The PhysioNet/Computers in Cardiology Challenge 2006: QT Interval Measurement [THIRD DRAFT]

GB Moody[1], H Koch[2], U Steinhoff[2]

[1]Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Physikalisch-Technische Bundesanstalt, Berlin, Germany

Abstract

Participants in the seventh annual PhysioNet/CinC Challenge developed and evaluated methods for measuring the QT interval, using the 549 records of the PTB Diagnostic ECG Database. Fifteen entrants entered sets of manually reviewed measurements, and the record-by-record medians of these defined the 549 "gold standard" reference QT measurements. Twenty-five entrants submitted sets of automatically-derived measurements. All entrants were allowed to omit records considered unreadable. Each entry received a score, calculated as the RMS error in milliseconds (relative to the reference QT measurements) divided by the fraction of records measured. The best scores for manual and automated entries were 6.67 ms and 16.34 ms respectively; typical scores were 10-20 ms for manual entries and 20-30 ms for automated entries. Significantly, a meta-entry derived from the medians of six automated entries achieved a score of 10.93 ms, better than all but four manual entries.

1. Introduction

Can the QT interval be measured by fully automated methods with an accuracy acceptable for clinical evaluations?

On first consideration, QT interval measurement in the ECG might seem a rather worked-out problem. In comparison with manual methods, automated methods offer advantages in terms of absolute repeatability of measurements, immunity from errors related to observer fatigue, lapses of attention, and transcription, as well as efficiency and cost considerations that permit either more extensive and rigorous testing for the same cost as manual methods, or more rapid testing at lower cost. An extensive literature documents a wide variety of approaches to the problem.

The recent adoption of ICH E14[1] by the US FDA, the EU's European Medicines Agency, Japan's National Institute of Health Services, and their counterparts in other nations, has drawn renewed attention to this question. ICH E14 is a set of guidelines for clinical evaluation of QT/QTc interval prolongation and proarrhythmic potential for non-antiarrhythmic drugs. Among much else, the guidelines endorse manual QT interval measurement for "thorough QT/QTc studies," and cite the need for further research before the use of fully automated methods can be accepted in these studies: "[T]he "thorough QT/QTc study" would warrant particularly careful attention to interval measurement. At present, this would usually involve the measurement by a few skilled readers (whether or not assisted by computer) operating from a centralized ECG laboratory. *If well-characterized data validating the use of fully-automated technologies become available, the recommendations in the guidance for the measurement of ECG intervals could be modified.*" [emphasis added]

Thus it is clear that regulatory agencies that have adopted ICH E14 are currently unconvinced of the reliability of automatic QT interval measurements. A major part of the motivation for this year's Challenge is to provide well-characterized data that might support modifications of the ICH E14 recommendations with respect to fully-automated methods.

The Challenge included separate divisions (and awards) for participants using manual and semi-automated methods (division 1), and fully automated methods (divisions 2 and 3). By comparing manually reviewed measurements with those obtained algorithmically from the same ECG recordings, we aimed to establish a firm basis for assessing the accuracy and reliability of fully automatic methods for QT interval measurement, as compared to the manual and semi-automated methods currently endorsed by the ICH E14 guidelines.

One of PhysioNet's major goals is to foster the creation and free dissemination of high-quality open-source software for research on clinically and scientifically interesting subjects[2]. Software contributed in the course of previous Challenges has stimulated new collaborations among its authors, and offers rare opportunities to compare the strengths of varied approaches objectively. Authors who submitted the source code for their fully automated algorithms were entered into division 3 (the "open source" division) of the Challenge.

2. Methods

The data used for this year's Challenge are the 549 recordings of the PTB Diagnostic ECG Database[3, 4], which was contributed to PhysioNet in September 2004 by its creators (Michael Oeff, Hans Koch, Ralf Bousseljot, and Dieter Kreiseler of the Physikalisch-Technische Bundesanstalt in Berlin). Each of these recordings contains 15 simultaneously recorded signals: the conventional 12 leads and the 3 Frank (XYZ) leads. Each of these is digitized at 1000 samples per second, with 16 bit resolution over a range of ±16.384 mV. The records come from 294 subjects (each represented by one to five records) with a broad range of ages and diagnoses. About 20% of the subjects are healthy controls. A detailed clinical summary accompanies each record. The records are typically about two minutes in length, with a small number of shorter records (none less than 30 seconds).

