Z-Score for Benchmarking Reader Competence in a Central ECG Laboratory

Gopi Krishna Panicker, B.H.M.S., P.G.D.C.R., Dilip R. Karnad, M.D.,
Rajesh Joshi, M.Sc., Sheetal Shetty, B.H.M.S., P.G.D.C.R., Niraj Vyas, M.B.B.S.,
P.G.D.C.R., Snehal Kothari, M.D., D.M., and Dhiraj Narula, M.D., D.M., M.R.C.P.

From the Quintiles ECG Services, Mumbai, India

Background: ECGs from thorough QT studies must be read in a central laboratory by trained experts. Standards of expertise are not presently defined. We, therefore, studied the use of Z-scores to define reader competence.

Methods: Two hundred ECGs were read by 24 experts and the mean and standard deviation (SD) of QT measurements calculated for each ECG. Z-scores $(|QT_{\text{reader}} - \text{mean } QT_{\text{experts}}| / SD_{\text{experts}})$ for each ECG and mean of absolute Z-scores of all ECGs read by a reader were calculated. The highest mean absolute Z-score of experts was considered the cutoff to define competence. Hundred of these standardized ECGs were used to assess performance of readers from the central laboratory.

Results: All experts had mean absolute Z-scores ≤ 1.5 . Using this cutoff, one of 28 experienced readers and 7 of 15 trainees had unacceptable Z-scores. After re-training, all achieved Z-scores < 1.5 . Comparing histograms of actual Z-scores of the 100 ECGs of readers with unacceptable scores with that of the reader with the best Z-score showed two patterns. Readers with histograms having a peak and tails similar to that of the best reader, but with leftward or rightward shift, consistently made shorter or longer QT measurements, respectively. A histogram with a flatter peak and wider tails, suggested that measurements were long in some ECGs and short in others.

Conclusion: Mean absolute Z-score is useful to assess competence for measuring the QT interval on ECGs. Analysis of histograms can pinpoint problems in QT measurements.

Ann Noninvasive Electrocardiol 2009;14(1):19–25

thorough QT/QTc study; electrocardiography; reader variability; quality assurance; phase I studies;
drug-induced QT prolongation

The International Conference on Harmonization (ICH) E14 guidelines recommend that all new drugs with systemic bioavailability should be subjected to a "thorough QT/QTc study."¹ The guidelines specify that any drug that prolongs the mean QTc interval by ≥ 5 ms with an upper one-sided 95% confidence bound of ≥ 10 ms would become a regulatory concern and would be subjected to stringent ECG analysis in subsequent clinical studies. The extent of QT prolongation to be detected in a thorough QT/QTc study is extremely small. A standard 12-lead ECG used in clinical practice is usually recorded at 25 mm/s. Thus, a 5 ms change would correspond to one-eighth of a millimetre on

the standard ECG. In order to achieve this high precision, the E14 guidelines stipulate the standards of ECG acquisition, collection, storage and interpretation of the ECG.¹

In a recent study, Viskin et al. found that while 80% of cardiologists specialized in cardiac electrophysiology were able to differentiate between normal ECGs and those with long QT syndrome, only 40% of noncardiologist physicians and 50% of cardiologists could do so.^{2,3} This is because although the onset of the QRS complex is sharp and well defined, the end of the T wave is not always clearly defined and usually merges gradually with the baseline.⁴ Thus, identification of the end of the

Address for reprints: Dr. Gopi Krishna Panicker, B.H.M.S., P.G.D.C.R., Quintiles ECG Services, 502 A, Leela Business Park, M.V. Road, Andheri (East), Mumbai 400059, India. Fax: +91 22 66950159; E-mail: gopi.panicker@quintiles.com

Conflict of Interest: Employment: Gopi Krishna Panicker, Rajesh Joshi, Sheetal Shetty, Niraj Vyas, and Snehal Kothari are employees of Quintiles ECG Services, Mumbai. Consultant or Advisory Role: Dilip Karnad and Dhiraj Narula. Stock Ownership: None. Honoraria: None. Research Funding: None. Expert Testimony: None. Other Remuneration: None.

