, Georgia Institute of Technology, USA

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) 2014 Challenge*, published in the concurrent special 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 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%.

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 (ECGs), blood pressure (BP), electroencephalogram (EEG), respiration (RESP), photoplethysmogram (PPG), and others (Saeed et al. (2011); Moody and Mark (1996); Welch et al. (1991); Terzano et al. (2001)).Because some of these signals carry information pertaining to the cardiovascular system (Fig. 1 and 2), the PhysioNet/Computing in Cardiology 2014 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, Pahlm and Sörnmo (1984); Hamilton and Tompkins (1986); Portet et al. (2005); Pan and Tompkins (1985); Kohler et al. (2002); Zong, Heldt, Moody and Mark (2003); Zong, Moody and Jiang (2003); Starmer et al. (1973); Okada (1979); Mendelson (1992); Chen et al. (2009); Chang et al. (2009); Liu et al. (2010); Li and Clifford (2012); Moody and Mark (1982)). 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) et al. 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 seven second 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. (2013)).

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) et al. used a Kalman filtering frameworks to robustly calculate heart rate from the ECG and the PPG pulsatile waveforms recorded from 85 four hour (or less) ‘monitoring periods’ (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) et al. 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. Li et al. (2008) solved this issue by including in non-linearly weighted signal quality indices. The authors used a database of 6000 hours 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 lack of reproducibility or consistency is 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 special 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 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 special issue are available in PhysioNet (http://physionet.org/challenge/2014/sources/).

2. Overview of the Challenge 2014

2.1. Data Description

The Challenge data set used for this special 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 minutes, but several records were shorter than 10 minutes. 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 (CO2), central venous pressure (CVP), electrocardiogram (ECG), electroencephalogram (EEG), electromyography (EMG), electrooculogram (EOG), pulmonary arterial pressure (PAP), general pressure (pressure), nasal or abdominal respiration (RESP), oxygen level (SO2), 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, 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:

$$Se_{gross}=100\cdot\frac{TP}{TP+FN} \quad (1)$$

$$PPV_{gross}=100\cdot\frac{TP}{TP+FP} \quad (2)$$

$$Se_{average}=\frac{100}{n}\cdot\sum_{i=1}^{n}\frac{TP_i}{TP_i+FN_i} \quad (3)$$

$$PPV_{average}=\frac{100}{n}\cdot\sum_{i=1}^{n}\frac{TP_i}{TP_i+FP_i} \quad (4)$$

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 average of these four statistics, equations (1), (3), (2), and (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 (Fig. 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 minutes (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 seconds 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 the 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 special 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 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 banned 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, 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 special 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 seconds 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 in the article.

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 3dB 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 non-linear 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 Kohler et al. (2002); Pahlm and Sörnmo (1984) 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, Heldt, Moody and Mark (2003). 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. 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. Gieraltowski 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, Heldt, Moody and Mark (2003); 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 Pimentel et al. (2014); Johnson 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 260ms by default, or if possible by the median delay between peaks on the ECG and ABP signals (calculated over 20 seconds). Johnson et al. (2014) shifted ABP peaks by 200ms by default, or by the average delay between ECG and ABP detections over 60 seconds. Interestingly, Vollmer (2014) had a dynamic delay across the signal, updated every 20 seconds, 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 40ms, 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. Gieraltowski et al. (2014) used a default delay of 280ms or averaged the delay for all ECG and ABP detections if they were available. Finally, Antink et al. (2014) used a default delay of 200ms 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, 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 $\left\lfloor \frac{D}{2} \right\rfloor +1$ 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 Special Issue

The top scores for entries graded on the revised hidden test data set is displayed in Fig. 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 special issue (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 ECGs 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 special 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.

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.

DeCooman 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 other approach did unfortunately performed better that the one with the automatic signal labelling procedure. They also suggested a majority voting approach for fusing the peaks 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 heartbeats 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) et al. 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) the authors 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.

