The Design of Group Sequential and Adaptive Clinical Trials

Christopher Jennison

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

Danish Society of Biopharmaceutical Statistics

Copenhagen, September 2019

Motivation: Phase 3 clinical trials

Phase III trials are conducted as the last stage in the drug development process.

Two positive studies are usually required to confirm that a new treatment is superior to the current standard treatment.

Regulators customarily require a hypothesis test to reach significance at the one-sided 2.5% level.

Studies may recruit hundreds, or even thousands, of subjects at a cost of as much as €10k to €50k per patient.

The time taken to reach a conclusion eats into the limited patent lifetime remaining to the company developing the drug.

Thus, there are strong incentives to reach an early conclusion for either a positive or negative decision.

Motivation: Interim monitoring

Clinical trials methodology can also be applied to animal trials and epidemiological studies, where there is similar motivation from

Ethics

Administration (accrual, compliance, ...)

Economics

to monitor the conduct of the trial and examine accumulating data.

Subjects should not be exposed to unsafe, ineffective or inferior treatments.

National and international guidelines for clinical trials call for interim analyses to be performed — and reported.

It is now standard practice for clinical trials to have a Data and Safety Monitoring Board (DSMB) to oversee the study and consider the option of early termination.

Motivation: Repeated hypothesis tests during a study

Suppose θ represents the difference in mean responses in a two-treatment comparison.

In a superiority trial, we wish to test H_0: $\theta \leq 0$ against $\theta > 0$.

If a test of H_0 is carried out at one-sided significance level $\alpha = 0.025$ on K occasions during the course of the trial, the *overall* type I error rate is:

Number of tests, K	Overall error rate	Number of tests, K	Overall error rate
1	0.025	10	0.097
2	0.042	20	0.124
3	0.054	100	0.190
5	0.071	∞	1.000

See Armitage et al. (*JRSS, A*, 1969).

Motivation: Adaptive clinical trial designs

Around the year 2000, there was a surge of interest in "adaptive" trials which allow changes in study design based on interim results.

An adaptive trial could:

- Route more patients to the treatment that seems to work best
- Drop treatments that don't seem to be effective
- Add more of the type of patients who react best to a particular treatment
- Merge two different phases of drug development into one trial

This represented a dramatic change from the philosophy of simple Phase III trials, designed to answer fully formulated questions through a pre-defined protocol and statistical analysis plan.

Time has shown what such designs can (and cannot) achieve.

Outline of course

1. **Group sequential tests (1)**

   Error spending tests

   Examples of group sequential designs:

      With a normal response,

      With a binary response,

      With a survival endpoint

2. **Group sequential tests (2)**

   Group sequential tests with a delayed response

   "Over-run" data after a group sequential test

   Inference on termination of a group sequential trial

Outline of course

3. Multiple testing procedures

Introduction to multiple testing

Graphical representation of multiple testing procedures

Combining multiple testing and group sequential tests

Testing a secondary endpoint after a group sequential test

4. Adaptive clinical trial designs (1)

Combination tests

Sample size re-estimation

Testing multiple hypotheses

Closed Testing Procedures

Outline of course

5. Adaptive clinical trial designs (2)

- Enrichment designs
- Seamless Phase 2/3 designs

6. Multi-armed group sequential trials

- Multi-armed multi-stage (MAMS) designs
- A survival trial with treatment selection
  - Avoiding type I error inflation
  - Assessing the benefits of an adaptive design

Part 1. Group sequential tests (1)

1.1. Introduction

1.2. Sequential distribution theory

1.3. Computations for group sequential tests

1.4. Benefits of group sequential testing

1.5. Error spending tests

1.6. Examples of group sequential designs:

With a normal response,

With a binary response,

With a survival endpoint

1.1 Group sequential tests: Introduction

Suppose a new treatment (Treatment A) is to be compared to a placebo or positive control (Treatment B) in a Phase III trial.

The treatment effect θ for the **primary endpoint** represents the advantage of Treatment A over Treatment B.

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

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

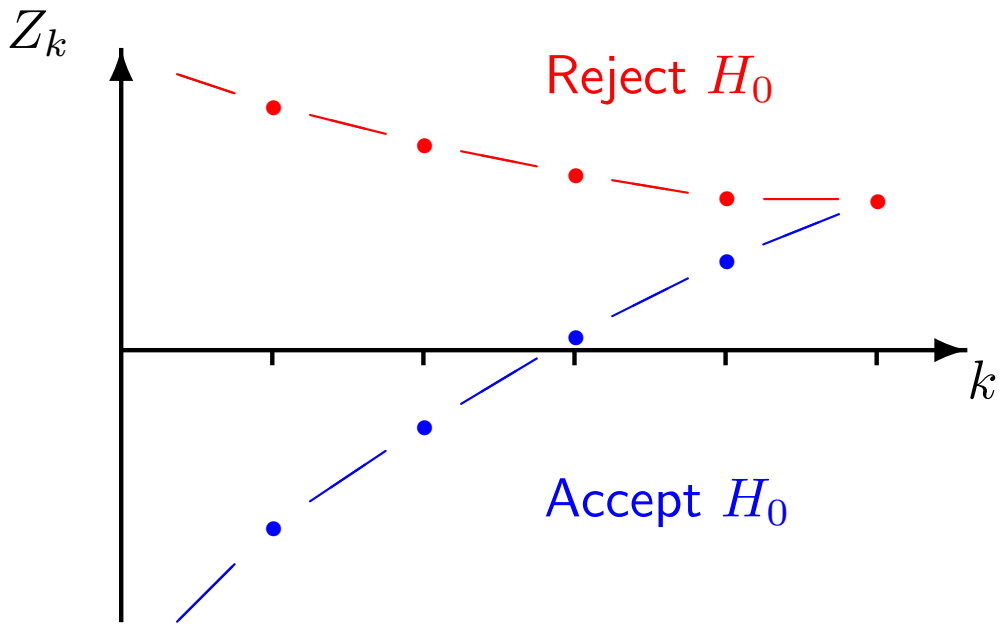
$$P_{\theta=0}\{\text{Reject } H_0\} = \alpha,$$

$$P_{\theta=\delta}\{\text{Reject } H_0\} = 1 - \beta.$$

In a group sequential trial, data are examined on a number of occasions to see if an early decision may be possible.

Group sequential tests

A typical boundary for a one-sided test, expressed in terms of standardised test statistics $Z_1, \ldots, Z_K$, has the form:



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

Crossing the lower boundary implies stopping for "futility" with acceptance of H_0.

Benefits of group sequential testing

Earlier decisions

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

Fewer patients recruited

Expected sample sizes for group sequential designs are, typically, around 60 to 70% of the fixed sample size for a trial with the same type I error rate and power.

Stopping failing trials early

Early stopping "for futility" can release resources to continue the development of other promising treatments.

1.2 Joint distribution of parameter estimates

Reference: Ch. 11 of *Group Sequential Methods with Applications to Clinical Trials*, Jennison & Turnbull, 2000 (hereafter, JT).

