Basic and advanced methods for group sequential trials

Marcel Wolbers & Kaspar Rufibach

Methods, Collaboration and Outreach group (MCO),
Data & Statistical Sciences, Roche Basel

February 13, 2023

Generated 2023-06-07 at 21:23:37.

Program for this course

- Overview of general **group-sequential** trial methodology.
- Extensions of the basic methodology.
- Case study for survival data.
- **Inference** in group-sequential designs.
- Testing of **multiple endpoints** in group-sequential designs.
- Hierarchical testing of primary and **secondary endpoints** in a group-sequential design.

rpact and adaptR

rpact (R Package for Adaptive Clinical Trials) is the **preferred software** at Roche PD DSS for the design and analysis of group-sequential and adaptive trials:

- Powerful, fully documented and validated package which implements all methods in Wassmer and Brannath (2016) and beyond.
- Development financially supported by a consortium of pharma companies including Roche.
- Continuously developed, includes a shiny app & methods for adaptive designs.

adaptR (`https://go.roche.com/adaptR`; requires VPN) is a **comprehensive Roche resource** for group-sequential and adaptive trial designs:

- Website contains training materials, rpact vignettes, rpact validation document, FAQs etc.

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Confirmatory group-sequential trials

- Trials with pre-planned interim analyses which allow to potentially stop the trial prematurely for **futility** or **efficacy**.
- **Type I error protection**, despite looking into the data multiple times.
- **Ethical**, **economic**, and **administrative** reasons for multiple looks.
- Extension: **Adaptive trials** which allow for general adapation of **design** and/or **analysis** features after interim analyses.

Standard references: Jennison and Turnbull (2000), Proschan et al. (2006), Wassmer and Brannath (2016).

Efficacy interim analyses

Idea: If primary endpoint result already convincing at interim. ⇒ Declare efficacy, unblind the trial and file results to regulatory agencies.

Challenge: **Multiple** opportunities to declare drug working ⇒ if drug does not work, multiple opportunities to commit type I error ⇒ need to adjust overall significance level α^*.

For efficacy interim analyses performed at equal information increments:

# analyses	1	2	5	10	∞
Probability of a type I error	0.05	0.08	0.14	0.19	1.00

Not computed via $1-(1-\alpha^*)^k$! We are repeatedly testing on **accumulating data**.

Efficacy interim: how to adjust α^*?

Bonferroni, i.e. use $\alpha^* = 0.025$ at interim and final analysis. Valid, but **conservative** - we can do better!

How? By exploiting **correlation** between test statistic at interim and final analysis.

Generic example:

- One interim at 50% of information.
- Overall two-sided significance overall significance level $\alpha^* = 0.05$.
- **Bonferroni**: $\alpha = 0.025$ at each analysis.
- **Pocock group-sequential design**: $\alpha = 0.02937$ at each analysis. **Increased** significance level, "easier" to be significant.

Split of α^* must not be 50:50 between analyses. Typically, make it much harder to stop early $\Rightarrow$ O'Brien-Fleming stopping boundary.

Agenda

1. Introduction
2. **Pocock's and O'Brien-Fleming's designs for normal data**
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

No interim – 2-sample z-test (σ known)

Group A: $X_{A1}, \ldots, X_{An} \ \sim \ N(\mu_A, \sigma^2)$ i.i.d.,
Group B: $X_{B1}, \ldots, X_{Bn} \ \sim \ N(\mu_B, \sigma^2)$ i.i.d..

Null hypothesis:

$$H_0 \ : \ \theta \quad = \quad \mu_A - \mu_B \ = \ 0.$$

Alternative:

$$H_1 : \theta \quad = \quad \mu_A - \mu_B \ \neq \ 0.$$

Standardized test statistic:

$$Z \ = \ \frac{1}{\sqrt{2n\sigma^2}}\Big(\sum_{i=1}^{n} X_{Ai} - \sum_{i=1}^{n} X_{Bi}\Big) \quad \sim \quad N\Big(\theta\sqrt{n/(2\sigma^2)}, 1\Big).$$

Reject H_0 if $|Z| \ \geq \ z_{1-\alpha^*/2}$ (=1.96 for $\alpha^* = 0.05$).

Pocock's design with one interim at 50% of information

Pocock (1977).

Accrue patients in two consecutive stages with $m = n/2$ patients per arm (trial total: $2n$ patients).

Standardized test statistic after first stage:

$$Z_1 = \frac{1}{\sqrt{2m\sigma^2}}\Big(\sum_{i=1}^{m} X_{Ai} - \sum_{i=1}^{m} X_{Bi}\Big).$$

Standardized test statistic at end of trial:

$$Z_2 = \frac{1}{\sqrt{2n\sigma^2}}\Big(\sum_{i=1}^{n} X_{Ai} - \sum_{i=1}^{n} X_{Bi}\Big).$$

Pocock's design: Reject H_0 if $|Z_k|$ exceeds **constant boundary** $C = C(K, \alpha^*)$.

K: number of planned analyses. α^*: overall significance level.

Pocock's design with one interim at 50% of information

Pocock's design: Reject H_0 if $|Z_k|$ exceeds **constant boundary** $C = C(K, \alpha^*)$.

After first stage – **interim** analysis:

- If $|Z_1| \ \geq \ C$ stop trial and reject H_0.
- Otherwise: continue trial, randomize $m = n/2$ additional patients per arm.

At end of trial, if it was not stopped prematurely – **final** analysis:

- If $|Z_2| \ \geq \ C$ reject H_0.
- Otherwise do not reject H_0.

Crucial question: Choice of critical value C so that this is a **level** α^* **test**?

Joint distribution of test statistics (for $m = n/2$)

Define **independent stagewise $Z-$statistics** Z_{stage1} and Z_{stage2} (including only data from each stage):

$$\begin{aligned} Z_{stage1} &= \frac{1}{\sqrt{2m\sigma^2}}\Big(\sum_{i=1}^{m} X_{Ai} - \sum_{i=1}^{m} X_{Bi}\Big) \\ Z_{stage2} &= \frac{1}{\sqrt{2m\sigma^2}}\Big(\sum_{i=m+1}^{n} X_{Ai} - \sum_{i=m+1}^{n} X_{Bi}\Big) \end{aligned}$$

Then the **full standardized test statistics** Z_1 and Z_2 (including all data up to each stage) satisfy:

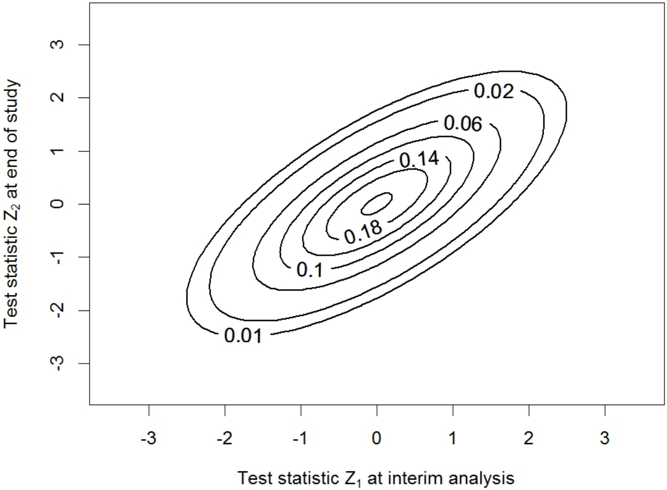
$$\begin{aligned} Z_1 &= Z_{stage1} \\ Z_2 &= 2^{-1/2} \cdot Z_{stage1} + 2^{-1/2} \cdot Z_{stage2} \end{aligned}$$

$\Rightarrow$ Joint distribution of Z_1 and Z_2 under H_0 is **multivariate normal**:

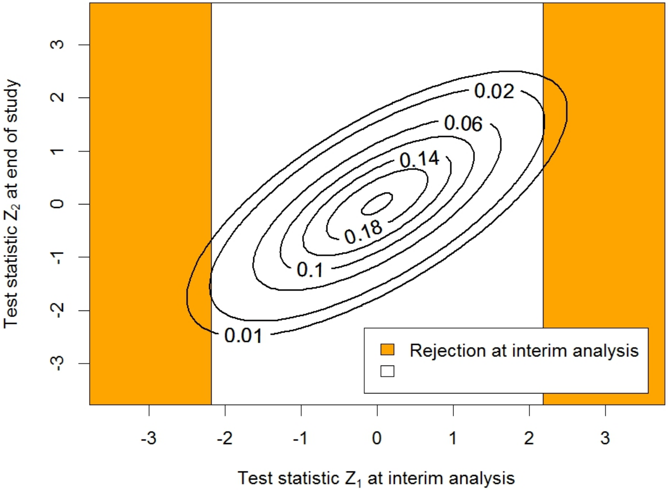
$$(Z_1, Z_2) \quad \sim \quad N\left(\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ \Sigma = \begin{pmatrix} 1 & 2^{-1/2} \\ 2^{-1/2} & 1 \end{pmatrix}\right).$$

Note: **Variance assumed known** in derivation.

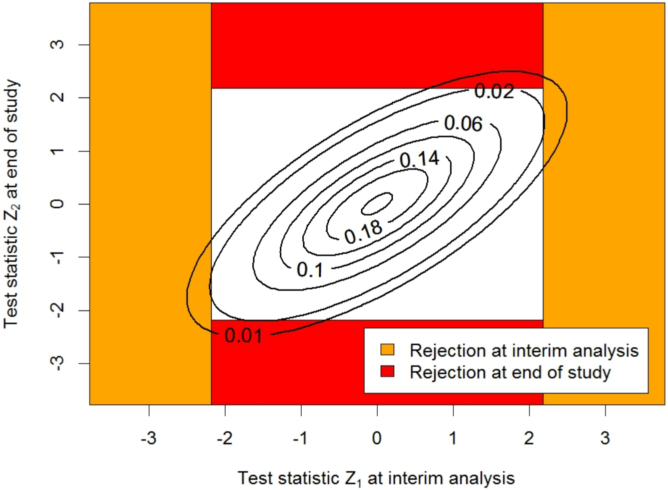
Joint density under H_0



Joint density under H_0



Joint density under H_0



Choice of critical value C

Choose C such that **family-wise error rate** (FWER) α^* is kept:

$$\begin{aligned} P_{H_0}(|Z_1| \geq C) + P_{H_0}(|Z_1| < C \text{ and } |Z_2| \geq C) &= P_{H_0}(|Z_1| \geq C \text{ or } |Z_2| \geq C) \\ &\leq \alpha^*. \end{aligned}$$

Computation using joint distribution and R package **mvtnorm** Genz et al. (2021):

```
> library(mvtnorm)
> alpha <- 0.05
> mu <- c(0, 0)
> Sigma <- matrix(c(1, 1 / sqrt(2), 1 / sqrt(2), 1), ncol = 2)
> q1 <- qmvnorm(1 - alpha / 2, mean = mu, sigma = Sigma)$quantile
> c(q1, 2 * (1 - pnorm(q1)))     ## quantile and significance level multivariate Normal

[1] 2.17831034 0.02938294

> q2 <- qnorm(1 - alpha / 4)
> c(q2, 2 * (1 - pnorm(q2)))     ## quantile and significance level Bonferroni

[1] 2.241403 0.025000
```

Jennison and Turnbull (2000): Chapter 19, "Numerical Computations for Group Sequential Tests".

Computation with R-package rpact

```
> library(rpact)
> ### Pocock design with one interim at 50% information
> design1 <- getDesignGroupSequential(typeOfDesign = "P",      # Pocock("P")
+                                     kMax = 2,                # Number of stages
+                                     sided = 2, alpha = 0.05, beta = 0.2)
> summary(design1)
```

```
Sequential analysis with a maximum of 2 looks (group sequential design)

Pocock design, two-sided overall significance level 5%, power 80%,
undefined endpoint, inflation factor 1.1104, ASN H1 0.8529, ASN H01 1.0388,
ASN H0 1.0941.

Stage                                 1      2
Information rate                    50%   100%
Efficacy boundary (z-value scale)  2.178  2.178
Stage Levels                      0.0147 0.0147
Cumulative alpha spent            0.0294 0.0500
Overall power                     0.4638 0.8000
```

General formulation of Pocock's design

Patients grouped into $k = 1, \ldots, K$ consecutive stages of equal size, m patients per arm in each stage.

Standardized test statistic after k-th stage:

$$Z_k \quad = \quad (2mk\sigma^2)^{-1/2}\Big(\sum_{i=1}^{mk} X_{Ai} - \sum_{i=1}^{mk} X_{Bi}\Big).$$

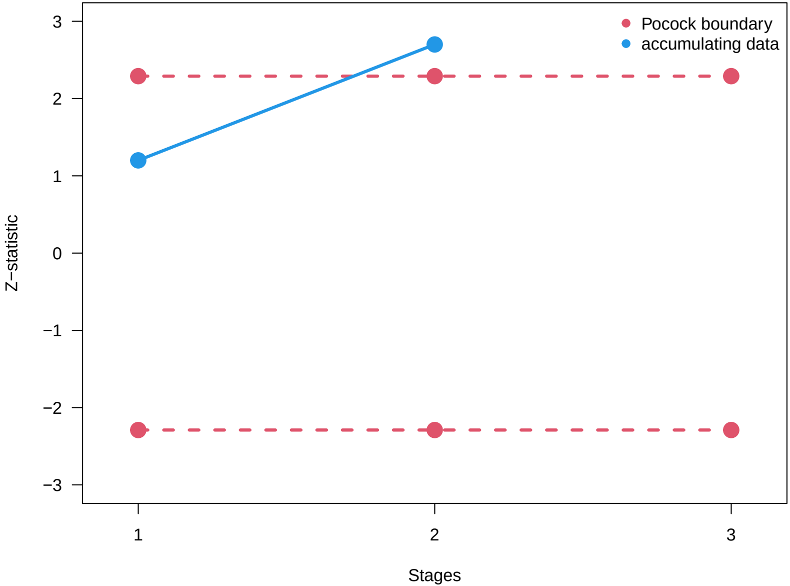
Pocock: Reject H_0 if Z_k crosses **constant** boundary C:

- At each stage: if $|Z_k| \geq C$ stop trial, reject H_0. Otherwise, continue to next stage.
- After final stage K: if $|Z_K| \geq C$ reject H_0, otherwise do not reject.

Choose C such that trial maintains **FWER**:

$$P_{H_0}(|Z_k| \geq C \text{ for at least one } k) \quad \leq \quad \alpha^*.$$

Possible trial result



Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Sample size for Pocock's design

Trial must have power $1 - \beta$ for specific alternative $H_1 \ : \ \mu_A - \mu_B \ = \ \theta$
$\Rightarrow$ choose sample size n such that

$$P_{H_1 : \mu_A - \mu_B \ = \ \theta}(|Z_k| \geq C \text{ for at least one } k) \quad \geq \quad 1 - \beta.$$

Facts:

- Sample size n: depends on α^*, β, **inversely proportional** to $(\theta/\sigma)^2$.
- Sample size n_{groupseq}: depends on **K**, α^*, β, **inversely proportional** to $(\theta/\sigma)^2$.

