# A QT Interval Inaccuracy Index (QTI) for Highly Automated TQT Studies

Mously D Diaw[1,2], Stéphane Papelier[1], Alexandre Durand-Salmon[1], Jacques Felblinger[2,3], Julien Oster[2,3]

[1] Cardiabase, Banook Group, Nancy, France
[2] IADI, U1254, Inserm, Université de Lorraine, Nancy, France
[3] CIC-IT 1433, Université de Lorraine, Inserm, CHRU de Nancy, Nancy, France

## Abstract

*Aims: Thorough QT studies (TQT) evaluate the QT prolonging effect of drugs and consequently their potential proarrhythmic risk. These studies require meticulous ECG analysis; automatic QT interval measurements are usually overread by experts and adjusted if necessary. Our study aimed to provide a QT Interval Inaccuracy index (QTI) to automatically identify inaccurate automated QT interval estimations.*

*Methods: 12-lead ECG recordings and their manual interval measurements were obtained from 2 TQT study databases in PhysioNet (ECGDMMLD and ECGRDVQ). We derived 268 single-lead features that might relate to the accuracy of automatic QT intervals computed by convolutional neural network (CNN) based QT estimators previously trained on our own ECG database. Using these features (inter-estimator QT variability metrics, CNN-derived ECG characteristics, estimator confidence levels), classification of accurate and inaccurate automatic QT intervals was performed with a regularized logistic regression algorithm trained on the ECGDMMLD database. The QTI was then defined for each ECG recording as the average probability of the QT being inaccurate across the 12 leads ($0 \leq QTI \leq 1$).*

*Results: With a QTI threshold of 0.5, 41% inaccurate automatic QT intervals were identified in the ECGRDVQ database. Better estimates of the drug-induced QTc changes were obtained by correcting these QT intervals. The QTc prolongation obtained with this semi-automated method differed from the manual one by on average 2.98 ms (std = 1.82 ms) across the 4 drugs studied, compared to 24.1 ms (std = 21.0 ms) for the fully automated method.*

*Conclusion: The QTI allows a more accurate and robust analysis of ECG signals from TQT studies. The proposed QTI technique is closely linked to the performance of the CNN-based QT estimator; improvement of the latter would require retraining of the QTI.*

## 1. Introduction

Heart-rate corrected QT (QTc) prolongation is the primary biomarker for assessing the risk of drug-induced *torsades de pointes* during Thorough QT (TQT) studies. Ten of thousands of ECG signals are analyzed during such studies, usually with a semi-automated method where every computer-based QT measurement is overread by ECG experts and manually adjusted if necessary. Fully automated methods are indeed not yet recommended due to reliability concerns especially when analyzing noisy or abnormal ECG signals including drug-induced T-wave morphology changes, low amplitude P or T waves and overlapping U waves.

The workload and cost of TQT studies could be significantly lowered if experts were only required to overread signals where automated QT interval measurement is challenging. Signal quality indices were proposed to automatically detect low-quality ECG signals that need expert review. The quality indices were derived from ECG morphology features or outlier interval measurements [1, 2]. This approach works on the assumption that the automated algorithm chosen for QT measurement is reliable on clean ECG signals, but this still requires extensive validation on drug-induced ECG abnormalities.

In this work, we propose an algorithm-centered approach that estimates the level of inaccuracy of automated QT measurements, denoted as the QT Inaccuracy Index (QTI). The QTI relies on recently suggested convolutional neural networks (CNN) based QT estimators [3]. Similarly to what has been proposed for the estimation of ECG quality by comparing two automated QRS detectors [4], we suggest classifying QT measurements as accurate or inaccurate by comparing the outputs of three CNN-based models, in addition to analyzing CNN-derived ECG morphology features. This method aims at automatically selecting the ECG signals with less accurate QT measurements to be reviewed by experts.

Table 1. Description of the databases.

| | Train | Validation | Test |
|---|---|---|---|
| | ECGDMMLD | ECGDMMLD | ECGRDVQ |
| Nbr Subjects | 17 | 5 | 22 |
| Nbr Leads[1] | 38,112 | 12,420 | 62,784 |

[1] Nbr Leads: 12xNbr of ECG recordings

## 2. Method

### 2.1. Data

Two publicly available databases, published on PhysioNet [5], were used in this study: (i) ECGDMMLD database [6] was used for training and validation purposes; (ii) while the ECGRDVQ database [7] was used as an external test database. Both databases consisted in prospective randomized placebo-controlled clinical trials testing the cardiotoxicity of four different drugs. They include 12-lead ECG signals and expert-validated QT measurements over several timepoints. A summary of both databases is given in Table 1.

