CrowdLabel: A Crowdsourcing Platform for Electrophysiology

Tingting Zhu[1], Joachim Behar[1], Tasos Papastylianou[1], Gari D. Clifford[1,2]

[1] Institute of Biomedical Engineering, University of Oxford, UK
[2] Department of Biomedical Informatics, Emory University, Georgia Institute of Technology, USA

Abstract

In foetal electrocardiographic monitoring, assessment of foetal QT (FQT) in identifying foetal hypoxia has been limited mainly due to the lack of available public databases with expert labels. Our proposed platform, CrowdLabel, a web-based open-source annotation system, was developed for crowdsourcing medical labels from multiple expert and/or non-expert annotators. We describe the platform and an example of use; to improve FQT estimation by creating reference labels against which automated algorithms can be benchmarked. A total of 501, 30s segments were extracted from 15 foetal ECG (FECG) recordings from a private database. 23 volunteers participated in the study and provided a total of 7,307 FQT annotations, which were aggregated using a probabilistic label aggregator (PLA). The best annotator identified by the PLA had a standard deviation in the change of FQT annotations of 13.35 ms and 35.52 ms when labelling FECG with 'very good' and 'poor' signal quality respectively. The PLA does not require any ground truth to identify the best annotator or annotations. Annotator accuracy was also shown to be a function of objective signal quality measures. The feasibility of the CrowdLabel annotation system for ECG crowdsourcing with an unknown ground truth, as well as the results of the first experiment conducted using such a platform have been demonstrated.

1. Introduction

For medical applications, the ground truth annotations against which an algorithm or treatment is evaluated is often ascertained through manual labels by clinical experts. However, significant intra- and inter- observer variability and various human biases limit accuracy [1]. The electrocardiogram (ECG) is a standard tool for assessing cardiovascular health. Disagreements in ECG diagnostic annotations may be due to intrinsic difficulties in interpreting the signals that are linked to the level of training or experience of the annotators [2]. Disagreements may be exacerbated by significant amounts of noise such as motion artefacts, electrode contact noise, and baseline drift [3].

In the context of foetal monitoring, cardiotocography is used in clinical practice to record foetal heart rate (FHR). FHR can be recorded using an invasive (scalp ECG) or non-invasive (ultrasound) transducer. Studies of the morphology and temporal parameters of the foetal ECG (FECG) can provide valuable information to assess the foetal well-being. It has been shown that shortening of the QT interval (measured using the scalp electrode) was associated with intrapartum hypoxia (resulting in metabolic acidosis) [4]. However, the scalp ECG is invasive, requires a certain degree of cervical dilatation and only one electrode placed on the foetus head can be used. An alternative solution could be to develop a non-invasive method for measuring Foetal QT (FQT) using abdominal ECG sensors. However, extraction of morphological parameters such as the QT interval or ST deviation is particularly challenging due to low amplitude of the FECG signal. Furthermore, the lack of available public databases with expert labels reduces the opportunities for the scientific community to contribute solutions, and compare between alternatives.

The PhysioNet data archive, Physiobank, offers over 50 collections of biosignals [5]. PhysioNet provides a lightweight signal viewer [6], LightWAVE (http://physionet.org/lightwave/), which replicates features of WAVE (Waveform and Annotation Viewer and Editor) for accessing Physiobank through any modern web browser.

In this work, we have developed an open-source web-based platform, CrowdLabel, which uses a customised version of LightWAVE, to crowdsource medical labels from biosignals, such as FQT annotations. Through the use of a probabilistic framework, we propose a methodology for improving QT interval estimation in FECGs using multiple annotators. The resultant reference labels can be used to facilitate benchmarking of automated FQT measurement algorithms.

2. CrowdLabel

The CrowdLabel annotation system was adapted from LightWAVE version 0.38. It consists of a user authenti-

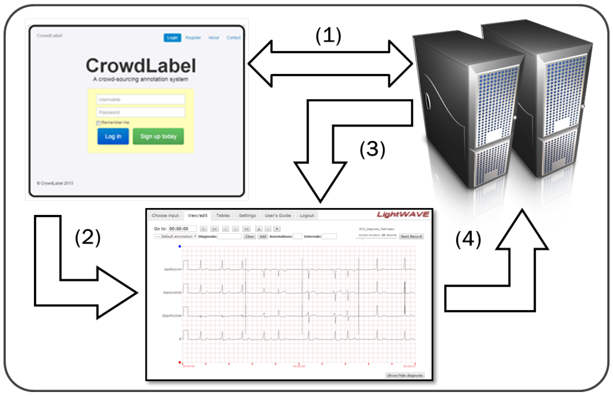


Figure 1. The main components of the CrowdLabel annotation system: (1) A user may log into the CrowdLabel authentication interface through HTTPS, where the user credentials are checked with a MySQL database (credentials are encrypted with MD5 and security checked for SQL injection attacks). (2) After gaining access, the user proceeds to the LightWAVE client interface, where user can specify the database name and record number to be loaded. (3) The user's provided information is submitted to the local database repository, and the corresponding medical data are loaded in SVG format dynamically either from the local private database or Physiobank repository. (4) Annotations provided by the user are submitted to the local server in a separate MySQL database and viewed after submission.