©2009, Copyright the Authors
Journal compilation ©2009, Wiley Periodicals, Inc.

T wave is subject to considerable variability.²⁻⁴ The E14 guidelines, therefore, recommend that ECGs must be read in a central ECG laboratory by a few skilled readers.¹ However, the guidelines raise several unanswered questions. Who can be termed a "skilled reader?" What is an acceptable level of accuracy that a "skilled reader" must achieve?

The accuracy of a reader could be calculated from the difference between the QT measurements made in a set of ECGs by a reader and the gold standard QT values. However, the problem is that there cannot be a gold standard for the QT interval on an ECG since it is a subjective measurement. In the absence of a gold standard, the average of measurements made by a group of experienced readers is usually assumed to be the gold standard. If the standard deviation of the readings made by the experts is known, the performance of a particular reader can then be quantified with respect to that of the experts using the standardized normal deviate or the Z-score.⁵ We, therefore, conducted this study to use the Z-score to assess performance of readers in a central ECG lab, to define what an unacceptable Z-score is, and to see if we could use these criteria to evaluate new readers before they were allowed to read ECGs from research studies.

METHODS

The study was conducted in the central ECG laboratory of Quintiles ECG Services, Mumbai, India, which serves as a centralized ECG laboratory for many Phase I to Phase III studies.

Study Group

The readers were grouped as expert readers, experienced readers and trainee readers.

Expert readers (n = 24) were physicians with a postgraduate degree in internal medicine with special training in reading ECGs or Board certified cardiologists who had done a minimum of 3 years of cardiology fellowship after obtaining the MD degree in internal medicine, were certified as specialist cardiologists by an Indian University and were Fellows of the American College of Cardiology. These physicians had experience of 3–20 years in managing patients with cardiac disorders including arrhythmias. In addition, each of them had read more than 2000 digital ECGs from drug studies on a digital platform.

Experienced readers (n = 28) were physicians with a graduate or postgraduate degree in medicine

with special training in reading ECGs and had been reading digital ECGs from drug studies for at least one year in the central ECG laboratory. Trainee readers (n = 15) were physicians with a graduate or postgraduate degree in medicine newly recruited for reading ECGs. All trainee readers joining the laboratory underwent a structured training program where they were trained to read ECGs using the computer software. After the initial training, they read a minimum of 500 ECGs, as specified by the American College of Cardiology/American Heart Association guidelines for defining clinical competence in electrocardiography.⁶

ECG RECORDING

Electrocardiographs used in this study were recorded using a digital ECG machine (Model Eli 250, Mortara Instrument Inc, Milwaukee, WI, USA) with a data acquisition rate of 1000 samples per second. The electrocardiograms were directly transferred electronically to the computer system in a FDA compliant XML (extensible markup language) file format using proprietary software provided by the manufacturer. Once the XML file was created, all the intervals were measured using digital on-screen calipers to manually place annotations on the ECG using CalECG Software version 1.3 (AMPS LLC, New York, NY, USA). A total of 200 ECGs from 200 different subjects were randomly selected from the ECG database of the Quintiles ECG laboratory.

ECG Reading

Throughout this study, placements of annotations on ECGs were done manually without any prior computer-determined placement of fiducial marks. The RR and QT intervals were measured on five consecutive beats in a single lead (typically lead II) and the mean RR and QT intervals from the 5 complexes were calculated and entered into the database. The start of the QT interval was defined as the first deflection of the QRS complex. The end of the QT interval was defined as the intersection of the descending part of the T wave (positive T wave) with the isoelectric line (threshold method). If the U wave interrupted the T wave before it returned to baseline, the QT interval was measured as the nadir between T and U waves. As is normal practice in the Quintiles ECG central laboratory, QT and RR intervals were measured in lead II. If the tracing in this lead was unsuitable, the order of

preference of alternative leads used for measurements was lead V_5 followed by V_4 and V_3 . A lead was rejected if muscle artifacts or electrical interference obscured the QRS complex or T waves, or the signal to noise ratio was >0.5 for the QRS complex. The choice of complexes for measurement in the select lead was left to the discretion of the readers.

