Physiol. Meas. **33** (2012) doi:10.1088/0967-3334/33/9/E01

EDITORIAL

Signal quality in cardiorespiratory monitoring

This focus issue of *Physiological Measurement* follows the 38th Annual International Computing in Cardiology (CinC) Conference, hosted in Hangzhou, China in September 2011 by Zhejiang University. Each year, the NIH-sponsored PhysioNet resource (http://physionet.org/) runs an open competition lasting several months, aimed at encouraging the development of solutions to an unsolved or poorly solved problem in biomedicine, in most cases making use of relevant clinical and experimental data provided freely by PhysioNet. Participants in these annual challenges discuss their diverse approaches to the Challenge problems during dedicated scientific sessions at CinC. The topics of these PhysioNet/CinC Challenges range from physiologic signal processing and analysis to forecasting and modelling clinically important events and processes.

In 2011, the PhysioNet/CinC Challenge was to develop an efficient algorithm able to run within a mobile phone, that can provide useful feedback in the process of acquiring a diagnostically usable 12-lead electrocardiogram (ECG). At a minimum, such an algorithm should indicate if an ECG is of adequate quality for interpretation, completing its analysis within a few seconds, while the patient is still present, so that another recording can be made immediately if needed.

The ECG is among the most useful tools for diagnosing cardiovascular diseases (CVD), the most frequent cause of death worldwide. Although both CVD and mobile phones are ubiquitous, adequate primary health care is not. Many rural populations around the world rely on clinics staffed by lay volunteers to identify those in need of secondary care by health care professionals in distant city hospitals. It is increasingly feasible to provide rural clinics with inexpensive medical instruments such as electrocardiographs that transmit digital ECGs to smart phones for storage and display. These devices extend the reach of diagnosticians to remote areas, but without some means of quality control, technology alone cannot deliver consistently usable information to those able to interpret it. Methods that improve the quality of data collected result in better usage of the scarcest resource, clinical expertise. The growing interest in mHealth to provide point-of-care diagnostics to underserved populations is also driving the desire to leverage the power of smart phones to insert intelligence into medical data acquisition.

PhysioNet provided a Challenge data set of 2000 12-lead ECG records, together with an open-source sample application able to run on an Android phone. (These remain freely available to interested readers at http://physionet.org/challenge/2011/.) The application was provided as a working example of a Challenge entry that can read an ECG and classify it as acceptable or unacceptable. Gold standard annotations (grades) for the ECGs were crowd-sourced from the public and invited experts. The annotators were also asked to rate their own expertise or experience level. In all, 8,327 grades were obtained; 1,733 ECGs were classified as acceptable or unacceptable, and 267 as indeterminate. In nearly all of the latter group, only a single grade was available; divided opinions were very rare. There was a high degree of self-consistency, consistency with other observers at the same and at different experience levels, and consistency with the reference classifications regardless of experience level. A random selection of half of the Challenge data set was designated as Set A, a training subset,

and participants were provided with the grades for these 1000 ECGs. The remaining records were divided at random into Set B, a public test subset (500 ECGs available for study, with grades withheld) and Set C, a hidden test subset (500 ECGs not available for study, used only by PhysioNet for testing submitted algorithms).

Each participant entered one or more of three Challenge events. In event 1, participants developed algorithms for classifying ECGs with respect to quality, and submitted their algorithms' classifications of Set B. Each entry was scored for accuracy (defined as the fraction of Acceptable and Unacceptable ECGs that were correctly classified, with Indeterminate ECGs excluded). In the other two events, participants submitted Java implementations of their algorithms to be used in the sample mobile application; these were tested in two reference mobile phones (in event 2, using Set B, with scoring as in event 1; and in Event 3 using Set C, with scoring based on a function of both accuracy and mobile phone processing speed). A total of 49 teams and individuals participated in the Challenge. Accuracies generally varied between 80% and 93% with average execution times of less than 2 seconds on the reference phones. [Many participants reported the accuracy of their methods as measured using training Set A; if these results are significantly better than those obtained using test Sets B and (especially) C, this is indicative of overfitting the training data.] A full description of the Challenge can be found in Silva *et al* (2011).

Problems of noise and transient drops in data quality are not just confined to diagnostic ECGs however, and issues of noise plague physiologic measurements of many types in hospitals, causing high levels of false alarms, (Aboukhalil *et al* 2008). Although many of the papers in this focus issue (presented in alphabetical order by the surname of the first author) relate specifically to the PhysioNet/CinC Challenge 2011, several address the broader question of signal quality metrics in cardiorespiratory monitoring.