Let $\widehat{\theta}_k$ denote the estimate of θ based on data at analysis k.

The information for θ at analysis k is

$$\mathcal{I}_k = \{\text{Var}(\widehat{\theta}_k)\}^{-1}, \quad k = 1, \ldots, K.$$

Canonical joint distribution of $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$

In many situations, $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ are approximately multivariate normal,

$$\widehat{\theta}_k \sim N(\theta, \mathcal{I}_k^{-1}), \quad k = 1, \ldots, K,$$

and

$$\text{Cov}(\widehat{\theta}_{k_1}, \widehat{\theta}_{k_2}) = \text{Var}(\widehat{\theta}_{k_2}) = \mathcal{I}_{k_2}^{-1} \quad \text{for } k_1 < k_2.$$

Sequential distribution theory

The joint distribution of $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ can be derived directly for:

θ a single normal mean,

$\theta = \mu_A - \mu_B$, comparing two normal means.

The canonical distribution also applies when θ is a parameter in:

a general normal linear model,

a general model fitted by maximum likelihood (large sample theory).

Thus, theory supports general comparisons, including:

crossover studies,

analysis of longitudinal data,

comparisons adjusted for covariates.

Canonical joint distribution of $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$

A single normal mean

Suppose $X_1, X_2, \ldots$ are independent $N(\theta, \sigma^2)$ responses.

For $n_1 < n_2$, define

$$\widehat{\theta}_1 = \frac{X_1 + \ldots + X_{n_1}}{n_1}, \quad \widehat{\theta}_2 = \frac{X_1 + \ldots + X_{n_1} + \ldots + X_{n_2}}{n_2}.$$

The joint distribution of $\widehat{\theta}_1$ and $\widehat{\theta}_2$ is bivariate normal.

Marginally

$$\widehat{\theta}_1 \sim N(\theta, {\mathcal{I}_1}^{-1}) \quad \text{and} \quad \widehat{\theta}_2 \sim N(\theta, {\mathcal{I}_2}^{-1}),$$

where

$$\mathcal{I}_1 = \frac{n_1}{\sigma^2} \quad \text{and} \quad \mathcal{I}_2 = \frac{n_2}{\sigma^2}.$$

Canonical joint distribution of $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$

It remains to check the covariance:

$$\begin{aligned}
\mathsf{Cov}(\widehat{\theta}_1, \widehat{\theta}_2) &= \mathsf{Cov}\left(\frac{X_1 + \ldots + X_{n_1}}{n_1}, \frac{X_1 + \ldots + X_{n_1} + \ldots + X_{n_2}}{n_2} \right) \\
&= \mathsf{Cov}\left(\frac{X_1 + \ldots + X_{n_1}}{n_1}, \frac{X_1 + \ldots + X_{n_1}}{n_2} \right) \\
&= \frac{1}{n_1\, n_2} \mathsf{Var}(X_1 + \ldots + X_{n_1}) \\
&= \frac{\sigma^2}{n_2} = \mathcal{I}_2^{-1} \\
&= \mathsf{Var}(\widehat{\theta}_2).
\end{aligned}$$

Canonical joint distribution of $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$

A two-treatment comparison

Suppose observations on Treatments A and B, respectively, are

$$X_{Ai} \sim N(\mu_A, \sigma^2) \quad \text{and} \quad X_{Bi} \sim N(\mu_B, \sigma^2),$$

and $\theta = \mu_A - \mu_B$.

At analysis k, with n_{Ak} observations on Treatment A and n_{Bk} on Treatment B,

$$\widehat{\theta}_k \;=\; \widehat{\mu}_{A,k} - \widehat{\mu}_{B,k} \;=\; \frac{1}{n_{Ak}} \sum_{i=1}^{n_{Ak}} X_{Ai} - \frac{1}{n_{Bk}} \sum_{i=1}^{n_{Bk}} X_{Bi}.$$

Exercise: Show that $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ have the canonical joint distribution, i.e., they are multivariate normal,

$$\widehat{\theta}_k \;\sim\; N(\theta, \mathcal{I}_k{}^{-1}), \quad \text{where} \quad \mathcal{I}_k = \{\sigma^2/n_{Ak} + \sigma^2/n_{Bk}\}^{-1}$$

and $\mathrm{Cov}(\widehat{\theta}_{k_1}, \widehat{\theta}_{k_2}) = \mathrm{Var}(\widehat{\theta}_{k_2})$ for $k_1 < k_2$.

Canonical joint distribution of $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$

All efficient estimators have the canonical covariance

That is, if $\widehat{\theta}_1$ at analysis 1 and $\widehat{\theta}_2$ at analysis 2 are efficient, unbiased estimators of θ, then

$$\mathsf{Cov}(\widehat{\theta}_1, \widehat{\theta}_2) = \mathsf{Var}(\widehat{\theta}_2).$$

Proof:

Suppose $\mathsf{Cov}(\widehat{\theta}_1, \widehat{\theta}_2) \neq \mathsf{Var}(\widehat{\theta}_2)$, so $\mathsf{Cov}(\widehat{\theta}_1 - \widehat{\theta}_2, \widehat{\theta}_2) \neq 0$.

Consider an unbiased estimator of the form $\widehat{\theta}_2^* = \widehat{\theta}_2 + \epsilon\,(\widehat{\theta}_1 - \widehat{\theta}_2)$.

For ϵ small and of the opposite sign to $\mathsf{Cov}(\widehat{\theta}_1 - \widehat{\theta}_2, \widehat{\theta}_2)$,

$$\mathsf{Var}(\widehat{\theta}_2^*) < \mathsf{Var}(\widehat{\theta}_2),$$

contradicting the assumption that $\widehat{\theta}_2$ is an efficient estimator of θ.

Canonical joint distribution of Z-statistics

In testing H_0: $\theta = 0$, the *standardised statistic* at analysis k is

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

For these statistics,

$(Z_1, \ldots, Z_K)$ is multivariate normal,

$Z_k \sim N(\theta\sqrt{\mathcal{I}_k}, 1), \quad k = 1, \ldots, K,$

$\mathsf{Cov}(Z_{k_1}, Z_{k_2}) = \sqrt{\mathcal{I}_{k_1}/\mathcal{I}_{k_2}}$ for $k_1 < k_2$.

Canonical joint distribution of score statistics

The *score statistics*, $S_k = Z_k\sqrt{\mathcal{I}_k}$, are also multivariate normal with

$$S_k \sim N(\theta\,\mathcal{I}_k, \mathcal{I}_k), \quad k = 1, \ldots, K.$$

The score statistics possess the "independent increments" property,

$$\text{Cov}(S_k - S_{k-1},\, S_{k'} - S_{k'-1}) = 0 \quad \text{for } k \neq k'.$$

It can be helpful to know that the score statistics behave as Brownian motion with drift θ observed at times $\mathcal{I}_1, \ldots, \mathcal{I}_K$.