In November 2005, the Challenge was announced on PhysioNet. Participants were asked to choose the first representative beat (not ectopic, and free of significant noise, artifact, and baseline wander in lead II) in each of the 549 recordings; to measure to the nearest millisecond the times of the PQ junction (Q-wave onset) and the end of the T-wave in lead II; and to submit these measurements for scoring. The difference between a T-end time and the corresponding PQ time was taken to be the QT interval. The PQ and T-end times were not used independently in this study.

Entrants in division 1 were permited to omit measurements for as many as half of the 549 records, an allowance that we hoped would make the task seem less formidable to would-be participants. Entrants in the other two divisions were allowed to omit no more than 5% of the records (but the choice of which measurements to omit, if any, had to be made by software and not by manual review). An algorithm that can provide a highly reliable measurement for 95% of cases, and that alerts the user to inspect the others manually, is much more useful than one that measures all cases but delivers plausible but random results without warning for 5% of cases. By permitting a small number of measurements to be omitted, we hoped to encourage the development of useful rather than blindly optimistic software.

2.1. Reference QT intervals

During the development of the CSE Database, a "median self-centering approach" was followed to evaluate the performance of different algorithms for measuring ECG waveforms[5]. The methods used in this Challenge for bootstrapping the determination of the reference QT intervals, and for scoring the entries, were inspired by this important and closely related work.

As originally envisioned, the reference QT interval for each of the 549 records was to have been defined as the median of all valid QT measurements for that record. To avoid undue influence from multiple submissions by a single participant, only the most recent entry from each participant in each division was considered valid. We refer to the reference QT intervals calculated in this way as the "silver standard."

The original plan involving the silver standard assumed (incorrectly) that we would not be able to attract sufficient participation in division 1 to establish a "gold standard," defined as for the silver standard but based on manually reviewed measurements only. In a number of cases, however, academic researchers entering divisions 2 and 3 had teamed with clinicians entering division 1 in order to obtain a set of manually-reviewed QT measurements to support development of automated methods. By mid-July 2006, it was clear that we had more than enough division 1 entries to serve as a basis for a gold standard, and participants were notified that the reference measurements would be based on division 1 entries only. In retrospect, it is clear that the initial lack of any reference QT measurements in the Challenge database was a powerful stimulus to create division 1 entries.

2.2. Ranking the entries

Once a few entries had been received, we derived an interim silver standard and began providing feedback to participants by scoring their entries. *Raw scores* were determined by the square root of the sum of the squares of the differences between each QT measurement submitted and the corresponding reference QT (the RMS QT error, in milliseconds). The raw score divided by the *yield* (the fraction of records measured) defined the *normalized score* used for ranking entries.

Entrants were permitted to revise their entries, and most did so. As noted, each entry replaced any previous entry. Normalized scores were provided for each entry received, in batches at intervals up to several weeks in the early summer, and more frequently as the final deadline approached. Final scores were sent to all participants in early September, shortly after the final deadline.

2.3. GQT: A simple QT estimator

The first author submitted an unofficial entry in division 3 as a test of the scoring software. This entry used a very simple algorithm, GQT, to achieve a final score of 27.06 ms. Surrounding the time t_{qrs} of each detected beat, GQT defines points $t_a = t_{qrs} - 200$ ms and $t_b = t_{qrs} + 500$

