Testing a secondary endpoint after a group sequential test

Christopher Jennison

Department of Mathematical Sciences,
University of Bath, UK
http://people.bath.ac.uk/mascj

University of Bath
February 2020
Outline of talk

1. A group sequential test for a primary endpoint

2. How should one test a secondary endpoint after a positive outcome for the primary endpoint?

3. Multiple testing procedures

4. Combining multiple testing and group sequential designs

5. Testing a secondary endpoint after a group sequential test
Suppose a new treatment (Treatment A) is to be compared to a placebo or positive control (Treatment B) in a Phase III trial.

The treatment effect $\theta$ for the primary endpoint represents the advantage of Treatment A over Treatment B.

If $\theta > 0$, Treatment A is more effective.

We wish to test the null hypothesis $H_0: \theta \leq 0$ against $\theta > 0$ with

\[ P_{\theta=0}\{\text{Reject } H_0\} = \alpha, \]
\[ P_{\theta=\delta}\{\text{Reject } H_0\} = 1 - \beta. \]

In a group sequential trial, data are examined on a number of occasions to see if an early decision may be possible.
A typical boundary for a one-sided test has the form:

\[ Z_k \]

Crossing the upper boundary leads to early stopping for a positive outcome, rejecting \( H_0 \) in favour of \( \theta > 0 \).

Crossing the lower boundary implies stopping for “futility” with acceptance of \( H_0 \).
Benefits of group sequential tests

**Earlier decisions**

Group sequential testing can speed up the process to introduce an effective new treatment.

**Fewer patients recruited**

Expected sample sizes for group sequential designs are, typically, around 70% of the fixed sample size for a trial with the same type I error rate and power.
2. Testing a secondary endpoint

In a trial of two treatments, A and B, a group sequential test is carried out on the primary endpoint, which has treatment effect $\theta_1$.

Suppose $H_1: \theta_1 \leq 0$ is rejected in favour of $\theta_1 > 0$.

The investigators wish to test whether Treatment A is also superior for a secondary endpoint, with treatment effect denoted by $\theta_2$.

Some familiarity with “gatekeeping” procedures for testing multiple hypotheses suggests it may be legitimate to pass on the type I error $\alpha = 0.025$ to a second hypothesis test.

As this test will be only conducted once, the investigators plan to carry out a fixed sample size, level $\alpha$ test of $H_2: \theta_2 \leq 0$ vs $\theta_2 > 0$ using the available data on the secondary endpoint.

**Is this approach to testing the two endpoints valid?**
Testing a secondary endpoint: Example

Suppose the primary endpoint is tested using a Pampallona & Tsiatis group sequential design with shape parameter $\Delta = 0$.

There are 4 analyses, type I error probability is $\alpha = 0.025$ and power is 0.8 at $\theta_1 = 1$.

This test has upper boundary:

$$Z_k = 3.90/\sqrt{k}$$

and lower boundary

$$Z_k = 1.48\sqrt{k} - 2.02/\sqrt{k},$$

where $k = 1, \ldots, 4$.

If the upper boundary is crossed, the secondary endpoint is tested in a level $\alpha$, fixed sample size test, using current data.
Testing a secondary endpoint: Example

The plot shows the probability of rejecting $H_2: \theta_2 \leq 0$, under $\theta_2 = 0$, when the secondary endpoint is tested as described above.

The two endpoints have correlation $\rho$. For modest values of $\rho$, the type I error rate for testing $H_2$ exceeds the nominal 0.025.

Hung, Wang and O’Neill (J. Biopharm. Statis., 2007) have noted this approach to testing a secondary endpoint is not valid.

So, how should the secondary endpoint be tested?
3. Multiple testing procedures

Our example had one primary and one secondary endpoint.

More generally, a clinical trial may involve

Co-primary endpoints

*Positive outcomes required for at least one endpoint*

*Positive outcomes required on all endpoints*

Secondary endpoints, tertiary endpoints, ... 

The trial may have

Multiple treatments,

Pre-defined sub-populations of patients.

If the trial is group sequential, each hypothesis may be tested on several occasions.
The familywise error rate

Suppose we have \( h \) null hypotheses, \( H_i: \theta_i \leq 0 \) for \( i = 1, \ldots, h \). After our analysis, we accept or reject each of these \( h \) hypotheses.