Survival data

The canonical joint distributions also arise for

a) estimates of a parameter in Cox's proportional hazards regression model,

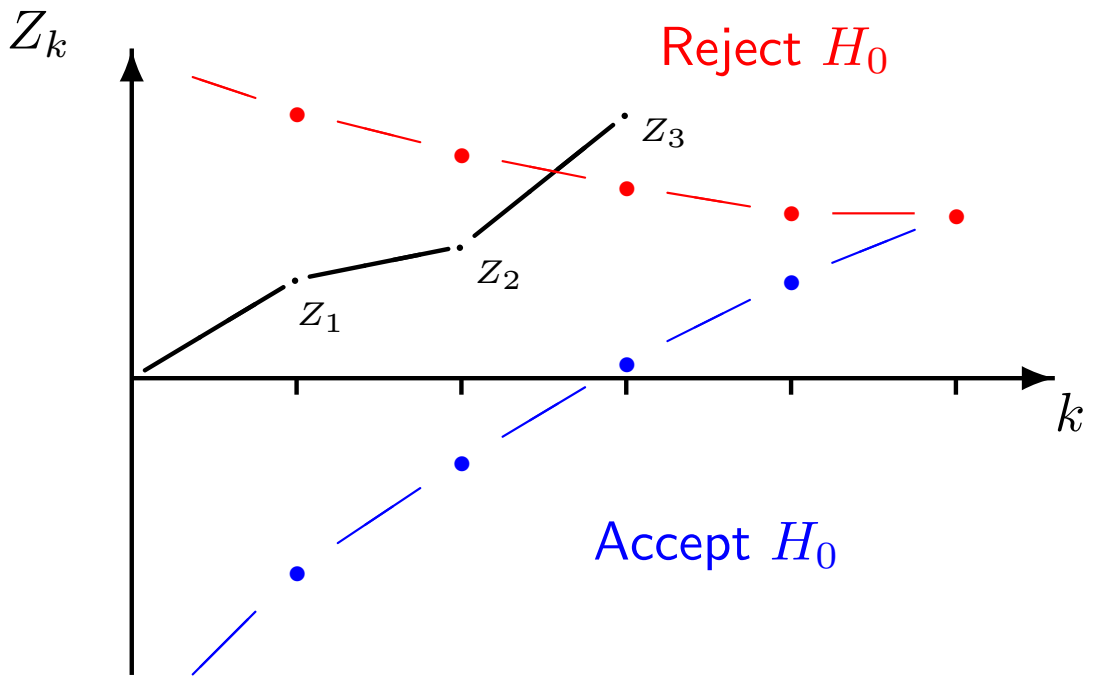
b) log-rank statistics for comparing two survival curves.

For survival data, observed information is roughly proportional to the number of failures.

The "error spending" approach can be used to define group sequential tests that can handle unpredictable and unevenly spaced information levels.

Reference: "Group-sequential analysis incorporating covariate information", Jennison & Turnbull (*JASA*, 1997).

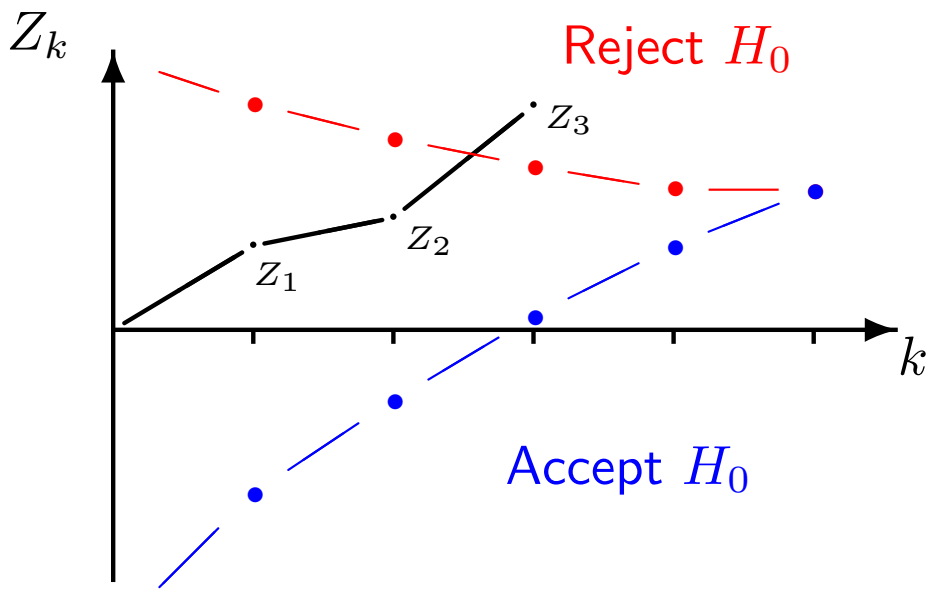
1.3 Computations for group sequential tests (GSTs)



In order to find $P_\theta\{\text{Reject } H_0\}$, etc., we need to calculate the probabilities of basic events such as

$$a_1 < Z_1 < b_1, \quad a_2 < Z_2 < b_2, \quad Z_3 > b_3.$$

Computations for group sequential tests



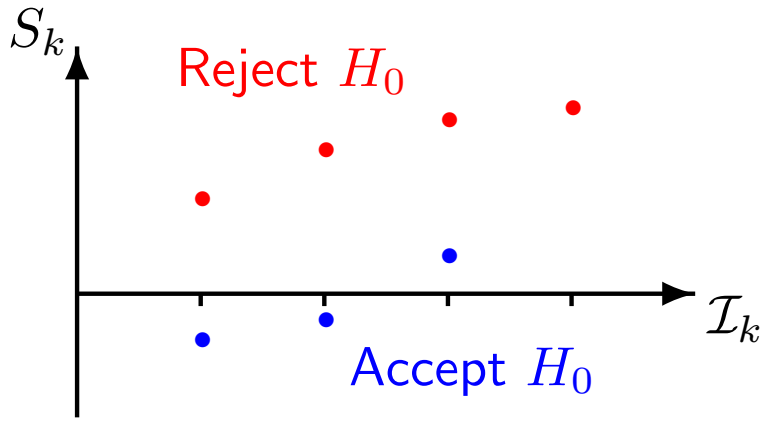
Probabilities such as $P_\theta\{a_1 < Z_1 < b_1,\ \ a_2 < Z_2 < b_2,\ \ Z_3 > b_3\}$ can be computed by repeated numerical integration (JT, Ch. 19).

Combining these probabilities yields type I error rate, power, expected sample size, etc., of a group sequential design.

Constants and group sizes can be chosen to define a test with a specific type I error probability and power.

One-sided tests: The Pampallona & Tsiatis family

To test H_0: $\theta \leq 0$ against the *one-sided* alternative $\theta > 0$ with type I error probability α and power $1 - \beta$ at $\theta = \delta$.