ms (or 200 ms before the next t_{qrs}, whichever is earlier). The time of R-peak (or nadir), t_r, is taken to be the time between t_a and t_b for which the amplitude is most different from that at t_a. The time of the PQ junction is taken to be the time between t_a and t_r - 40 ms at which the amplitude is most different from a linear interpolation between the samples at t_a and t_r. GQT similarly finds the J-point (the end of the QRS) at the time t_j between $t_r + 40$ ms and $t_r + 120$ ms at which the amplitude is most different from a line connecting the samples at t_r and t_b. Using the same strategy, GQT finds the T-peak by searching between t_j and t_b; if the T amplitude (the difference in amplitude between t_j and the T-peak) is less than 100 μV, the search for the T-end is skipped. Otherwise, GQT finds the T-end by searching up to 200 ms following the T-peak for the amplitude most different from a line connecting the samples at the T-peak and t_b; since this estimate tends to be about 10 ms early compared with manual measurements, the estimated time of the T-end is set 10 ms after the point found in this way. Finally, any cases in which the QT interval is too short (less than 290 ms) or long (more than 475 ms) are rejected, since they are likely to be incorrect. (It is therefore important to examine all rejected cases, since any with pathologically short or long QT intervals will be among them.) For the purposes of this Challenge, this process is repeated for each beat, and applied separately to leads II, aVF, and V3. To choose a representative beat, GQT determines the median QT measurement from all of those measured (in all three leads), selects a beat having this measurement, and reports the PQ and T-end times for that beat.

The use of leads aVF and V3 in this process is motivated by the observation that even when the signal quality in lead II is poor, the typical QT interval in (an ideal) lead II can be predicted with reasonable accuracy from QT intervals measured in aVF and V3.

2.4. Meta-6: an even simpler algorithm

The large number of division 2 and 3 entries might suggest, as it did to us, that if varying methods work well, an algorithm drawing on the strengths of several good methods might work even better. With this in mind, algorithm "Meta-6" uses the outputs of the three best-performing algorithms in each of divisions 2 and 3, representing a diversity of successful techniques. "Meta-6" rejects any records rejected by more than one of these six base algorithms, and also rejects those for which the base algorithms disagree most; the QT intervals of the remaining records are estimated as the medians of the measurements obtained by the base algorithms.

3. Results

Fifteen participants submitted manually-reviewed entries. Although they were required to measure only half of the 549 records, all but three measured 95% or more; as a consequence, at least nine and usually twelve or more pairs of manually-reviewed PQ and T-end measurements were available for all but one of the records, and these were used to obtain the "gold standard" reference QT intervals. The lone exception was record "patient285/s0544_re", which did not contain recognizable ECG signals.

The Challenge attracted 28 participants in divisions 2 and 3, four of whom also entered division 1, and 24 of whom present papers in this volume describing their methods.

Division 1: The best score achieved by a manually-reviewed entry was 6.67 ms (raw score: 6.65 ms, yield: 0.998), submitted by Mariano Llamedo Soria of the Universidad Tecnologica Nacional FRBA, Buenos Aires, Argentina. In all, seven entries in division 1 received scores below 20 ms.

Division 2: The best score achieved by an automated method was 16.34 ms (raw score: 15.53 ms, yield: 0.951), submitted by Dieter Hayn of ARC Seibersdorf Research GmbH, Graz, Austria, who describes his approach elsewhere in this volume. This entry achieved seventh place overall.

Division 3: The best score achieved by an open source automated method was 17.33 ms (raw score: 17.30 ms, yield: 0.998), by Yuriy Chesnokov of the Unilever Centre for Molecular Science Informatics, Cambridge University, whose paper describing his approach also appears in this volume. This entry achieved ninth place overall, and second among all automated methods.

In all, three entrants in divisions 2 and 3 earned scores below 20 ms, and six more received scores below 30 ms. Additional scores are posted on the Challenge web site (http://physionet.org/challenge/2006/).

The "Meta-6" algorithm achieved a score of 10.93 ms (raw score: 10.39, yield: 0.951), which would have earned it fifth place overall had it been an official entry, by combining results from the winners of divisions 2 and 3 with those of Juan Pablo Martinez (University of Zaragoza, Spain), Joel Xue (GE Healthcare, Milwaukee, USA), Ivaylo Christov (Centre of Biomedical Engineering, Bulgarian Academy of Sciences, Sofia, Bulgaria) and algorithm GQT (described above).