Standardization of ECGs

The expert readers read each of the 200 test ECGs. The mean and SD of the QT measurements made by the 24 readers was calculated for each ECG. The mean QT of the 24 measurements for each ECG was considered as the gold standard QT for that ECG. These ECGs were to be used as standardized ECGs for Z-score evaluation. Once the set of standardized ECGs was developed, these ECGs were used to test reader competence.

Testing Competence Using the Z-Score

The experienced readers were assigned a set of 100 ECGs, which were randomly selected from the 200 standardized ECGs, for measurement of QT and RR intervals. Another set of 100 randomly selected standardized ECGs was assigned to the trainee readers.

The Z-score is a standardized score expressing the difference of an individual raw score with the population mean, using the standard deviation as a unit.⁵ The Z-score for each ECG read by the reader was calculated as

$$Z_{\text{ECG}} = \frac{QT_{\text{reader}} - \text{mean}QT_{\text{expert}}}{SD_{\text{expert}}}.$$

The mean absolute Z-score for each individual reader was calculated as the average of the Z-score for all ECGs with positive and negative signs ignored:

$$Z_{\text{mean abs}} = \frac{Z_{\text{ECG1}} + Z_{\text{ECG2}} + \dots + Z_{\text{ECGn}}}{n}.$$

Mean absolute Z-scores were calculated for each of the expert readers, and for readers who were to be tested for competence. The highest mean absolute Z-score from the expert group was used as the cutoff value to define competence. Those with a Z-score below the cutoff value were considered competent to read ECGs from clinical trials.

RESULTS

The twenty-four expert readers read a set of 200 ECGs to create the database of standardized ECGs. Of these 200 ECGs, the lowest mean of QT interval was 315.8 ms and the maximum was 462.8 ms and the mean QT of the 200 ECGs was 389.6 ms. QTcB values ranged from 312 to 506 ms and QTcF from 315 to 509 ms. The standard deviations of QT measurements of the 24 experts for the 200 ECGs ranged from 0.7 ms to 36.8 ms, with a mean of 5.4 ms. The mean absolute Z-scores of each of the expert readers was calculated. The Z-scores of all experts was less than 1.5 (Fig. 1A). We, therefore, decided that this value would be the appropriate cutoff for an acceptable Z-score to define a competent ECG reader.

Use of Z-Scores to Assess Performance of All Readers in the Laboratory

A set of 100 ECGs was randomly selected from the 200 standardized ECGs to assess the performance of 28 readers from the ECG laboratory. The mean absolute Z-score for each of the 28 readers was obtained. The difference in milliseconds between the measured QT values and the gold standard QT for all ECG read by each reader were also recorded. Since the mean absolute Z-score was an average of the absolute Z-scores of each ECG, the distribution was positively skewed with a peak between 0.62 and 0.75 (Fig. 1B). Of the 28 readers, only one reader had a mean absolute Z-score greater than 1.5.

Use of Z-Scores to Assess Performance of New Readers

During a 1 year period, 15 new readers were selected to join the laboratory. The distribution of the mean absolute Z-scores of these readers were widely scattered when compared to those of experienced readers (Fig. 1C). Only eight of the 15 readers had acceptable Z-scores of ≤ 1.5 after initial training. The seven readers with Z-scores > 1.5 underwent further training. After retraining, 5 readers achieved Z-scores of ≤ 1.5 . Two readers (Reader A and Reader B) required further training before they could achieve acceptable Z-scores.

To further evaluate the performance of readers with unacceptable Z-scores, we plotted the histograms of the actual Z-scores for the 100 ECGs

Figure 1. Distribution of mean absolute Z-scores of QT interval measurements by readers (A) Z-scores of expert readers (n = 24) whose measurements were used to develop a cutoff value for an acceptable Z-score. Based on their performance, a Z-score of ≤ 1.5 was considered to be acceptable. (B) Z-scores of trained readers (n = 28) working in the laboratory—one reader had Z score > 1.5 . (C) Z-scores of new readers (n = 15) while undergoing training were widely scattered and only 8 readers had Z-scores ≤ 1.5 .

for each of these readers and of the reader with the best Z-score (Fig. 2). The histogram of the best reader had a peak at 0 and relatively short symmetrical tails. The histogram for Reader A (Z-score =