For the P & T test with parameter Δ, boundaries on the score statistic scale are

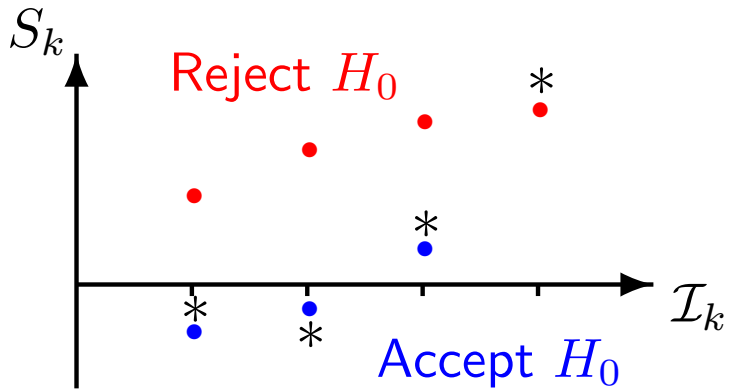
$$a_k = \mathcal{I}_k \, \delta - C_2 \, \mathcal{I}_k^{\Delta}, \quad b_k = C_1 \, \mathcal{I}_k^{\Delta}.$$

The computational methods described above can be used to find C_1, C_2 and $\mathcal{I}_K$ such that the test has the specified error rates.

Reference: Pampallona & Tsiatis (*JSPI*, 1994).

One-sided tests with a non-binding futility boundary

Regulators are not always convinced a trial monitoring committee will abide by the stopping boundary specified in the protocol.



The sample path shown above leads to rejection of H_0. Since such paths are not included in type I error calculations, the true type I error rate is under-estimated.

If a futility boundary is deemed to be *non-binding*, the type I error rate should be computed ignoring the futility boundary.

However, investigators will wish to know power and expected sample size when the futility boundary *is* obeyed.

1.4 Benefits of group sequential testing

In order to test H_0: $\theta \leq 0$ against $\theta > 0$ with type I error probability α and power $1-\beta$ at $\theta = \delta$, a fixed sample size study needs information

$$\mathcal{I}_{fix} = \frac{\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2}{\delta^2},$$

where Φ is the standard normal CDF.

Information is (roughly) proportional to sample size in many clinical trial settings.

A GST with K analyses will need to be able to continue to a maximum information level $\mathcal{I}_K$, greater than $\mathcal{I}_{fix}$.

On average, the GST can stop earlier than this and expected information on termination, $\mathbb{E}_\theta(\mathcal{I})$, will be considerably less than $\mathcal{I}_{fix}$, especially under extreme values of θ.

We call $R = \mathcal{I}_K / \mathcal{I}_{fix}$ the *inflation factor* of a group sequential test.

Optimal group sequential tests

We can seek a GST that minimises expected information $\mathbb{E}_\theta(\mathcal{I})$ under certain values of the treatment effect, θ, with a given number of analyses K and inflation factor R.

Eales & Jennison (*Biometrika*, 1992) and Barber & Jennison (*Biometrika*, 2002) optimise designs for criteria of the form

$$\sum_i w_i \, \mathbb{E}_{\theta_i}(\mathcal{I}) \quad \text{or} \quad \int f(\theta) \, \mathbb{E}_\theta(\mathcal{I}) \, d\theta,$$

where f is a normal density.

These optimised designs could be used in their own right.

They also serve as benchmarks for other methods which may have additional useful features (e.g., error spending tests).

Computing optimal group sequential tests

In optimising a GST, we create a Bayes sequential decision problem, placing a prior on θ and defining costs for sampling and for making incorrect decisions.

Such a problem can be solved rapidly by dynamic programming.

We then search for the combination of prior and costs such that the solution to the (unconstrained) Bayes decision problem has the specified frequentist error rates α at $\theta = 0$ and β at $\theta = \delta$.

The resulting design solves both the Bayes decision problem and the original frequentist problem.

NB: Although the Bayes decision problem is introduced as a computational device, this derivation demonstrates that an efficient frequentist design should also be a good Bayesian procedure.

Benefits of group sequential testing

One-sided GSTs with binding futility boundaries, minimising $\{\mathbb{E}_0(\mathcal{I}) + \mathbb{E}_\delta(\mathcal{I})\}/2$ for K equally sized groups, $\alpha = 0.025$, $1 - \beta = 0.9$ and $\mathcal{I}_{max} = R\mathcal{I}_{fix}$.

Minimum values of $\{\mathbb{E}_0(\mathcal{I}) + \mathbb{E}_\delta(\mathcal{I})\}/2$, as a percentage of $\mathcal{I}_{fix}$

	R					
K	1.01	1.05	1.1	1.2	1.3	*Minimum over R*
2	80.8	74.7	**73.2**	73.7	75.8	73.0 at R=1.13
3	76.2	69.3	66.6	**65.1**	65.2	65.0 at R=1.23
5	72.2	65.2	62.2	59.8	**59.0**	58.8 at R=1.38
10	69.2	62.2	59.0	56.3	**55.1**	54.2 at R=1.6
20	67.8	60.6	57.5	54.6	**53.3**	51.7 at R=1.8

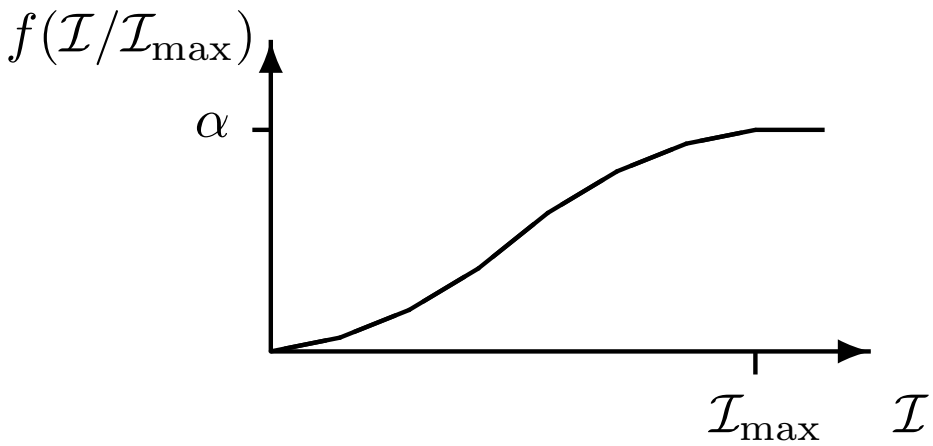
Note: $\mathbb{E}(\mathcal{I}) \searrow$ as $K \nearrow$ but with diminishing returns,
$\mathbb{E}(\mathcal{I}) \searrow$ as $R \nearrow$ up to a point.

1.5 Error spending tests (JT Ch. 7)

When the sequence $\mathcal{I}_1, \mathcal{I}_2, \ldots$ is unpredictable, a group sequential design must adapt to observed information levels.