4. Discussion and conclusions

The large number of participants (39 in all, from 17 nations) demonstrated a high level of interest in this problem, and the diversity of the successful automated QT measurement methods suggests that many roads lead to good QT interval estimates.

Previous studies have shown that inter-observer variability in estimates of QT intervals, even between human experts, can be large[6]. The central issue in "thorough QT/QTc studies," however, is detection and measurement of *changes* in repolarization in response to a drug, where consistency (intra-observer variability) is paramount. The immunity of automated methods to fatigue, attention lapses, and transcription errors may be significant advantages relative to manual methods. Future investigations making use of ECGs from drug studies and examining QTc (for which errors might be distributed differently than QT) would be valuable complements to the Challenge.

The asymmetry of the process of developing the "gold standard," which is determined entirely by manual measurements, might be avoided, permitting an assessment of automated methods that does not proceed from an a priori assumption of the superiority of manual measurements. For example, we might weight each entry's influence on a new standard by a function of its score determined using the existing standard, iterating until the weights stabilize.

The current study supports the conclusion that manual QT interval measurements may have a root mean squared error below 10 ms, but only two of fifteen manually reviewed entries achieved this level of accuracy, which was roughly twice as accurate as the best of the 28 automated methods entered in the Challenge. At least three automated methods demonstrated better accuracy than the majority of manually reviewed entries, however, so it is also clear that a well-designed algorithm is likely to produce results comparable to those that can be expected from manually reviewed measurements.

Significantly, the "Meta-6" results demonstrate that the diverse approaches employed by the best of the automated methods are to a useful degree complementary, by achieving an accuracy much better than any individual division 2 or 3 entry, better than 70% of the division 1 entries, and nearly matching the best of the manually reviewed entries. This result points the way to significant improvements in automated QT interval estimation by exploiting the diverse strengths of the constituent algorithms.

Independent of these considerations, the QT interval measurements obtained as a product of this Challenge will support further research aimed at designing robust measures of repolarization characteristics. The difficulty of making accurate measurements of QT intervals on individual beats is a hindrance to studies of variability of repolarization (for example, in exercise). By creating a collection of reference QT intervals that have been measured by many expert observers and automated methods, the Challenge can be a starting point for future studies of alternative measurements that may be possible to derive reliably even when QT intervals cannot be determined directly.

Acknowledgements

PhysioNet/Computers in Cardiology Challenges are conducted using the facilities of PhysioNet, a public service of the Research Resource for Complex Physiologic Signals, which is supported by a grant from the National Center for Research Resources of the US National Institutes of Health (P41 RR13622). The Challenge awards are funded by Computers in Cardiology. Finally, we thank all of the Challenge participants, many of whom describe their approaches in this volume.

References

[1] ICH. E14 Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Non-Antiarrhythmic Drugs. 2005. http://www.fda.gov/cber/gdlns/iche14qtc.htm.

[2] 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 (June 13);101(23):e215–e220. Circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215.

[3] Bousseljot R, Kreiseler D, Schnabel A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomedizinische Technik 1995;40 Ergänzungsband I:S 317.

[4] Kreiseler D, Bousseljot R. Automatisierte EKG-Auswertung mit Hilfe der EKG-Signaldatenbank CARDIODAT der PTB. Biomedizinische Technik 1995;40 Ergänzungsband I:S 319.

[5] Willems J, Arnaud P, van Bemmel J, Bourdillon P, Brohet C, Dalla Volta S, Andersen J, Degani R, Denis B, Demeester M, et al. Assessment of the performance of electrocardiographic computer programs with the use of a reference data base. Circulation 1985;71(3):523–534.

[6] McLaughlin NB, Campbell RW, Murray A. Accuracy of four automatic QT measurement techniques in cardiac patients and healthy subjects. Heart 1996;76:422–426.

See http://physionet.org/challenge/2006/ for access to the PTB database, the QT measurements, the sources for the division 3 entries, and additional information about the Challenge.

Address for correspondence:

George B. Moody
MIT Room E25-505A, Cambridge, MA 02139 USA
george@mit.edu