A testing procedure’s **familywise error rate** under a set of values \( \theta = (\theta_1, \ldots, \theta_h) \) is

\[
Pr_{\theta}\{\text{Reject } H_i \text{ for some } i \text{ with } \theta_i \leq 0}\]

\[
= Pr_{\theta}\{\text{Reject at least one true } H_i\}.
\]

The familywise error rate is controlled **strongly** at level \( \alpha \) if this error rate is at most \( \alpha \) for all possible combinations of \( \theta_i \) values.

Then

\[
Pr\{\text{Reject any true } H_i\} \leq \alpha \text{ for all } (\theta_1, \ldots, \theta_h).
\]
Suppose we test $h$ null hypotheses, each at significance level $\alpha/h$.

If all $h$ null hypotheses are true,

$$Pr\{\text{Reject at least one of } H_1 \ldots H_h\} \leq Pr\{\text{Reject } H_1\} + \ldots + Pr\{\text{Reject } H_h\} = h \frac{\alpha}{h} = \alpha.$$

If only some of the $h$ null hypotheses are true,

$$Pr\{\text{Reject at least one true } H_i\} < \alpha.$$

So we have strong control of the familywise error rate.

We start by considering applications in fixed sample size study designs...
Example: A Bonferroni test with co-primary endpoints

A trial compares a new treatment against control with respect to:

- **Endpoint 1, Core MACE** (*Major Adverse Cardiac Event* — CV-related death, nonfatal stroke, or nonfatal MI)
- **Endpoint 2, Expanded MACE** (Core MACE plus hospitalization for unstable angina or coronary revascularization).

Type I error probability $\alpha = 0.025$ is divided between the endpoints.

With $Z$-statistics $Z_1$ and $Z_2$ for endpoints 1 and 2,

- An effect on Core MACE is declared if
  \[ Z_1 > \Phi^{-1}(1 - \alpha/2) = 2.24, \]
- An effect on Expanded MACE is declared if
  \[ Z_2 > \Phi^{-1}(1 - \alpha/2) = 2.24. \]
This Bonferroni procedure can be represented graphically as:

\[ \frac{\alpha}{2} \quad | \quad \frac{\alpha}{2} \]

\[ 1 \quad 2 \]

There is a positive correlation between the two tests, due to the common aspects of the two endpoints.

Hence, familywise type I error is protected conservatively.

Power when \( H_1 \) and \( H_2 \) are false can be increased by “recycling” type I error after one or other hypothesis is rejected.
The Holm procedure is a version of the Bonferroni procedure that “recycles” error probability after rejecting $H_1$ or $H_2$.

This method can be represented as:

If $H_1$ is rejected at level $\alpha/2$, we pass that error probability to $H_2$ and test this hypothesis at level $\alpha$.

If $H_2$ is rejected at level $\alpha/2$, we pass that error probability to $H_1$ and test this hypothesis at level $\alpha$. 
Proof that FWER is protected

If $H_1$ and $H_2$ are both true,

\[
\begin{align*}
\text{FWER} &= Pr\{\text{Reject } H_1 \text{ or } H_2\} \\
&\leq Pr\{Z_1 > \Phi^{-1}(1 - \alpha/2)\} + Pr\{Z_2 > \Phi^{-1}(1 - \alpha/2)\} \\
&\leq \alpha/2 + \alpha/2 = \alpha.
\end{align*}
\]

If $H_1$ is true and $H_2$ is false,

\[
\begin{align*}
\text{FWER} &= Pr\{\text{Reject } H_1\} \leq Pr\{Z_1 > \Phi^{-1}(1 - \alpha)\} = \alpha.
\end{align*}
\]

$H_2$ is true and $H_1$ false: Similar to $H_1$ true and $H_2$ false.

$H_1$ and $H_2$ both false: A type I error cannot be made.
A hierarchical testing or “gatekeeping” procedure

Consider a trial where

The null hypothesis $H_1$ concerns the primary endpoint,

The null hypothesis $H_2$ relates to a secondary endpoint,

and $H_2$ will only be tested if $H_1$ has already been rejected.

First, we test $H_1$ at significance level $\alpha$.

If $H_1$ is rejected, we continue and test $H_2$ at significance level $\alpha$. 