Consequence:

- Sufficient to tabulate **sample size inflation factor** $R(K, \alpha^*, \beta) = n_{\textbf{groupseq}}/n$, see e.g. Jennison and Turnbull (2000) (Table 2.2).
- Alternative: Let **rpact** do the work.

Example: Cholesterol reduction trial

RCT of a new drug (A) vs. “standard of care” (treatment B).

Primary endpoint: Reduction in cholesterol 4 weeks after randomization.

Assumptions:

- **Normal** data, **known variance** $\sigma^2 = 0.5$.
- $\alpha^* = 0.05$, 90% power to detect a difference between arms of $\theta = 0.3$ mmol/l.
- **Pocock** design with **K=3**.

Sample size calculation for conventional trial without interim analysis gives n=**234** subjects in total (rounded).

Jennison and Turnbull (2000) (Table 2.2) gives **sample size inflation factor** $R(K = 3, \alpha^* = 0.05, \beta = 0.9) = \mathbf{1.151}$.

$\Rightarrow$ **Sample size for Pocock with** $K = 3$: $n_{groupseq} = 1.151 \cdot n \approx 1.151 \cdot 234 \approx \mathbf{270}$.

Sample size for cholesterol trial with rpact

```
> design2 <- getDesignGroupSequential(typeOfDesign="P", kMax=3, sided=2, alpha=0.05, beta=0.1)
> ss2 <- getSampleSizeMeans(design=design2, alternative=0.3, stDev=sqrt(0.5), normalApproximation=TRUE)
> summary(ss2)
```

```
Sample size calculation for a continuous endpoint

Sequential analysis with a maximum of 3 looks (group sequential design), overall
significance level 5% (two-sided).
The sample size was calculated for a two-sample t-test (normal approximation),
H0: mu(1) - mu(2) = 0, H1: effect = 0.3, standard deviation = 0.707, power 90%.

Stage                                         1      2      3
Information rate                          33.3%  66.7%   100%
Efficacy boundary (z-value scale)         2.289  2.289  2.289
Overall power                            0.3890 0.7311 0.9000
Expected number of subjects               168.4
Number of subjects                         89.6  179.1  268.7
Cumulative alpha spent                   0.0221 0.0379 0.0500
Two-sided local significance level       0.0221 0.0221 0.0221
Lower efficacy boundary (t)              -0.342 -0.242 -0.198
Upper efficacy boundary (t)               0.342  0.242  0.198
Exit probability for efficacy (under H0) 0.0221 0.0159
Exit probability for efficacy (under H1) 0.3890 0.3421

Legend:
  (t): treatment effect scale
```

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Local significance levels $\neq$ exit probabilities under H_0

Group sequential trial with **global (overall) significance level** α^* which rejects H_0 after stage k if $Z_k \geq C_k$ (with $C_k = C$ for Pocock's design).

Local significance level α_k corresponding to stage k satisfies

$$\alpha_k \quad = \quad P_{H_0}(|Z_k| \geq C_k) \;=\; 2(1 - \Phi(C_k)).$$

Exit probability π_k **under** H_0 or **error spent** at stage k:

$$\pi_k \quad = \quad P_{H_0}(|Z_1| \leq C_1, \ldots, |Z_{k-1}| \leq C_{k-1}, |Z_k| > C_k).$$

Which of these sum up to α^*?

$$\sum_{k=1}^{K} \alpha_k \;>\; \alpha^* \qquad \sum_{k=1}^{K} \pi_k \;=\; \alpha^*.$$

Slud and Wei (1982), Jennison and Turnbull (2000) (p. 147).

Exit probabilities under H_0 and H_1 for the cholesterol trial

k	exit probability under H_0	exit probability under H_1
1	0.022	0.39
2	0.016	0.34
3	0.012	0.17
Sum	0.05 (type I error)	0.90 (power)

Note that the probability to enter final stage K is
$1 - (0.022 + 0.016) = 0.962$ under H_0 and
$1 - (0.39 + 0.34) = 0.27$ under H_1.

Expected sample sizes for cholesterol trial

Expected sample size: mean sample size when boundary is crossed or trial proceeds to last analysis.

Single stage trial: **234** patients are accrued in any case.
Sample sizes Pocock 3-stage trial:

- If trial runs to end, i.e. **maximal** sample size: **270**.
- **Expected under H_0**:

$$0.022 \cdot 90 + 0.016 \cdot 180 + (1 - (0.022 + 0.16)) \cdot 270 \quad = \quad \mathbf{265}.$$

- **Expected under H_1**:

$$0.39 \cdot 90 + 0.34 \cdot 180 + (1 - (0.39 + 0.34)) \cdot 270 \quad = \quad \mathbf{170}.$$

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

O'Brien-Fleming's design

Cholesterol trial (Pocock's design): compare p-value at the end to $\alpha = 0.022$.

- Difficult to communicate to clinicians.
- If trial stopped early, effect size might not be overwhelming. Balance with amount of follow-up.

O'Brien-Fleming: Similar to Pocock, **but** different stopping boundaries: H_0 rejected if

$$|Z_k| \geq D\sqrt{\mathbf{K/k}}.$$

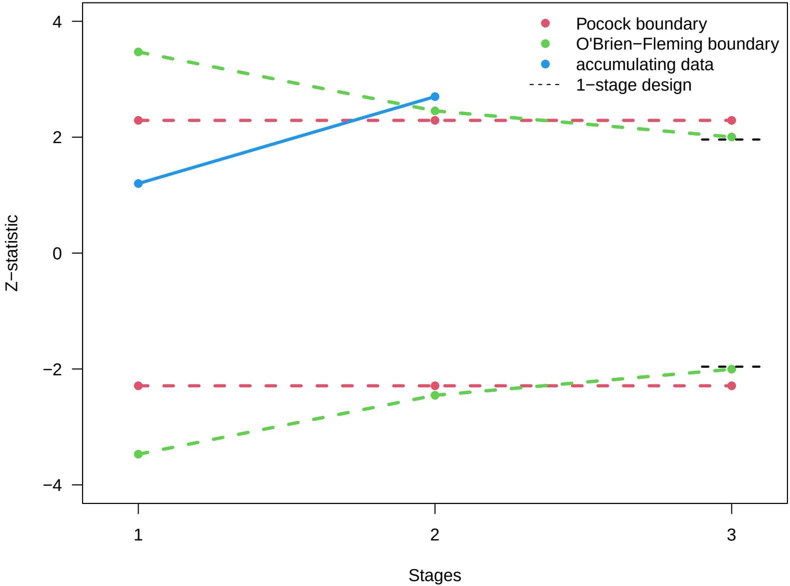
Find D to keep FWER α^* as in Pocock's design.

For $K = 3$, the boundaries are:

k	**Pocock**			**O'Brien-Fleming**		
	critical value	significance level	exit probability	critical value	significance level	exit probability
1	2.29	0.022	0.022	3.47	0.0005	0.0005
2	2.29	0.022	0.016	2.45	0.014	0.014
3	2.29	0.022	0.012	2.00	0.045	0.036

Maximal sample size for cholesterol trial: 234 (single-stage), 238 (O'Brien-Fleming).

Pocock vs. O’Brien-Fleming boundaries



Features of designs

O'Brien-Fleming: Most frequently used design.

- **Conservative at early interims**: only stops if evidence overwhelming.
- Final analysis almost at nominal α^* (0.045 in cholesterol example).
- Maximal sample size only slightly larger than in single-stage design (238 vs. 234).

Pocock:

- Equal (low) level for all interims (0.022 in cholesterol example).
- Higher probability to stop trial early $\Rightarrow$ expected sample size considerably lower than for single-stage.
- **However**: maximal sample size substantially larger compared to single-stage (270 vs. 234).

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Limitations

Limitations of the designs introduced so far:

- Interims must be performed exactly at **equal information increments**.
- Total number of analyses K **pre-specified**.
- So far only 2-sample z-test (**normal distribution, known variance**) discussed where joint distribution of Z-statistics at interim analyses can be derived exactly.

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. **Extensions of the basic methods**
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Non-normal data - Theory

Let $\widehat{\theta}_k$ denote the **treatment effect estimate** of the true effect θ based on data at stage k from any statistical model.

The **information** for θ at stage k is defined as $\mathcal{I}_k = (\mathrm{Var}(\widehat{\theta}_k))^{-1}$.

For testing $H_0 : \theta = 0$, the corresponding **standardized Z-statistic** at analysis k is

$$Z_k \quad = \quad \frac{\widehat{\theta}_k}{\sqrt{\mathrm{Var}(\widehat{\theta}_k)}} = \widehat{\theta}_k \sqrt{\mathcal{I}_k}.$$

Fact: In many settings, the standardized Z-statistics $(Z_1, \ldots, Z_K)$ have approximately the **canonical joint distribution**:

- $(Z_1, \ldots, Z_K)$ is multivariate normal.
- $Z_k \sim N(\theta\sqrt{\mathcal{I}_k}, 1)$ for $k = 1, \ldots, K$.
- $Cov(Z_{k_1}, Z_{k_2}) = \sqrt{\mathcal{I}_{k_1}/\mathcal{I}_{k_2}}$ for $k_1 < k_2$.

Jennison and Turnbull (2000): Section 3.5 and Chapter 11.

Non-normal data - practice

Theory of group-sequential trials:

- **Applicable to standardized Z-statistics resulting from any model**,
- as long as they have canonical joint distribution.

Canonical joint distribution applies if treatment effect θ estimated based on a general model fitted by **maximum likelihood** (follows from large sample theory).

In particular: theory applies to standard statistical methods for

- normal,
- binary,
- survival,
- longitudinal data Jennison and Turnbull (2000) (p. 67),

including their stratified and covariate-adjusted versions.

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Information fraction and general idea

Pocock, O'Brien-Fleming:

- Interim analysis at equal **information increments**.
- Total number of analysis K **pre-specified**.

In practice, difficult to exactly adhere to the pre-planned timing of interim analyses.

Information fraction t_k at stage k (relative to max. information at final stage K):

- General definition: $t_k = \mathcal{I}_k / \mathcal{I}_{\max}$.
- Normal or binomial endpoint: $t_k = n_k / N_{\max}$ ($n_k =$ number of patients with an outcome by stage k).
- Survival endpoints: $t_k = d_k / d_{\max}$ ($d_k =$ number of events by stage k).

Idea of Lan and DeMets (1983): Define **α-spending function** of the information fraction t_k which describes how type I error is distributed amongst interim analyses.

α-spending function

α-spending function:

- Non-decreasing function of the information fraction.
- $\alpha : [0, 1] \to [0, \alpha^*]$, where α^* is overall significance level of trial.
- At interim k one can spend $\alpha(t_k) - \alpha(t_{k-1})$ of total type I error.

Determine boundaries $C_1, \ldots, C_k$ for standardized test statistics $Z_1, \ldots, Z_K$ at information fractions $t_1, \ldots, t_k$ such that for the **exit probabilities**

$$\begin{aligned} \pi_1 &= P_{H_0}(|Z_1| \geq C_1) &= \alpha(t_1), \\ \pi_k &= P_{H_0}(|Z_1| < C_1, \ldots, |Z_{k-1}| < C_{k-1}, |Z_k| \geq C_k) &= \alpha(t_k) - \alpha(t_{k-1}). \end{aligned}$$

Pocock, O'Brien-Fleming have (approximate) formulation in terms of α-spending functions.

Key advantage:

- **Timing** and **number** of interims can be changed **during trial**.
- Changes **NOT** allowed to be "motivated" based on interim results!
- adaptR tutorial; requires VPN.

Examples of α-spending functions

Pocock (approximately, `typeOfDesign = "asP"`):

$$\alpha_1(t) \quad = \quad \min\{\alpha^*, \alpha^* \log(1 + (e - 1)t)\}.$$

O'Brien-Fleming (approximately, `typeOfDesign = "asOF"`):

$$\alpha_2(t) \quad = \quad \min\{\alpha^*, 2 - 2\Phi(z_{1-\alpha^*/2}/t^{-1/2})\}.$$

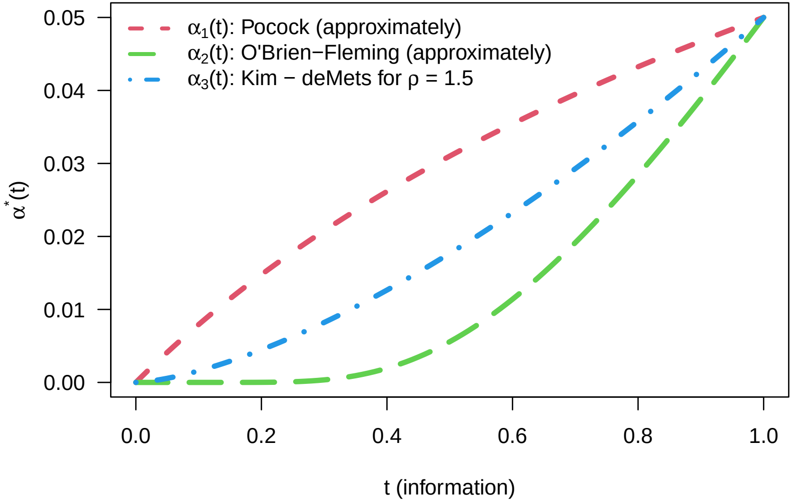
Power-family of Kim and DeMets (1987) (`typeOfDesign = "asKD"`):

$$\alpha_3(t) \quad = \quad \min\{\alpha^*, \alpha^* t^\rho\}.$$

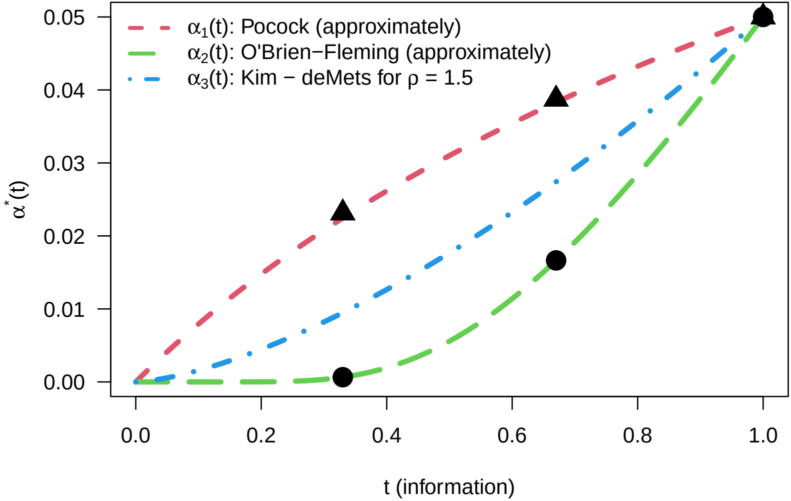
More examples in Section 3.3 of Wassmer and Brannath (2016).

rpact: Specify in `getDesignGroupSequential`.