Lan & DeMets (*Biometrika*, 1983) introduced "error spending" tests of H_0: $\theta = 0$ against $\theta \neq 0$.

Maximum information design with spending function $f(\mathcal{I}/\mathcal{I}_{\max})$



The boundary at analysis k is set to give cumulative type I error probability $f(\mathcal{I}_k/\mathcal{I}_{\max})$.

If $\mathcal{I}_{\max}$ is reached without rejecting H_0, then H_0, is accepted.

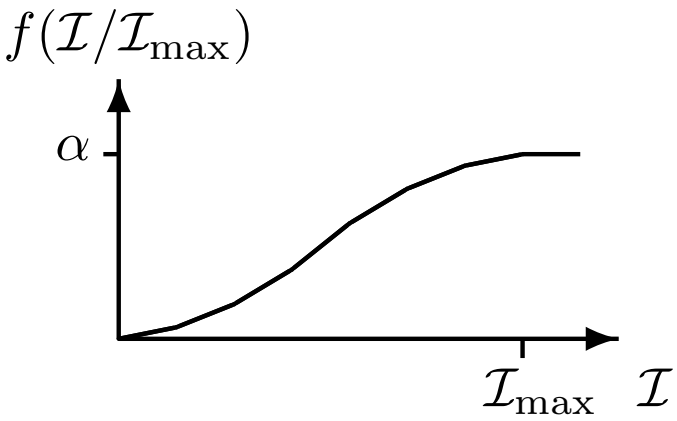
One-sided error spending tests

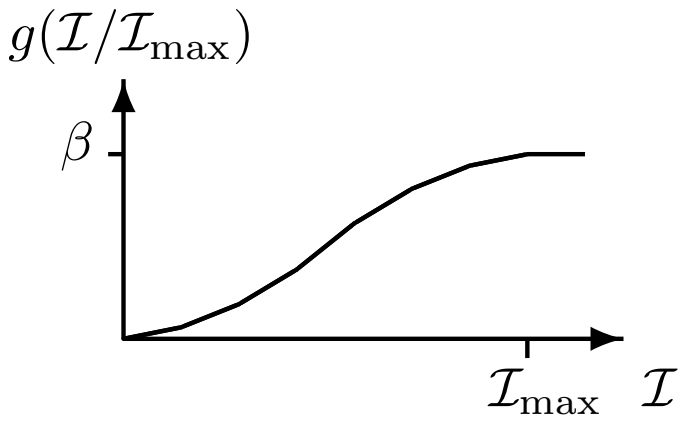
For a one-sided test of H_0: $\theta \leq 0$ against $\theta > 0$ with

Type I error probability α at $\theta = 0$,

Type II error probability β at $\theta = \delta$,

we need two error spending functions.





Type I error probability α is spent according to the function $f(\mathcal{I}/\mathcal{I}_{\max})$, and type II error probability β according to $g(\mathcal{I}/\mathcal{I}_{\max})$.

One-sided error-spending tests

Analysis 1:

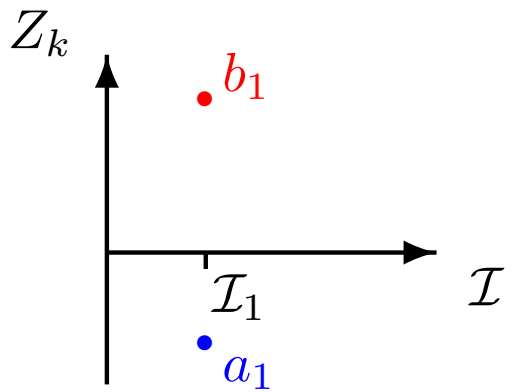
Observed information $\mathcal{I}_1$.

Reject H_0 if $Z_1 > b_1$, where

$$P_{\theta=0}\{Z_1 > b_1\} = f(\mathcal{I}_1/\mathcal{I}_{\max}).$$

Accept H_0 if $Z_1 < a_1$, where

$$P_{\theta=\delta}\{Z_1 < a_1\} = g(\mathcal{I}_1/\mathcal{I}_{\max}).$$



One-sided error-spending tests

Analysis 2: Observed information $\mathcal{I}_2$

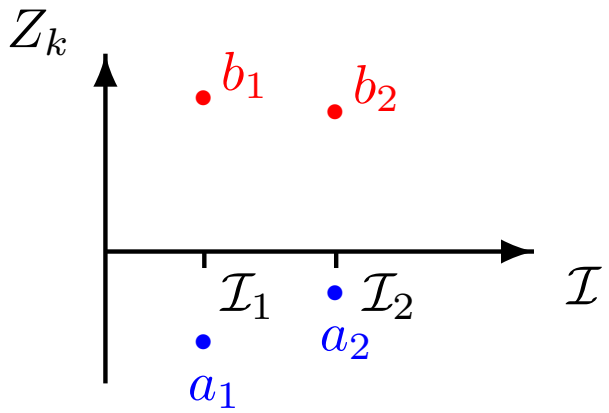
Reject H_0 if $Z_2 > b_2$, where

$$P_{\theta=0}\{a_1 < Z_1 < b_1, Z_2 > b_2\} = f(\mathcal{I}_2/\mathcal{I}_{\max}) - f(\mathcal{I}_1/\mathcal{I}_{\max})$$

— *note that, for now, we assume the futility boundary is binding.*

Accept H_0 if $Z_2 < a_2$, where

$$P_{\theta=\delta}\{a_1 < Z_1 < b_1, Z_2 < a_2\} = g(\mathcal{I}_2/\mathcal{I}_{\max}) - g(\mathcal{I}_1/\mathcal{I}_{\max}).$$



One-sided error-spending tests

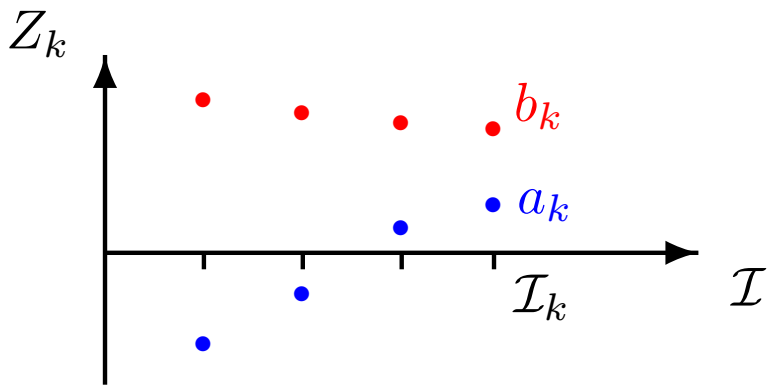
Analysis k: Observed information $\mathcal{I}_k$

