Bayesian Voting of Multiple Annotators for Improved QT Interval Estimation

Tingting Zhu, Alistair E W Johnson, Joachim Behar, Gari D Clifford

Department of Engineering Science, University of Oxford, Oxford, UK

Abstract

Human bias and significant intra- and inter- observer variance exist in electrocardiogram QT interval evaluation. A Bayesian approach (BA) with an informative prior, that combines measures from multiple humans or algorithms as well as contextual information (such as heart rate and signal quality) was developed for inferring the $true$ QT length. The developed method is compared to the mean and median voting approaches by computing the root-mean-square (RMS) error between the computed QT lengths and the reference annotations provided by the 2006 PhysioNet/Computing in Cardiology Challenge. The BA with features can reduces the human RMS error of QT estimates to 6.04ms and 13.97ms for automated algorithms, out-performing the results in the Challenge of 6.67ms and 16.34ms respectively. For three annotators, the BA had a 10.7% improvement over the next best voting strategy for manual annotations, and 14.4% for automated algorithms. For large numbers of annotators, the BA estimates became approximately equal to the best-performing annotator.

1. Introduction

The Electrocardiogram (ECG) is a standard and powerful 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[1]. Disagreements may be exacerbated by significant amounts of noise such as motion artefacts, electrode contact noise, and baseline drift[2]. In this study, the potential for improving QT interval estimation in ECGs is explored using multiple annotators in a Bayesian voting framework.

The QT interval is a marker for ventricular depolarisation and repolarisation of the cardiac muscle cells[2]. Prolongation of the QT interval serves as an important risk factor for arrhythmias and sudden cardiac death[3]. In addition to the high costs for manual annotations, previous studies have reported significant intra- and inter- observer variability in QT annotations, ranging from 10 to 30ms[4, 5]. Manual annotators also appear to underestimate the true T wave end-point due to various T wave morphologies and different noise sources[6]. In comparison to manual annotators, automated algorithms offer time efficiency, repeatability, and cost-savings benefits.

One of the major challenges for automated annotation of the ECG is the substantial discrepancy in the QT interval estimation when compared to manual methods[7]. As there is no recommendation in guidance for regulating automated algorithms, it is difficult to assess the acceptability of automated annotation algorithms. The PhysioNet/Computing in Cardiology (PCinC) 2006 Challenge strove for accurate QT interval measurements on a publicly available database, with the best automated algorithm achieving an error of 16.34ms[8].

In situations where the ground truth is not readily available, it is common to have multiple different annotators evaluate the data to provide aggregate labels. Simple methods like the mean and median methods generally perform well only if there are a large number of annotators available. A more effective and less biased probabilistic approach, combining annotations using expectation-maximization (EM), was first proposed by Dawid and Skene for error measurements in patient records based on the results from multiple responses without a gold standard[9]. The crowd-sourcing EM method proposed by Raykar *et al.*[10] was similar to the model proposed by Dawid and Skene[9]. It extended the original method by jointly estimating the annotation labels with a regression model. In the context of QT estimation, a crowd-sourcing algorithm was proposed by Moody[8] at the PCinC 2006 Challenge. The "Meta-6" algorithm combined the strengths of the three best-performing annotators from both the open source and closed source automated algorithms. It excluded records rejected by more than one of these six algorithms as well as those mostly disagreed. The QT intervals were estimated from the remaining records by taking the medians of the measurements. When compared to the reference QT interval, the "Meta-6" algorithm achieved a root mean squared error of 10.93ms.

The crowd-sourcing EM algorithm proposed by Raykar *et al.*[10] is applied to QT interval estimation to assess the feasibility and potential of the Probabilistic EM Algorithm (PEMA) in a Bayesian framework to improve QT interval estimation. As differentiating the physiologic changes from noise is one of the major challenges in observing sin-

gle lead QT dynamicity (i.e. Beat specific QT changes), a surrogate for the heart rate and signal quality is jointly estimated with the annotations. These results were compared to the mean and the median voting strategies.

2. Methods

2.1. Multivariate regression

In this study, it is assumed that we have N records from R annotators. Let $\mathbf{D} = [\mathbf{x}_i, y_i^{j=1}, \cdots, y_i^R]_{i=1}^N$, where $\mathbf{x}_i$ is a feature vector for the ith observation, and y^j corresponds to the given annotation provided by the jth annotator whereas y_i represents the true unknown annotation. In the context of QT interval analysis, we assume that y_i^j is a noisy version of y_i (i.e. the true length of the QT interval), which can be described using a Gaussian model. The precision of the jth annotator, λ^j, is defined as the inverse-variance in the model. Furthermore, y_i can be predicted using a multivariate regression model as $y_i = \mathbf{w}^T\mathbf{x}_i + \epsilon$, where $\mathbf{w}$ is the regression coefficients and ϵ is a zero-mean Gaussian noise.