cation interface, a LightWAVE client, a back-end server database which mirrors Physiobank, and a local server which stores annotations provided by users (see Figure 1). CrowdLabel is a standalone platform which was written in PHP, HTML5, JavaScript, JQuery, JQuery User Interface, and uses Scalable Vector Graphics (SVG). The configuration of the back-end server follows the LightWAVE's infrastructure, which utilises a common gateway interface application to retrieve data from local data repository (including PhysioBank and private databases) and delivers them dynamically upon request generated by the LightWAVE client. The user authentication interface and the LightWAVE client can run within any modern web browser and does not require installation on the users computer.

In additional to interval and/or point labelling, CrowdLabel provides the opportunity for the annotator to provide an assessment of signal quality the user's confidence level annotation functionality. The signal quality is divided into five grades with an additional 'Skip' option depending on the quality of the annotated interval (see Figure 2): (A) An outstanding recording with no visible noise or artefacts; such an ECG may be difficult to interpret for intrinsic reasons, but not technical ones. (B) A good recording with transient artefacts 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. (F) An unacceptably poor recording that cannot be interpreted with confidence because of significant technical flaws. In terms of confidence level, there are five levels of confidence: $\gtrsim$90% (Extremely confident); 60$\sim$90% (Mostly confident); 33$\sim$60% (Unsure); $\lesssim$33% (Not at all confident); Skip (No response).

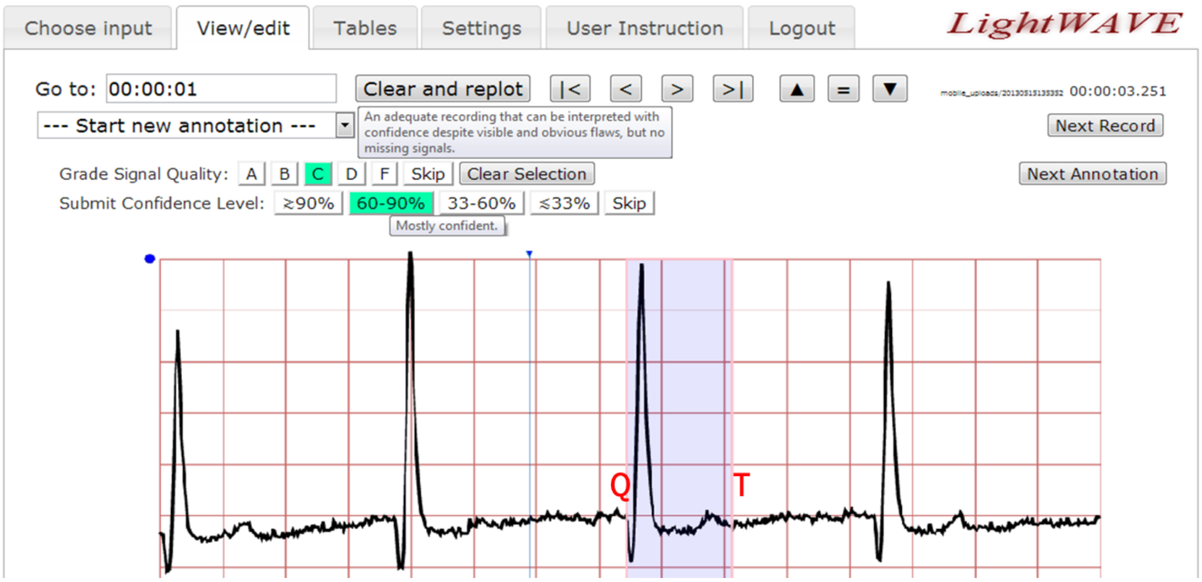
3. The Probabilistic Framework

Within our previous study [7] on adult QT annotations using the Physikalisch-Technische Bundesanstalt Diagnostic ECG Database [8], the results have shown that using a probabilistic label aggregator (PLA) with beat specific signal quality and heart rate (bHRSQIs) feature based on the Expectation Maximization algorithm outperformed annotations provided by the best human/algorithmic annotator. PLA can therefore provide an improved 'gold standard' for QT annotation tasks even when ground truth is not readily available. Here we have deployed PLA with bHRSQIs feature for aggregating FQT annotations, weighting annotators based on their precision in an unsupervised manner.