Examples of α-spending functions



Examples of α-spending functions



Protocol specification for group-sequential trials with α-spending functions

Pre-specify in protocol:

- **Maximal information**: sample size (targeted number of patients or events).
- **Planned timing** of interim analyses (information fractions).
- Chosen **α-spending function**.
- **Boundary values**: local significance levels or Z-values in case of exact adherence to planned interim timing. Approximate boundary values on treatment effect scale (MDD's) also useful.
- Statement that boundary values will be **re-calculated** as per α-spending function in case of deviations from planned timing.

Example: IMbrave150

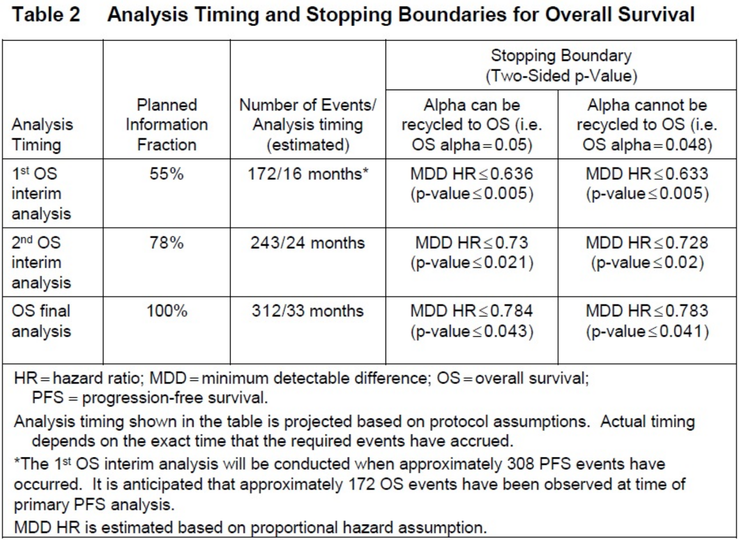


Table 2 Analysis Timing and Stopping Boundaries for Overall Survival

Analysis Timing	Planned Information Fraction	Number of Events/ Analysis timing (estimated)	Stopping Boundary (Two-Sided p-Value)	
			Alpha can be recycled to OS (i.e. OS alpha = 0.05)	Alpha cannot be recycled to OS (i.e. OS alpha = 0.048)
1st OS interim analysis	55%	172/16 months*	MDD HR ≤ 0.636 (p-value ≤ 0.005)	MDD HR ≤ 0.633 (p-value ≤ 0.005)
2nd OS interim analysis	78%	243/24 months	MDD HR ≤ 0.73 (p-value ≤ 0.021)	MDD HR ≤ 0.728 (p-value ≤ 0.02)
OS final analysis	100%	312/33 months	MDD HR ≤ 0.784 (p-value ≤ 0.043)	MDD HR ≤ 0.783 (p-value ≤ 0.041)

HR = hazard ratio; MDD = minimum detectable difference; OS = overall survival; PFS = progression-free survival.

Analysis timing shown in the table is projected based on protocol assumptions. Actual timing depends on the exact time that the required events have accrued.

*The 1st OS interim analysis will be conducted when approximately 308 PFS events have occurred. It is anticipated that approximately 172 OS events have been observed at time of primary PFS analysis.

MDD HR is estimated based on proportional hazard assumption.

link to SAP

Example: IMbrave150

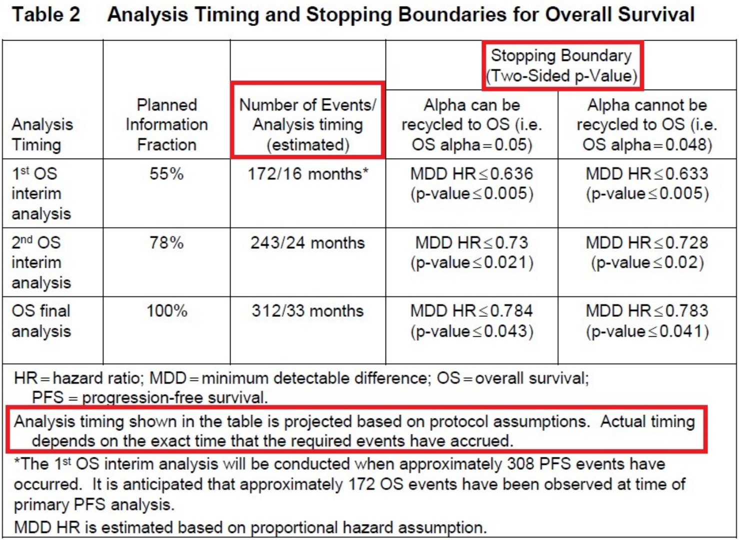


Table 2 Analysis Timing and Stopping Boundaries for Overall Survival

			Stopping Boundary (Two-Sided p-Value)	
Analysis Timing	Planned Information Fraction	Number of Events/ Analysis timing (estimated)	Alpha can be recycled to OS (i.e. OS alpha = 0.05)	Alpha cannot be recycled to OS (i.e. OS alpha = 0.048)
1st OS interim analysis	55%	172/16 months*	MDD HR $\leq$ 0.636 (p-value $\leq$ 0.005)	MDD HR $\leq$ 0.633 (p-value $\leq$ 0.005)
2nd OS interim analysis	78%	243/24 months	MDD HR $\leq$ 0.73 (p-value $\leq$ 0.021)	MDD HR $\leq$ 0.728 (p-value $\leq$ 0.02)
OS final analysis	100%	312/33 months	MDD HR $\leq$ 0.784 (p-value $\leq$ 0.043)	MDD HR $\leq$ 0.783 (p-value $\leq$ 0.041)

HR = hazard ratio; MDD = minimum detectable difference; OS = overall survival; PFS = progression-free survival.

Analysis timing shown in the table is projected based on protocol assumptions. Actual timing depends on the exact time that the required events have accrued.

*The 1st OS interim analysis will be conducted when approximately 308 PFS events have occurred. It is anticipated that approximately 172 OS events have been observed at time of primary PFS analysis.

MDD HR is estimated based on proportional hazard assumption.

link to SAP **O'Brien-Fleming** ⇒ just above table in SAP. rpact exercise

Random under- or over-running

Timing of final analysis may randomly occur at a sligthly lower (**under-running**) or higher (**over-running**) information fraction than planned, i.e. $t_K < 1$ or $t_K > 1$.

Adjustment:

- Re-set information fraction at final analysis to 1. Re-calculate information fractions at interim analysis relative to the actually observed final information.
- Use the α that has actually been spent for interim analyses; spend all remaining α at the final analysis.
- Use updated information fractions and α-spending to calculate new boundary C_K.

Features:

- **Type I error protected** if timing of final analysis **not** "motivated" by interim results!
- **Power below or above its target** when there is under- or over-running.

Wassmer and Brannath (2016): Chapter 3.3, "The α-Spending Function Approach". Worked-out example:

`https://vignettes.rpact.org/html/rpact_boundary_update_example.html`.

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Futility interim

Stop trial early if drug does not work.

We look into data multiple times. Still, no adjustment of overall significance level α^* needed. **Why?**

- Why do we adjust α^* when doing multiple testing?
- Because of **inflated** type I error: declare drug as working although it does not work.
- Stopping a trial for futility: we claim drug does **not** work. No type I error possible - no adjustment of α^* needed.

No free lunch: occassionally, a trial for a working drug will be stopped for futility ⇒ adding futility analysis **reduces study power**.

Stopping for futility

Informal futility boundary:

- Formal interim boundaries for efficacy only.
- Informal rules for futility stopping (e.g. if estimate in ”wrong direction”, low conditional power, no signal in ”early” secondary endpoints, etc.).
- Stopping for futility **does not inflate the type I error rate** but **reduces power**.

Non-binding formal futility boundary:

- Formal interim boundaries for efficacy (α-spending) and futility (fixed bounds or β-spending).
- Type I error protected even if futility boundary is ignored.
- Sample size calculation (conservatively) assumes that futility boundary will be obeyed ⇒ **need sample size increase to maintain power.**

Binding (”mandatory”) futility boundary:

- Type I error only protected if futility boundary is adhered to.
- **Not recommended:** Power gain usually small, IDMC ”forced” to stop trial, not accepted by health authorities.

Cholesterol example without futility interim analysis

```
> design2 <- getDesignGroupSequential(typeOfDesign = "P", kMax = 3, sided = 2, alpha = 0.05, beta = 0.1)
> ss2 <- getSampleSizeMeans(design = design2, alternative = 0.3, stDev = sqrt(0.5),
+                           normalApproximation = TRUE)
> summary(ss2)
```

```
Sample size calculation for a continuous endpoint

Sequential analysis with a maximum of 3 looks (group sequential design), overall
significance level 5% (two-sided).
The sample size was calculated for a two-sample t-test (normal approximation),
H0: mu(1) - mu(2) = 0, H1: effect = 0.3, standard deviation = 0.707, power 90%.

Stage                                         1      2      3
Information rate                          33.3%  66.7%   100%
Efficacy boundary (z-value scale)         2.289  2.289  2.289
Overall power                            0.3890 0.7311 0.9000
Expected number of subjects               168.4
Number of subjects                         89.6  179.1  268.7
Cumulative alpha spent                   0.0221 0.0379 0.0500
Two-sided local significance level       0.0221 0.0221 0.0221
Lower efficacy boundary (t)              -0.342 -0.242 -0.198
Upper efficacy boundary (t)               0.342  0.242  0.198
Exit probability for efficacy (under H0) 0.0221 0.0159
Exit probability for efficacy (under H1) 0.3890 0.3421

Legend:
  (t): treatment effect scale
```

Cholesterol example with non-binding futility interim analysis

```
> ## futility boundary at Z = 0 for both interims (futility for estimates in "wrong" direction)
> design3 <- getDesignGroupSequential(typeOfDesign = "P", kMax = 3, sided = 1, alpha = 0.025,
+                                     beta = 0.1, futilityBounds = c(0, 0), bindingFutility = FALSE)
> ss3 <- getSampleSizeMeans(design = design3, alternative = 0.3, stDev = sqrt(0.5),
+                           normalApproximation = TRUE); summary(ss3)

Sample size calculation for a continuous endpoint

Sequential analysis with a maximum of 3 looks (group sequential design), overall
significance level 2.5% (one-sided).
The sample size was calculated for a two-sample t-test (normal approximation),
H0: mu(1) - mu(2) = 0, H1: effect = 0.3, standard deviation = 0.707, power 90%.

Stage                                         1      2      3
Information rate                          33.3%  66.7%   100%
Efficacy boundary (z-value scale)         2.289  2.289  2.289
Futility boundary (z-value scale)             0      0
Overall power                            0.4011 0.7439 0.9000
Expected number of subjects               167.5
Number of subjects                         92.4  184.8  277.1
Cumulative alpha spent                   0.0110 0.0190 0.0250
One-sided local significance level       0.0110 0.0110 0.0110
Efficacy boundary (t)                     0.337  0.238  0.194
Futility boundary (t)                         0      0
Overall exit probability (under H0)      0.5110 0.1328
Overall exit probability (under H1)      0.4218 0.3436
Exit probability for efficacy (under H0) 0.0110 0.0079
Exit probability for efficacy (under H1) 0.4011 0.3428
```

Cholesterol example: comparison of design without and with non-binding futility interim

```
> ## Design without futility interim
> ceiling(t(as.data.frame(ss2)[1, c("maxNumberOfSubjects", "expectedNumberOfSubjectsH0",
+                                   "expectedNumberOfSubjectsH1")]))

                              1
maxNumberOfSubjects         269
expectedNumberOfSubjectsH0  264
expectedNumberOfSubjectsH1  169
```

```
> ## Design with futility interim
> ceiling(t(as.data.frame(ss3)[1, c("maxNumberOfSubjects", "expectedNumberOfSubjectsH0",
+                                   "expectedNumberOfSubjectsH1")]))

                              1
maxNumberOfSubjects         278
expectedNumberOfSubjectsH0  171
expectedNumberOfSubjectsH1  168
```

Non-binding futility boundary **slightly increases maximal sample size** for 90% power but **dramatically reduces expected sample size under H0**.

More on futility analyses in toolbox course, Module 1.

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

What does *stopping a trial for efficacy* mean?

Stopping for efficacy - not an automatic decision!

Decision to prematurely stop trial ⇒ **not based on statistical criteria alone**:

- **Robust** and clinically convincing. Sensitivity analyses.
- Data should be sufficiently **mature**, i.e. have enough follow up: new drug might be more effective early, but not in the long run.
- All patients should have received treatment: if not ⇒ ethical imperative to allow for cross-over of control patients ⇒ makes estimation of long-term effect estimates, e.g. overall survival, difficult.

Studies stopped too early for success might not have accumulated sufficient safety information, regulators are more concerned with safety than efficacy.

Van Norman (2019)

What does *stopping a trial for efficacy* mean?

Statistically:

- **Reject null hypothesis** of ”no effect of drug” in hypothesis test.
- (Typically) Unblind trial and **file**.

Operationally:

- Trial continues as before: patients finish treatment, remain on assessment schedule.
- Data collection might be reduced: IRC-PFS only necessary for approval - that's done!
- Other **efficacy and safety** data remains important: survival follow-up, long-term follow-up of primary endpoint and safety. We will keep taking **follow-up snapshots**!

Is there a *correct* snapshot?

Which one is the "true", "correct" snapshot then?

None - they serve **different purposes**!

- **Efficacy interim snapshot**: decision about hypothesis. You can only reject a hypothesis once!
- Results cannot get "more significant" with follow-up snapshots, even if p-value becomes smaller!
- **Follow-up snapshots**: more data $\Rightarrow$ more precise estimate of effect size. Give effect estimate and confidence interval.

Risk of bias through **unblinding**!

If we do not stop at efficacy interim? **Trial can still be a success**! PTS goes down though.

What does *stopping a trial for futility* mean?

What does *stopping a trial for futility* mean?

Low probability you reject null hypothesis at final analysis ⇒ stop trial now.

- **Save resources**. Maybe not for this trial (often lots of $$$ already spent), but may reallocate resources.
- **Prevent further exposure** of patients to new therapy.
- Inform other programs.

If we do not stop at futility interim? **Trial can still be a failure**! PTS goes up!

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Survival data

Effect of interest: log hazard ratio $\theta = \log \lambda$.

