A Statistical View on ICH E14

Georg Ferber
Novartis Pharma AG, Basel
APF-Tagung, Heidelberg, 17/18 Nov 2005
Overview

Development of the Thorough QT/QTc Study (TQS) in ICH E14
  - Basic problems of the TQS
  - Formulation as Intersection-Union Test

Developments in response to ICH E14
  - Work done to understand the problem
    • Bias
    • Power
  - Optimising design and analysis

Past and future contributions
The Thorough QT/QT<sub>c</sub> Study – Initial version

The requirement for a placebo and active controlled "intense phase I trial" appeared already in the first version of ICH E14. However, the statistical concept was rather vague:

QT/QTc interval data should be presented in terms of means, standard deviations, ranges, and confidence intervals. Clinical trials that investigate the QT/QTc interval prolongation potential of a drug should have sufficient power \((i.e., \geq 80\%)\) to detect modest mean differences between treatment groups \((e.g., 4-5 \text{ msec at Emax QT/QTc})\)....

(3.4 Statistical Considerations of the version of 28-Jan-2003)
Formulating the statistical concept – Osaka version

The formulation as non-inferiority problem (although never explicit) appeared in the Osaka version (12-Nov-2003)

... a negative ‘thorough QT/QT<sub>c</sub> study’ is one where the placebo-subtracted difference for the drug on the QT/QT<sub>c</sub> interval is around 5 ms, with a 95% confidence interval that excludes an effect >7.5 ms. This upper bound is chosen to reflect the uncertainty related to the variability of repeated measurement.

... mean changes from baseline in the observed QT/QT<sub>c</sub> interval are presented as time-matched control and treatment group values (...). The maximal mean change in the QT/QT<sub>c</sub> interval and its timing should be reported.
... a negative ‘thorough QT/QTc study’ is one in which the upper bound of the 95% one-sided confidence interval for the largest time-matched mean effect of the drug on the QTc interval excludes 10 ms. This definition is chosen to provide reasonable assurance that the mean effect of the study drug on the QT/QTc interval is not greater than around 5 ms.
Is this a statistically reasonable criterion?

Naive statistical interpretation:

\[ X = (X_1, \ldots, X_T)^T \] – vector of time matched mean differences
\[ Z := \text{Max}\{X_i\} \text{ is our random variable of interest} \]

Determine a (one sided 95 %) CI for Z.

• Problem 1: What do we know about the distribution of X?
• Problem 2: Even if we assume a benign distribution of X, what do we know about distribution of Z?
The problem is poorly understood

Do we have a time series model for QT measurements?
• e. g. does a AR(1) model match the data?
• how robust are results against deviations from a model?

What reasonable assumptions would be needed to solve the problem of a CI for $Z := \text{Max}\{X_i\}$
• theoretically?
• using simulations?
Comment from a Colleague

I wholeheartedly agree that it is, to say the least, "problematic" for the guideline to ask statisticians analyzing QT data to have to deal with "an unresolved problem", without at least telling them that it's unsolved...
An alternative, pragmatic formulation

The Max-of-Mean-Difference condition will be met if for each timepoint, the confidence interval for the mean difference between the drug and placebo (baseline-adjusted) excludes an effect >10 ms.

In this (somewhat more stringent) version on an Intersection-Union Test the problem is amenable to conventional multiple noninferiority hypothesis testing.
What does the IUT do?

- It opens a way to analyse the TQS according to the guidance
- It is conservative in the sense that it protects public health
- It does not remove the bias inherent in the endpoint definition
- In particular, the sponsor is still punished for adding a timepoint
- A realistic power estimate is difficult
- Simple sample size calculations can lead to unrealistically large trials.
In the meantime ...

A growing body of insight is being accrued on

- the size and properties of the bias introduced
- the power of the intersection-union test under realistic conditions

Results are mainly based on simulations that start with real data.
Bias

The bias introduced is largest if the difference between active and placebo is constant over time, in particular also, if there is no drug induced QT prolongation.

In this case, simulations come up with values up to 3 ms.

The bias will be negligible if there is a pronounced peak in QT prolongation that is captured only by one timepoint.

The bias is not necessarily larger for a trial with a larger number of timepoints, but adding a timepoint without changing the other timepoints will never reduce the bias.
Power

A lower bound of the power of a TQS can be obtained as a product of the power-values for each timepoint.

As these power values depend on the true QT prolongation, knowledge about PK and PD of the drug can be used to get a more realistic bound for power.

Obviously, power also depends on the correlation between values as different timepoints, but it seems difficult to take this into account.

Alternatively, power can be estimated based on bootstrapping from the placebo data of a similar trial and an assumed QTc-effect.
Dependence on number of timepoints

Ways out that have been considered

- Concentration-response modelling
  - Workshop with FDA in summer
    interesting, but not accepted as replacement for Max-Mean analysis

- Modelling of time course of QT/QTc
  use of model based CI
  - adding a timepoint should improve quality of fit
  - risk of introducing a bias in direction of a false negative finding
Design considerations

For routine drug development, the design goal should be to obtain a reasonably powered TQS

• with a low number of subjects exposed
• at a low cost.
Design optimisation

Minimise residual variance

- by optimising the study design
- by optimising the endpoint
- by choosing the appropriate analysis
Optimising the study design

• when can we use a cross-over design and what is the gain?
  – Interestingly, it seems that there are more parallel group designs than one would expect.

• are replicates cost-effective?
  – Needs to be explored and depends on the number of timepoints, in general seems to be the case.

• what is the best baseline?
  – Time matched
  – Pre-dose
  – Average of pre-dose day
  – For cross-over: do we need a baseline at all?
Optimising the endpoint

Choice of correction method for heart rate

• Fixed correction (Fridericia seems best choice)
• Individual correction
  risk of introducing
  – additional variability, if not enough baseline and predose data available (no problem with Holter data)
  – bias, if baseline data do not span the whole range of heart rates encountered under drug
• Correction based on random effect model may reduce extra variability.
Choosing the appropriate analysis

Confidence intervals can be based on

- Pointwise t-tests
- Pointwise ANCOVA
  - including baseline
  - including further explanatory variables
- A hierarchical repeated measures model for all timepoints.
What did statisticians contribute to ICH E14?

- Raising awareness for statistical problems
- Clarify statistical interpretation of guidance text (e.g. for the endpoint of the TQS)
- Provide data (both real and simulated) to obtain realistic criteria
- Make sure that methods exist to follow the guideline

- Contributions to other questions were considered secondary to the above
  - definition of baseline
  - preferred correction method
  - crossover vs. parallel group design
Main contributors

Initial drive came from PhRMA and its QT statistical expert team.

Chairs: Marilyn Agin, Pfizer, Scott Patterson, GSK

Since 2005: Statistics discussion group for ICH E14

Members from all 6 ICH parties

Chairs (de facto): Stella Machado, FDA, Marilyn Agin, Pfizer

European members:

Rob Hemmings (EMEA)
Harry Southworth (Astra Zeneca)
Georg Ferber (Novartis)