2.2. The EM algorithm

The likelihood of the parameter $\theta = \{\mathbf{w}, \lambda\}$ for a given $\mathbf{D}$ can be solved using the EM algorithm proposed by Dempster *et al.*[11] in two-step iterative process: i) The E-step estimates the expected true annotations for all records, $\hat{\mathbf{y}}$, each is a weighted sum of the provided annotations from all annotators and their precisions:

$$\hat{\mathbf{y}} = \frac{\sum_{j=1}^R \lambda^j \mathbf{y}^j}{\sum_{j=1}^R \lambda^j} \tag{1}$$

ii) The M-step is based on the current estimation of $\hat{\mathbf{y}}$ and $\mathbf{D}$, then the model parameters such as the precision and regression coefficient, λ and $\mathbf{w}$, can be calculated by solving the zero gradient of the log-likelihood respectively:

$$\mathbf{w} = (\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T)^{-1} \sum_{i=1}^N \mathbf{x}_i\hat{\mathbf{y}}_i \tag{2}$$

$$1/\lambda^j = \frac{1}{N}\sum_{i=1}^N (y_i^j - \mathbf{w}^T\mathbf{x}_i)^2 \tag{3}$$

Note: $\mathbf{w}$ may be solved using a standard linear regression model with features, $\mathbf{x}$, except it is trained with $\hat{\mathbf{y}}$ as the predicted ground truth. In this instance, $\hat{\mathbf{y}}$ is a precision weighted mean of the response from all the annotators.

When features, $\mathbf{x}$, are not available, the precision may be solved as the least square difference between the actual ground truth and the predicted one.

An equal λ as a prior among all annotators is assumed to initialise the PEMA, and the initial guess of the expected QT annotation is set to be $\hat{\mathbf{y}} = \frac{1}{R}\sum_{j=1}^R \mathbf{y}^j$. Then the E-step and M-step can be iterated until convergence of λ. As a result, the PEMA establishes a weighted sum of annotations estimating the expected true annotations, as well as providing the precision of each annotator.

2.3. Data description

The data were drawn from the QT interval annotations generated from participants in the PCinC 2006 Challenge[8]. There were two categories of annotations: manual and automated. A total of 89 entries including revised submissions, with 38,621 annotations were considered: 20 human annotators in Division 1, 48 automated algorithms in Division 2 (closed source), and 21 in Division 3 (open source). Division 4 had a total of 69 automated algorithms as it combined annotators from Division 2 and 3. A single record, "patient285/s0544re", was excluded as it did not contain any recognisable ECG signals. Annotations for 548 records of the Physikalisch-Technische Bundesanstalt Diagnostic ECG Database (PTBDB) were processed using the PEMA, mean, and median voting strategies. The competition score for each entry was calculated from the root mean square (RMS) difference between the submitted and the reference QT intervals, weighted by the number of records attempted. The reference annotations were generated from Division 1's entries as detailed in[8].

2.4. Beat detection and feature extraction

The Lead II ECG was digitised at 1000 samples per second, with 16 bit resolution, over a range of $\pm$16.384mV. The records in the PTBDB came from 294 subjects, each represented by one to five recordings. Each record was up to 2 minutes in length. QRS waves were detected using an open source QRS detector, *eplimited*[12].

The first 5-second of each record was considered in this study as this was where the majority of annotations occurred, and it also reduced large inter-beat variations. An example of a 5-second segment is shown in Fig. 1. In each record, the beat specific heart rates were estimated through the R-R intervals and the square root of R-R intervals together with signal quality metrics proposed in[13] were used as features for the PEMA.

To account for inter-beat variations, each annotated representative beat was located and a median of the five preceding R-R intervals was assigned as a feature to each annotator in the individual record. If the annotated beat occurred at the beginning of the record, the first five beats were considered to estimate the median R-R interval.

The signal quality indices (SQI) were measured in a 5-second window, overlapped by four seconds to account for the QT dynamicity. These SQI features provide extra information on the signal quality of the annotated beat in a 5-second window. An example of a noisy ECG signal is shown in Fig. 1, large inter-observer variation (i.e. 100ms) occurs among two human annotators who had annotated on the same ECG beat.

2.5. Methodology of validation

The PEMA was applied to each individual division to estimate a weighted sum of annotations and the cor-

responding RMS errors were compared with the best-performing scores that were published in the Challenge. In addition, the PEMA was compared to the median and mean voting strategies when selecting different number of annotators. A sub-sampling method was performed by randomly selecting K annotators 100 times, and this was repeated with K varied from 1 to the maximum number of annotators in each division. The mean and standard error ($\mu \pm \sigma_\mu$ms) of the RMS error of the PEMA, the mean, and the median were calculated and assessed. The average percentage of records with full annotation was also measured when selecting different number of annotators.