1 $H_1$, primary endpoint

2 $H_2$, secondary endpoint
Proof that FWER is protected

Subject $H_1$ is true.

A family-wise error occurs if $H_1$ is rejected (whether or not $H_2$ is also rejected). So

$$\text{FWER} = Pr\{\text{Reject } H_1\} = Pr\{Z_1 > \Phi^{-1}(1 - \alpha)\} = \alpha.$$  

If $H_1$ is false and $H_2$ is true,

$$\text{FWER} = Pr\{\text{Reject } H_1 \text{ and then reject } H_2\} \leq Pr\{Z_2 > \Phi^{-1}(1 - \alpha)\} = \alpha.$$  

If $H_1$ and $H_2$ are both false, a type I error cannot be made.
Testing co-primary and secondary endpoints

The figure below represents a testing procedure that starts with a Bonferroni test of $H_1$ and $H_2$.

Then, if either $H_1$ or $H_2$ is rejected, the associated type I error is passed on to the test of $H_3$.

We can prove there is strong control of FWER at level $\alpha$ by considering all combinations of $H_1$, $H_2$ and $H_3$ being True or False.
We can add more “recycling” to the previous testing procedure.

\[ \alpha/2 \quad | \quad \alpha/2 \]

1 \quad 2

\[ H_1, H_2: \text{ co-primary endpoints} \]

3 \quad \text{secondary endpoint}

The additional lines in the graph indicate that

If \( P_1 \leq \alpha/2 \) and \( P_3 \leq \alpha/2 \), then \( H_2 \) is tested at level \( \alpha \),

If \( P_2 \leq \alpha/2 \) and \( P_3 \leq \alpha/2 \), then \( H_1 \) is tested at level \( \alpha \).
We may prefer to gain maximum power for tests of co-primary endpoints before testing a secondary endpoint. To do this, we recycle type I error probability between $H_1$ and $H_2$ before allocating any error probability to $H_3$.

A graphical representation is:

$H_1, H_2$: co-primary endpoints

$H_3$: secondary endpoint

Half of the type I error probability is cycled through $H_1$, $H_2$ and on to $H_3$.

The other half is cycled through $H_2$, $H_1$ and on to $H_3$. 
More complex procedures: General methodology

As we add more options, and get more creative, we can produce some quite complex procedures.

Two papers, published simultaneously, describe an elegant way to describe complex multiple testing procedures.

“A recycling framework for the construction of Bonferroni-based multiple tests” by Burman, Sonesson and Guilbaud, Statistics in Medicine, 2009.

“A graphical approach to sequentially rejective multiple test procedures” by Bretz, Maurer, Brannath and Posch, Statistics in Medicine, 2009.

These procedures are closed testing procedures in which the tests of intersection hypotheses are weighted Bonferroni tests.

It is implicit in their method of construction that these procedures provide strong control of the FWER.
A figure from Burman et al. (2009)

The following diagrams illustrate the graphical representations of multiple testing procedures used by Burman et al.

(a) and (b) A parallel gatekeeping procedure

(c) and (d) A fallback procedure
A figure from Bretz et al. (2009)

And here is an example of a graphical representation of a procedure as defined by Bretz et al.

Figure 3. Graphical illustration of the Bonferroni–Holm procedure with $m=3$ hypotheses and initial allocation $\alpha=(\alpha/3, \alpha/3, \alpha/3)$. 
A figure from Bretz et al. (2009)

And here is an example of a graphical representation of a procedure as defined by Bretz et al.

Figure 3. Graphical illustration of the Bonferroni–Holm procedure with $m=3$ hypotheses and initial allocation $\alpha=(\alpha/3, \alpha/3, \alpha/3)$.

Question: How can we apply such a procedure in a group sequential trial?
4. Multiple testing procedures and group sequential designs

Maurer & Bretz (Statist. in Biopharm. Research, 2013) explain how to carry out tests of multiple hypothesis in a group sequential trial with strong control of FWER.

Consider a multiple testing procedure for hypotheses $H_1, \ldots, H_h$ that involves testing $H_1, \ldots, H_h$ at different significance levels, possibly increasing these levels after other hypotheses are rejected.

Define group sequential tests of each hypothesis with type I error rates equal to the various significance levels that may be applied.