Test: logrank.

Sequence of logrank test statistics has **approximately** canonical joint distribution.

- Conditional on observed variance at each interim ⇒ conditional on observed **number of events**.
- Why "approximately"? Logrank test statistic only approximately normal even without group-sequential.
- Generalizes to **Cox regression**.

What if we use another test than logrank (e.g. to increase power under non-proportional hazards)?

- Use test statistic that has canonical joint Normal distribution.
- Weighted logrank potentially fails to have this, if entry is staggered (increments are NOT independent, info fraction not proportional to #events).
- Hasegawa (2016), Roychoudhury et al. (2021), Wang et al. (2021), Course by Keaven Andersen.

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
    - Pocock's design
    - Sample size for Pocock's designs
    - Exit probabilities and expected sample size
    - O'Brien-Fleming's design
    - Limitations
3. Extensions of the basic methods
    - Non-normal data
    - Flexible timing of interim analyses
    - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. **Case study: GALLIUM randomized clinical trial**
7. Inference in group-sequential designs
    - Illustration of the problem
    - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
    - Introduction
    - Assumption-free methods which control type I error
    - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
    - Type I error inflation of naive stagewise hierarchical strategy
    - Methods which protect type I error
10. Take home messages

GALLIUM

- **Population**: Treatment-naive follicular lymphoma (FL) patients.
- **Comparison**: Rituximab + chemotherapy vs. **Obinutuzumab** + chemotherapy.
- Rituximab: Rituxan, Mabthera. Obinutuzumab: Gazyva(ro).
- **Phase III, 1:1 randomized, open-label** clinical trial.
- Primary endpoint: investigator-assessed **progression-free survival**.
- 1202 patients.

Marcus et al. (2017), NEJM.

Gallium: planned interim analyses

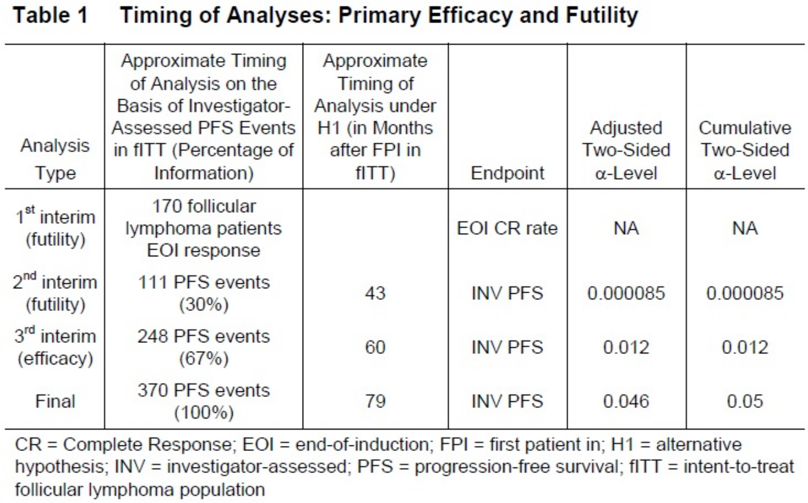


Table 1 Timing of Analyses: Primary Efficacy and Futility

Analysis Type	Approximate Timing of Analysis on the Basis of Investigator-Assessed PFS Events in fITT (Percentage of Information)	Approximate Timing of Analysis under H1 (in Months after FPI in fITT)	Endpoint	Adjusted Two-Sided α-Level	Cumulative Two-Sided α-Level
1st interim (futility)	170 follicular lymphoma patients EOI response		EOI CR rate	NA	NA
2nd interim (futility)	111 PFS events (30%)	43	INV PFS	0.000085	0.000085
3rd interim (efficacy)	248 PFS events (67%)	60	INV PFS	0.012	0.012
Final	370 PFS events (100%)	79	INV PFS	0.046	0.05

CR = Complete Response; EOI = end-of-induction; FPI = first patient in; H1 = alternative hypothesis; INV = investigator-assessed; PFS = progression-free survival; fITT = intent-to-treat follicular lymphoma population

Described in the statistical analysis plan ⇒ **pre-specification**.

Recommendation to add **MDD**'s to table.

Gallium: conducted interim analyses

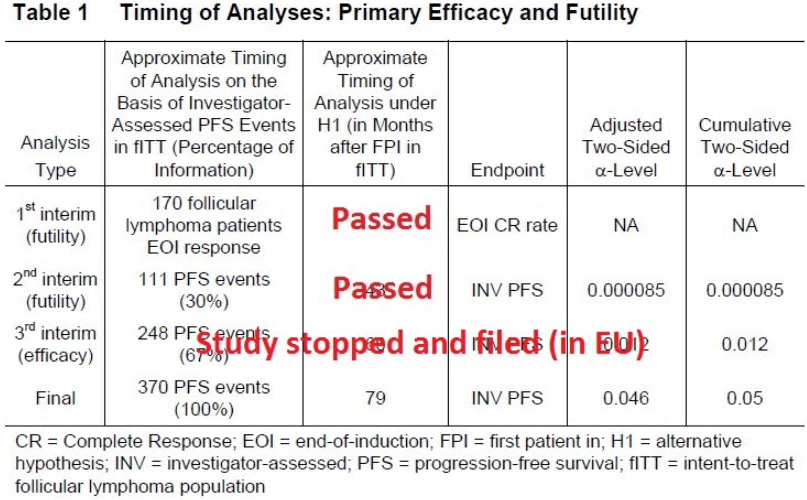


Table 1 Timing of Analyses: Primary Efficacy and Futility

Analysis Type	Approximate Timing of Analysis on the Basis of Investigator-Assessed PFS Events in fITT (Percentage of Information)	Approximate Timing of Analysis under H1 (in Months after FPI in fITT)	Endpoint	Adjusted Two-Sided α-Level	Cumulative Two-Sided α-Level
1st interim (futility)	170 follicular lymphoma patients EOI response	Passed	EOI CR rate	NA	NA
2nd interim (futility)	111 PFS events (30%)	Passed 43	INV PFS	0.000085	0.000085
3rd interim (efficacy)	248 PFS events (67%)	60	INV PFS	0.012	0.012
Final	370 PFS events (100%)	79	INV PFS	0.046	0.05

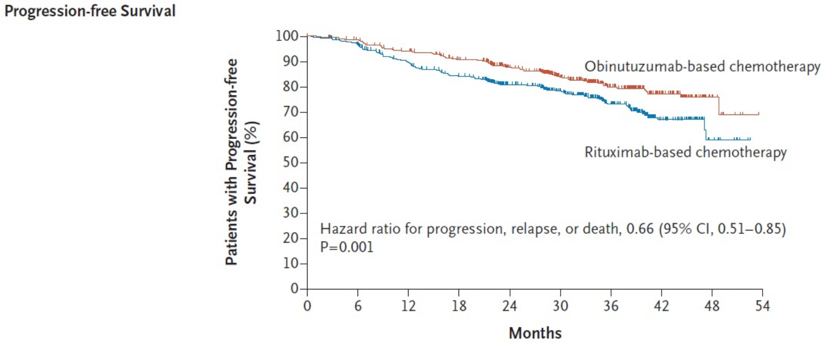
Study stopped and filed (in EU)

CR = Complete Response; EOI = end-of-induction; FPI = first patient in; H1 = alternative hypothesis; INV = investigator-assessed; PFS = progression-free survival; fITT = intent-to-treat follicular lymphoma population

Described in the statistical analysis plan ⇒ **pre-specification**.

Recommendation to add **MDD**'s to table.

GALLIUM: results at efficacy interim


Progression-free Survival

No. at Risk

Obinutuzumab-based chemotherapy	601	570	536	502	405	278	168	75	13	0
Rituximab-based chemotherapy	601	562	505	463	378	266	160	68	10	0

How to perform inference accounting for early stopping?

Futility interim analyses

1st futility interim on **response proportions**.

2nd futility interim on **primary endpoint PFS** with futility boundary: hazard ratio $= 1$.

Distribution of effect estimate $\widehat{\theta} = \log(\text{HR})$ at futility interim: $\widehat{\theta} \sim N(\theta, \sqrt{4/111})$.

θ: log of considered hazard ratio:

- Under H_0: $\log(1) = 0$,
- under H_1, i.e. used to power study: $\log(0.741)$.

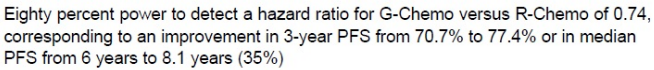
Stopping probabilities: $P_{H_0}(\widehat{\theta} > 0)$ and $P_{H_1}(\widehat{\theta} > 0)$.

Why spend α^* at futility interim?

FDA does not trust you that you continue trial if you see overwhelmingly positive result at futility interim.

Gallium: initial assumption vs. MDD at efficacy interim

Initial assumption used for sample size computation:



Eighty percent power to detect a hazard ratio for G-Chemo versus R-Chemo of 0.74, corresponding to an improvement in 3-year PFS from 70.7% to 77.4% or in median PFS from 6 years to 8.1 years (35%)

Minimal detectable difference (MDD):

- Largest observed hazard ratio for which Gallium will **just be significant**, i.e. give a *p*-value of α.
- MDD is **analysis-dependent**:
  - Significance level α different at interim and final.
  - MDD depends on standard error $\Rightarrow$ number of events analysis is performed at.
- Efficacy interim: $\alpha = 0.012, d = 248 \Rightarrow \text{MDD} = \mathbf{0.728}$.
- Final analysis: $\alpha = 0.046, d = 370 \Rightarrow \text{MDD} = \mathbf{0.813}$.
- Compare MDDs to **0.74** used for powering:
  - MDDs say something about **null hypothesis**.
  - Effect for powering is specification of **alternative** hypothesis.

Expected number of events

Single-stage design: 349 needed in any case.

Probability to stop after respective stage with one **futility** and one **efficacy** interim:

Analysis	# events	No effect, i.e.	Effect size to
		under H_0	have 80% power
futility interim	111	0.500	0.066
efficacy interim	248	0.006	0.434
final	370	(1 - 0.500 - 0.006)	(1 - 0.066 - 0.434)
		= 0.494	= 0.500

Expected number of events (**ignoring** interim on response!):

- Under H_0: $0.500 \cdot 111 + 0.006 \cdot 248 + 0.494 \cdot 370 =$ **240**.
- Under H_1: $0.066 \cdot 111 + 0.434 \cdot 248 + 0.500 \cdot 370 =$ **300**.

Conclusions: compared to single-stage design,

- if H_1 is true, group-sequential needs on **average** 349 - 300 = 49 or 100 · (1 - 300 / 349)% = 14.1% less events to show same effect.
- if H_0 is true, group-sequential needs on **average** 349 - 240 = 109 or 100 · (1 - 240 / 349)% = 31.3% less events.

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. **Inference in group-sequential designs**
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Inference in group-sequential trials

So far: focus on how to **design** group-sequential **hypothesis test**.

Once data has been collected ⇒ we want more than just "reject/do not reject H_0":

- point estimates,
- confidence intervals,
- p-values.

Two types of inference:

1. After decision on H_0 (reject/do not reject) at interim or final.
2. Estimate of treatment effect at an interim or final, irrespective of decision.
   - **Repeated** confidence intervals, p-values, etc. Simply invert GSD test at each interim.
   - **Blinded** trial: maybe of interest to iDMC, but not sponsor.

   Monitoring tool.

In pivotal trials, 1) above is relevant.

What is the problem?

Selective nature of sampling procedure: samples leading to termination at interim are "random highs", "extreme by nature". Wassmer and Brannath (2016)

Analysis time for a group-sequential trial:

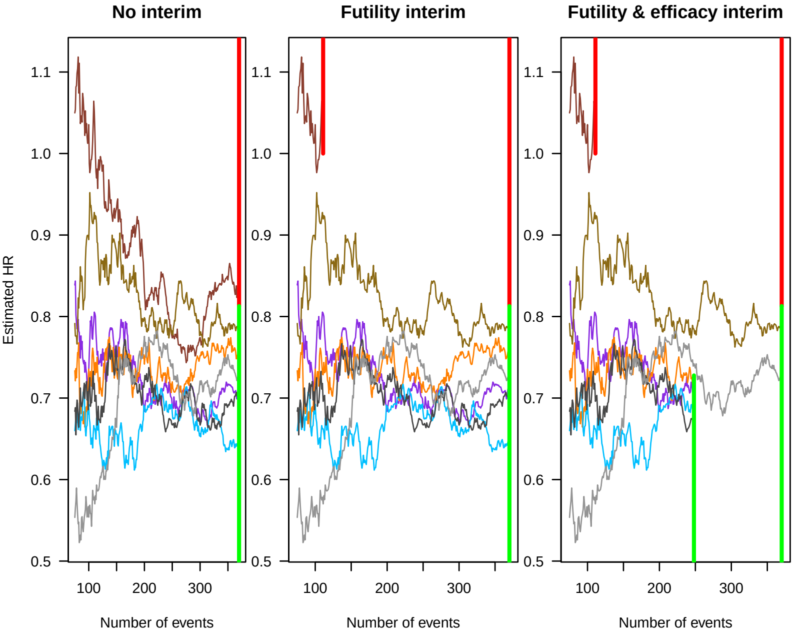
- when stopping boundary is crossed ($k < K$) or
- when final stage is reached ($k = K$).

Standard but **naive** inference: Just "take what you have" once you stop:

- point estimates are **biased**,
- confidence intervals **may not have right coverage**.
- Since stopping boundaries do not influence derivation of maximum likelihood estimate $\Rightarrow$ naive estimates remain to be MLE, but **lose other properties**, such as e.g. (asymptotic) unbiasedness.

Why biased? How much? Relevant? When?

Why? We truncate trajectories at interim(s)



Two types of bias

Conditional bias: $E(\widehat{\theta} \mid \text{stop at stage } k) - \theta_{true}$

- Bias of an estimate in a trial **conditional on the stage at which the trial stopped**.
- Sampling distribution of estimate has sub-density restricted to rejection region.
- Ohman Strickland and Casella (2003), Fan et al. (2004), Freidlin and Korn (2009), Shimura et al. (2017) (time-to-event).