Figure 2. Histogram of Z-scores of ECGs read by a reader with an unacceptable mean absolute Z-score and the reader with the best mean absolute Z-score. (A) The histogram of the reader (Reader D) with the best Z score (grey bars) and that of a reader (Reader A) with an unacceptably high Z-score (black bars). Reader A's histogram had a similar shape to that of Reader D, but was shifted to the left (negative Z-score). Reader D measured the QT interval close to the mean QT while Reader A consistently measured the QT interval to be shorter than the mean QT. (B) The histogram of Reader D and that of another reader (Reader B) with an unacceptably high Z score (black bars). Reader B's histogram was flatter and had a greater spread than that of Reader D. Reader B was inconsistent and measured the QT interval to be longer or shorter than the mean QT.

1.52) had a similar peak and symmetrical tails, but the entire histogram was shifted to the left so that the peak was at -1.0 (Fig. 2) suggesting that the QT measurements were consistently short compared to the gold standard. In contrast, the histogram of Reader B (Z-score = 1.59) had a shorter peak at 1.5 and wide tails on either side (Fig. 2), indicating excessively long QT measurements in some ECGs and excessively short QT measurements in others.

Figure 3. The effect of ECG quality on inter-reader variability of QT interval measurement and the standard deviation of measurements made by a group of experts. Note that the reader's QT measurement is 10 ms shorter than the mean QT of the expert group for both ECGs. However, the ECG in Panel A had a clear T wave offset and hence a smaller standard deviation than the ECG in Panel B, which had a larger standard deviation. Therefore, a 10 ms difference was unacceptable for Panel A (Z-score = 3.33) but was acceptable for Panel B (Z-score = 1.43).

DISCUSSION

The Z-score is probably best suited for assessing performance of readers in a central ECG laboratory for several reasons. First, there is no gold standard for the annotations placed on an ECG tracing to measure the QT interval.² It is therefore logical to compare their annotations with those placed by persons with acceptable levels of competence. Z-scores are commonly used for interpretation of psychological tests or bone densitometry data where again there is no gold standard.^{5,7,8} Another reason which makes the Z-score an attractive method for application to ECG readers is because limits of acceptability (difference between the reader's measurement and the mean QT measurements of experts) are not fixed and may differ from one ECG to another based on how clearly the end of

the T wave is demarcated (Fig. 3).⁹ In the present study, we used the Z-score to evaluate the accuracy of the QT measurements made by ECG readers. In psychological testing, performance of a subject on many test items has to be combined to obtain a composite evaluation.⁷ Likewise, the mean absolute Z-score was calculated for each reader as the average of the absolute Z-scores (Z-scores without the positive or negative sign) of all ECGs read, since an excessively long QT measurement was as unacceptable as an excessively short measurement. A low mean absolute Z-score therefore implies that QT measurements were close to the mean of the experts; a high score indicates large differences between measurements by the reader and the experts.

In order to assign Z-scores, we first had to have a database of ECGs which had been read by many

experts, so that each ECG was standardized (i.e., had a known mean and standard deviation of QT measurements placed by experts). ECGs from this database could be randomly selected to test individual readers. A set of 200 standardized ECGs was developed. We, then, had to define a cutoff value of the mean absolute Z-score, which would define an acceptable level of competence. We found that none of the expert readers had a Z-score of >1.5 . We, therefore, selected 1.5 as the cutoff value; only readers with a mean absolute Z-score of ≤ 1.5 would be considered competent. While this cutoff value is quite arbitrary, we reasoned that each reader in the laboratory should be as accurate in their QT measurements as the experts; hence the value of 1.5 was selected.

On evaluation of the readers already working in the centralized ECG laboratory, only one of 28 readers had a Z-score greater than 1.5. This reader was identified for retraining. However, the real value of this method was revealed when 15 new recruits were subjected to this test. All new readers in our laboratory are subjected to a six week training program. They also read 500 ECGs, as specified by the American College of Cardiology/American Heart Association guidelines for defining clinical competence in electrocardiography.⁶ Despite this, 7 of the 15 readers had Z-scores above 1.5 and required further training. Two readers required even more extensive training before they could achieve an acceptable Z-score.