Gieraltowski 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 followed was is 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 special 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, 42511, 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 2014 database do 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 a rhythm to another (see for example (Behar et al., 2013) 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 are changing 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[‡]).

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

[‡] http://www.physionet.org/physiotools/random-search/

technology where multiple ECG channels can be recorded along other pulsatile signals such as the PPG. The publications in this special issue have shown an improvement in the range of 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

Antink, CH.; Bruser, C.; Leonhardt, S. Computing in Cardiology Conference (CinC), 2014. IEEE; 2014. Multimodal sensor fusion of cardiac signals via blind deconvolution: A source-filter approach; p. 805-808.

Antink CH, Brüser C, Leonhardt S. Detection of heart beats in multimodal data: A robust beat-to-beat interval estimation approach. Physiological Measurement. 2015; x(x):x.

Behar, J.; Johnson, AE.; Oster, J.; Clifford, G. An echo state neural network for foetal ECG extraction optimised by random search. Machine Learning for Clinical Data Analysis and Healthcare NIPS Workshop; Lake Tahoe, USA. 2013b.

Behar J, Oster J, Clifford GD. Combining and benchmarking methods of foetal ECG extraction without maternal or scalp electrode data. Physiological Measurement. 2014; 35(8):1569. [PubMed: 25069410]

Behar J, Oster J, Li Q, Clifford GD. ECG signal quality during arrhythmia and its application to false alarm reduction. Biomedical Engineering, IEEE Transactions on. 2013; 60(6):1660–1666.

Chang KM, Chang KM, et al. Pulse rate derivation and its correlation with heart rate. Journal of Medicine in Biology and Engineering. 2009; 29(3):132–7.

Chen, L.; Reisner, AT.; Reifman, J. Engineering in Medicine and Biology Society, 2009 EMBC 2009 Annual International Conference of the IEEE. IEEE; 2009. Automated beat onset and peak detection algorithm for field-collected photoplethysmograms; p. 5689-5692.

Clifford, GD. PhD thesis. Dept of Engineering Science, University of Oxford; 2002. Signal processing methods for heart rate variability.

Clifford GD, Silva I, Behar J, Moody GB. Non-invasive fetal ECG analysis. Physiological Measurement. 2014; 35(8):1521. [PubMed: 25071093]

Clifford, G.; Silva, I.; Moody, B.; Li, Q.; Kella, D.; Shahin, A.; Kooistra, T.; Perry, D.; Mark, R. Computing in Cardiology. IEEE; 2015. The PhysioNet/Computing in Cardiology Challenge 2015: Reducing False Arrhythmia Alarms in the ICU. **URL**: http://physionet.org/challenge/2015/

De Cooman T, Goovaerts G, Varon C, Widjaja D, Van Huffel S. Heart beat detection in multimodal data using signal recognition and beat location estimation. Computing in Cardiology Conference (CinC), 2014. 2014; 41:257–260.

DeCooman T, Goovaerts G, Varon C, Widjaja D, Willemen T, Huffel SV. Heart beat detection in multimodal data using automatic signal type recognition. Physiological Measurement. 2015; x(x):x.

Eaton, JW.; Bateman, D.; Hauberg, S. GNU Octave version 3.0.1 manual: a high-level interactive language for numerical computations. CreateSpace Independent Publishing Platform; 2009.

Feldman JM, Ebrahim MH, Bar-Kana I. Robust sensor fusion improves heart rate estimation: clinical evaluation. Journal of Clinical Monitoring. 1997; 13(6):379–384. [PubMed: 9495290]

Galeotti L, Scully CG, Vicente J, Johannesen L, Strauss DG. Robust algorithm to locate heartbeats from multiple physiological waveforms by individual signal detector voting. Physiological Measurement. 2015; x(x):x.

Gieraltowski J, Ciuchciński K, Grzegorczyk I, Kośna K, Soliński M, Podziemski P. RS slope detection algorithm for extraction of heart rate from noisy, multimodal recordings. Physiological Measurement. 2015; x(x):x.