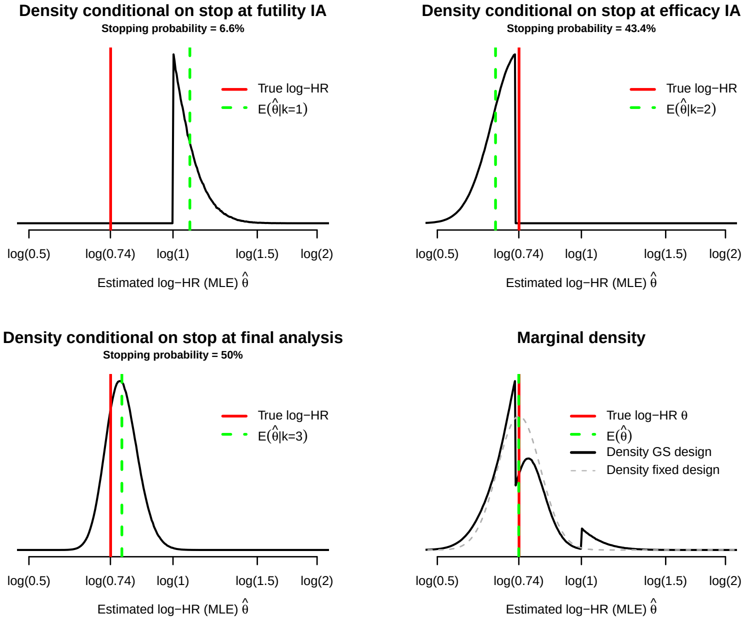
Unconditional or **global** bias: $E(\widehat{\theta}) - \theta_{true}$

- Bias of estimate if **trial is repeated many times**.
- Density of estimate is mixture of conditional (truncated Normal) densities.
- Fan and Demets (2006).

The sizes of both biases **depend on true parameter**.

The conditional bias can be severe at every stage, but they typically "average out" to much smaller unconditional bias ⇒ Gallium example below.

Densities of the MLE for Gallium under H1 (HR = 0.74)



Density conditional on stop at futility IA

Stopping probability = 6.6%

Density conditional on stop at efficacy IA

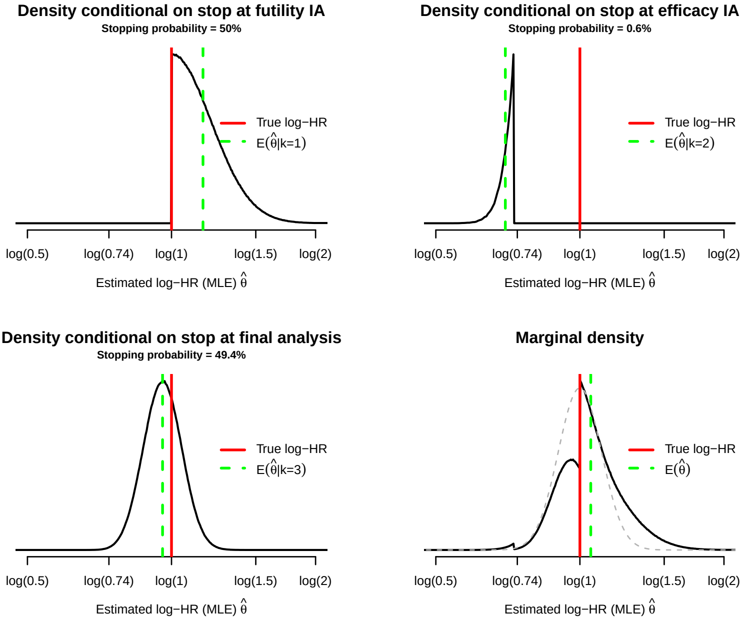
Stopping probability = 43.4%

Density conditional on stop at final analysis

Stopping probability = 50%

Marginal density

Densities of the MLE for Gallium under H0 (HR = 1)



Density conditional on stop at futility IA

Stopping probability = 50%

Density conditional on stop at efficacy IA

Stopping probability = 0.6%

Density conditional on stop at final analysis

Stopping probability = 49.4%

Marginal density

How large is bias in practice?

Based on **simulation studies:**

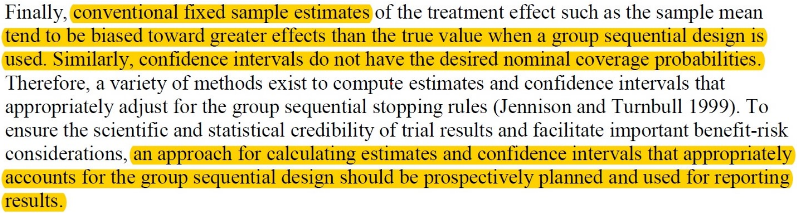
> *For trials with a well-designed interim-monitoring plan, stopping after 50% or more events had been collected has a* **negligible impact** *on estimation.*

Freidlin and Korn (2009)

> *Group sequential designs with stopping rules seek to minimize exposure of patients to a disfavored therapy and speed dissemination of results, and* **such designs do not lead to materially biased estimates***. . . . Superiority demonstrated in a randomized clinical trial stopping early and designed with appropriate statistical stopping rules is* **likely a valid inference***, even if the estimate may be slightly inflated.*

Wang et al. (2016)

FDA guidance



Finally, conventional fixed sample estimates of the treatment effect such as the sample mean tend to be biased toward greater effects than the true value when a group sequential design is used. Similarly, confidence intervals do not have the desired nominal coverage probabilities. Therefore, a variety of methods exist to compute estimates and confidence intervals that appropriately adjust for the group sequential stopping rules (Jennison and Turnbull 1999). To ensure the scientific and statistical credibility of trial results and facilitate important benefit-risk considerations, an approach for calculating estimates and confidence intervals that appropriately accounts for the group sequential design should be prospectively planned and used for reporting results.

FDA guidance on "Adaptive Designs for Clinical Trials of Drugs and Biologics"

U.S. Food and Drug Administration (2019).

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Inference upon early stopping or completion of a group-sequential trial

Several methods proposed, see e.g. Wassmer and Brannath (2016) (Chapter 4) and Grayling and Wason (2022).

Few available in standard software.

p-values

Sample space Ω of group-sequential design for survival endpoint:

- all pairs (k, θ) with k: stage, θ: log hazard ratio,
- such that θ is extreme enough to reject H_0 at stage k.

Observed value: (k^*, θ^*).

p-value:

$$p := P_{\theta=0}\Big(\text{obtain } (k, \theta) \text{ as or more extreme than } (k^*, \theta^*)\Big).$$

"extreme" ⇒ needs ordering of Ω ⇒ Ω is **2-dimensional** ⇒ has no **intrinsic ordering**!

Stage-wise ordering, MLE ordering, Likelihood ratio ordering, Score test ordering.

Stage-wise ordering

Comparing pairs (k_1, θ_1) and (k_2, θ_2) according to **stage-wise ordering**:

- Rejection of H_0 at same stage k: comparison of θ_1 and θ_2 determines order.
- Rejection of H_0 at different stages: comparison of k_1 and k_2 determines order; **rejection at earlier stage** is considered more extreme.

Key features of **stage-wise ordering**:

- $p \leq \alpha^*$ if and only if H_0 is rejected: **internal consistency**,
- p-value does **not depend** on information levels or group sizes beyond stopping stage ⇒ stage-wise ordering naturally fits with **error-spending approach**.

Other orderings may not have last property.

Focus: methods based on **stage-wise ordering** of group-sequential sample space:

- Good **theoretical properties**.
- **Available in rpact.**

Point estimation and confidence intervals based on stage-wise ordering

Point estimator:

- **Median unbiased estimator:** Upper limit of a one-sided 50% confidence interval of the form $(-\infty, l_{0.5})$.

Confidence intervals:

- Derive $(1-\alpha)$ confidence interval via **test-inversion**, i.e. collect all parameter values for which test would not reject at level α.

Note: If H_0 is rejected at the first interim, then estimates, confidence intervals, and p-values based on the stage-wise ordering are **identical** to results from standard (naive) inference.

Global (unconditional) bias adjusted estimator

Defined as the solution $\widehat{\theta}_{adj}$ of equation (4.14) in Wassmer and Brannath (2016) (page 98):

$$\widehat{\theta}_{adj} = \widehat{\theta}_{MLE} - B_{\widehat{\theta}_{adj}}(\widehat{\theta}_{MLE})$$

where $B_{\widehat{\theta}_{adj}}(\widehat{\theta}_{MLE})$ corresponds to the unconditional bias of the MLE assuming that $\widehat{\theta}_{adj}$ is the true population parameter.

The bias is equal to the estimator minus its expected value. Therefore, $\widehat{\theta}_{adj}$ is the **parameter for which the expected value of the MLE is equal to the observed $\widehat{\theta}_{MLE}$**.

Estimator was the **overall winner in a simulation study** by Grayling and Wason (2022) if marginal and conditional bias as well as mean-squared error are all important.

Bias evaluation and hence the estimator **depends on information levels or group sizes beyond the stopping stage** which is somewhat undesirable.

Methods for associated confidence intervals and p-values are less established.

Conditional bias adjusted estimates for a trial stopped at stage *k*

Defined as the solution $\widehat{\theta}_{adj}^{(k)}$ of the equation:

$$\widehat{\theta}_{adj}^{(k)} = \widehat{\theta}_{MLE}^{(k)} - B_{\widehat{\theta}_{adj}^{(k)}}(\widehat{\theta}_{MLE}^{(k)})$$

where $B_{\widehat{\theta}_{adj}^{(k)}}(\widehat{\theta}_{MLE}^{(k)})$ corresponds to the conditional bias of the MLE assuming that $\widehat{\theta}_{adj}$ is the true population parameter.

Can reduce the conditional bias but showed **large (conditional and unconditional) bias and large mean-squared error** for some simulation scenarios in Grayling and Wason (2022).

Leads to large **attenuation of the treatment** effect if trial stops at an interim analyses and the MLE is close to the boundary.

Application to the Gallium trial

Gallium stopped at the efficacy interim after 245 of 370 (66.2%) events because the observed *p*-value of 0.0012 was below the local significance level of 0.012.

Estimator	Hazard ratio estimate
Standard (naive)	0.66 (95% CI 0.51 to 0.85)
Median unbiased	0.66 (95% CI 0.52 to 0.85)
Global bias adjusted	0.67
Conditional bias adjusted	0.74

The **median unbiased** and the **global bias adjusted** estimates are **almost identical** to the **standard estimate.**

The **conditional bias adjusted** estimate is **more attenuated** towards 1.

Summary: Inference in group-sequential trials

Standard (naive) analysis can lead to treatment effect **over-estimation** if the trial is stopped early for efficacy.

Bias is typically **not substantial** if stopping occurs after $> 50\%$ of information.

Adjusted inference based on the **stage-wise ordering** (**median unbiased estimator**, CI, p-value) is fully compatible with the hypothesis test and is implemented in **rpact**.

The **global bias adjusted estimator** is often a good choice if marginal and conditional bias as well as mean-squared error are important.

The **conditional bias adjusted estimator** may reduce conditional bias but can be associated with large bias and mean-squared error and a large attenuation of the treatment effect estimate.

Recommendations and worked example in FAQ on adaptR: link (requires VPN).

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. **Multiple endpoints in group-sequential trials**
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Motivation

Sponsor may want to have possibility for **multiple potential regulatory claims**:

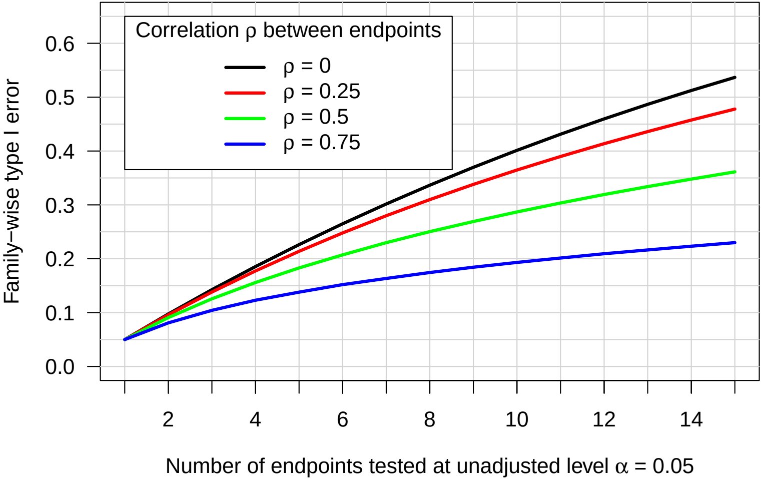
- Multiple endpoints.
- Multiple comparisons ($>$2 treatment arms).
- Multiple analysis times (group-sequential trials).
- Multiple populations.

⇒ Risk of inflation of type I error.

Health authorities mandate **type I error protection** across all endpoints for which a claim is sought.

⇒ **Which methods** should be used to control type I error in a multiplicity scenario?

Type I error inflation without multiplicity adjustment



FDA draft guidance “Multiple Endpoints in Clinical Trials”

Type I error should be controlled for entire trial, i.e. for all endpoints for which a claim is sought ⇒ **strong control of the family-wise error rate (FWER)**.

This includes controlling the type I error rate within and between the primary and secondary endpoint families.

Acceptable methods:

- **Assumption-free**: Bonferroni, Holm, fixed-sequence, fallback, gatekeeping, etc. More generally: **graphical procedures** (based on closed testing).
- **Weak assumptions (if satisfied)**: Hochberg, prospective alpha-allocation, truncated Hochberg.

U.S. Food and Drug Administration (2017)

Strong control of the family-wise error rate (FWER)

Examine family of null hypotheses (e.g. for K endpoints or K treatment comparisons): $H_{0,1}, \ldots, H_{0,K}$.

- Some of the $H_{0,i}$ may be true, some may be false.

Test procedure **strongly controls FWER** at global level α^* if:

- Probability of ≥ 1 false rejection(s) of any of the true $H_{0,i}$ is $\leq \alpha^*$.
- This holds regardless of which and how many of the $H_{0,i}$ are true.
- Requires demonstration of alpha control for every possible configuration of $H_{0,i}$'s being true.

Weak FWER control only requires type I error control if **all** $H_{0,i}$ are true $\Rightarrow$ typically considered insufficient.

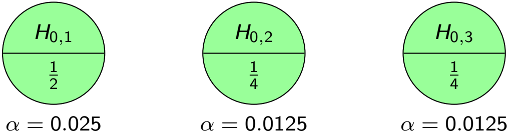
From now on, "control of type I error" always refers to strong FWER control.

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
    - Pocock's design
    - Sample size for Pocock's designs
    - Exit probabilities and expected sample size
    - O'Brien-Fleming's design
    - Limitations
3. Extensions of the basic methods
    - Non-normal data
    - Flexible timing of interim analyses
    - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
    - Illustration of the problem
    - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
    - Introduction
    - **Assumption-free methods which control type I error**
    - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
    - Type I error inflation of naive stagewise hierarchical strategy
    - Methods which protect type I error
10. Take home messages

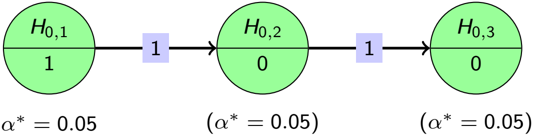
Weighted Bonferroni test

- Assign weights w_i with $0 \leq w_i \leq 1$ and $\sum_{i=1}^{K} w_i = 1$ to the K null hypotheses.
- Test each null hypothesis $H_{0,i}$ at adjusted level $w_i \cdot \alpha^*$.



