Multi-arm Group Sequential Designs
with a Simultaneous Stopping Rule

Susanne Urach, Martin Posch
ICODOE 2016 Memphis, Tennessee, USA

This project has received funding from the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement number FP HEALTH 2013-603160.
Objectives of multi-arm multi-stage trials

**Aim:** Comparison of several treatments to a common control

**Compared to separate, fixed sample two-armed trials**
- less patients needed than for separate controlled clinical trials
- larger number of patients are randomised to experimental treatments
- possibility to stop early for efficacy or futility

**Objective:** Identify all treatments that are superior to control

**Objective:** Identify at least one treatment that is superior to control

Which stopping rule?
Multi-arm multi-stage trials

Design setup: group sequential Dunnett test

- Comparison of two treatments to a control
- Normal endpoints, variance known
- One sided tests: $H_A : \mu_A - \mu_C \leq 0$ and $H_B : \mu_B - \mu_C \leq 0$
- Control of the Family Wise Error Rate (FWER) = 0.025
- Two stage group sequential trial: one interim analysis at $\frac{N_{\text{max}}}{2}$
- $Z_{A,i}, Z_{B,i}$ are the cumulative z-statistics at stage $i=1,2$
Classical group sequential Dunnett tests with “separate stopping”
Objective: Identify all treatments that are superior to control

“separate stopping rule”: Treatment arms, for which a stopping boundary is crossed, stop.

E.g.:

\[ H_B \] is rejected at interim

\[ A \] can go on and is tested again at the end

Magirr, Jaki, Whitehead (2012)
Classical group sequential Dunnett tests with “separate stopping”

Closed group sequential tests

Local group sequential tests for $H_A \cap H_B$ and $H_A, H_B$ are needed!!!

A hypothesis is rejected with FWER $\alpha$ if the intersection hypothesis and the corresponding elementary hypothesis are rejected locally at level $\alpha$. 

$$H_A \cap H_B$$

$$H_A$$

$$H_B$$

$$H_A$$

$$H_B$$

$$H_A \cap H_B$$
Classical group sequential Dunnett tests with “separate stopping”

Closed group sequential tests

Reject if $\max(Z_{A,1}, Z_{B,1}) > u_1$ or $\max(Z_{A,2}, Z_{B,2}) > u_2$

Reject if $Z_{A,1} > v_1$ or $Z_{A,2} > v_2$

Reject if $Z_{B,1} > v_1$ or $Z_{B,2} > v_2$

$u_1, u_2$...global boundaries
$v_1, v_2$...elementary boundaries

Koenig, Brannath, Bretz and Posch (2008)
Xi, Tamhane (2015)
Maurer, Bretz (2013)
Group sequential Dunnett tests with “simultaneous stopping”
"simultaneous stopping rule": If at least one rejection boundary is crossed, the whole trial stops.

**Objective:** Identify at least one treatment that is superior to control

If, e.g., $H_B$ is rejected at interim then the trial is stopped:
Simultaneous versus Separate Stopping

- The **FWER** is controlled when using the boundaries of the separate stopping design.

- The **expected sample size (ESS)** is lower compared to separate stopping designs.

- The **power to reject**
  - **any** null hypothesis is the **same** as for separate stopping designs.
  - **both** null hypotheses is **lower** than for separate stopping designs.

→ **Trade-off between ESS and conjunctive power**
Can one \textit{relax the interim boundaries} when stopping simultaneously?

How large is the impact on \textit{ESS and power} when stopping simultaneously or separately?

How to \textit{optimize} the critical boundaries for either stopping rule?
Question 1: Relaxation of interim boundaries?

For simultaneous stopping:

- The boundaries $u_1$, $u_2$ for the local test of $H_A \cap H_B$ cannot be relaxed.
- The boundaries $v_1$, $v_2$ for the local test of $H_j$ can be relaxed.

Intuitive explanation

If, e.g., $H_B$ is rejected at interim, but $H_A$ not, $H_A$ is no longer tested at the final analysis and not all $\alpha$ is spent.

It’s possible to choose improved boundaries for the elementary tests.

(similar as for group sequential multiple endpoint tests in Tamhane, Metha, Liu 2010).
What changes when stopping simultaneously?

Example: O’Brien Fleming boundaries

Reject if $\max(Z_{A,1}, Z_{B,1}) > u_1$ or $\max(Z_{A,2}, Z_{B,2}) > u_2$

$u_1 = 3.14, u_2 = 2.22$

Reject if $Z_{A,1} > v_1$ or $Z_{A,2} > v_2$

$\nu_1 = 2.80, \nu_2 = 1.98$

Reject if $Z_{B,1} > v_1$ or $Z_{B,2} > v_2$

$\nu_1 = 2.80, \nu_2 = 1.98$
What changes when stopping simultaneously?
Example: O’Brien Fleming boundaries

Reject if \( \max(Z_{A,1}, Z_{B,1}) > u_1 \) or \( \max(Z_{A,2}, Z_{B,2}) > u_2 \)

\[ u_1 = 3.14, \quad u_2 = 2.22 \]

Reject if \( Z_{A,1} > \nu_1 \) or \( Z_{A,2} > \nu_2 \)

\[ \nu_1 = 2.80, \quad \nu_2 = 1.98 \]

Reject if \( Z_{B,1} > \nu_1 \) or \( Z_{B,2} > \nu_2 \)

\[ \nu_1 = 2.80, \quad \nu_2 = 1.98 \]

For simultaneous stopping there is no second stage test if one of the null hypotheses can already be rejected at interim.
FWER for simultaneous stopping if only $H_A$ holds ($\delta_A = 0$)
Question 1: Relaxation of interim boundaries?