Chen and Yang (2012) approached the Challenge problem of screening the quality of 12-lead ECGs by applying the inverse Dower transform to obtain 3-lead VCGs that they classified using multiscale recurrence analysis with self-organising maps to identify time-frequency features associated with poor and good quality ECG. They report 95.25% accuracy on the training data (Set A) and 90.0% on the independent test data (Set B).

Clifford *et al* (2012) extended their original work in the Challenge, which used a series of signal quality metrics (based on morphological, statistical and spectral characteristics) and a support vector machine or multilayer perceptron neural network. The modifications included 1) labelling of and training on single leads, 2) upsampling the noisy data using the noise stress test database, 3) varying the window size and 4) testing their system on arrhythmic data. A classification accuracy of 98% on the training data (Set A) and 97% on the test data (Set B) was achieved. Reducing the window size led to a moderate drop in accuracy by $< 1\%$ per second removed, although this may be partially attributed to the transience of the noise. Tests on arrhythmic data led to a drop in accuracy to 93% indicating that algorithms may need retraining for some arrhythmias.

Di Marco *et al* (2012) addressed ECG quality assessment by identifying baseline drift, flat line, QRS-artifact, spurious spikes, amplitude step changes, and other noise, using a time-frequency approach. Classification was based on cascaded single-condition decision rules tested levels of contaminants against classification thresholds. A supervised learning approach was also taken to combining the thresholds. The authors found that their cascaded heuristic threshold algorithm performed best with an accuracy of 91.40% on the Challenge test data (Set B).

Hayn *et al* (2012) explored the use of four criteria; a no signal detector, a spike detector, a lead crossing point analysis and a measure of the robustness of QRS detection. An accuracy of 93.3% was achieved on the training data (Set A) and 91.6% on the test data (Set B). A

simplified version of their algorithm (omitting the robustness measure) was the winning entry in event 3 of the Challenge. Scores for this algorithm for events 2 and 3 (both run on a smart phone) were 0.834 (Set B) and 0.873 (Set C) respectively.

Jekova *et al* (2012) aimed to identify four major sources of ECG quality disruption: missing signal or reduced energy of the QRS complex above 4Hz; presence of high amplitude and steep artifacts above 1Hz; baseline drift at frequencies below 1Hz; power-line interference in a band $\pm$2Hz around its central frequency; and high-frequency and electromyographic noises above 20Hz. The authors introduced 13 adjustable thresholds for amplitude and slope criteria, and reported the sensitivity (Se) and specificity (Sp) of their methods for detecting unacceptable ECGs; by adjusting thresholds, they obtained results ranging from Se = 98.7%, Sp = 80.9% to Se = 81.8%, Sp = 97.8% on the Challenge training data (Set A).

Johannesen and Galeotti (2012) described a two stage approach to screening ECGs, first identifying missing signals, large voltage shifts, and saturation, then quantifying baseline wave, mains frequency, and muscle noise using average template matching. The authors achieved a classification accuracy of 92.3% on the Challenge training data (Set A) and 90.0% on the test data (Set B).

Li and Clifford (2012) extended the method of Clifford *et al* described above to obtain a signal quality metric for the photoplethysmogram (PPG). The features were based on cross correlation with a local beat template, stretched in both a linear and nonlinear manner then combined using a multi-layer perceptron neural network. An expert-labelled database of 1055 segments of PPG, each 6 seconds long, recorded from 104 separate critical care admissions during both normal and verified arrhythmic events, was used to train and test their algorithms. An accuracy of 97.5% on the training set and 95.2% on test set was reported.

Monasterio *et al* (2012) took a novel approach to identifying the quality of a data segment by combining both physiological and signal quality features in a machine learning framework, using multiple cardiovascular signals (ECG, PPG and respiratory waveforms). The aim of the research was to classify desaturations in neonates as true or false. A total of 1616 desaturation events from 27 neonatal admissions were annotated by two independent reviewers as true (physiologically relevant) or false (noise-related). The patients were divided into two independent groups for training and validation, and a series of signal quality and physiological metrics (such as gradient of heart rate and respiration rate) were estimated. A support vector machine was trained to use 13 of these features to classify the events as true or false. An accuracy of 100% was achieved during training, and a sensitivity of 86%, a specificity of 91%, and an accuracy of 90% was achieved in the test set.

Redmond *et al* (2012) used three annotators to manually annotate 300 short single-lead ECG recordings to identify movement artifact, QRS locations and signal quality, with overreading to reconcile differences in order to obtain a gold standard three-level quality index (good, average, or bad). Template-based and signal morphology-based features were then presented to a Parzen-window supervised statistical classifier model, which achieved a three-level classification accuracy of 78.7% when using fully automated preprocessing algorithms to remove gross motion artifact and detect QRS locations. The authors note that this accuracy is similar to the human inter-scorer agreement.