Gieraltowski JJ, Ciuchcinski K, Grzegorczyk I, Kosna K, Solinski M, Podziemski P. Heart rate variability discovery: Algorithm for detection of heart rate from noisy, multimodal recordings. Computing in Cardiology Conference (CinC), 2014. 2014; 41:253–256.

Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation. 2000; 101(23):e215–e220. [PubMed: 10851218]

Gritzali F, Frangakis G, Papakonstantinou G. Detection of the P and T waves in an ECG. Computers and Biomedical Research. 1989; 22(1):83–91. [PubMed: 2914427]

Hamilton, P. Computers in Cardiology, 2002. IEEE; 2002. Open source ECG analysis; p. 101-104.

Hamilton PS, Tompkins WJ. Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. Biomedical Engineering, IEEE Transactions on. 1986; (12):1157–1165.

Johannesen L, Vicente J, Scully CG, Galeotti L, Strauss DG. Robust algorithm to locate heart beats from multiple physiological waveforms. Computing in Cardiology Conference (CinC), 2014. 2014; 41:277–280.

Johnson AE, Behar J, Andreotti F, Clifford GD, Oster J. R-peak estimation using multimodal lead switching. Computing in Cardiology Conference (CinC), 2014. 2014; 41:281–284.

Johnson AE, Behar J, Andreotti F, Clifford GD, Oster J. Multimodal heart beat detection using signal quality indices. Physiological Measurement. 2015; x(x):x.

Kohler BU, Hennig C, Orglmeister R. The principles of software QRS detection. Engineering in Medicine and Biology Magazine, IEEE. 2002; 21(1):42–57.

Li Q, Clifford G. Dynamic time warping and machine learning for signal quality assessment of pulsatile signals. Physiological Measurement. 2012; 33(9):1491. [PubMed: 22902950]

Li Q, Mark RG, Clifford GD. Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter. Physiological Measurement. 2008; 29(1):15. [PubMed: 18175857]

Li Q, Mark RG, Clifford GD, et al. Artificial arterial blood pressure artifact models and an evaluation of a robust blood pressure and heart rate estimator. Biomedical Engineering Online. 2009; 8(1):13. [PubMed: 19586547]

Liu SH, Chang KM, Fu TH. Heart rate extraction from photoplethysmogram on fuzzy logic discriminator. Engineering Applications of Artificial Intelligence. 2010; 23(6):968–977.

Mendelson Y. Pulse oximetry: theory and applications for noninvasive monitoring. Clinical Chemistry. 1992; 38(9):1601–1607. [PubMed: 1525987]

Mollakazemi JM, Atyabi SA, Ghaffari A. Heart beat detection using multimodal data coupling method. Physiological Measurement. 2015; x(x):x.

Moody GB, Mark RG. Development and evaluation of a 2-lead ECG analysis program. Computers in Cardiology. 1982; 9:39–44.

Moody, GB.; Mark, RG. Computers in Cardiology, 1996. IEEE; 1996. A database to support development and evaluation of intelligent intensive care monitoring; p. 657-660.

Moody GB, Moody B, Silva I. Robust Detection of Heart Beats in Multimodal Data: the PhysioNet/Computing in Cardiology Challenge 2014. Computing in Cardiology Conference (CinC), 2014. 2014; 41

Nemati S, Malhotra A, Clifford GD. Data fusion for improved respiration rate estimation. EURASIP Journal on Advances in Signal Processing. 2010; 2010:10.

Okada M. A digital filter for the QRS complex detection. Biomedical Engineering, IEEE Transactions on. 1979; (12):700–703.

Pahlm O, Sörnmo L. Software QRS detection in ambulatory monitoring-a review. Medical and Biological Engineering and Computing. 1984; 22(4):289–297. [PubMed: 6379330]

Pan J, Tompkins WJ. A real-time QRS detection algorithm. Biomedical Engineering, IEEE Transactions on. 1985; (3):230–236.