Hierarchical fixed sequence procedure

- Hypotheses ordered in pre-specified sequence (e.g. by clinical relevance or likelihood of success).
- Null hypothesis $H_{0,i}$ can only be tested and therefore rejected by testing it at the unadjusted level α^* if all hypotheses with higher priority can also be rejected.

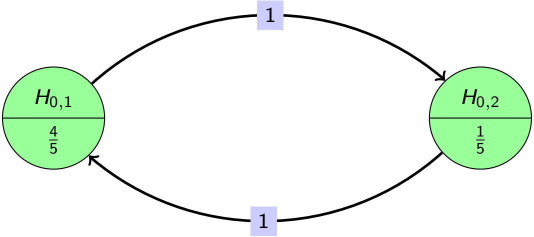


Graph based multiple testing procedures

Visualization of a wide range of multiple testing procedures as graphs:

- **Vertices** correspond to **null hypotheses** to which **initial weights** w_i are assigned.
- **Weighted edges** describe how type I error is passed on if a null hypothesis is rejected.

Procedure independently developed by Bretz et al. (2009) and Burman et al. (2009), implemented in **R package gMCP** (Rohmeyer and Klinglmueller (2018)), and **endorsed by the FDA** (U.S. Food and Drug Administration (2017)).



Graphical representation of a weighted Bonferroni-Holm test.

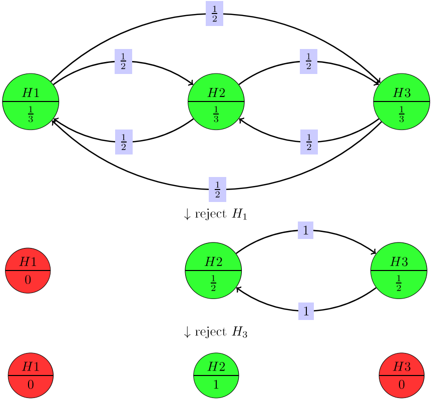
Graph based multiple testing procedures

Iterative analysis (performed by software for complex cases):

1. Check if any null hypothesis $H_{0,i}$ can be rejected based on the weights, i.e. if any p-value $p_i \leq w_i \cdot \alpha$.
2. If no: none of the (remaining) null hypothesis can be rejected. STOP.
3. If yes: Reject null hypothesis $H_{0,i}$ and **update the graph**, i.e.
   - Remove vertex i.
   - Update weights and edges according to mathematical formula (Bretz et al. (2009)) which
     - recycles the alpha allocated to that test according to the edge weights and
     - updates the edge weights to avoid α^* being recycled to already rejected null hypotheses.
   - Go back to Step 1 with the updated graph, repeat the process.

It can be shown that the procedure is a consonant **closed testing procedure**, i.e. **type I error is controlled**.

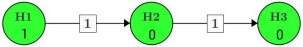
Illustration of analysis and graph updates



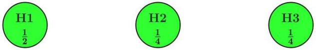
Example showing how two null hypotheses can be rejected at level $\alpha^* = 0.05$ with *p*-values $p_1 = 0.01$, $p_2 = 0.07$, and $p_3 = 0.02$ (taken from vignette of gMCP).

Fall-back procedure and variations as graphs

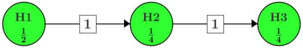
Hierarchical Fixed Sequence



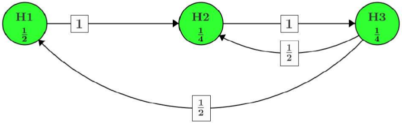
Weighted Bonferroni



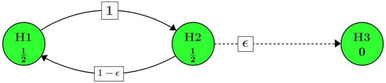
Fall-back



Improved fall-back

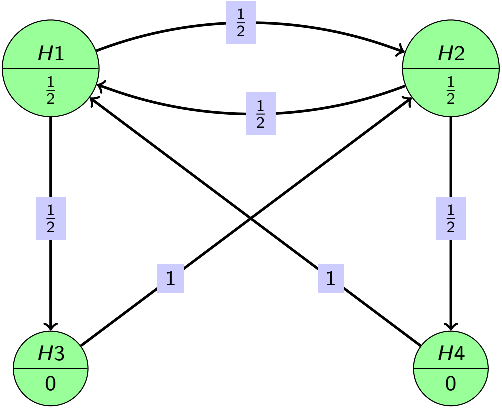


Fall-back with ε-edges representing a very small initial weight
(i.e. rejection of H_3 only possible after both H_1 and H_2 have been rejected)



One option for a setting with two primary and two secondary hypotheses

Example taken from Maurer et al. (2011).



⇒ **Simulations** under a range of scenarios are essential to understand the operating characteristics of any complex scheme.

Graph based multiple testing procedures and group-sequential trials

Graph based methods using the weighted Bonferroni intersection test can also be applied to group-sequential designs (Maurer and Bretz (2013)):

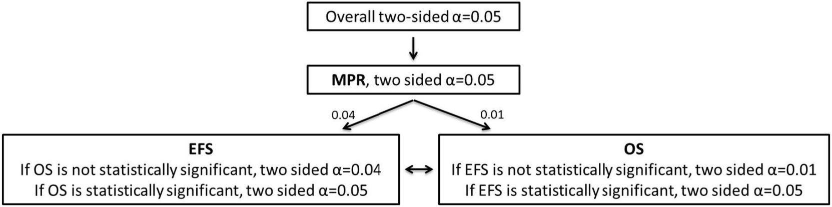
- **Replace standard tests** at level $w_i \cdot \alpha^*$ **with corresponding group-sequential tests** at the same overall level using α-spending functions.
- If null hypothesis $H_{0,i}$ of the group-sequential test for endpoint i (at overall level $w_i \cdot \alpha^*$) can be rejected at stage $k \Rightarrow$ update the graph and the weights for other endpoint as $w_j \to w_j'$ as before $\Rightarrow$ group-sequential tests for other endpoints j can be tested at updated overall level $w_j' \cdot \alpha^*$ **at stage k and later stages**.

Method is valid under mild monotonicity conditions on the α-spending functions which are satisfied for all standard spending families (e.g. O'Brien-Fleming, Pocock, power function).

Caveat: Method requires definition of group-sequential boundaries at overall levels $w_i \cdot \alpha^*$ for all initial and updated weights w_i, i.e. it is not correct to just multiply the local significance levels from a group-sequential test with overall level α^* with w_i!

Example: IMpower 030 trial

- Atezolizumab vs placebo in combination with platinum-based chemotherapy in patients with resectable stage II, IIIa, or select IIIB NSCLC, Peters et al. (2019).
- Primary endpoint: MPR (major pathological response) rate at the time of surgical resection.
- Key secondary endpoints: EFS, OS.
- Two interim analysis and a final analysis planned for EFS and OS (at the same timepoints). First interim analysis conducted at the timing of the primary analysis of MPR. Approximate O'Brien-Fleming α-spending.
- Type I error control plan:



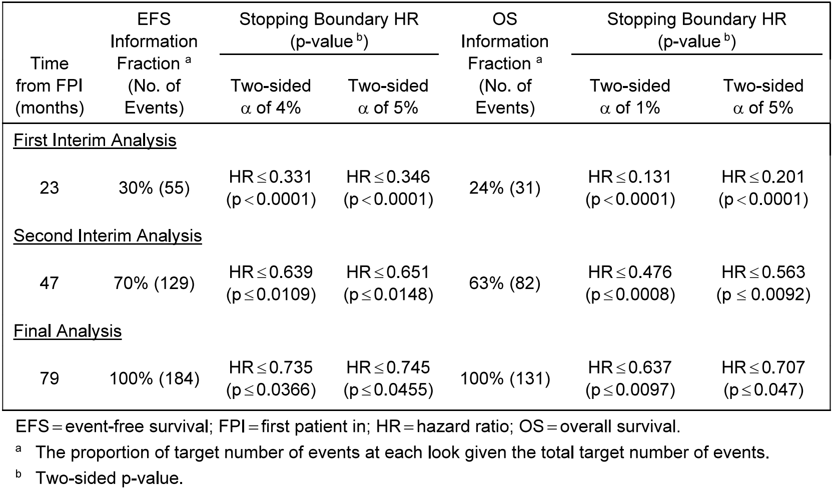
Overall two-sided α=0.05

MPR, two sided α=0.05

EFS
If OS is not statistically significant, two sided α=0.04
If OS is statistically significant, two sided α=0.05

OS
If EFS is not statistically significant, two sided α=0.01
If EFS is statistically significant, two sided α=0.05

IMpower030: Analysis timing and stopping boundary for key secondary endpoints EFS and OS



Time from FPI (months)	EFS Information Fraction [a] (No. of Events)	Stopping Boundary HR (p-value [b])		OS Information Fraction [a] (No. of Events)	Stopping Boundary HR (p-value [b])	
		Two-sided α of 4%	Two-sided α of 5%		Two-sided α of 1%	Two-sided α of 5%
First Interim Analysis						
23	30% (55)	HR ≤ 0.331 (p < 0.0001)	HR ≤ 0.346 (p < 0.0001)	24% (31)	HR ≤ 0.131 (p < 0.0001)	HR ≤ 0.201 (p < 0.0001)
Second Interim Analysis						
47	70% (129)	HR ≤ 0.639 (p ≤ 0.0109)	HR ≤ 0.651 (p ≤ 0.0148)	63% (82)	HR ≤ 0.476 (p ≤ 0.0008)	HR ≤ 0.563 (p ≤ 0.0092)
Final Analysis						
79	100% (184)	HR ≤ 0.735 (p ≤ 0.0366)	HR ≤ 0.745 (p ≤ 0.0455)	100% (131)	HR ≤ 0.637 (p ≤ 0.0097)	HR ≤ 0.707 (p ≤ 0.047)

EFS = event-free survival; FPI = first patient in; HR = hazard ratio; OS = overall survival.

[a] The proportion of target number of events at each look given the total target number of events.

[b] Two-sided p-value.

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Assumption-free settings and settings with known correlations

- **Assumption-free methods** control type 1 error regardless of the correlation between endpoints or tests.
    - They will frequently be **conservative** in the case of no correlation or positive correlation.
- We can **increase power by making assumptions**, e.g. of no or positive correlation.
    - E.g. Bonferroni-Holm → Hochberg test (not discussed here).
    - Need to be able to robustly guarantee the assumptions.
- In some cases, a **correlation structure** is **enforced by the design** and way the data was collected.
    - This can **increase power further**.

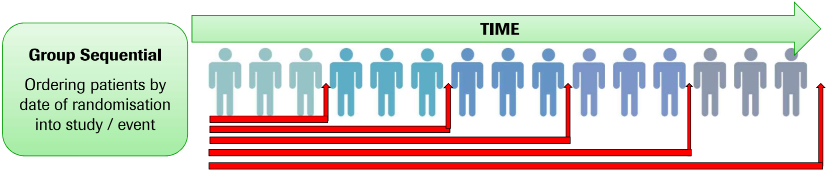
Multiple testing procedures exploiting known correlations

Three settings where correlation structure between endpoints is implied by trial design and can be exploited to increase power while protecting type I error:

- Multi-arm trials with shared control arm: **Dunnett type tests** (not discussed here).
- Multiple analyses of the data over time: **Group-sequential trials**.
- Same endpoint in nested subgroup(s) and full population: **Mathematically equivalent to group-sequential trials.**

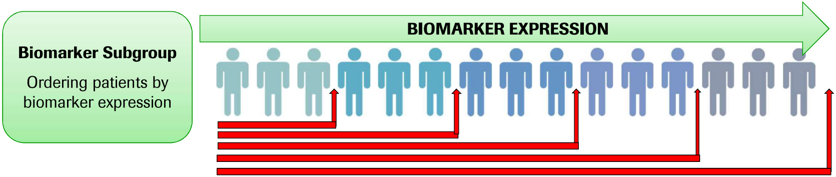
More complex correlation structures arise and can be exloited if these design features are **combined**, e.g. group-sequential trials in a subgroup and the full population.

Mathematical equivalence between group-sequential designs and biomarker subgroup studies



Group Sequential

Ordering patients by date of randomisation into study / event



Biomarker Subgroup

Ordering patients by biomarker expression

Approximate joint distribution of standardized Z-statistics in the full population and a subgroup

Denote the information fraction (size) of the subgroup S relative to the full population F by τ.

Under the joint global null hypothesis, the joint distribution of Z_S and Z_F is **multivariate normal** (Spiessens and Debois (2010)):

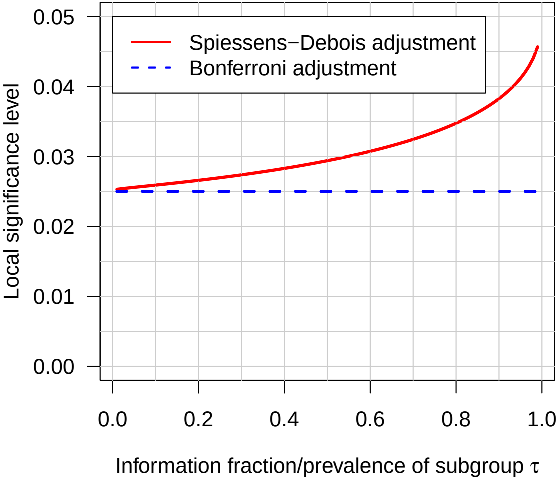
$$(Z_S, Z_F) \quad \sim \quad N\left(\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \, \Sigma = \begin{pmatrix} 1 & \sqrt{\tau} \\ \sqrt{\tau} & 1 \end{pmatrix}\right).$$

For example, if $\tau = 0.5$, the correlation is $\sqrt{0.5} = 0.71$.

$\Rightarrow$ **Software for group-sequential trials** can be used to determine critical values for the global null hypothesis (which is relevant if the two populations are co-primary).

Local significance levels for co-primary populations S and F

Shown are local significance levels for equal critical values for S and F, i.e. a Pocock-type adjustment, depending on τ.



Exploit correlation between PFS and OS using illness-death multistate model

Co-primary endpoints PFS and OS.

Typical setup: plan group-sequential design **independently** for PFS and OS.

- OS can only have PH under very unrealistic conditions.
- Ignores correlation between PFS and OS.

RAAN capstone project:

- Use illness-death model to tweak critical values and number of events using simulation.
- Exploit correlation between PFS and OS.
- Properly model NPH for OS.
- Potential for quite **huge power gains** for OS. But also power loss! paper, R package simIDM on CRAN.

Summary for multiple endpoints section

Multiple co-primary endpoints provide **multiple opportunities for a significant result** ("hedging one's bets"). Depending on scenarios, this can **increase power** compared to a single primary endpoint.

