Drug Saf (2013) 36:389–392
DOI 10.1007/s40264-013-0056-4

COMMENTARY

Choice of Baseline in Parallel Thorough QT Studies

Claus Graff

Published online: 26 April 2013

Although the ICH E14 Guidance [1] was adopted in 2005 there is still much debate on the appropriate baseline needed to adjust for the variable nature of the QT interval and its heart rate corrected value (QTc) in the parallel Thorough QT (TQT) study. Numerous reports have documented the influence of circadian rhythm [2], food ingestion [3], sleep [4], and autonomic tone [5] on the QT/QTc interval. In the absence of baseline measurements, these factors can make it difficult to interpret QT effects of the study drug.

In this issue of *Drug Safety* Dr. Zhang and colleagues [6] advocate the use of a time-matched baseline in parallel TQT studies to maximize the precision and accuracy of point estimates for QTc effects. Their recommendation is based on analyses of commonly used baseline correction methods using a large data set of parallel TQT studies submitted to the FDA for statistical review by sponsors.

One assumption in the choice of baseline is that it should have no influence on the magnitude of QT effects (high accuracy) and it should minimize the uncertainty of the effects (high precision). Both accuracy and precision are important for the interpretation of a TQT trial. As a result, there is much focus on the implications of using different baseline adjustments in the parallel TQT study and understanding the role each baseline definition can have in determining the outcome of the trial as either positive or negative.

The distinction between a positive or negative TQT study is based on the upper bound of the 95 % one-sided confidence interval (CI) for the largest time-matched mean difference in QTc between drug and placebo (baseline adjusted). If these (double-delta) ΔΔQTc values exclude 10 ms at all study times the result is a negative TQT study [1]. Intensive ECG monitoring is then hardly ever required in subsequent trials (phase II/III). Conversely, drugs for which an effect exceeding 10 ms cannot be excluded almost always require additional QT interval monitoring in target patients including dose-concentration effects, outlier analysis, changes in mean QTc values, analysis of QTc in subgroups of interest, and identification of individuals who develop a markedly prolonged QTc.

The need for additional ECG monitoring in late phase trials can add millions to drug development and may ultimately pose financial limitations on the number of medications which can be developed [7]. The importance of correctly adjusting for baseline in the TQT study is therefore evident.

1 Adjusting for Baseline in the Parallel TQT Study

A commonly used baseline for each study arm is the time-matched baseline. In this design, QT measurements are taken at exactly the same time-points on the day prior to the beginning of treatment as on the treatment day and then subtracted from all post-dose values. The assumption is that the within-group diurnal patterns in the placebo and active treatment groups are stable and will be adjusted when calculating the placebo adjusted change from baseline (double delta) ΔΔQTc for the study drug.

Alternatively a time-averaged baseline can be used for each study arm where all baseline values (usually recorded at time points matching the on-treatment recordings for the treatment arm) are averaged to give a single baseline value

C. Graff (✉)
Medical Informatics Group (MI), Department of Health Science and Technology, Aalborg University,
Fredrik Bajers Vej 7 C1-202, 9220 Aalborg, Denmark
e-mail: cgraff@hst.aau.dk

which is subtracted from all post-dose values. With this approach it is assumed that the between-group diurnal patterns in the placebo and active treatment groups are similar and will be adjusted when calculating ΔΔQTc for the study drug.

Although not recommended for the parallel design, a pre-dose baseline can also be calculated for each study arm (usually as the average value of baseline QT measurements obtained at multiple time points prior to dosing on the day of treatment) to give a single baseline value for each study arm which is subtracted from all post-dose values. The assumption is that between-group diurnal patterns are similar.

The current recommendation (per the ICH E14 Guidance [1]) for baseline measurements of QT intervals in parallel TQT studies is to obtain one full day of measurements per study arm on the day before dosing at time points matching on-treatment recordings. Within this framework, both time-matched and time-averaged baselines are valid designs.

