Improving the Quality of ECGs Collected Using Mobile Phones: The PhysioNet/Computing in Cardiology Challenge 2011

Ikaro Silva[1,2], George B Moody[1], Leo Celi[1,2]

[1]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, USA
[2]Sana Mobile, Cambridge, MA, USA

Abstract

The aim of the PhysioNet/Computing in Cardiology Challenge 2011 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 useful 12-lead ECG recordings. PhysioNet provided a large set of ECG records for use in the Challenge, together with an open-source sample application that can run on an Android phone, and can classify ECGs as acceptable or unacceptable. A total of 49 teams and individuals participated in challenge, which entailed three events. In event 1, participants developed algorithms for classifying ECGs with respect to quality, and submitted their algorithms' classifications of 500 ECGs, obtaining 89-93% accuracy using variety of methods. In event 2, participants submitted Java implementations of their algorithms to be used in the sample mobile application; we tested these in two reference mobile phones using the same data set and scoring method as in event 1, obtaining 80-91% accuracy. Event 3 was similar to event 2, but was conducted using a set of ECGs not available for study by the participants, and the scoring was a function of both accuracy and mobile phone processing speed; in this event, similar levels of accuracy were achieved with average execution times of less than 2 seconds on the reference phones.

1. Introduction

In addition to the burden of communicable diseases such as malaria, tuberculosis and HIV, developing countries are facing a steady growth in the prevalence of chronic, non-communicable diseases, including heart disease and cancer. Mobile health, or the use of cellphones to support clinical care, provides an opportunity to expand the reach of quality health-care to address both types of disease burdens even in the most remote villages. Cellphones are used more than any other modern technology throughout the developing world[1,2]. The ITU estimates that in 2010, there were 5.3 billion mobile subscribers (77% of the world's population), with 67.6 mobile phones per 100 persons in the developing world, and 116.1 in developed nations[3]. It is no surprise that mobile health is being touted as the biggest breakthrough in health systems improvement in developing nations[4]. The positive potential for mobile health is huge, but not without risks. If expanded and decentralized access to health care results in an increase in the demand for expert diagnosis, and the quality of the data needing interpretation is not maintained, a loss of efficiency will follow. If the capacity of the health care system to provide timely expert interpretation is exceeded, the result may be missed opportunities and a net decrease in the number of patients served, even as the patient population increases. Rigorous quality control is thus essential, not only for accurate diagnosis, but also to preserve, and if possible to enhance, the efficiency and capacity of the health care system to serve its patients.

Compact and inexpensive battery-powered ECG recorders can transmit diagnostic ECGs via Bluetooth to mobile phones, which can relay them to experts for interpretation. In developing nations, where the experts are concentrated in urban hospitals, this technology can permit underserved rural populations to benefit from otherwise inaccessible expertise. If this possibility is to become reality, however, it will be necessary for health care providers in underserved regions to become proficient in collecting high-quality ECGs, to avoid the risk of saturating the experts' capacity. Furthermore, since expert interpretation may not be immediately available, it is important to obtain a recording that can be interpreted without waiting for an expert opinion on its quality, since it may be difficult to obtain another ECG on another day from a patient who may live far from a clinic.

For these reasons, the aim of the PhysioNet/Computing in Cardiology Challenge 2011 is to encourage the development of software that can run in a mobile phone, recording an ECG and providing useful feedback about its quality. Ideally, the software should be able to indicate within a few seconds, while the patient is still present, if the ECG is of adequate quality for interpretation, or if another record-

ISSN 0276–6574

ing should be made. The software should identify common problems (such poor skin-electrode contact, external electrical interference, and artifact resulting from patient motion) and either compensate for these deficiencies or provide guidance for correcting them. Within the context of this challenge, however, submissions were scored only with respect to how well their quality assessments of specific test ECGs predicted human assessments of quality, and (in one event) the time required for the algorithm to make a classification.

2. Methods

2.1. Challenge data set

The data used for the PhysioNet/CINC 2011 Challenge consisted of 2,000 twelve-lead ECGs, each 10 seconds long, with standard diagnostic bandwidth (0.05-100 Hz). The 12 leads (I, II, III, aVR, aVF, aVL, V1, V2, V3, V4, V5, and V6) were obtained simultaneously; each was recorded at 500 samples per second, 16 bits per sample, with 5 μV resolution.