FWER for simultaneous stopping if only $H_A$ holds ($\delta_A = 0$)

O'Brien Fleming

- improved group sequential boundaries
- classical group sequential boundaries

Type I error rate vs. $\delta_B$

The graph compares the type I error rate for improved and classical group sequential boundaries as a function of $\delta_B$. The O'Brien Fleming approach is illustrated with these boundaries, showing how the type I error rate changes with different values of $\delta_B$. The graph highlights the effectiveness of the improved boundaries in controlling the type I error rate compared to the classical approach.
Question 2: Impact on ESS and power?

For $\alpha = 0.025$ and $\delta_A = \delta_B = 0.5$

Conjunctive Power = Power to reject both false hypotheses
Disjunctive Power = Power to reject at least one false hypothesis

<table>
<thead>
<tr>
<th>Boundaries $u_i$ for $H_1 \cap H_2$</th>
<th>separate stopping rule</th>
<th>simultaneous stopping rule</th>
<th>improved simultan.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boundary $u_1$ for $H_1 \cap H_2$</td>
<td>$u_1 = 3.14, u_2 = 2.22$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interim boundary $v_1$</td>
<td>2.80</td>
<td>2.80</td>
<td>2.08</td>
</tr>
<tr>
<td>Final boundary $v_2$</td>
<td>1.98</td>
<td>1.98</td>
<td>1.98</td>
</tr>
<tr>
<td>Maximum $\alpha$ for test of $H_j$</td>
<td>0.025</td>
<td>0.019</td>
<td>0.025</td>
</tr>
<tr>
<td>Disj. power</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>N</td>
<td>324</td>
<td>324</td>
<td>324</td>
</tr>
<tr>
<td>ESS</td>
<td>230</td>
<td>205</td>
<td>205</td>
</tr>
<tr>
<td>Conj. power</td>
<td>0.89</td>
<td>0.69</td>
<td>0.76</td>
</tr>
</tbody>
</table>
Question 2: Impact on ESS and power?
For $\alpha = 0.025$ and $\delta_A = \delta_B = 0.5$

Conjunctive Power = Power to reject both false hypotheses
Disjunctive Power = Power to reject at least one false hypothesis

<table>
<thead>
<tr>
<th>Boundaries $u_i$ for $H_1 \cap H_2$</th>
<th>separate stopping rule</th>
<th>simultaneous stopping rule</th>
<th>improved simultan.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interim boundary $v_1$</td>
<td>2.80</td>
<td>2.80</td>
<td>2.08</td>
</tr>
<tr>
<td>Final boundary $v_2$</td>
<td>1.98</td>
<td>1.98</td>
<td>1.98</td>
</tr>
<tr>
<td>Maximum $\alpha$ for test of $H_j$</td>
<td>0.025</td>
<td>0.019</td>
<td>0.025</td>
</tr>
<tr>
<td>Disj. power</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>N</td>
<td>324</td>
<td>324</td>
<td>324</td>
</tr>
<tr>
<td>ESS</td>
<td>230</td>
<td>205</td>
<td>205</td>
</tr>
<tr>
<td>Conj. power</td>
<td>0.89</td>
<td>0.69</td>
<td>0.76</td>
</tr>
</tbody>
</table>
Optimized multi-arm multi-stage designs
Optimized designs

For $\alpha = 0.025$ and $\delta_A = \delta_B = 0.5$.

<table>
<thead>
<tr>
<th>Design</th>
<th>“Separate stopping”</th>
<th>“Simultaneous stopping”</th>
<th>“Improved simult. stopping”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boundaries</td>
<td>group sequential</td>
<td>group sequential</td>
<td>improved group sequential</td>
</tr>
<tr>
<td>Stopping rule</td>
<td>separate stopping rule</td>
<td>simultaneous stopping rule</td>
<td>simultaneous stopping rule</td>
</tr>
</tbody>
</table>
### Optimized designs

For $\alpha = 0.025$ and $\delta_A = \delta_B = 0.5$.