At each analysis, conduct tests of $H_1, \ldots, H_h$ using the boundary points of their group sequential tests for the current analysis.

In doing this, follow the testing hierarchy and “re-cycling rules” to determine the type I error rate of each hypothesis testing boundary.

Stop the study when key conclusions have been reached.
For group sequential implementation of the above multiple testing procedure, we need

GSTs at levels $\alpha/3$, $\alpha/2$ and $\alpha$

for each of the hypotheses, $H_1$, $H_2$ and $H_3$. 

Figure 3. Graphical illustration of the Bonferroni–Holm procedure with $m=3$ hypotheses and initial allocation $\alpha=(\alpha/3, \alpha/3, \alpha/3)$. 
5. Testing a secondary endpoint after a sequential test

A correct gatekeeping procedure

We discussed a group sequential trial comparing the effects of two treatments with on a primary endpoint. Then, if a positive result is obtained, a secondary endpoint is tested.

In Maurer & Bretz’s scheme, we need to specify a level $\alpha$ group sequential test for the secondary endpoint: this test of $H_2$ will be applied whenever the trial terminates.

The group sequential test of $H_1$ determines the stopping time for the trial

The group sequential test of $H_2$ is used for the secondary analysis if and when $H_1$ is rejected.
A correct gatekeeping procedure

Let $Z_{1,1}, \ldots, Z_{1,K}$ be $Z$-statistics for testing $H_1: \theta_1 \leq 0$ at analyses $1, \ldots, K$.

The group sequential test of $H_1$ stops at analysis $k$ to

Reject $H_1$ if $Z_{1,k} \geq b_k$,

Accept $H_1$ if $Z_{1,k} < a_k$.

Boundary values for the test of $H_1$ control the type I error rate at level $\alpha$ under $\theta_1 = 0$, i.e.,

$$\sum_{k=1}^{K} Pr\{Z_{1,1} \in (a_1, b_1), \ldots, Z_{1,k-1} \in (a_{k-1}, b_{k-1}), Z_{1,k} > b_k\} = \alpha.$$ 

Suppose this GST stops to reject $H_1$ at analysis $k^*$ \ldots
A correct gatekeeping procedure

Let $Z_{2,1}, \ldots, Z_{2,K}$ be $Z$-statistics for testing $H_2: \theta_2 \leq 0$.

The level $\alpha$ group sequential test of $H_2$ rejects $H_2$ at analysis $k$ if $Z_{2,k} \geq c_k$, where under $\theta_2 = 0$

$$\sum_{k=1}^{K} Pr\{Z_{2,1} < c_1, \ldots, Z_{2,k-1} < c_{k-1}, Z_{2,k} > c_k\} = \alpha. \quad (1)$$

(The trial’s stopping rule is based on the primary endpoint, so we do not need a lower boundary for early acceptance of $H_2$.)

When the GST of $H_1$ has rejected $H_1$ at analysis $k^*$, we reject $H_2$ if $Z_{2,k^*} \geq c_{k^*}$.

A gatekeeping procedure could reject $H_2$ if

$$Z_{2,k} \geq c_k \quad \text{for any } k \in \{1, \ldots, K\},$$

so the FWER is protected conservatively.
In a trial comparing two treatments, denote the treatment effects on the primary and secondary endpoints by $\theta_1$ and $\theta_2$.

Suppose the trial is conducted group sequentially, using a Pampallona & Tsiatis test with $\Delta = 0$ for the primary endpoint.

There are 4 analyses, $\alpha = 0.025$ and power is 0.8 at $\theta_1 = 1$.

If $H_1: \theta_1 \leq 0$ is rejected for the primary endpoint at analysis $k^*$, we test the secondary endpoint: we reject $H_2: \theta_2 \leq 0$ if

$$Z_{2,k^*} \geq c_{k^*}.$$ 

We consider two options for this test of $H_2$. 
Example: Testing primary and secondary endpoints

We consider two options for the group sequential test of $H_2$.