2.2. ECG human annotations

The ECGs were manually annotated by a group of 23 volunteer annotators, who identified themselves as 2 cardiologists, 1 (non-cardiologist) physician, 5 ECG analysts, 5 others with some experience reading ECGs, and 10 volunteers who had never read ECGs previously. Each annotator used a web browser to view and grade a random sequence of ECGs from the Challenge data set. We were able to estimate intra-observer variability, since most of the annotators graded a few of the ECGs more than once as a result of the random selection process.

The annotators were asked to give an overall assessment of each selected 12-lead ECG (i.e. all 12 leads, not each signal or portion of a signal individually), by assigning one of five possible letter grades to it: *A* (an outstanding recording with no visible noise or artifact; such an ECG may be difficult to interpret for intrinsic reasons, but not technical ones); *B* (a good recording with transient artifact or low-level noise that does not interfere with interpretation; all leads recorded well); *C* (an adequate recording that can be interpreted with confidence despite visible and obvious flaws, but no missing signals); *D* (a poor recording that may be interpretable with difficulty, or an otherwise good recording with one or more missing signals); or *F* (an unacceptably poor recording that cannot be interpreted with confidence because of significant technical flaws). Each grade represented the observer's assessment of the entire ECG record (10 seconds and 12 channels), as an overall measure of quality.

The letter grades were assigned these numerical values: A = 0.95, B = 0.85, C = 0.75, D = 0.6 and F = 0. For each ECG, we calculated the average of all grades, and gave it a *reference quality classification* of *Acceptable* (if two or more grades were available, the average grade ≥ 0.7, and no more than one grade was F), *Unacceptable* (if two or more grades were available, and the average grade < 0.7), or *Indeterminate* (otherwise; see figure 1).

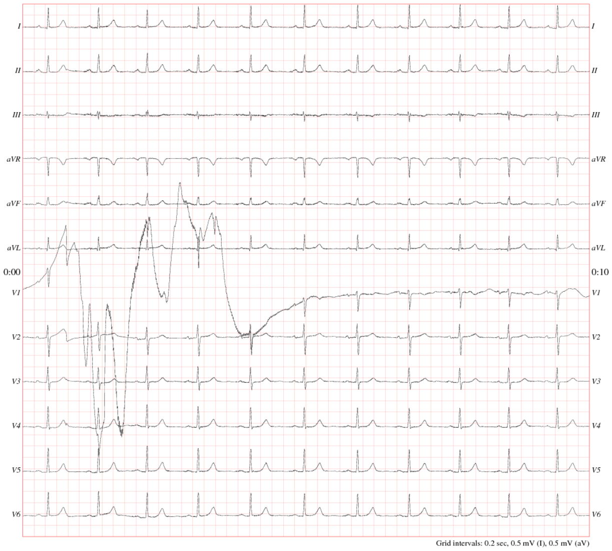

Figure 1. An *Indeterminate* ECG. Most of those who graded it gave it a C or better, but one gave it an F. Cases such as this one were not used to calculate scores in any Challenge event.

Each ECG was assigned randomly to one of three groups (training set A: 1000 ECGs, for which preliminary classifications were provided in April, and final classifications in July; test set B: 500 ECGs used in events 1 and 2, for which classifications were withheld; and set C: 500 ECGs used in event 3 but not available to participants).

2.3. Scoring

For events 1 and 2, the score for each entry was the fraction of correctly classified Acceptable and Unacceptable ECGs in set B (Indeterminate ECGs were excluded). In event 1, participants were ranked by the best final score obtained in up to five attempts.

In events 2 and 3, each participant submitted a single entry consisting of a Java module to be incorporated into an Android mobile application provided by PhysioNet. We tested each entry using test sets B (for event 2) and C (for event 3) by running it on two mobile phones running Android 2.1: a Motorola Defy (for ranking the entries in a controlled environment on a phone without network service or optional applications, but with floating-point hardware), and an HTC Hero 200 (not used for ranking the entries, but to estimate real-world performance using a typical phone with several applications installed, network service, without floating-point hardware).

The scoring for event 3 entries was calculated using the function

$$score_{event3} = accuracy \cdot e^{-(t-0.5)/10} \tag{1}$$

where t is the execution time (in seconds) on the Android phone and $accuracy$ is the percentage of correctly identified Acceptable and Unacceptable ECGs (as in events 1 and 2, but using test set C). The first time constant, 0.5 seconds, was chosen to reflect an ideal target speed time; thus the exponential function improves the score of an entry that requires less than 0.5 seconds on average. The second time constant, 10 seconds, reflects the length of an ECG; getting the last 10% in accuracy is not worth more than 10 seconds of execution time if it takes only 10 seconds to record another ECG.