Graphical methods are useful to describe multiple testing strategies and are endorsed by regulators.

Exploiting structural correlations leads to further power gains.

Agenda

1 Introduction

2 Pocock's and O'Brien-Fleming's designs for normal data
- Pocock's design
- Sample size for Pocock's designs
- Exit probabilities and expected sample size
- O'Brien-Fleming's design
- Limitations

3 Extensions of the basic methods
- Non-normal data
- Flexible timing of interim analyses
- Stopping for futility

4 Operational aspects of group-sequential designs

5 Group-sequential designs for survival data

6 Case study: GALLIUM randomized clinical trial

7 Inference in group-sequential designs
- Illustration of the problem
- Adjusted estimates, confidence intervals, and p-values

8 Multiple endpoints in group-sequential trials
- Introduction
- Assumption-free methods which control type I error
- Multiple testing procedures exploiting known or structural correlations

9 Hierarchical testing of primary and secondary endpoints in a group-sequential trial
- Type I error inflation of naive stagewise hierarchical strategy
- Methods which protect type I error

10 Take home messages

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Setup

Primary and secondary endpoint with corresponding null hypotheses $H_{0,p}$ and $H_{0,s}$.

Primary endpoint tested with a group-sequential design.

Possible testing approaches for secondary endpoint using the primary endpoint as a **gatekeeper** (Glimm et al. (2010)):

- **Stagewise hierarchical**: If primary endpoint reaches significance ⇒ test secondary endpoint at the same stage (i.e. only once).
- **Overall hierarchical**: If primary endpoint reaches significance ⇒ test secondary endpoint at the same stage and also at later stages (assuming trial continues despite significance for the primary endpoint).

Goal: choose group-sequential strategy and boundaries such that FWER $=$ P(reject $\geq$ 1 true $H_{0,i}$) $\leq \alpha^*$.

Overall and stagewise hierarchical approaches

Stagewise hierarchical

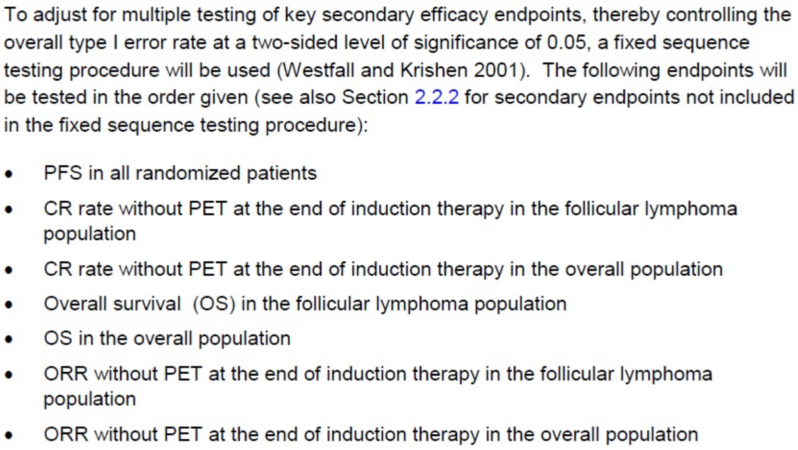
- Typical strategy, e.g. for ORR and PFS or PFS and PRO endpoints.
- Only **one look** at secondary endpoints allowed ⇒ Can we do better than defining a group-sequential boundary for secondary endpoints?
- The primary **focus of this section.**

Overall hierarchical

- Useful if secondary endpoint is expected to reach maturity later (e.g. PFS and OS).
- **Multiple looks** at secondary endpoint possible ⇒ FWER only protected if **group-sequential boundary** is also defined for the secondary endpoint.

Note: Endpoints for which final data for all patients is already available at the first interim analysis (e.g. ORR in some settings) do not need type I error adjustment.

GALLIUM SAP - naïve implementation of stagewise hierarchical approach



To adjust for multiple testing of key secondary efficacy endpoints, thereby controlling the overall type I error rate at a two-sided level of significance of 0.05, a fixed sequence testing procedure will be used (Westfall and Krishen 2001). The following endpoints will be tested in the order given (see also Section 2.2.2 for secondary endpoints not included in the fixed sequence testing procedure):

- PFS in all randomized patients
- CR rate without PET at the end of induction therapy in the follicular lymphoma population
- CR rate without PET at the end of induction therapy in the overall population
- Overall survival (OS) in the follicular lymphoma population
- OS in the overall population
- ORR without PET at the end of induction therapy in the follicular lymphoma population
- ORR without PET at the end of induction therapy in the overall population

Naïve stagewise hierarchical approach

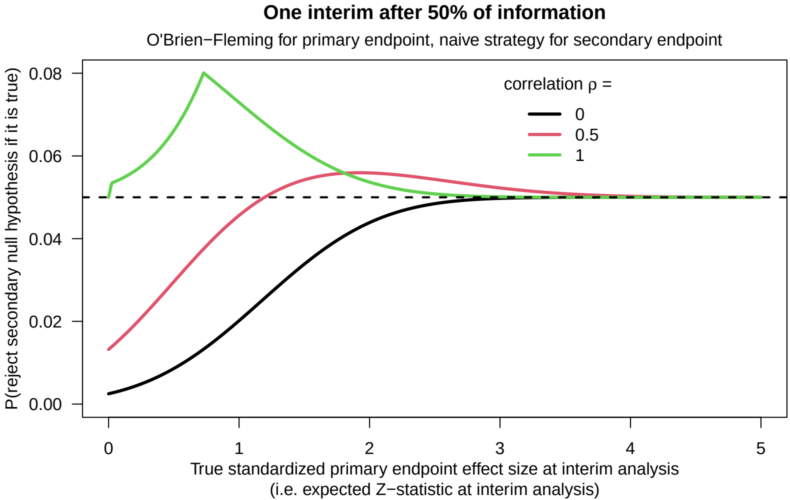
1. Test primary endpoint within group-sequential design.
2. If significant ⇒ test secondary endpoint at $\alpha = 0.05$ at the same stage.

Issues with naïve approach discussed in Hung et al. (2007), Glimm et al. (2010), Tamhane et al. (2010).

Glimm et al. (2010) analytically derive upper bound for type I error of secondary endpoint test assuming a bivariate normal distribution and one interim after 50% information:

- Maximal FWER = **0.08**.
- Achieved for correlation $\rho = 1$.
- Maximal possible inflation does not depend on error-spending function of primary endpoint (but the 1ry EP effect size where it occurs does!). Power in general does!

Secondary endpoint type I error for naïve strategy



Similar to Figure 1 in Tamhane et al. (2010).

Why is there no inflation
for uncorrelated endpoints, even
for the naïve strategy?

No inflation for uncorrelated endpoints

Endpoints uncorrelated, bivariate normal ⇒ endpoints even **independent**.

This implies:

$$P(\text{reject } H_{0,s} \cap \text{reject } H_{0,p}) \quad = \quad P(\text{reject } H_{0,p}) \cdot P(\text{reject } H_{0,s}).$$

If primary endpoint effect size increases $\Rightarrow P(\text{reject } H_{0,p}) \to 1$.

So:

$$P(\text{reject } H_{0,p}) \cdot P(\text{reject } H_{0,s}) \quad \to \quad P(\text{reject } H_{0,s}) \ = \ \alpha.$$

Why do we have FWER inflation
for the naïve strategy, if
endpoints are positively correlated?

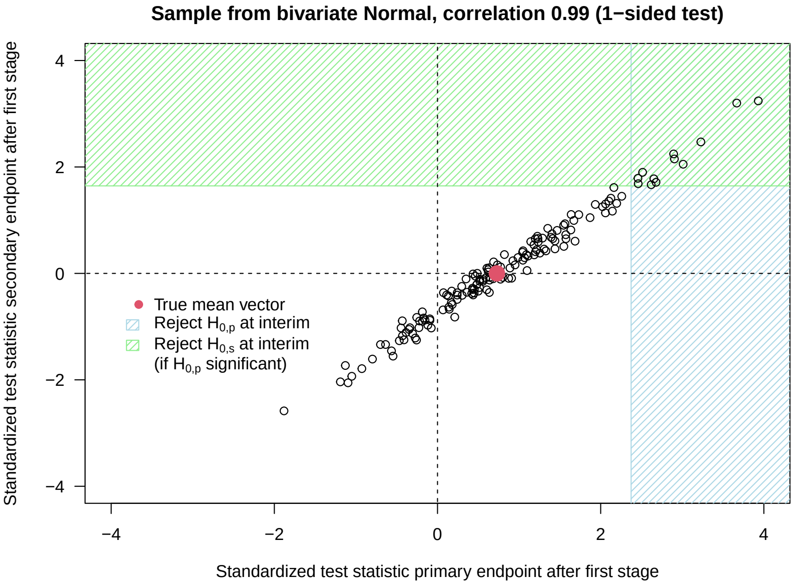
FWER inflation for naïve strategy and $\rho > 0$

Consider extreme case $\rho = 1$. Assume that **true standardized primary endpoint effect size is 0.73** after stage 1 but **true secondary effect size is 0**.

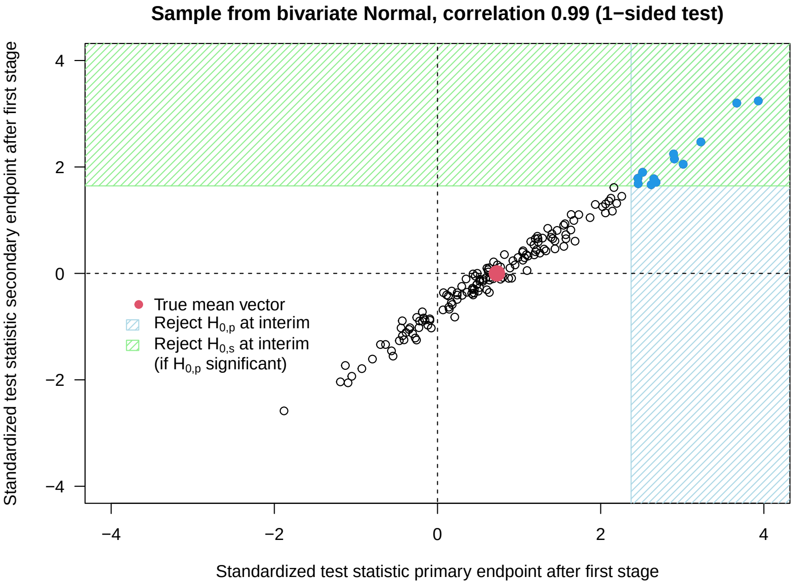
- If we stop at interim ⇒ **random high** of primary endpoint test statistic trajectory. For this effect size, this happens with probability **0.05**.
- Due to perfect correlation ⇒ random high of primary endpoint is also random high of secondary endpoint test statistic ⇒ $H_{0,s}$ also rejected with probability **0.05**.
- If we run to final ⇒ same phenomena, probability to be significant for primary **0.22**.
- Probability to be significant for secondary at final analysis **0.03**.

$\mathbf{0.05} + \mathbf{0.03} = \mathbf{0.08} > 0.05$.

Joint distribution of test statistics at interim - $\rho \approx 1$



Joint distribution of test statistics at interim - $\rho \approx 1$



Why does inflation decrease again to α with increasing primary endpoint effect size?

What happens if primary endpoint effect size increases, for $\rho > 0$?

If primary endpoint effect size increases,

- we stop more often at the interim, and ultimately,
- probability to stop at interim becomes 1.

Secondary endpoint is then tested **only once** at α.

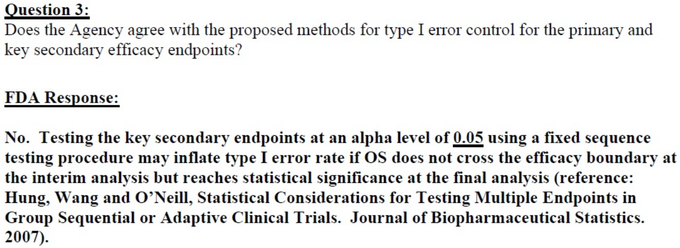
FDA position ... is evolving

Hung et al. (2007): paper by 3 FDA authors,

> *...when trials are conducted using a group-sequential design with interim analyses or can be extended using an adaptive design with an increase of sample size or total number of events, this conventional hierarchical testing strategy may violate the closure principle and the* **overall type I error rate may not be controlled in the strong sense***.*

27th March 2013: FDA commented on Gallium SAP. No mentioning of hierarchical testing.

FDA comment on CO39385 https://clinicaltrials.gov/ct2/show/NCT03016312, 10th March 2017:



Question 3:
Does the Agency agree with the proposed methods for type I error control for the primary and key secondary efficacy endpoints?

FDA Response:

No. Testing the key secondary endpoints at an alpha level of 0.05 using a fixed sequence testing procedure may inflate type I error rate if OS does not cross the efficacy boundary at the interim analysis but reaches statistical significance at the final analysis (reference: Hung, Wang and O'Neill, Statistical Considerations for Testing Multiple Endpoints in Group Sequential or Adaptive Clinical Trials. Journal of Biopharmaceutical Statistics. 2007).

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Which approach **does** protect overall type I error for stagewise hierarchical strategy?

FWER is protected if a **group-sequential boundary** is also defined for the secondary endpoint.

Which boundary to chose?

Primary endpoint:

- Choice of error-spending not affected by secondary endpoint ⇒ typically choose **O'Brien-Fleming**.
- Choice of primary boundary guided by usual **group-sequential design considerations**.

Secondary endpoint:

- Stopping depends on primary endpoint ⇒ not sensible to make bar high at interim using O'Brien-Fleming.
- Choice of secondary boundary more guided by **power considerations**.
- Use more balanced spending function, typically **Pocock**.

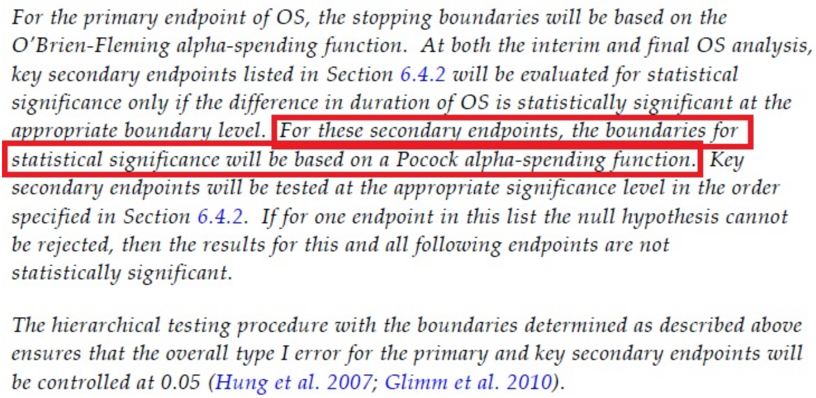
Uniformly optimal minimax strategy

Pocock boundaries for secondary:

- **Uniformly optimal minimax strategy**, i.e. optimal **minimax** allocation of α for hierarchical test: minimal significance level to be used at either interim or final is maximized.
- Uniformly more powerful than Bonferroni correcting over number of analyses.
- There does not exist a uniformly more powerful strategy.
- "Uniformly": over all correlations.