Find a_k and b_k to satisfy

$$P_{\theta=0}\{a_1 < Z_1 < b_1, \ldots, a_{k-1} < Z_{k-1} < b_{k-1}, Z_k > b_k\}$$
$$= f(\mathcal{I}_k/\mathcal{I}_{\max}) - f(\mathcal{I}_{k-1}/\mathcal{I}_{\max}),$$

and

$$P_{\theta=\delta}\{a_1 < Z_1 < b_1, \ldots, a_{k-1} < Z_{k-1} < b_{k-1}, Z_k < a_k\}$$
$$= g(\mathcal{I}_k/\mathcal{I}_{\max}) - g(\mathcal{I}_{k-1}/\mathcal{I}_{\max}).$$

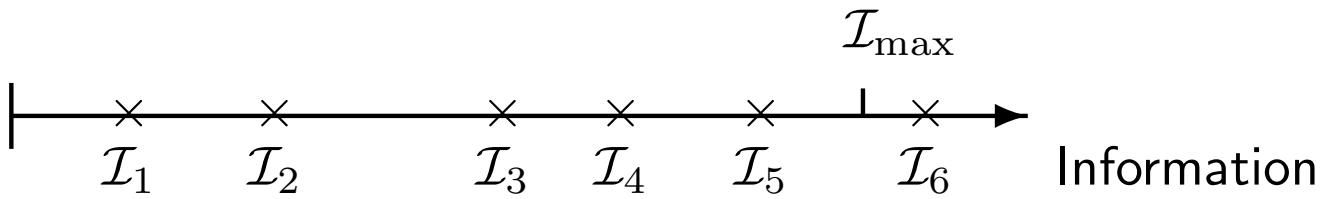


Remarks on error spending tests

1. Computation of (a_k, b_k) does **not** depend on future information levels, $\mathcal{I}_{k+1}$, $\mathcal{I}_{k+2}$, $\ldots$.

2. A "maximum information design" continues until a boundary is crossed or an analysis with $\mathcal{I}_k \geq \mathcal{I}_{\max}$ is reached.

If necessary, patient accrual can be extended to reach $\mathcal{I}_{\max}$.



3. If a maximum of K analyses is specified, the study terminates at analysis K with $f(\mathcal{I}_K/\mathcal{I}_{\max})$ defined to be α.

Then, b_K is chosen to give cumulative type I error probability α and we set $a_K = b_K$.

Remarks on error spending tests

4. The value of $\mathcal{I}_{\max}$ can be chosen so that boundaries converge at the final analysis when, say,

$$\mathcal{I}_k = (k/K)\,\mathcal{I}_{\max}, \quad k = 1, \ldots, K.$$

5. In a one-sided test with ρ-family error spending function, type I error probability is spent as

$$f(\mathcal{I}/\mathcal{I}_{\max}) = \alpha \, \min\{1, \, (\mathcal{I}/\mathcal{I}_{\max})^{\rho}\}$$

and type II error probability as

$$g(\mathcal{I}/\mathcal{I}_{\max}) = \beta \, \min\{1, \, (\mathcal{I}/\mathcal{I}_{\max})^{\rho}\}.$$

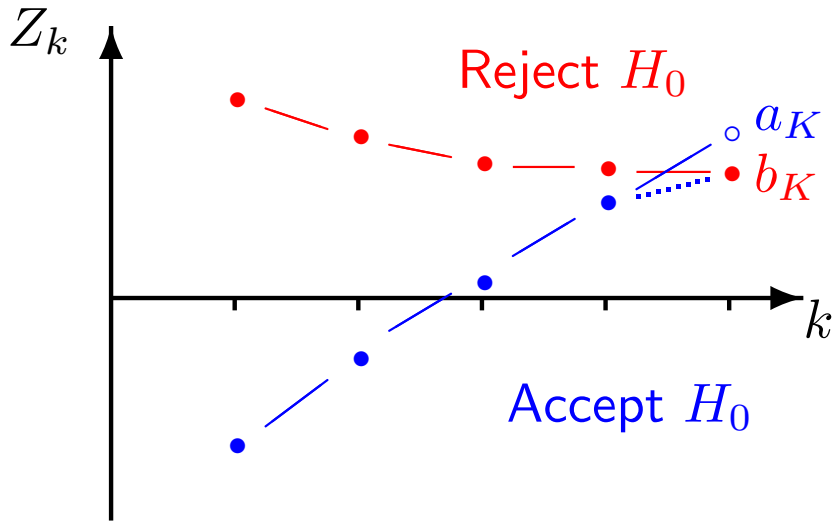
The value of ρ determines the inflation factor $R = \mathcal{I}_{\max}/\mathcal{I}_{fix}$.

Barber & Jennison (*Biometrika*, 2002) show the ρ-family provides tests with excellent efficiency for a given number of analyses K and inflation factor R.

Error spending tests: Over-running

The final analysis of a one-sided error spending test needs care.

If $\mathcal{I}_K > \mathcal{I}_{\max}$, solving for a_K and b_K is likely to give $a_K > b_K$.



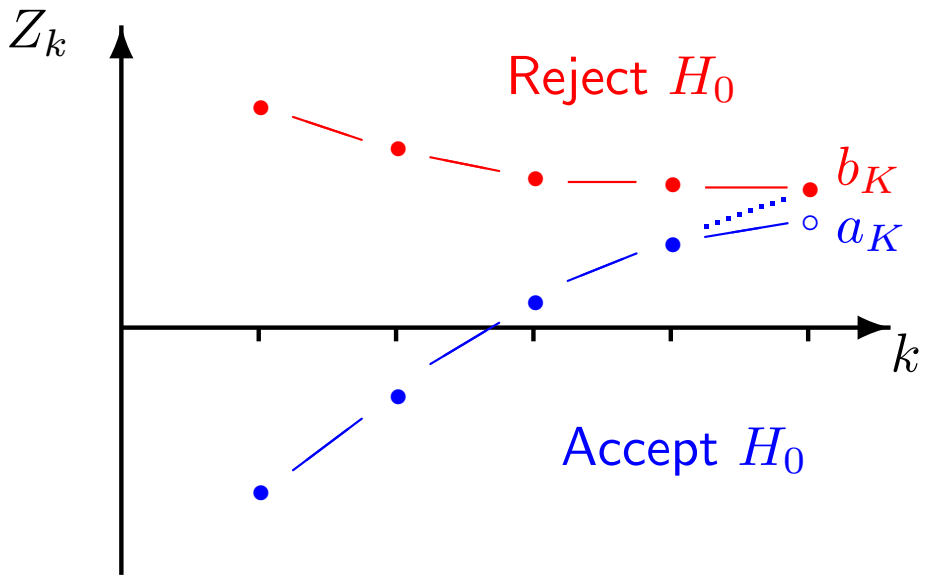
The value calculated for b_K guarantees type I error probability equal to α. So, reduce a_K to b_K — and gain extra power.

Even if $\mathcal{I}_K = \mathcal{I}_{\max}$, one may find $a_K > b_K$ if information levels deviate from the equally spaced values (say) used in setting $\mathcal{I}_{\max}$.