4. Data Description

A total of 23 researchers (doctoral students and postdoctoral researchers) participated in our study to label raw FQT intervals. The data were collected from a private database, which contains 15 records of healthy foetuses. Their FECGs were extracted by placing an invasive electrode on the foetus' head (fs = 1kHz), and records were divided into 501, 30-second segments. The segments were presented randomly to annotators and only one FQT interval was allowed to be annotated per segment. Prior to annotating FQT intervals, annotators were given minimal training through the tutorial on how to label a QT interval by the mean of a walk-through video, live demonstration, and handouts. In order to motivate participation, all participants were provided a chance to win a selection of rewards, such as a restaurant meal or sweets, depending on the number of segments annotated by the deadline that we set. An example of FQT annotation is shown in Figure 2. In addition to FQT label, each annotator was asked to grade the signal quality and his/her confidence level for their selected annotation.

In this study, we compared the root-mean-square error (RMSE) of each annotator with the PLA aggregated result for all FQT segments and measured the precision (i.e. 1/variance) of each annotator. In addition, we analysed the RMSE of each annotator by penalising annotators with respect to the number of segments they skipped. Furthermore, we compared the performance of the PLA rated best, medium, and worst annotators based on their precision values, and measured their FQT annotations on different signal quality of FECG segments.



Go to: 00:00:01

--- Start new annotation ---

An adequate recording that can be interpreted with confidence despite visible and obvious flaws, but no missing signals.

mobile_uploads/20130515135352 00:00:03.251

Grade Signal Quality: C

Submit Confidence Level: 60-90%

Mostly confident.

Figure 2. Example of a FQT annotation using our interface. The shaded blue rectangle represents a QT annotation on a single beat, which starts at the beginning of the Q wave and ends at the end of the T wave.

5. Results and Discussion

Two weeks after the launch of the study, 7,883 annotations were collected from 23 annotators of which 576 of them were annotated as ‘Skip’. The number of annotations performed by each participant is given in Figure 3. The remaining 7,307 FQT annotations were aggregated using the PLA with bHRSQIs feature. The first two participants who had finished annotating all segments were awarded restaurant meals, and those had annotated more than 50% segments were also rewarded sweets.

The participants were ranked based on the RMSE computed between their annotations and the aggregated annotations generated by the PLA (taken as being the ‘silver truth’). Furthermore, the RMSE of each annotator with and without penalising the amount of segments being annotated were estimated and shown in Figure 4.

The variance of each annotator was further estimated and compared. Figure 5 shows the distribution of results of the change of FQT annotations (△FQT) for the PLA rated annotators on segments of three patients with signal qualities of type very good (grade A and B), medium (grade C), and bad (grade D and F). The annotator with a lower variance (σ^2) indicates high consistency and hence higher precision. Three annotators had similar σ of △FQTs when the signal quality was very good (see the top plot in Figure 5):13.35 ms, 19.24 ms, and 18.39 ms for best, medium, and worst annotator respectively. When annotating bad signal quality segments, the σ of △FQT was much worse than those with very good signal quality (35.52 ms, 62.96 ms, and 75.65 ms). Nevertheless, the PLA selected best annotator has proved to have the least variance across different quality of segments.

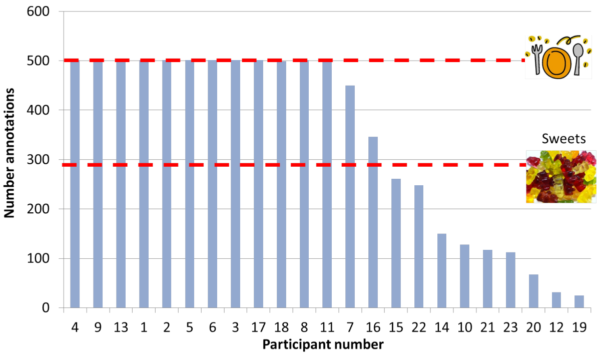


Figure 3. Ranking of the number of annotations provided by each of the 23 participants.

6. Conclusion and Future Work

We have demonstrated a proof-of-concept crowdsourcing methodology to generate accurate FQT annotations using CrowdLabel. The PLA can be used in the back-end system to aggregate annotations in an unsupervised setting. In the future, CrowdLabel aims to be used as a training system for annotating 12-lead ECGs. The self-rated signal quality and and confidence level can be used to provide feedback on the annotators’ progress in learning.