**A: Pocock boundary for $H_2$**

$$c_k = 2.361, \ k = 1, \ldots, 4.$$  

**B: OBF boundary for $H_2$**

$$c_k = 2.024 \sqrt{\frac{4}{k}}, \ k = 1, \ldots, 4.$$  

Note: The O’Brien & Fleming boundary requires a very high value of $Z_{2,k^*}$ to reject $H_2$ if the GST of $H_1$ stops at the first analysis.
Type I error probability for testing $H_2$

**A: Pocock boundary for $H_2$**

**B: OBF boundary for $H_2$**

Type I error probabilities are calculated under $\theta_2 = 0$, but they also depend on $\theta_1$ and the correlation, $\rho$, between the primary and secondary endpoints.

The OBF test of $H_2$ is particularly conservative when $\theta_1$ is large.
Power for testing $H_2$, $\rho = 0.25$

**A: Pocock boundary for $H_2$**

**B: OBF boundary for $H_2$**

Results are shown for the case that the variance of the secondary response is 0.5 times that for the primary response.

Power is shown as a function of $\theta_2$ for selected values of $\theta_1$.

The Pocock boundary for $H_2$ deals better with the trial’s uncertain termination time — which depends significantly on the value of $\theta_1$. 
Power for testing $H_2$, $\rho = 0.5$

A: Pocock boundary for $H_2$

B: OBF boundary for $H_2$

Results are shown for the case that the variance of the secondary response is 0.5 times that for the primary response.

Power is shown as a function of $\theta_2$ for selected values of $\theta_1$.

Again, the Pocock boundary for $H_2$ deals better with the trial’s uncertain termination time — which depends significantly on $\theta_1$. 
Testing a secondary endpoint: Further options

Conservatism in the overall procedure arises because the test of $H_1$ may stop at analysis $k^*$ when $Z_{2,k^*} < c_{k^*}$, but

$$Z_{2,k} \geq c_k \quad \text{for some} \quad k < k^* \quad \text{or} \quad k > k^*.$$

There are options for reducing conservatism and increasing power:

1. Reject $H_2$ if $Z_{2,k} \geq c_k$ for some $k < k^*$, even though $Z_{2,k^*} < c_{k^*}$. However, ignoring more recent data (and not using the sufficient statistic for $\theta_2$) may detract from the credibility of this decision.

2. Continue the trial to see if $Z_{2,k} \geq c_k$ at a future analysis. However, if the primary endpoint is observed for future subjects, the positive result on the primary endpoint could be “lost”.

Several authors have considered option (2), retaining a positive outcome for $H_1$, whatever the additional information about $\theta_1$. 
1. There are methods available to test multiple hypotheses in a group sequential design AND control the overall type I error probability.

2. Graphical representations (SiM papers, 2009) can help investigators to select — and understand — an appropriate multiple testing procedure.

3. There are many multiple testing schemes to choose from. The most suitable choice will depend on the importance to investigators of rejecting each null hypothesis and the likelihood of each null hypothesis being true or false.

4. When testing multiple hypotheses in a group sequential trial design, the key point is to use GSTs as the “testing rules” in the multiple testing scheme: if this is not done correctly, FWER may be inflated.
Testing a secondary endpoint: Further options

3. If the worst case scenario, in which a procedure’s maximum FWER occurs, can be identified then, the procedure may be calibrated so the FWER is equal to the specified level $\alpha$ in this scenario. See:

**Glimm, Maurer & Bretz (Stat. in Med., 2010)** Hierarchical testing of multiple endpoints in group-sequential trials.

**Tamhane, Mehta & Liu (Biometrics, 2010)** Testing a primary and a secondary endpoint in a group sequential design.

**Tamhane, Wu & Mehta (Stat. in Med., 2012)** Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (I) unknown correlation between endpoints.

**Tamhane, Gou, Jennison, Mehta & Curto (Biometrics, 2018)** A gatekeeping procedure to test a primary and a secondary endpoint in a group sequential design with multiple interim looks.
Tang & Geller (Biometrics, 1999) Closed testing procedures for group sequential clinical trials with multiple endpoints.

Ye, Liu & Yao (Statist. in Med., 2012) A group sequential Holm procedure with multiple primary endpoints.

Maurer & Bretz (Statist. in Biopharm. Research, 2013) Multiple testing in group sequential trials using graphical approaches.

Li, Wang, Luo, Grechko & Jennison (Biometrical Journal, 2018) Improved two-stage group sequential procedures for testing a secondary endpoint after the primary endpoint achieves significance.