Pangerc U, Jager F. Robust detection of heart beats in multimodal data using integer multiplier digital filters and morphological algorithms. Computing in Cardiology Conference (CinC), 2014. 2014; 41:285–288.

Pangerc U, Jager F. Robust detection of heart beats in multimodal records using slope- and peak-sensitive band-pass filters. Physiological Measurement. 2015; x(x):x.

Pimentel MA, Santos MD, Springer DB, Clifford GD. Hidden semi-markov model-based heartbeat detection using multimodal data and signal quality indices. Computing in Cardiology Conference (CinC), 2014. 2014; 41:553–556.

Pimentel MA, Santos MD, Springer DB, Clifford GD. Heartbeat detection in multimodal physiological data using a hidden semi-Markov model and signal quality indices. Physiological Measurement. 2015; x(x):x.

Portet F, Hernández AI, Carrault G. Evaluation of real-time QRS detection algorithms in variable contexts. Medical and Biological Engineering and Computing. 2005; 43(3):379–385. [PubMed: 16035227]

Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. Critical Care Medicine. 2011; 39(5):952. [PubMed: 21283005]

Sasaki Y. The truth of the F-measure. Teaching, Tutorial materials, Version: 26th October. 2007

Silva, I.; Behar, J.; Sameni, R.; Zhu, T.; Oster, J.; Clifford, GD.; Moody, GB. Computing in Cardiology. IEEE; 2013. Noninvasive Fetal ECG: the PhysioNet/Computing in Cardiology Challenge 2013.

Silva I, Moody G. An open-source toolbox for analysing and processing PhysioNet databases in MATLAB and Octave. Journal of Open Research Software. 2014; 2(1):e27. [PubMed: 26525081]

Starmer CF, McHale PA, Greenfield JC. Processing of arterial pressure waves with a digital computer. Computers and Biomedical Research. 1973; 6(1):90–96. [PubMed: 4695393]

Tarassenko, L.; Townsend, N. System and method for acquiring data. US Patent 6,839,659. 2005. URL: http://www.google.bj/patents/US6839659

Terzano MG, Parrino L, Sherieri A, Chervin R, Chokroverty S, Guilleminault C, Hirshkowitz M, Mahowald M, Moldofsky H, Rosa A, et al. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. Sleep Medicine. 2001; 2(6):537–553. [PubMed: 14592270]

Vollmer M. Robust detection of heart beats using dynamic thresholds and moving windows. Computing in Cardiology Conference (CinC), 2014. 2014; 41:569–572.

Welch J, Ford P, Teplick R, Rubsamen R. The Massachusetts General Hospital-Marquette Foundation hemodynamic and electrocardiographic database– comprehensive collection of critical care waveforms. Clinical Monitoring. 1991; 7(1):96–97.

Yu C, Liu Z, McKenna T, Reisner AT, Reifman J. A method for automatic identification of reliable heart rates calculated from ECG and PPG waveforms. Journal of the American Medical Informatics Association. 2006; 13(3):309–320. [PubMed: 16501184]

Zong, W.; Heldt, T.; Moody, G.; Mark, R. Computers in Cardiology, 2003. IEEE; 2003. An open-source algorithm to detect onset of arterial blood pressure pulses; p. 259-262.

Zong, W.; Moody, G.; Jiang, D. Computers in Cardiology, 2003. IEEE; 2003. A robust open-source algorithm to detect onset and duration of QRS complexes; p. 737-740.