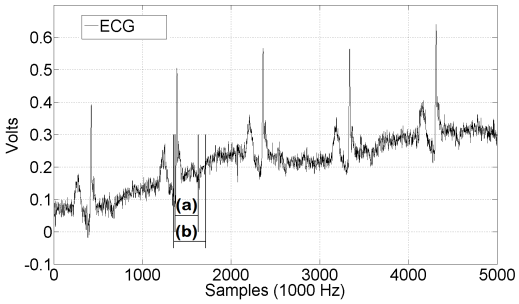

Figure 1. An example of a noisy 5-second ECG segment. The QT interval was considered to be 260ms by one human annotator (a) and 361ms by a second human annotator (b).

3. Results

The minimum RMS error in estimating the QT interval of the PEMA with and without features were estimated and compared as shown in Table 1. The PEMA RMS errors without features were significantly larger than those with features at 95% level of confidence. The PEMA using the beat-specific heart rate and SQI as features provided better RMS results as compared to the results of using other type of features.

The PEMA RMS error was 6.04ms when considering all human annotators in Division 1. It outperformed the mean voting strategy for all number of annotators but was worse than the median voting approach (RMS = 4.71ms for 20 annotators) after nine annotators. Using 15 out of 20 annotators (RMS = 6.62$\pm$0.98ms) for the PEMA achieved a similar error as the best score (RMS = 6.67ms) provided in the Challenge. In Division 2, the PEMA consistently outperformed the other two approaches and achieved a minimum RMS of 14.58ms when considering all 48 annotators. Seventeen annotators (RMS = 15.68$\pm$1.83ms) from Division 2 would be required for the PEMA to generate a similar RMS error as the best-performing annotator (RMS = 16.34ms). In Division 3, the PEMA continued to outperform the other two approaches and achieved a RMS of 16.58ms when considering 21 annotators. It also achieved less RMS error when compared with the best-performing annotator (RMS = 17.33ms) in this division. The PEMA had an RMS error of 13.97ms on Division 4 (see Fig. 2).

Table 1. Minimum RMS errors in ms of the PEMA without features (NF) and with HR, SQI or both.

Division (annotators)	NF	SQI	HR	HR+SQI
1 (20)	6.87	6.65	6.22	6.04
2 (48)	15.03	14.85	14.61	14.58
3 (21)	18.87	17.80	16.66	16.58
4 (69)	14.74	14.24	14.12	13.97

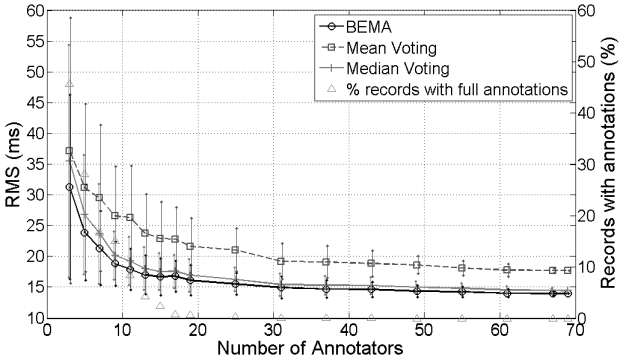

Figure 2. The mean and standard error of the RMS results of using the PEMA with features, median and mean voting are shown for the automated entry. The x-axis indicates selection of number of annotators. These plots were generated from 100 random subsets of annotators.

4. Discussion

In this study, the percentage of records with a complete set of annotations for all annotators was inversely proportional to the number of annotators. Randomly selecting fewer numbers of annotators without considering their annotations could lead to reduction in the average percentage of records with annotations (see Fig. 2). When increasing number of annotators further, the average percentage of records which were fully annotated by all participating annotators decreased to zero.

The PEMA with additional SQI and HR features produced a significantly smaller RMS error as compared to those using the features of heart rate alone, when selecting maximum number of annotators in all divisions (see Table 1). In all divisions, errors were significantly lower ($p < 0.05$ using a two-sided paired t-test) for HR, SQI, and HR+SQI than for NF, and for HR+SQI than for either HR or SQI individually.

Commonly, only three annotators are used to identify the QT interval length, and they may collaborate for labelling the QT interval. This type of direct collaboration often incorrectly weights one annotator and is more bias and less effective. In the situation where the annotators operate independently, the PEMA provides a 10.7% improvement over the next best voting strategy for manual annotations, and 14.4% for automated annotations (see Fig. 2). Therefore, combining human annotators using the PEMA

could potentially provide an optimal 'gold standard' in QT estimations even in the case when the ground truth is not available.