### 2.2. Automatic QT measurements

In a previous study [3], we trained three CNN-based QT estimators: (1) a U-Net model, (2) a residual neural network (KanResWide) adapted from Hicks et al. [8] and (3) a basic CNN with attention mechanism (AttnCNN). These models take as input a single-lead ECG heartbeat (cf. Figure 1.A). The QT output space was discretized in 2 ms sub-intervals or classes and the models first predict the QT interval class before computing the QT interval as the mid-point of the predicted sub-interval. The models also output the probability of belonging to that sub-interval, which can be interpreted as their inherent confidence level. In order to compute the automatic QT intervals for the ECGDMMLD and ECGRDVQ databases with these models, we first computed average beats for each of the 12 leads in the ECG recordings.

We will denote $\mathbf{QT}_{i,j}$, i $\in \{1,2,3\}$ , j $\in \{1,..,5\}$ the QT measured by Model 1 (U-Net), Model 2 (KanResWide) or Model 3 (AttnCNN) yielded by the $j_{th}$ fold of our 5-fold cross-validation experiment and $\mathbf{CL}_{i,j}$ the corresponding confidence level. The automatic QT against which is compared the manual QT is $QT_{automatic} = QT_{U-Net} = \frac{1}{5}\sum_{j=1}^{5} QT_{1,j}$. U-Net was chosen because it localizes both the QRS onset and the T offset while KanResWide and AttnCNN only output a QT interval estimate.

### 2.3. QT Interval Inaccuracy labeling

We want to identify less accurate QT measurements requiring manual review by an expert. To this aim, we labeled the available data with binary labels: 1 if the absolute difference between the actual QT interval and the automatic QT interval (U-Net) is above 15 ms and 0 otherwise

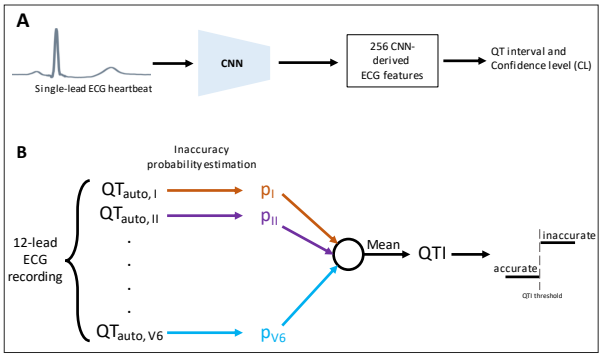


Figure 1. **A,** Single-lead automatic QT measurement with CNN-based models. **B,** The QTI for a given 12-lead ECG recording is defined as the mean probability that the automatic single-lead QT interval is inaccurate.

(1: inaccurate QT, 0: accurate). The 15 ms QT threshold was chosen as a reasonable middle ground of the reported limits of agreement (5-25 ms) between several ECG readers for expert adjudication [7, 9].

### 2.4. Feature engineering

In order to build the QTI, a total of 268 features were derived from the outputs of the 15 (3x5) CNN-based QT estimators. These features consist in the following:

**(i) U-Net confidence levels**: $\mathrm{CL}_{1,1-5}$ sorted in the ascending order.

**(ii) U-Net extracted ECG features**: The U-Net model computes 256 global features, which give a compact representation of the ECG template. In this work, we used the features computed by the model from the $1^{st}$ fold.

**(iii) QT variability metrics**: 7 variability metrics were computed from the automated QT intervals and defined as follows:

- the global QT standard deviation (SD)

$$\sigma_g = SD_{i,j}(QT_{i,j}, i \in \{1,2,3\}, j \in \{1,..,5\})$$

- the QT standard deviation for each model across folds

$$\sigma_i = SD_j(QT_{i,j}, j \in \{1,..,5\}), i \in \{1,2,3\}$$

- the mean dissensus (disagreement between models)

$$d = \frac{1}{15}\sum_{i=1}^{3}\sum_{j=1}^{5}\frac{(QT_{i,j} - meanQT)^2}{maxQT - minQT}$$

with *meanQT*, *maxQT* and *minQT* the mean, maximum and minimum of all $QT_{i,j}, i \in \{1,2,3\}, j \in \{1,..,5\}$.

- the mean absolute difference between the QT intervals estimated by U-Net and KanResWide

$$\epsilon_{1,2} = \frac{1}{5}\sum_{j=1}^{5}|QT_{1,j} - QT_{2,j}|$$

- the mean absolute difference between the QT intervals estimated by AttnCNN and KanResWide

$$\epsilon_{3,2} = \frac{1}{5}\sum_{j=1}^{5}|QT_{3,j} - QT_{2,j}|$$

## 2.5. Model building

**LASSO Logistic Regression:** We implemented a penalized logistic regression classifier with LASSO (L1 regularization), which allows to only keep the most relevant features amongst the 268 and avoid overfitting the model. The model was trained with Python's Scikit-learn library. The weights associated with each label were adjusted following their distribution in the training set (balanced mode).