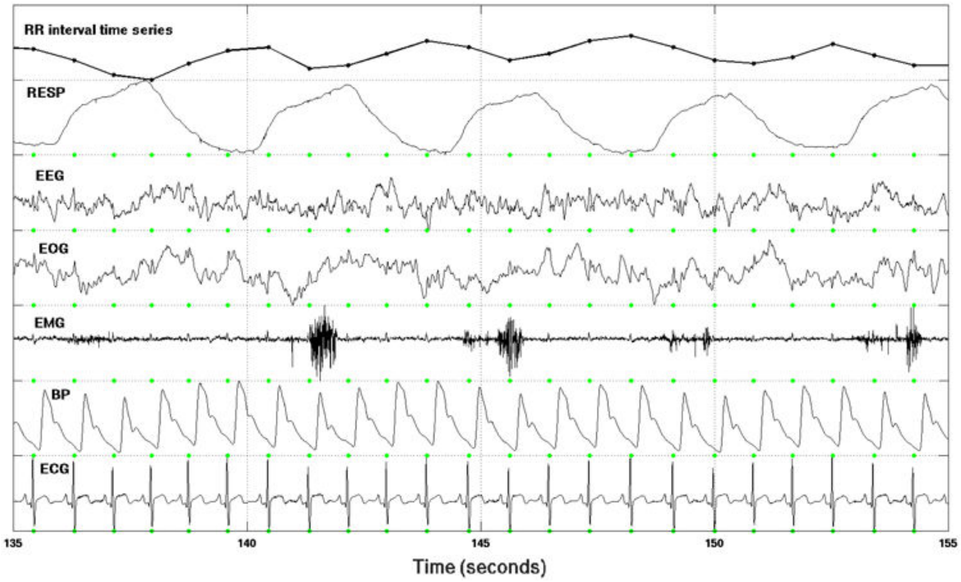

Figure 1.

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.

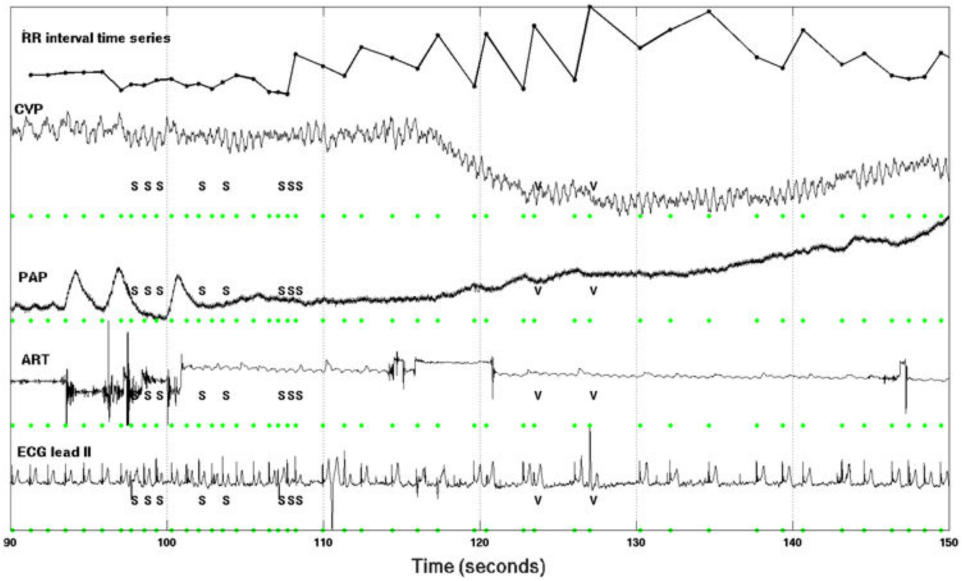

Figure 2.

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.