Error spending tests: Under-running

A final value $\mathcal{I}_K < \mathcal{I}_{\max}$ may arise when the last planned analysis is reached, e.g., at a maximum follow-up time in a survival study.

Then, solving for a_K and b_K is likely to give $a_K < b_K$.



Again, the value calculated for b_K gives type I error probability α.

So increase a_K to b_K — and attained power will be below $1 - \beta$.

One-sided error-spending tests: Non-binding futility

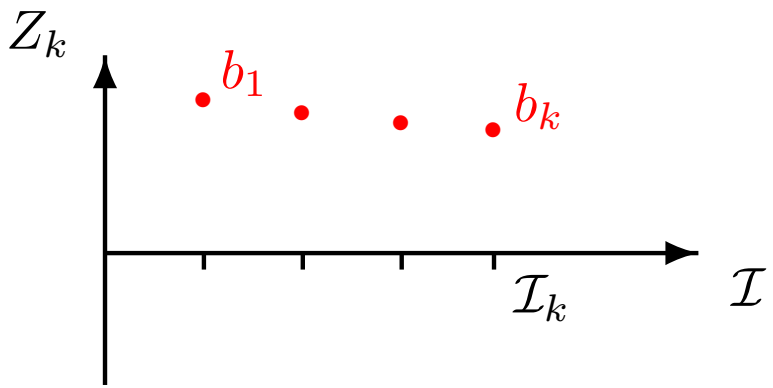
If the futility boundary is treated as non-binding, computation of the efficacy boundary only involves the type I error spending function $f(\mathcal{I}/\mathcal{I}_{\max})$.

Boundary values, b_1, b_2, $\ldots$, are calculated as the trial proceeds.

Analysis k: Observed information $\mathcal{I}_k$

Reject H_0 if $Z_k > b_k$, where

$$P_{\theta=0}\{Z_1 < b_1,\ \ldots\ ,\ Z_{k-1} < b_{k-1},\ Z_k > b_k\}$$
$$= f(\mathcal{I}_k/\mathcal{I}_{\max}) - f(\mathcal{I}_{k-1}/\mathcal{I}_{\max}).$$



One-sided error-spending tests: Non-binding futility

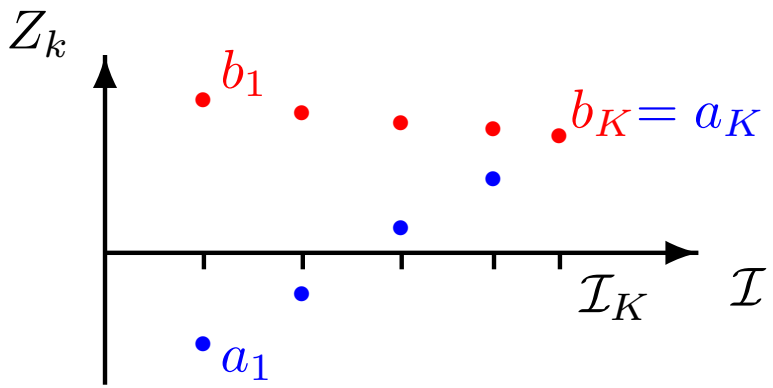
A futility boundary can be added through a type II error spending function $g(\mathcal{I}/\mathcal{I}_{\max})$.

For $k = 1, \ldots, K-1$*:*

At analysis k with observed information $\mathcal{I}_k$, set a_k to satisfy

$$P_{\theta=\delta}\{a_1 < Z_1 < b_1, \ldots, a_{k-1} < Z_{k-1} < b_{k-1}, Z_k < a_k\}$$
$$= g(\mathcal{I}_k/\mathcal{I}_{\max}) - g(\mathcal{I}_{k-1}/\mathcal{I}_{\max}).$$

For $k = K$*:* Set $a_K = b_K$.



1.6 (a) An error spending test with normal response

Consider a two-treatment comparison with responses

$$X_{Ai} \sim N(\mu_A, \sigma^2) \quad \text{on Treatment A,}$$

$$X_{Bi} \sim N(\mu_B, \sigma^2) \quad \text{on Treatment B.}$$

Setting $\theta = \mu_A - \mu_B$, we wish to test H_0: $\theta \leq 0$ against $\theta > 0$ with

Type I error rate $\alpha = 0.025$,

Power $1 - \beta = 0.9$ at $\theta = \delta = 0.4$.

We shall apply a ρ-family error spending design with $\rho = 2$, spending type I error probability as

$$f(\mathcal{I}/\mathcal{I}_{\max}) = \alpha \, \min\{1, \, (\mathcal{I}/\mathcal{I}_{\max})^2\}$$

and type II error probability as

$$g(\mathcal{I}/\mathcal{I}_{\max}) = \beta \, \min\{1, \, (\mathcal{I}/\mathcal{I}_{\max})^2\}.$$

A one-sided test with a non-binding futility boundary

Information

Suppose it is known that $\sigma^2 = 0.64$. (This is, of course, an unusual assumption — we consider the case of unknown σ^2 at the end of this Section.)

With total numbers of observations n_A on Treatment A and n_B on Treatment B, the estimated treatment effect has variance

$$\text{Var}(\widehat{\theta}) = \left(\frac{1}{n_A} + \frac{1}{n_B}\right)\sigma^2 = \left(\frac{1}{n_A} + \frac{1}{n_B}\right)0.64$$

and the Fisher information for θ is

$$\mathcal{I} = \{\text{Var}(\widehat{\theta})\}^{-1}.$$

It is this *information* that appears in the error spending functions.

Applying a ρ-family error spending test

A fixed sample trial with power 0.9 at $\theta = 0.4$ needs information

$$\mathcal{I}_{fix} = \frac{\{\Phi^{-1}(0.975) + \Phi^{-1}(0.9)\})^2}{0.4^2} = 65.7.$$

The ρ-family error spending test with $\rho = 2$, with 5 equally spaced analyses, and a **non-binding** futility boundary has an inflation factor $R = 1.133$.

Thus, this design needs $\mathcal{I}_{\max} = 1.133 \times 65.7 = 74.39$ to satisfy type I error and power requirements.

Information level $\mathcal{I}_{\max} = 74.39$ will be achieved by sample sizes

$$n_A = n_B = 95.$$

Thus, recruitment should be organised to reach this target at the 5th analysis.

Applying a ρ-family error spending test

Suppose we observe $\widehat{\theta}_1 = 0.10$ at analysis 1 based on $n_A = n_B = 20$ observations per treatment. Thus,

$$\text{Var}(\widehat{\theta}_1) = \left(\frac{1}{20} + \frac{1}{20}\right) 0.64 = 0.064$$

and the Fisher information for θ at this analysis is

$$\mathcal{I}_1 = 0.064^{-1} = 15.6.$$

Since $\mathcal{I}_{\max} = 74.39$, the type I and II error probabilities to be spent are

$$f(\mathcal{I}_1/\mathcal{I}_{\max}) = 0.025\,(15.6/74.39)^2 = 0.00110,$$

$$g(\mathcal{I}_1/\mathcal{I}_{\max}) = 0.1\,(15.6/74.39)^2 = 0.00440.$$

It follows that boundary values are $a_1 = -1.038$ and $b_1 = 3.061$ on the Z-scale.