**QTI definition:** As illustrated on Figure 1.B, in order to have a single prediction for each 12-lead ECG recording, the probabilities outputted by the logistic regression algorithm for each lead were averaged to give the QTI.

**Classification metrics:** The $F_2$ score was chosen as the main metric for this task.

$$F_2 = \frac{1+\beta^2}{\frac{\beta^2}{Recall} + \frac{1}{Precision}}, \beta = 2$$

Indeed, we aimed to reduce the number of false negatives $F_N$, i.e. the number of inaccurate QT measurements not reviewed by an expert, and therefore maximize the recall $T_P/(T_P + F_N)$ hence why it has stronger weight than the precision $T_P/(T_P + F_P)$.

**LASSO parameter tuning and QTI calibration:** The automatic QT measurements were classified based on a QTI threshold that we optimized along with the LASSO regularization strength ($\frac{1}{\lambda}$) by conducting a grid-search on the validation set. The parameters yielding the highest $F_2$ score were retained.

## 2.6. Evaluation of semi-automated TQT study

For the ECGRDVQ database, the automatic QT intervals with a QTI above threshold were replaced by the manual QT intervals. The placebo-corrected QTc changes estimated with this proposed semi-automated method were then compared to the manual and fully automated methods.

# 3. Results

## 3.1. Classification

The highest $F_2$ score was obtained on the validation set with a QTI threshold of 0.4 and $\lambda$ = 0.01. In Table 2, we reported the scores ($F_2$, recall and precision) obtained with these optimal parameters on both the validation and test sets. As the QTI threshold increases, the precision improves on the validation set while the recall decreases.

Table 2. Scores obtained on the validation and test sets after grid-search (QTI threshold = 0.4, $\lambda$ = 0.01).

| | $F_2$ | Recall | Precision |
|---|---|---|---|
| Validation | 0.798 | 0.953 | 0.484 |
| Test | 0.612 | 0.816 | 0.306 |
| Test (QTI thresh. = 0.5) | 0.639 | 0.776 | 0.375 |

LASSO selected the 21 most relevant features: 20 U-Net extracted ECG features and the highest U-Net CL.

## 3.2. Analysis of the TQT studies

Table 3 reports the percentage of inaccurate QT measurements computed with the predicted labels at different QTI thresholds and with the true labels (ground truth) for each drug in the ECGRDVQ database (Placebo, Dofetilide, Quinidine, Ranolazine and Verapamil). Except for Dofetilide, these drugs were not studied in the ECGDMMLD database. The drugs with a more significant QT prolonging effect have the highest percentage of ECGs to be reviewed (Dofetilide and Quinidine). Figure 2 shows that more inaccurate QT estimates are missed for the other 3 drugs (up to 88% for the placebo) but their difference from the manual intervals remains within acceptable ranges. The review of these ECGs give better estimates of the drug-induced QTc changes. For instance, with a QTI threshold of 0.5, the QTc prolongation (peak QTc change) computed with this semi-automated method and the manual method differ by 2.98 $\pm$ 1.82 ms on average across the 4 drugs compared to 24.1 $\pm$ 21.0 ms with the fully automated method. This is illustrated on Figure 3, which shows the time profiles obtained for the Dofetilide and Ranolazine studies with each method.

Table 3. Percentage of inaccurate QT measurements for the placebo and the 4 drugs in the ECGRDVQ database.

| | Placebo | Dofe. | Quin. | Rano. | Vera. | **Global ↓** |
|---|---|---|---|---|---|---|
| QTI threshold: 0.4 | 33% | 72% | 79% | 47% | 39% | **54%** |
| 0.5 | 19% | 63% | 70% | 30% | 26% | **41%** |
| Ground truth | 7% | 31% | 48% | 6% | 7% | **20%** |

# 4. Discussion and conclusion

The QTI, defined for each ECG recording as the mean probability of inaccuracy of the single-lead automatic QT measurements, allowed to automatically select ECGs to be manually reviewed by experts. The proposed approach is independent of the drug being assessed in the clinical study.

Different threshold settings could be implemented by either using the optimal QTI threshold found by grid-search

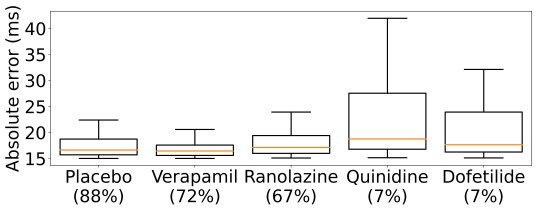


Figure 2. Difference between automatic QT intervals incorrectly labelled as accurate (QTI threshold = 0.5) and manual measurements.