Plotting the histogram of the Z-scores of the 100 ECGs of a reader with a score >1.5 along with that of reader with the best Z-score proved to be extremely useful to identify the reason why the reader had a high Z-score. One reader was consistently measuring the QT intervals shorter than the mean QT. The histogram of this reader showed a tall peak and small tails which were similar to those of the best reader (Fig. 2A), but the entire histogram was shifted to the left (toward negative values of Z). On the other hand, another new reader had a histogram with a shorter peak, and the tails were spread wider than that of the best reader (Fig. 2B). This reader was inconsistent in measuring QT intervals and the measurements were significantly higher or lower than the mean QT. Identifying these patterns greatly facilitated retraining of readers with a high mean absolute Z-score.

Ideally, the Z-score is used to compare the performance of an individual against that of a population.¹⁰ For ECG annotations, the population would

consist of expert trained cardiologists in central ECG laboratories from all over the world. However, this being a formidable proposition, we used the group of cardiologists associated with our central laboratory to create a set of standardized ECGs. Since the purpose of this study was to devise a method of defining competence of ECG readers for our central laboratory, this choice of experts is justifiable. However, if our results are to be generalized to other centres or laboratories, it would be necessary to involve expert readers from multiple global locations in the process of standardization of ECGs.

CONCLUSION

As more new drugs are subjected to thorough QT studies, there is a need for central ECG laboratories to expand ECG reading capabilities by adding new readers. Our experience suggests that without stringent training, new readers can increase variability in QT interval measurement. There are no standard methods to benchmark performance of individual readers in a central ECG laboratory, nor are there specified levels of competence or accuracy to define trained readers. We, therefore, believe that there is a need for an easy, reliable and robust method of defining reader competence. The Z-score method meets this requirement of benchmarking reader competence, provided a set of standardized ECGs is painstakingly developed. Analysis of histograms of the Z-scores of a reader with an unacceptable mean absolute Z-score can pin-point problems in QT measurements.

REFERENCES

1. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. The clinical evaluation of QT/QTc interval prolongation and proarrhythmic potential for non-antiarrhythmic drugs: E14. International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, Geneva, Switzerland, 2005. Available at <http://www.ich.org/LOB/media/MEDIA1476.pdf>.
2. Viskin S, Rosovski U, Sands AJ, et al. Inaccurate electrocardiographic interpretation of long QT: The majority of physicians cannot recognize a long QT when they see one. *Heart Rhythm* 2005.
3. Bai R, Yan GX. Accurate interpretation of the QT interval: A vital task that remains unaccomplished. *Heart Rhythm* 2005;2:575-577.
4. Malik M, Batchvarov VN. Measurement, interpretation and clinical potential of QT dispersion. *J Am Coll Cardiol* 2000;36:1749-1766.
5. Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research*, 4th Edition, Malden, MA, Blackwell, 2002, p. 78.

6. de Groot JC, de Leeuw F-E, Oudkerk M, et al. Cerebral white matter lesions and cognitive function: The Rotterdam Scan Study. *Ann Neurol* 2000;47:145–151.
7. Ellis KJ, Shypailo RJ, Hardin DS, et al. Z score prediction model for assessment of bone mineral content in pediatric diseases. *J Bone Miner Res* 2001;16:1658–1664.
8. Shetty S, Khan M, Salvi S, et al. Do ECG characteristics predict variability in QT measurements in clinical trials? *Indian Heart J* 2005;57:302.
9. Kadish AH, Buxton AE, Kennedy HL, et al. ACC/AHA clinical competence statement on electrocardiography and ambulatory electrocardiography: A report of the American College of Cardiology/American Heart Association/American College of Physicians—American Society of Internal Medicine Task Force on Clinical Competence (ACC/AHA Committee to Develop a Clinical Competence Statement on Electrocardiography and Ambulatory Electrocardiography). *J Am Coll Cardiol* 2001;38:2091–100.
10. Crawford JR, Howell DC. Comparing an individual's test scores against norms derived from small samples. *Clin Neuropsychol* 1998;12:482–486.