One rationale often used to justify the time-matched baseline design is to correct for circadian rhythms in QT intervals and eliminate the potential bias. However, contrary to popular belief, a time-averaged baseline can provide similar diurnal mean effect correction for treatment comparisons [8]. In a simulation study of type I error rates (false negative TQT) and statistical power [9] the time-averaged baseline had the highest power in comparison with the time-matched and pre-dose average designs. It was even concluded that a time-averaged design may be suitable for all TQT trials. When six statistical models were compared for both the primary hypothesis and the assay sensitivity test it was found that an ANCOVA model using the time-averaged baseline should be preferred in general, unless regulatory authorities mandate the use of a time-matched baseline [10]. In contrast, a time-matched baseline is recommended by Zhang et al. [6] and recently it was demonstrated that an ANCOVA model with change from time-matched baseline as the outcome and both the time-matched and time-averaged baselines as covariates was more efficient and robust compared to use of either baseline alone [11]. It is clear we do not have a generally accepted definition of baseline in the TQT study.

2 Dealing with Baseline Imbalance in QT Intervals between Groups

It is also clear that imbalance in the baseline QT/QTc between groups confounds the assessment of treatment effects due to regression towards the mean. An imbalance in mean QT values cannot be ignored when analyzing change from baseline in the TQT study, especially in the parallel group design where two treatment groups may have different baseline values purely by chance. Some investigators believe they can allow for imbalance between groups when change from baseline (ΔQT/ΔQTc) is used but the differential effects of regression towards the mean between groups cannot be cancelled out by change scores. Ignoring baseline imbalance in the analysis will lead to bias in the estimates of QT effects. For example, it was observed that subjects with high QTc values on baseline appear to have a smaller change from baseline when switched to sertindole treatment while subjects with a low baseline value have a higher change from baseline [12]. This observation is merely the result of regression towards the mean.

Randomization will on average produce groups that are comparable in terms of baseline characteristics. Still, baseline imbalance will occur by chance in TQT studies and it is inevitable that differences in the mean QT interval may exist between groups. In this case a simple unpaired t-test or ANOVA analysis based on post-treatment values or change from baseline would possibly fail to detect a treatment difference or even conclude a difference in the wrong direction. In general, a conditional test (ANCOVA) which takes account of the actual observed imbalance should be used.

However, the ICH E14 document [1] does not provide unequivocal guidance on the preferred method for dealing with potential baseline imbalance. Neither does the October 2012 Guidance for Industry questions and answers document [13] in which it is simply stated that baseline data should be taken into account in the statistical analysis without specifying a preferred method.

3 Influence of Baseline on the Accuracy and Precision of QT Effects

Maximizing the precision of the QTc effects is imperative because any increase in variability will result in a wider confidence interval and increased risk of type II error (false positive TQT). It has already been established theoretically that the time-averaged baseline method has the highest precision for QTc effects [8] and this method was also the most precise in the analysis by Zhang et al. [6] In other words, when conditions change between baseline and treatment in a TQT study, the time-averaged method will have more narrow confidence bands for QT estimates compared to the time-matched method.

Several factors may change during the course of a study, especially in the parallel design, and we do not fully understand their influence on the precision of QT estimates. For example, relatively little research has addressed the stability of diurnal patterns in individual subjects across