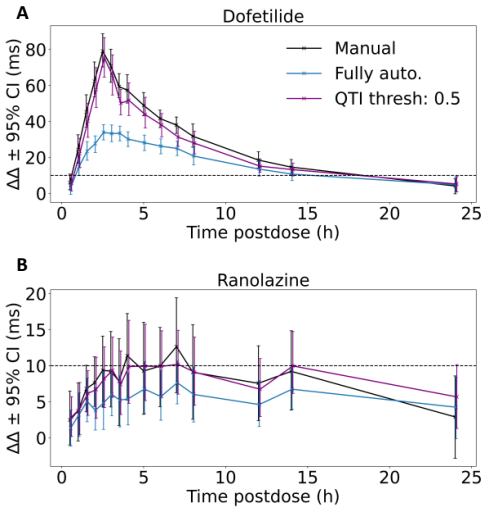


Figure 3. Placebo-corrected QTc changes estimated with 3 different QT measurement methods: manual, fully automated and semi-automated (QTI threshold = 0.5). The horizontal line represents the 10 ms regulatory threshold.

or by pre-defining a percentage of ECGs for manual review. The first approach could lead to higher burden on the cardiologist experts even though they could pass quickly over mislabeled QT measurements because the U-Net model provides delineation of the QRS onset and T offset. On the other hand, a too low pre-defined percentage of ECGs to be manually reviewed is time-saving but might lead to a less accurate TQT study analysis.

With a QTI of 0.5, the drugs with small to no QT prolonging effect in the ECGRDVQ database have ∼20-30% inaccurate QT measurements. Similar percentages were reported for expert review of manual QT measurements due to disagreement between the initial ECG readers [9]. For the drugs with a more significant effect, the percentage of inaccurate QT intervals is much higher. To lower this percentage, the performance of the CNN-based QT estimators should be improved on drug-induced T wave abnormalities. Indeed, automatic QT measurement is challenging in TQT studies mostly due to these abnormalities, which motivated us to build a QTI in the first place [3]. However, this correlation between QT inaccuracy and QT prolongation might lead to the selection of features heavily related to drug-induced morphology changes. Further investigation of the CNN-derived ECG characteristics selected by LASSO could help clarify this point.

The QTI is limited by its dependence to the 3 CNN-based QT estimators, although given the LASSO selected features it could be built using the U-Net model alone. But overall, it is a promising tool to conduct accurate and cost-effective AI-assisted TQT studies.

## References

[1] Panicker GK, Karnad DR, Kadam P, Badilini F, Damle A, Kothari S. Detecting moxifloxacin-induced QTc prolongation in thorough QT and early clinical phase studies using a highly automated ECG analysis approach. British Journal of Pharmacology 2016;173(8):1373–1380.

[2] Vaglio M, Isola L, Gates G, Badilini F. Use of ECG quality metrics in clinical trials. CinC Sep. 2010;173(8):505–508.

[3] Diaw MD, Papelier S, Durand-Salmon A, Felblinger J, Oster J. AI-Assisted QT Measurements for Highly Automated Drug Safety Studies. IEEE TBME Apr. 2022;Under review.

[4] Behar J, Oster J, Li Q, Clifford GD. ECG signal quality during arrhythmia and its application to false alarm reduction. IEEE TBME 2011;60(6):1660–1666.

[5] Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov PC, Mark R, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[6] Johannesen L, Vicente J, Mason JW, Erato C, Sanabria C, Waite-Labott K, Hong M, Lin J, Guo P, Mutlib A, et al. Late sodium current block for drug-induced long QT syndrome: results from a prospective clinical trial. Clinical Pharmacology Therapeutics 2016;99(2):214–223.

[7] Johannesen L, Vicente J, Mason JW, Sanabria C, Waite-Labott K, Hong M, Guo P, Lin J, Sørensen JS, Galeotti L, et al. Differentiating Drug-Induced Multichannel Block on the Electrocardiogram: Randomized Study of Dofetilide, Quinidine, Ranolazine, and Verapamil. Clinical Pharmacology Therapeutics 2014;96(5):549–558.

[8] Hicks SA, Isaksen JL, Thambawita V, Ghouse J, Ahlberg G, Linneberg A, Grarup N, Strümke I, Ellervik C, Olesen MS, et al. Explaining Deep Neural Networks for Knowledge Discovery in Electrocardiogram Analysis. Scientific reports 2021;11(1):1–11.

[9] Camm AJ, Yap YG, Malik M. Measurement of QT interval and repolarization assessment. Acquired long QT syndrome 2004;24–59.

Address for correspondence:

Julien Oster
Laboratoire IADI (Inserm U1254)
Bâtiment Recherche, CHRU Nancy-Brabois, Rue du Morvan
54500 Vandoeuvre, France
julien.oster@inserm.fr