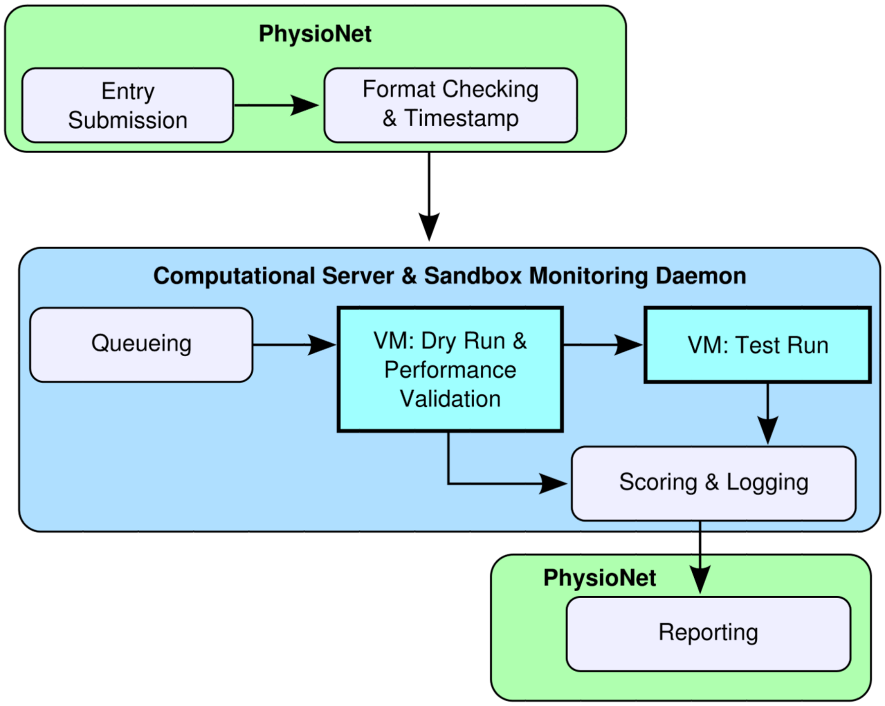
PhysioNet
Entry Submission
Format Checking & Timestamp
Computational Server & Sandbox Monitoring Daemon
Queueing
VM: Dry Run & Performance Validation
VM: Test Run
Scoring & Logging
PhysioNet
Reporting

Figure 3.

Diagram describing the process for automatic evaluation of Challenge entries.

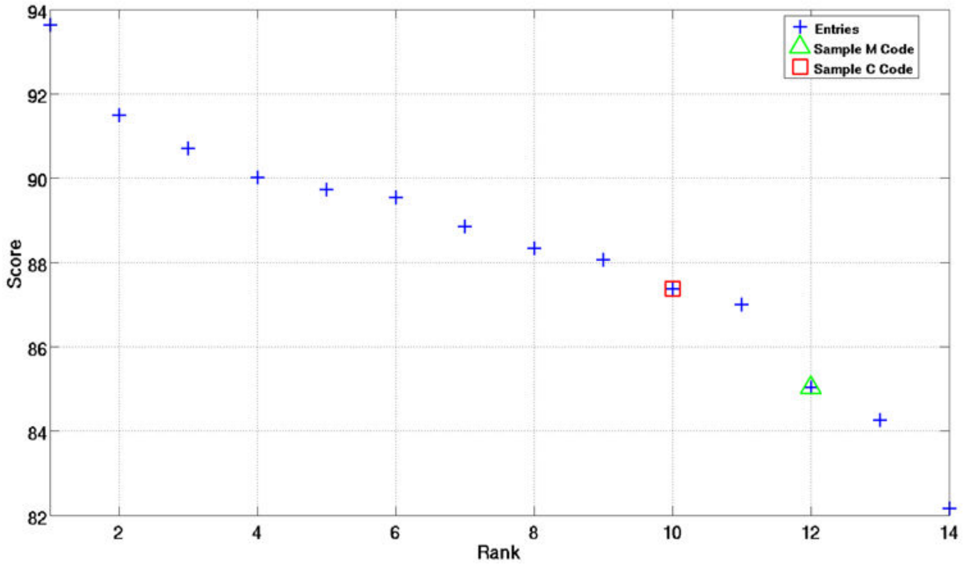

Figure 4.

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 (2).

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	CO2	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	SO2	1	23
Stroke Volume	SV	1	23

Table 2

Results for entries submitted during Phase III of the challenge on a 300 record test set. The sample entries were created by the Challenge organisers and are described in Moody et al. (2014).

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

Table 3

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

Special 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