Discussion in Glimm et al. (2010), Tamhane et al. (2010), and Tamhane et al. (2018).

Implementation in CO39385



For the primary endpoint of OS, the stopping boundaries will be based on the O'Brien-Fleming alpha-spending function. At both the interim and final OS analysis, key secondary endpoints listed in Section 6.4.2 will be evaluated for statistical significance only if the difference in duration of OS is statistically significant at the appropriate boundary level. For these secondary endpoints, the boundaries for statistical significance will be based on a Pocock alpha-spending function. Key secondary endpoints will be tested at the appropriate significance level in the order specified in Section 6.4.2. If for one endpoint in this list the null hypothesis cannot be rejected, then the results for this and all following endpoints are not statistically significant.

The hierarchical testing procedure with the boundaries determined as described above ensures that the overall type I error for the primary and key secondary endpoints will be controlled at 0.05 (Hung et al. 2007; Glimm et al. 2010).

From CO39385 protocol.

Further refinements - method 1

Li et al. (2018) suggest to reject the null hypothesis for the secondary endpoint at stage k if:

1. The group-sequential boundary for the primary endpoint is crossed at stage k (i.e. gatekeeper $H_{0,p}$ is rejected at level α).
2. The group-sequential boundary for the secondary endpoint has been crossed at any stage ℓ with $\ell \leq k$ (i.e. $H_{0,s}$ is rejected at level α).
3. The raw p-value for the secondary endpoint at stage k is $\leq \alpha$.

Items 1 and 2 are required for **type I error control**, item 3 is added to ensure the **"credibility" of the secondary endpoint finding**.

Item 2 "looks back" to results from earlier stages ℓ to **gain power** for the secondary endpoint.

Further refinements - method 2

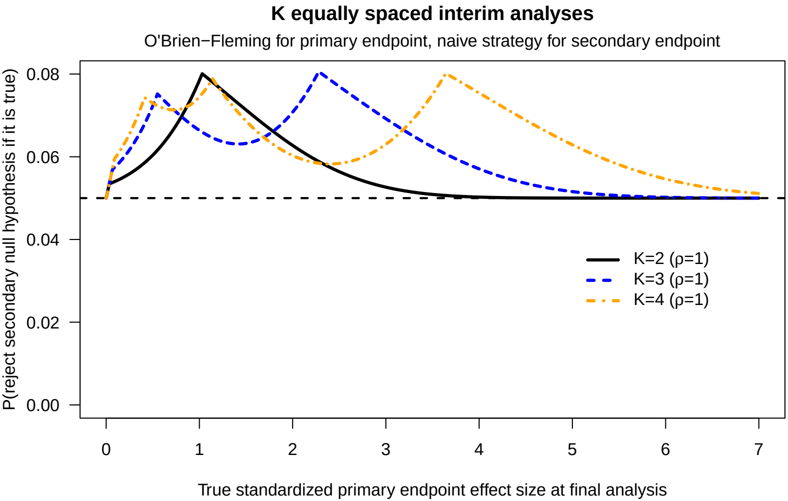
Tamhane et al. (2010) and Tamhane et al. (2018) make the following observations:

- For >1 interim analyses and $\rho = 1$, maximal secondary type I error of the naive strategy increases only minimally compared to $K = 2$. In contrast, the local significance levels of the Pocock boundary decrease considerably for a larger number of interim analyses.
- Maximal secondary type I error inflation of the naïve strategy is less pronounced if correlation between endpoints can be bounded away from 1.
  $\Rightarrow$ Exploit this if correlation between endpoints is known, e.g. if the same endpoint is evaluated in a subgroup and in the full population.

In these settings, the constant **Pocock boundary for the secondary endpoint can be refined leading to increased local significance levels without type I error inflation**.

For **exploratory R-code** see this **FAQ on adaptR**: link (requires VPN).

Secondary endpoint type I error for naïve strategy for correlation $\rho = 1$ and >1 interim analyses



Refined boundaries

Using the method in Tamhane et al. (2018) for K equally spaced analyses, primary O'Brien-Fleming and secondary Pocock (constant) boundary.

Secondary boundary (**Z-scale**, one-sided $\alpha = 0.025$):

		Refined boundary		
K	Original boundary	($\rho = 0$)	($\rho = 0.5$)	($\rho = 1$; worst case)
2	2.179	1.960	2.018	2.179 (no refinement)
3	2.290	1.960	2.024	2.181
4	2.362	1.960	2.022	2.179

Secondary boundary (corresponding two-sided **local significance level scale**):

		Refined boundary		
K	Original boundary	($\rho = 0$)	($\rho = 0.5$)	($\rho = 1$; worst case)
2	0.029	0.050	0.044	0.029 (no refinement)
3	0.022	0.050	0.043	0.029
4	0.018	0.050	0.043	0.029

Recommendations for stagewise hierarchical approach

- Naïve testing of secondary endpoint inflates type I error in the group-sequential setting.
- Pocock (constant) boundary recommended for secondary endpoint.
- Refinement method 1 (Li et al. (2018)) can be used in case of only one interim analysis and unknown correlation.
- Refinement method 2 (Tamhane et al. (2018)) can be used in case of >1 interim analysis or known correlation between endpoints.
- Considerable refinements are possible if correlation between endpoints can be bounded away from 1.

Recommendations for overall hierarchical approach

- Optimal stopping boundary for secondary endpoint depends on circumstances. (Topic less discussed in the literature.)
- If secondary endpoint reaches maturity later (e.g. PFS and OS), O'Brien-Fleming boundaries for secondary endpoint may still be sensible.
- Refinement method 1 (Li et al. (2018)) is also valid here but method 2 (Tamhane et al. (2018)) cannot be applied.

Future: potentially explore correlation between PFS and OS using illness-death multistate model, see Slide 93.

Agenda

1. Introduction
2. Pocock's and O'Brien-Fleming's designs for normal data
   - Pocock's design
   - Sample size for Pocock's designs
   - Exit probabilities and expected sample size
   - O'Brien-Fleming's design
   - Limitations
3. Extensions of the basic methods
   - Non-normal data
   - Flexible timing of interim analyses
   - Stopping for futility
4. Operational aspects of group-sequential designs
5. Group-sequential designs for survival data
6. Case study: GALLIUM randomized clinical trial
7. Inference in group-sequential designs
   - Illustration of the problem
   - Adjusted estimates, confidence intervals, and p-values
8. Multiple endpoints in group-sequential trials
   - Introduction
   - Assumption-free methods which control type I error
   - Multiple testing procedures exploiting known or structural correlations
9. Hierarchical testing of primary and secondary endpoints in a group-sequential trial
   - Type I error inflation of naive stagewise hierarchical strategy
   - Methods which protect type I error
10. Take home messages

Take home messages

- Group-sequential designs improve on Bonferroni by **exploiting** correlation.
- Compared to single stage design: Maximal sample size if running to final analysis larger, but **average** sample size lower.
- α-spending: Total information pre-specified, but **timing and number** of interim analysis can be flexibly chosen.
- **rpact** can be used for planning and analyzing group-sequential (and adaptive) trials.
- Conventional inference is **biased** in group-sequential trials. However, for late efficacy interim in typical clinical trial situation, bias is **negligible**.
- Naive testing of secondary endpoint in **stagewise hierarchical** framework inflates type I error. Recommended approach: use Pocock group-sequential test (or refinements) for secondary endpoints.
- **Graphical methods** are useful for situations with multiple endpoints.
- Check out **adaptR** (`https://go.roche.com/adaptR`; requires VPN), a **comprehensive Roche resource** for group-sequential and adaptive trial designs.

Thank you very much for your attention.

References I

- Bretz, F., Maurer, W., Brannath, W. and Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Stat Med* **28** 586–604.

- Burman, C. F., Sonesson, C. and Guilbaud, O. (2009). A recycling framework for the construction of Bonferroni-based multiple tests. *Stat Med* **28** 739–761.

- Fan, X. and Demets, D. L. (2006). Conditional and unconditional confidence intervals following a group sequential test. *Journal of Biopharmaceutical Statistics* **16** 107–122. PMID: 16440840. . (`https://doi.org/10.1080/10543400500406595`

- Fan, X. F., DeMets, D. L. and Lan, K. K. G. (2004). Conditional bias of point estimates following a group sequential test. *Journal of Biopharmaceutical Statistics* **14** 505–530. PMID: 15206542. . (`https://doi.org/10.1081/BIP-120037195`

- Freidlin, B. and Korn, E. L. (2009). Stopping clinical trials early for benefit: impact on estimation. *Clin Trials* **6** 119–125.

- Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. and Hothorn, T. (2021). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.1-3. . (`https://CRAN.R-project.org/package=mvtnorm`

- Glimm, E., Maurer, W. and Bretz, F. (2010). Hierarchical testing of multiple endpoints in group-sequential trials. *Stat. Med.* **29** 219–228.

- Grayling, M. J. and Wason, J. M. (2022). Point estimation following a two-stage group sequential trial. *Statistical Methods in Medical Research* 09622802221137745.

- Hasegawa, T. (2016). Group sequential monitoring based on the weighted log-rank test statistic with the fleming-harrington class of weights in cancer vaccine studies. *Pharmaceutical statistics* **15** 412–419.

References II

- Hung, H. M. J., Wang, S.-J. and O'Neill, R. (2007). Statistical considerations for testing multiple endpoints in group sequential or adaptive clinical trials. *J. Biopharm. Statist.* **17** 1201–1210.

- Jennison, C. and Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Chapman & Hall/CRC, Boca Raton, FL.

- Kim, K. and DeMets, D. L. (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* **74** 149–154.

- Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70** 659–663.

- Li, H., Wang, J., Luo, X., Grechko, J. and Jennison, C. (2018). Improved two-stage group sequential procedures for testing a secondary endpoint after the primary endpoint achieves significance. *Biometrical Journal* **60** 893–902.
  . (https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201700231

- Marcus, R., Davies, A., Ando, K., Klapper, W., Opat, S., Owen, C., Phillips, E., Sangha, R., Schlag, R., Seymour, J. F., Townsend, W., Trneny, M., Wenger, M., Fingerle-Rowson, G., Rufibach, K., Moore, T., Herold, M. and Hiddemann, W. (2017). Obinutuzumab for the First-Line Treatment of Follicular Lymphoma. *N. Engl. J. Med.* **377** 1331–1344.

- Maurer, W. and Bretz, F. (2013). Multiple testing in group sequential trials using graphical approaches. *Statistics in Biopharmaceutical Research* **5** 311–320.

- Maurer, W., Glimm, E. and Bretz, F. (2011). Multiple and repeated testing of primary, coprimary, and secondary hypotheses. *Statistics in Biopharmaceutical Research* **3** 336–352.

References III

- Ohman Strickland, P. A. and Casella, G. (2003). Conditional inference following group sequential testing. *Biometrical Journal* **45** 515–526.
  . (http://dx.doi.org/10.1002/bimj.200390029

- Peters, S., Kim, A. W., Solomon, B., Gandara, D. R., Dziadziuszko, R., Brunelli, A., Garassino, M. C., Reck, M., Wang, L., To, I., Sun, S. W., Gitlitz, B. J., Sandler, A. and Rizvi, N. (2019). IMpower030: Phase III study evaluating neoadjuvant treatment of resectable stage II-IIIB non-small cell lung cancer (NSCLC) with atezolizumab (atezo) + chemotherapy. *Annals of Oncology* **30**. Mdz064.014.
  . (https://doi.org/10.1093/annonc/mdz064.014

- Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64** 191–199.

- Proschan, M., Lan, K. and Wittes, J. (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach*. Springer, New York.

- Rohmeyer, K. and Klinglmueller, F. (2018). *gMCP: An R Package for Graph Based Multiple Comparison Procedures*. R package version 0.8-14.

- Roychoudhury, S., Anderson, K. M., Ye, J. and Mukhopadhyay, P. (2021). Robust design and analysis of clinical trials with nonproportional hazards: A straw man guidance from a cross-pharma working group. *Statistics in Biopharmaceutical Research* **0** 1–15.
  . (https://doi.org/10.1080/19466315.2021.1874507

- Shimura, M., Gosho, M. and Hirakawa, A. (2017). Comparison of conditional bias-adjusted estimators for interim analysis in clinical trials with survival data. *Stat Med* **36** 2067–2080.

References IV

- Slud, E. and Wei, L. (1982). 2-Sample Repeated Significance Tests Based on the Modified Wilcoxon Statistic. *Journal of the American Statistical Association* **77** 862–868.
- Spiessens, B. and Debois, M. (2010). Adjusted significance levels for subgroup analyses in clinical trials. *Contemp Clin Trials* **31** 647–656.
- Tamhane, A. C., Gou, J., Jennison, C., Mehta, C. R. and Curto, T. (2018). A gatekeeping procedure to test a primary and a secondary endpoint in a group sequential design with multiple interim looks. *Biometrics* **74** 40–48.
- Tamhane, A. C., Mehta, C. R. and Liu, L. (2010). Testing a primary and a secondary endpoint in a group sequential design. *Biometrics* **66** 1174–1184.
- U.S. Food and Drug Administration (2017). *Guidance for Industry: Multiple Endpoints in Clinical Trials*. Draft guidance.
- U.S. Food and Drug Administration (2019). *Guidance for Industry: Adaptive Designs for Clinical Trials of Drugs and Biologics*.
- Van Norman, G. A. (2019). Phase ii trials in drug development and adaptive trial design. *JACC: Basic to Translational Science* **4** 428–437.
  . (https://www.sciencedirect.com/science/article/pii/S2452302X19300658
- Wang, H., Rosner, G. L. and Goodman, S. N. (2016). Quantifying over-estimation in early stopped clinical trials and the ”freezing effect” on subsequent research. *Clin Trials* **13** 621–631.
- Wang, L., Luo, X. and Zheng, C. (2021). A simulation-free group sequential design with max-combo tests in the presence of non-proportional hazards. *Pharmaceutical Statistics* **20** 879–897.
  . (https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.2116

References V

- Wassmer, G. and Brannath, W. (2016). *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer.

Doing now what patients need next

R version and packages used to generate these slides:

R version: R version 4.2.2 (2022-10-31 ucrt)

Base packages: stats / graphics / grDevices / utils / datasets / methods / base

Other packages: nloptr / MASS / rpact / gsDesign / mvtnorm / reporttools / xtable

This document was generated on 2023-06-07 at 21:23:37.