3. Results

A total of 8,327 grades were obtained. In all 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, such as in Figure 1, were very rare. Table 1 summarizes the consistency of the annotators' grades as a function of experience level, showing 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.

In event 1, the top scores were obtained by the team of Xia et al.[5] with 0.932, followed closely by Li and Clifford[6] with 0.926, and 7 other participants who all scored 0.9 or better. In event 2, Xia et al. also had the top result of 0.914, Moody[7] scored 0.896, and others scored between 0.833 and 0.880. In event 3, Hayn et al.[8], with 0.873, and Chudacek et al.[9], with 0.872, achieved the best results, with others scoring between 0.791 and 0.845.

Figure 2 shows the average processing time on a mobile phone vs accuracy for the algorithms submitted to Event 3. Several of the most accurate algorithms require 0.5 seconds or less on the reference (Motorola) phone, or about 1.5 seconds or less on the control (HTC) phone; more accurate results were not obtained by longer-running entries, demonstrating that good agreement with the reference quality classifications can be achieved within reasonable processing times, even on a phone that is running other applications, has a network connection, and lacks floating-point hardware.

Table 1. Consistency of grades, by experience level. *Intra*: mean intra-observer consistency (the fraction of grades given by the same annotator to the same ECG at different times that agree with each other); *Inter*: mean inter-observer consistency (the fraction of grades that agree with those given by others with the same experience level, excluding ECGs with fewer than 3 grades); *Accuracy*: the fraction of grades consistent with reference quality classifications (A, B, and C are consistent with Acceptable, and D and F are consistent with Unacceptable; Indeterminate ECGs and those graded by fewer than 3 annotators are excluded). Random: Monte Carlo simulation of 1000 runs of random grades on the same ECGs.

Level	Intra	Inter	Accuracy
None	0.945	0.946	0.929
Some	0.947	0.916	0.977
Analyst	0.971	0.948	0.938
Cardiologist	0.980		
Physician	0.952		
All	0.954	0.949	0.921
Random	0.75	0.72	0.69

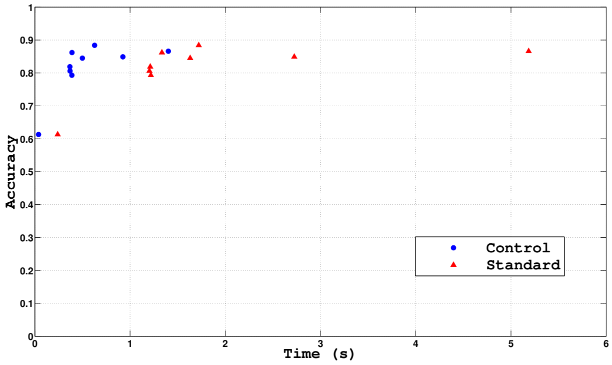

Figure 2. Execution time vs. accuracy of event 3 entries.

4. Discussion and conclusions

The top competitors employed a variety of techniques, using a wide range of features including entropy [5], higher order moments[6], filtering residues[10], signal-to-noise ratio[8, 11], regularity[12], and intra-lead information[6, 12, 13]. The classification methods used in the challenge included decision trees[14, 15], support vector machines[6, 16], fuzzy logic[17], and heuristic rules[7, 18].

A difficult task in the challenge was detection of electrode misplacement. This was not explicitly defined as a criterion for rejecting an ECG that could be interpreted with confidence. While it is possible that the human annotators with little experience ignored or were not aware of electrode misplacements resulting in lead reversals, the accurate detection of such reversals is not without some difficulties. Although 149 records (about 10% of the records available to the competitors) were identified as having likely electrode misplacement by one or more participants, the inter-observer consistency of these identifications was very low (at most 23%). An independent algorithm for electrode reversal detection was run by the PhysioNet organizers on the dataset, and comparison of the algorithm with the submitted list from the competitors yielded a consistency of 36% at most, and a false detection rate of at least 63%. In addition, from a small intersection of the records submitted by the competitors and the records graded by the physician expert, who had a strong background in ECG analysis, none of the three records were classified as unacceptable by the expert. The difficulty in detecting misplaced electrodes is compounded by the fact that certain clinical conditions, such as right ventricular hypertrophy or right axis deviation, can yield abnormal electrical vector projections when electrodes are placed acurately[19].