Xia *et al* (2012a) report on their approach to the Challenge problem, which won events 1 and 2. Twelve signal quality heuristics were developed and calculated for each of the 12 ECG leads, yielding a 12 by 12 matrix. The elements were then summed and thresholded to provide a classification for a given 12 lead ECG. After optimisation of the threshold, the authors achieved an accuracy of 95%, with a sensitivity of 88% and specificity of 97%.

In the final paper of this focus issue, three of the authors of the previous paper (Xia, Garcia, and Zhao) Xia *et al* (2012b) focused on detection of electrode misplacement using a series of

ECG features and a multilayer perceptron neural network. In the best case, with clean ECGs from training data, they were able to detect LA/LL misplacement with 87.4% accuracy, and all other misplacements with 98.4% accuracy. Noisy ECGs, and those containing arrhythmias, presented a more difficult challenge, and the authors note that accuracy may be poor on test data, suggesting the need for a more extensive data set for development and testing of electrode misplacement detection methods.

It is noteworthy that the use of shared data sets in many of the papers in this focus issue permits the reader to make objective comparisons of the performance of the methods described in them. These comparisons can point the way to further advances toward assessment of quality in ECGs and other physiologic signals. Seven Challenge participants, including the authors of several of the papers included here, have also contributed their algorithms as open-source software for further study; these can be found at http://physionet.org/challenge/2011/sources/.

As noted on PhysioNet, 'The annual PhysioNet/CinC Challenges seek to provide stimulating yet friendly competitions, while at the same time offering both specialists and non-specialists alike opportunities to make progress on significant open problems whose solutions may be of profound clinical value. The use of shared data provided via PhysioNet makes it possible for participants to work independently toward a common objective. At CinC, participants can make meaningful results-based comparisons of their methods; lively and well-informed discussions are the norm at scientific sessions dedicated to these challenges. Discovery of the complementary strengths of diverse approaches to a problem when coupled with deep understanding of that problem frequently sparks new collaborations and opportunities for further study.... It is especially significant that many of those who have participated in these challenges would not otherwise have had access to the data needed to study these topics.'

PhysioNet invites readers of *Physiological Measurement* to participate in its Challenge series, and to identify and develop topics for future Challenges.

Gari D Clifford, *Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK*
George B Moody, *Harvard/MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA, USA*

Guest Editors

References

Aboukhalil A, Nielsen L, Saeed M, Mark R G and Clifford G 2008 Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform *J. Biomed. Inform.* **41** 442–51

Chen Y and Yang H 2012 Self-organized neural network for the quality control of 12-lead ECG signals *Physiol. Meas.* **33** 1399–1418

Clifford G D, Behar J, Li Q and Rezek I 2012 Signal quality indices and data fusion for determining acceptability of electrocardiograms collected in noisy ambulatory environments *Physiol. Meas.* **33** 1419–33

Di Marco L Y, Duan W, Bojarnejad M, Zheng D, King, Murray A and Langley P 2012 Evaluation of an algorithm based on single-condition decision rules for binary classification of 12-lead ambulatory ECG recording quality *Physiol. Meas.* **33** 1435–48

Hayn D, Jammerbund B and Schreier G 2012 QRS detection based ECG quality assessment *Physiol. Meas.* **33** 1449–61

Jekova I, Krasteva V, Christov I and Abächerli R 2012 Threshold-based system for noise detection in multilead ECG recordings *Physiol. Meas.* **33** 1463–77

Johannesen J and Galeotti L 2012 Automatic ECG quality scoring methodology: mimicking human annotators *Physiol. Meas.* **33** 1479–89

Li Q and Clifford G D 2012 Dynamic time warping and machine learning for signal quality assessment of pulsatile signals *Physiol. Meas.* **33** 1491–1501

Monasterio V, Burgess F and Clifford G D 2012 Robust classification of neonatal apnoea-related desaturations *Physiol. Meas.* **33** 1503–16

Redmond S J, Xie Y, Chang D, Basilakis J and Lovell N H 2012 Electrocardiogram signal quality measures for unsupervised telehealth environments *Physiol. Meas.* **33** 1517–33

Silva I, Moody G B and Celi L 2011 Improving the quality of ECGs collected using mobile phones: the PhysioNet/Computing in Cardiology Challenge 2011 *Comput. Cardiol.* **38** 1273–76

Xia H, Garcia G, Bains J, Wortham D and Zhao X 2012a Matrix of regularity for improving the quality of ECGs *Physiol. Meas.* **33** 1535–48

Xia H, Garcia G and Zhao X 2012b Automatic detection of ECG electrode misplacement: a tale of two algorithms *Physiol. Meas.* **33** 1549–61