The current model we use for the PLA does not consider the bias of each annotator (i.e. the annotator can be pre-

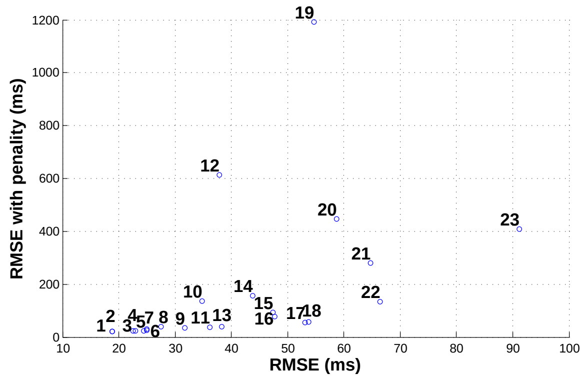


Figure 4. RMSE vs. RMSE with penalty for each annotator; the RMSE was computed between their annotations and the aggregated annotations generated by the PLA using bHRSQIs feature (taken as being the 'silver truth'). The participant numbers are labelled on the plot.

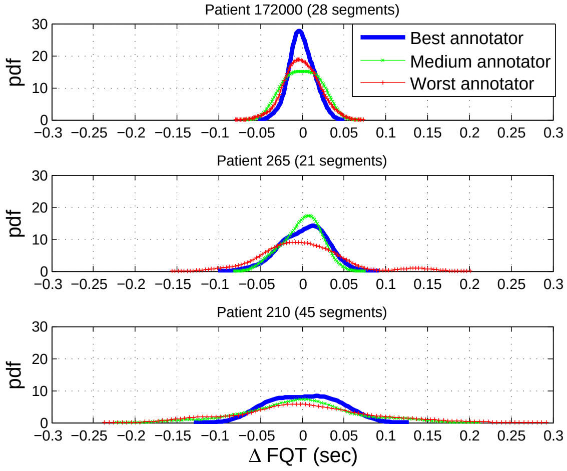


Figure 5. The variance of FQT annotations (△FQT) for the PLA selected annotators are plotted for segments with very good, medium, and bad signal quality from top to bottom respectively.

cise but always consistently over- or under- estimates the QT intervals). Figure 6 shows that participant number 10, for example, always over-estimates the FQT annotations, whereas participants number 18, 22, and 23 always underestimate the FQT annotations. By correcting the offset of each annotator, including a bias term estimation into the PLA model might lead to further improvement in estimating the 'true' FQT.

Acknowledgements

TZ and TP are supported by the RCUK Digital Economy Programme grant number EP/G036861/1. TZ is also supported by China Mobile Research Institute. JB is supported by the UK EPSRC, the Balliol French Anderson Scholarship Fund, and MindChild Medical Inc. The authors would like to thank all study participants and George Moody for generous help and LightWAVE source.

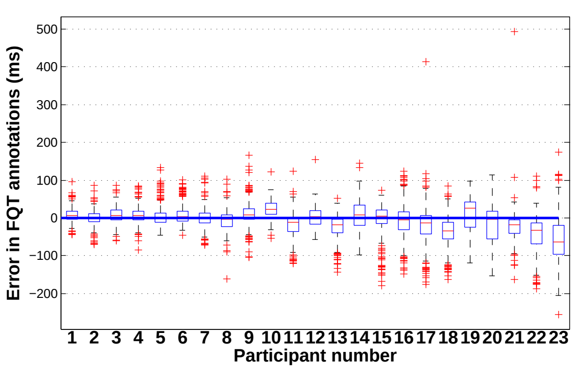


Figure 6. A boxplot of the error of submitted FQT annotations when compared with results of the PLA aggregation.

References

[1] Brady WJ, O'Connor RE. Interpretation of the electrocardiogram: clinical correlation suggested. Eur Heart Jour 2008; 29(1):1–3.

[2] Salerno SM, Alguire PC, Waxman HS. Competency in interpretation of 12-lead electrocardiograms: a summary and appraisal of published evidence. Ann Intern Med 2003; 138(9):751–760.

[3] Clifford GD, Azuaje F, McSharry PE. Advanced Methods and Tools for ECG Analysis. Engineering in Medicine and Biology. Norwood, MA, USA: Artech House, 2006.

[4] Oudijk MA, Kwee A, Visser GH, Blad S, Meijboom EJ, Rosn KG. The effects of intrapartum hypoxia on the fetal QT interval. BJOG IntJ Obstet Gy 2004;111(7):656–660.

[5] 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.

[6] Moody G. Lightwave: Waveform and annotation viewing and editing in aweb browser. In Computing in Cardiology Conference. 2013; 17–20.

[7] Zhu T, Johnson AE, Behar J, Clifford GD. Crowd-sourced annotation of ecg signals using contextual information. Ann Biomed Eng 2014;42(4):871–884.

[8] Bousseljot R Kreiseler D SA. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB uber das Internet. Biomedizinische Technik 1995;40(1):317–318.

Address for correspondence:

Tingting Zhu
Dept. of Engineering Science,
Univeristy of Oxford
tingting.zhu@eng.ox.ac.uk