Overall, the PhysioNet/Computing in Cardiology Challenge 2011 shows promising results for fast and accurate ECG quality control on a mobile platform. The open-source Java code and data will remain available for those interested in improving or implementing the algorithms, and is a step toward extending the reach of high-quality health care affordably and efficiently using mobile phone technology.

Acknowledgements

The authors would like to thank the volunteer annotators, the GSM Association for its generous support of this year's Challenge awards and data collection activities, the Board and membership of Computing in Cardiology for their continuing and strong support of the Challenge series, Gari Clifford and his students for valuable feedback, and all of the participants in this year's Challenge. This work was funded by NIH cooperative agreement U01-008577 and NIH grant R01-EB001659.

References

[1] Sutherland E. Counting mobile phones, SIM cards, and customers. Africa 2008 April;1–10.

[2] Lester R, Gelmon L, Plummer F. Cell phones: tightening the communication gap in resource-limited antiretroviral programmes? AIDS 2006;20(17):2242–4.

[3] International Telecommunications Union. Key global telecom indicators for the world telecommunication service sector. `http://www.itu.int/ITU-D/ict/statistics/at_glance/KeyTelecom.html`, 2011.

[4] Gerber T, Olazabal V, Brown K, Mendez P. An agenda for action on global e-health. Health Affairs 2010;29(2):238–8.

[5] Xia H, McBride J, Sullivan A, Bock TD, Bains J, Wortham D, Zhao X. A multistage computer test algorithm for improving the quality of ECGs. Computing in Cardiology 2011;38.

[6] Li Q, Clifford G. Signal quality indices and data fusion for determining acceptability of electrocardiograms collected in noisy ambulatory environments. Computing in Cardiology 2011;38.

[7] Moody BE. A rule-based method for ECG quality control. Computing in Cardiology 2011;38.

[8] Hayn D, Jammerbund B, Schreier G. Real-time visualization of signal quality during mobile ECG recording. Computing in Cardiology 2011;38.

[9] Chudacek V, Zach L, Kuzilek J, Spilka J, Lhotska L. Simple scoring system for ECG signal quality assessment on Android platform. Computing in Cardiology 2011;38.

[10] Ho T, Chen X. PhysioNet Challenge 2011: Improving the quality of electrocardiography data collected using real time QRS-complex and T-wave detection. Computing in Cardiology 2011;38.

[11] Johannesen L. Assessment of ECG quality on an Android platform. Computing in Cardiology 2011;38.

[12] Kalkstein N, Kinar Y, Na'aman M, Neumark N, Akiva P. Classification of quality for ECG readings. Computing in Cardiology 2011;38.

[13] Maan A, Zwet E, Oliveira Martens S, Man S, Schalij M, Wall E, Swenne C. Matrix multiplication and baseline analysis methods to classify ECGs in the 2011 Physionet/CinC challenge. Computing in Cardiology 2011;38.

[14] Zaunseder S, Huhle R, Malberg H. CinC challenge assessing the usability of ECG by ensemble decision trees. Computing in Cardiology 2011;38.

[15] Baumgartner B, Mendoza A, Sprunk N, Knoll A. ECG quality rating for mobile devices. Computing in Cardiology 2011;38.

[16] Kuzilek J, Huptych M, Chudacek V, Spilka J, Lhotska L. Data driven approach to ECG signal quality assessment using multistep SVM classification. Computing in Cardiology 2011;38.

[17] Chiang Y, Hsu W, Liu S, Jiang Z, Jia J, Li Y, Li W, Wu J. Incorporating a priori knowledge into hidden Markov models for inadequate ECG detection. Computing in Cardiology 2011;38.

[18] Langley P, Marco L, King S, Maria C, Duan W, Bojarnejad M, Wang K, Zheng D, Allen J, Murray A. An algorithm for assessment of ECG quality acquired via mobile telephone. Computing in Cardiology 2011;38.

[19] Dubin D. Rapid Interpretation of EKGs. Sixth edition. Cover Pub Co, 2000.

Address for correspondence:

Ikaro Silva
45 Carleton St., Building E25-505
Cambridge, MA, 02139, USA
ikaro@mit.edu