The reference QT intervals in this study were provided based on bootstrapping the median of up to 15 annotators, which explained the case when increasing the number of annotators the PEMA performed slightly worse than the median voting strategy. Furthermore, human annotators generally underestimate the QT intervals as mentioned in[6]. Therefore, the errors provided in the automated entries were always larger than those in the manual entries.

Among the entries for automated algorithms, the RMS error in Division 4 estimated using the PEMA cannot be compared directly with the results of the "Meta-6" algorithm (RMS = 10.93ms). This is because out of 548 records, the "Meta-6" algorithm had excluded approximately 26 records from the PTBDB based on arbitrary disagreements between annotators, whereas the PEMA considered all records that were available.

5. Conclusion

The accuracy of estimating the QT interval for a channel of ECG (Lead II) using multiple independent annotators was analysed and compared in this study using different voting strategies. The PEMA was shown to consistently outperform the median voting algorithms for less than nine human annotators and for any number of automated algorithms. For large numbers of annotators, the PEMA estimates became approximately equal to the best-performing annotator, even though the identity of the best annotator was unknown. In addition, it outperformed the mean voting strategy in all circumstances. Therefore, the PEMA has the potential to provide a more realistic reference when no gold standard exists. That is, since the PEMA approach works by comparing inter-annotator variations, no absolute reference data is required. The RMS error of 13.97ms when combining multiple algorithms (69 in total), is the lowest so far reported in the literature for non-human QT estimation. The PEMA with the beat-specific heart rate and SQI features not only addresses the issue of large inter- and intra- observer variation, but also provides an improvement in QT estimation as compared to the median and mean voting strategies when there are few annotators. In particular, the use of features provides contextual information so that the PEMA can learn how varying physiological and noise conditions affect each annotator differently. This approach, or incorporating context into weighting of annotators, is likely to have a larger impact for noisier data sets or annotators with a variety of specialisations or skill levels.

Acknowledgments

TZ and AJ acknowledge the support of the RCUK Digital Economy Programme grant number EP/G036861/1 (Oxford Centre for Doctoral Training in Healthcare Innovation). TZ also acknowledges the support of China Mobile Research Institute. JB is supported by the UK EPSRC and the Balliol French Anderson Scholarship Fund.

References

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

[2] Clifford GD, Azuaje F, McSharry PE. Advanced Methods and Tools for ECG Analysis, volume 1 of Engineering in Medicine and Biology. 1 edition. Norwood, MA, USA: Artech House, 2006.

[3] ICH. Guidance for Industry E14: Clinical Evaluation of QT/ QTc Interval Prolongation and Proarrhythmic Potential for Non- Antiarrhythmic Drugs. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformat%ion/Guidances/ucm073153.pdf.

[4] Ehlert FA, Goldberger JJ, Rosenthal JE, Kadish AH. Relation between QT and RR intervals during exercise testing in atrial fibrillation. Am J Cardiol Aug 1992;70(3):332–338.

[5] Christov I, Dotsinsky I, Simova I, Prokopova R, Trendafilova E, Naydenov S. Dataset of manually measured QT intervals in the electrocardiogram. Biomed Eng Online 2006;5:31.

[6] Clifford GD, Villarroel MC. Model-based determination of QT intervals. In Computers in Cardiology. 2006; 357 –360.

[7] Marek Malik. Errors and misconceptions in ECG measurement used for the detection of drug induced QT interval prolongation. J Electrocardiol 2004;37, Supplement:25 – 33. ISSN 0022-0736.

[8] Moody GB, Koch H, Steinhoff U. The PhysioNet/ Computers in Cardiology Challenge 2006: QT interval measurement. In Computers in Cardiology. 2006; 313 – 316.

[9] Dawid AP, Skene AM. Maximum likelihood estimation of observer error-rates using the EM algorithm. J R Stat Soc Ser C Appl Stat 1979;28(1):20–28.

[10] Raykar VC, Yu S, Zhao LH, Valadez GH, Florin C, Bogoni L, Moy L, Blei D. Learning from crowds. J Mach Learn Res 2010;1297–1322.

[11] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Series B 1977;39(1):1–38.

[12] Hamilton PS, Tompkins WJ. Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. IEEE Trans Biomed Eng 1986;(12):1157–1165.

[13] Clifford GD, Behar J, Li Q, Rezek I. Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms. Physiol Meas 2012;33(9):1419–1433.

Address for correspondence:

Tingting Zhu
IBME, Old Road Campus, Oxford, OX3 7DQ, UK
tingting.zhu@eng.ox.ac.uk