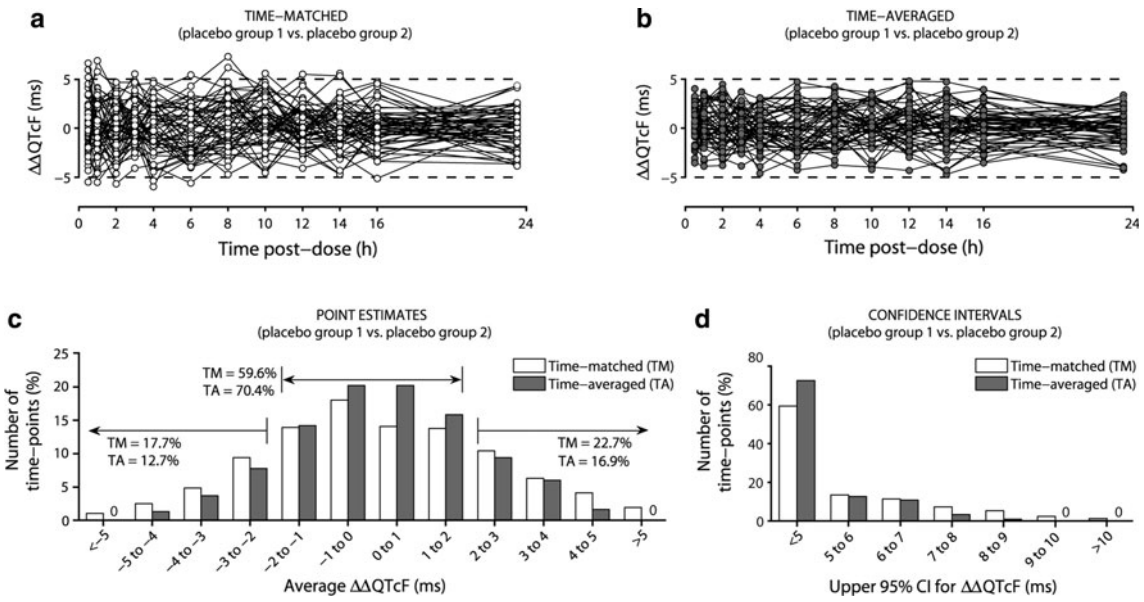

Fig. 1 Difference in mean Fridericia corrected QT effects (QTcF) between two populations of equal size randomly sampled 57 times (number of studies analyzed by Zhang et al. [6]) from the placebo arm of a single parallel TQT study with 62 subjects in the arm. **a** Using the time-matched baseline method. **b** Using the time-averaged baseline method. **c** Distribution of point estimates for QTcF effects. **d** Magnitude of upper 95 % one-sided confidence limits for point estimates of QTcF effects

contiguous days. The conventional belief (and an inherent assumption in the time-matched design) is that there is no sequence effect, in other words, that the diurnal pattern is stable. However, this condition is not met in TQT studies. Individual subjects can show substantial variability across study days [14] and the variability of mean QTc effects is increased when the time between baseline and on-treatment measurements increases [15]. Systematic decrements in mean QTc values can also be observed between two days, even in periods preceding drug treatment [14]. The opposite has also been found, an apparent increase in the magnitude of QTc change from baseline with longer interval between the baseline and on-drug measurements [15].

Clearly, the stability of diurnal patterns cannot be assessed by observing QTc fluctuations on a single baseline day, nor by comparison of QTc values between groups on baseline, both central analyses in the Zhang et al. study [6]. Moreover, the implications of such comparisons for the primary hypothesis in a TQT study are not obvious. It is also worth remembering that if 57 independent hypotheses are tested (the number of TQT trials in the Zhang et al. study [6]) at a significance level of 0.05, the likelihood of finding a difference in QTc between two groups for a particular baseline method is substantial (95 %; [1–0.95^{57}]). This does not imply however, that we understand why one method should be preferred over the other.

Although the study by Zhang et al. [6] validates one particular method over the other, it is important to emphasize that the assessment of bias and accuracy may be complicated by the absence of an accepted definition of baseline for the TQT study. Failure to recognize this condition can lead to significant misinterpretation of the observed differences between methods.

Instead, one may wish to address an issue of more concern to sponsors of TQT studies—the type II error rate. For example, it would be possible to divide a placebo group (both baseline and treatment days) into two random groups and test the hypothesis of no treatment effect. This is shown in Fig. 1 for 57 random divisions of the placebo arm (number of studies analyzed by Zhang et al. [6]) of a recently conducted parallel TQT study [16]. The effects on the Fridericia corrected QT interval (ΔΔQTcF) are noticeably different for the two baseline methods. Therefore, keeping in mind the 10 ms threshold of regulatory concern (upper 95 % one-sided CI for ΔΔQTc) it is clear that the choice of baseline is an important consideration in parallel TQT trials. It is also clear that, if two sponsors conduct a TQT in an identical manner but choose to use different baseline methods, the trial outcomes and type II error rates could be very different. For example, a time-matched baseline but not a time-averaged baseline would result in several type II errors in the illustration above. It is of course understood that the average parallel TQT study includes more than 31 subjects per study arm, and the data in Fig. 1 is only meant to demonstrate one type of analysis which is much needed in order to arrive at a definitive conclusion about the appropriate baseline to use in parallel TQT studies.