<table>
<thead>
<tr>
<th>Design</th>
<th>“Separate stopping”</th>
<th>“Simultaneous stopping”</th>
<th>“Improved simult. stopping”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boundaries</td>
<td>group sequential</td>
<td>group sequential</td>
<td>improved group sequential</td>
</tr>
<tr>
<td>Stopping rule</td>
<td>separate stopping rule</td>
<td>simultaneous stopping rule</td>
<td>simultaneous stopping rule</td>
</tr>
<tr>
<td>$N_{max}$</td>
<td></td>
<td></td>
<td>chosen to achieve disjunctive power of 0.9</td>
</tr>
<tr>
<td>Obj. function to optimize $u_1, u_2$</td>
<td>expected sample size</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Optimized designs

For $\alpha = 0.025$ and $\delta_A = \delta_B = 0.5$.

<table>
<thead>
<tr>
<th>Design</th>
<th>“Separate stopping”</th>
<th>“Simultaneous stopping”</th>
<th>“Improved simult. stopping”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boundaries</td>
<td>group sequential</td>
<td>group sequential</td>
<td>improved group sequential</td>
</tr>
<tr>
<td>Stopping rule</td>
<td>separate stopping rule</td>
<td>simultaneous stopping rule</td>
<td>simultaneous stopping rule</td>
</tr>
<tr>
<td>$N_{max}$</td>
<td>chosen to achieve disjunctive power of 0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obj. function to optimize $u_1, u_2$</td>
<td>expected sample size</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obj. function to optimize $v_1, v_2$</td>
<td>expected sample size</td>
<td>conjunctive power</td>
<td></td>
</tr>
</tbody>
</table>
Optimized boundaries $\delta_A = 0.5, \delta_B = 0.5$

<table>
<thead>
<tr>
<th></th>
<th>separate</th>
<th>simultaneous</th>
<th>improved simult.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>2.47</td>
<td>2.41</td>
<td>2.41</td>
</tr>
<tr>
<td>$u_2$</td>
<td>2.38</td>
<td>2.43</td>
<td>2.43</td>
</tr>
<tr>
<td>$v_1$</td>
<td>2.05</td>
<td>2.06</td>
<td>2.00</td>
</tr>
<tr>
<td>$v_2$</td>
<td>2.38</td>
<td>2.37</td>
<td>2.06</td>
</tr>
<tr>
<td>conj. power</td>
<td>0.85</td>
<td>0.71</td>
<td>0.76</td>
</tr>
<tr>
<td>ESS</td>
<td>225</td>
<td>205</td>
<td>205</td>
</tr>
<tr>
<td>$N_{max}$</td>
<td>318</td>
<td>324</td>
<td>324</td>
</tr>
</tbody>
</table>
Power to reject both null hypotheses

Power to reject at least one false hypothesis = 90% for all designs.
Optimal expected sample size (ESS)

ESS reduction between 8% and 16%.
Unknown variance: Extension to the t test

- P-value approach: z-score boundaries are converted to p-value boundaries and then applied to t-test p-values
- Simulation of t-statistics for p-value approach (optimized for $\delta_A = \delta_B = 1$) for $\sigma = 1$.

<table>
<thead>
<tr>
<th>Design</th>
<th>N</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>separate</td>
<td>8</td>
<td>0.0259</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>0.0257</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.0251</td>
</tr>
<tr>
<td>improved</td>
<td>8</td>
<td>0.0261</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>0.0258</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.0250</td>
</tr>
</tbody>
</table>
Summary

- The **optimal design** depends on the type of objective:
  - Reject all hypotheses
  - Reject at least one hypothesis

- **Simultaneous stopping compared to separate stopping** leads to
  - lower expected sample size
  - the same power to reject any hypothesis
  - lower power to reject both hypotheses

**Improved boundaries** can be used to regain some of the power to reject both null hypotheses.

- **Limitation:** If improved boundaries are used, the simultaneous stopping rule must be adhered to!

- **Extensions:**
  - more treatment arms, stopping for futility
  - optimal choice of first stage sample size/allocation ratio
References

- Thall et al. (1989): one treatment continues, futility stopping, two stages, power comparisons under LFC
- Follmann et al. (1994): Pocock and OBF MAMS designs, Dunnett and Tukey generalisations, several stages
- Stallard & Todd (2003): only one treatment is taken forward, several stages, power comparisons
- Stallard & Friede (2008): stagewise prespecified number of treatments
- **Magirr, Jaki, Whitehead (2012)**: FWER of generalised Dunnett
- Koenig, Brannath, Bretz (2008): closure principle for Dunnett test, adaptive Dunnett test
- **Magirr, Stallard, Jaki (2014)**: Flexible sequential designs
- Di Scala & Glimm (2011): Time to event endpoints
- **Wason & Jaki (2012)**: Optimal MAMS designs
- **Tamhane & Xi (2013)**: multiple hypotheses and closure principle
- Maurer & Bretz (2013): Multiple testing using graphical approaches
FWER inflation when $u_1^* = z_{1-\alpha} = 1.96$

![Graph showing the relationship between $\alpha$ and $\delta_B$. The graph is labeled as "Alpha inflation OBF design" and shows a peak at $\delta_B$ around 1.5 with $\alpha$ values ranging from 0.01 to 0.03. The dashed line represents a threshold for $\alpha$.](image)
Difference in expected sample size: OBF design
Difference in conjunctive power: OBF design