Acknowledgments No sources of funding were used in the preparation of this commentary and the author has no conflict of interest to declare.

References

1. E14 Clinical evaluation of QT/QTc interval prolongation and proarrhythmic potential for non-antiarrhythmic drugs. Guidance for industry (2005) online. Available from URL: http://www.fda.gov/RegulatoryInformation/Guidances/ucm129335.htm. Accessed 27 Mar 2013.
2. Smetana P, Batchvarov V, Hnatkova K, et al. Circadian rhythm of the corrected QT interval: impact of different heart rate correction models. Pacing Clin Electrophysiol. 2003;26:383–6.
3. Taubel J, Wong AH, Naseem A, et al. Shortening of the QT interval after food can be used to demonstrate assay sensitivity in thorough QT studies. J Clin Pharmacol. 2012;52:1558–65.
4. Browne KF, Prystowsky E, Heger JJ, et al. Prolongation of the Q-T interval in man during sleep. Am J Cardiol. 1983;52:55–9.
5. Ahnve S, Vallin H. Influence of heart rate and inhibition of autonomic tone on the QT interval. Circulation. 1982;65:435–9.
6. Zhang J, Dang Q, Malik M. Baseline correction in parallel thorough QT studies. Drug Saf 2013.
7. Fermini B, Fossa AA. The impact of drug-induced QT interval prolongation on drug discovery and development. Nat Rev Drug Discov. 2003;2:439–47.
8. Meng Z, Quan H, Fan L, et al. Use of the average baseline versus the time-matched baseline in parallel group thorough QT/QTc studies. J Biopharm Stat. 2010;20:665–82.
9. Sethuraman V, Sun Q. Impact of baseline ECG collection on the planning, analysis and interpretation of 'thorough' QT trials. Pharm Stat. 2009;8:113–24.
10. Sun GG, Quan H, kringle R, et al. Comparison of statistical models adjusting for baseline in the analysis of parallel-group thorough QT/QTc studies. J Biopharm Stat. 2012;22:438–62.
11. Lu K. An efficient and robust analysis of covariance model for baseline adjustment in parallel-group thorough QT/QTc studies. Statist Med 2012. doi: 10.1002/sim.5614 (Epub ahead of print).
12. Nielsen J, Graff C, Hardahl T, et al. Sertindole causes distinct electrocardiographic T-wave morphology changes. Eur Neuropsychopharmacol. 2009;19:702–7.
13. E14 Clinical evaluation of QT/QTc interval prolongation and proarrhythmic potential for non-antiarrhythmic drugs. Guidance for industry. Questions and answers (R1) 2012 (online). Available from URL: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm073161.pdf. Accessed 29 Mar 2013.
14. Beasley CM Jr, Benson C, Xia JQ, et al. Systematic decrements in QTc between the first and second day of contiguous daily ECG recordings under controlled conditions. Pacing Clin Electrophysiol. 2011;34:1116–27.
15. Hollister AS, Montague TH. Statistical analysis plans for ECG Data: controlling the intrinsic and extrinsic variability in QT data. In: Morganroth J, Gussak I, editors. Cardiac safety of noncardiac drugs. Totowa: Humana Press; 2005. p. 239–57.
16. Matz J, Graff C, Vainio PJ, et al. Effect of nalmefene 20 and 80 mg on the corrected QT interval and T-wave morphology: a randomized, double-blind, parallel-group, placebo- and moxifloxacin-controlled, single-centre study. Clin Drug Investig. 2011;31:1–13.