Applying a ρ-family error spending test

Applying the stopping boundary at the first analysis

The standard error of $\widehat{\theta}_1$ is $0.064^{1/2} = 0.253$.

Hence

$$Z_1 = \frac{\widehat{\theta}_1}{s.e.\,(\widehat{\theta}_1)} = \frac{0.10}{0.253} = 0.395.$$

The boundary values are $a_1 = -1.038$ and $b_1 = 3.061$.

Since $a_1 < Z_1 < b_1$, the trial continues to the next analysis.

Applying the stopping boundary at subsequent analyses

Successive analyses proceed along the same lines until a boundary is crossed or the final analysis is reached.

Applying a ρ-family error spending test

After further analyses, suppose the cumulative sample sizes and information levels $\mathcal{I}_k$ are as recorded below.

Analysis k	*Cumulative sample size* $n_A + n_B$	$\mathcal{I}_k$	*Boundary* a_k	b_k
1	40	15.6	-1.038	3.061
2	80	31.2	0.072	2.681
3	120	46.9	0.887	2.436
4	164	64.1	1.653	2.213
5	190	74.2	2.135	2.135

The test with a **non-binding** futility boundary, has critical values a_k and b_k as shown.

If the futility boundary is actually obeyed, the attained type I error rate is 0.023 and power 0.898 is achieved when $\theta = 0.4$.

Applying a ρ-family error spending test

If the observed treatment effect estimates are $\widehat{\theta}_1 = 0.10$, $\widehat{\theta}_2 = 0.06$, $\widehat{\theta}_3 = 0.21$, and $\widehat{\theta}_4 = 0.31$, then the trial stops to reject H_0 at analysis 4.

Analysis k	$\mathcal{I}_k$	*Boundary* a_k	*Boundary* b_k	$\hat{\theta}_k$	*s.e.* $(\hat{\theta}_k)$	Z_k
1	15.6	-1.038	3.061	0.10	0.253	0.395
2	31.2	0.072	2.681	0.06	0.179	0.335
3	46.9	0.887	2.436	0.21	0.146	1.438
4	64.1	1.653	2.213	0.31	0.125	2.481
5	—	—	—	—	—	—

In this case, $\mathcal{I}_5$ and $\widehat{\theta}_5$ are not observed.

An error spending test with a binding futility boundary

Suppose the same trial is conducted with a **binding** futility boundary — using the same f and g, and with $\mathcal{I}_{max} = 74.39$.

Then, we have:

Analysis k	*Cumulative sample size* $n_A + n_B$	$\mathcal{I}_k$	*Boundary* a_k	b_k
1	40	15.6	-1.038	3.061
2	80	31.2	0.072	2.681
3	120	46.9	0.887	2.436
4	164	64.1	1.653	**2.203**
5	190	74.2	**2.044**	**2.044**

The upper boundary is now lower at analyses 4 and 5.

With a binding futility boundary, the lower efficacy boundary gives higher power: when $\theta = 0.4$, the power is 0.905.

GSTs for normal data with unknown variance

We can modify the preceding methods to deal with unknown σ^2.

Plans are made assuming an initial estimate of σ^2, but updated as new estimates of σ^2 become available.

In order to test H_0: $\theta \leq 0$ vs $\theta > 0$ with type I error rate α and power $1-\beta$ at $\theta = \delta$, a fixed sample test requires

$$\mathcal{I}_{fix} = \frac{\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\alpha)\}^2}{\delta^2}$$

and for an error spending design with inflation factor R, the target information level is $\mathcal{I}_{\max} = R\mathcal{I}_{fix}$.

Thus we modify sample size during the study, aiming to achieve

$$\left\{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)\hat{\sigma}^2\right\}^{-1} = R\mathcal{I}_{fix}$$

at the final analysis.

GSTs for normal data with unknown variance

In an error spending test, information is spent as a function of the estimated information at each analysis, based on the current $\hat{\sigma}^2$.

A boundary $\{a_k, b_k\}$ for test statistics $\{Z_k\}$ is calculated following the error spending approach, as in the case of known σ^2.

This is then converted to a boundary for t-statistics $\{T_k\}$ by preserving the significance level of each boundary point.

At analysis k, define

$$p_{1k} = 1 - \Phi(a_k) \quad \text{and} \quad p_{2k} = 1 - \Phi(b_k).$$

If T_k has ν_k degrees of freedom, let $\tilde{a}_k$ and $\tilde{b}_k$ be such that

$$P(T_{\nu_k} > \tilde{b}_k) = p_{2k} \quad \text{and} \quad P(T_{\nu_k} > \tilde{a}_k) = p_{1k}.$$

The test stops to reject H_0 if $T_k > \tilde{b}_k$ and to accept H_0 if $T_k < \tilde{a}_k$.

East: Demonstration, Normal data

We wish to plan a two treatment comparison, testing for *superiority*, i.e., testing

$$H_0\text{: } \theta \leq 0 \quad \text{vs} \quad \theta > 0,$$

where $\theta = \mu_A - \mu_B$.

We shall consider how to:

- Design a one-sided, error spending test with

  Type I error probability, $\alpha = 0.025$,

  Power $1 - \beta = 0.9$ at $\theta = 0.4$ when $\sigma^2 = 0.64$,

- Apply the test,

- Include a binding or non-binding futility boundary.

1.6 (b) An error spending test with binary data

Treatment for heart failure

A new treatment is to be compared to the current standard.

The primary endpoint

is re-admission to hospital (or death) within 30 days.

The current treatment

has a re-admission rate of 25%.

Testing for superiority

It is hoped the new treatment will reduce re-admissions to 20%.

Denote re-admission probabilities by p_t and p_c on the new treatment and control.

To establish superiority of the new treatment, we carry out a test of H_0: $p_t \geq p_c$ against $p_t < p_c$ — hoping to reject H_0.

Binary example: The testing problem

Setting $\theta = p_c - p_t$, we wish to test H_0: $\theta \leq 0$ against $\theta > 0$ with

Type I error rate $\alpha = 0.025$ at $\theta = 0$,

Power $1 - \beta = 0.9$ when $\theta = \delta = 0.05$.

Let

$n_t, y_t =$ Numbers of subjects, re-admissions on the treatment arm,

$n_c, y_c =$ Numbers of subjects, re-admissions on the control arm,

$\hat{p}_c = y_c/n_c, \quad \hat{p}_t = y_t/n_t.$

For large n_t and n_c we have, approximately,

$$\widehat{\theta} = \hat{p}_c - \hat{p}_t \sim N\left(\theta, \frac{p_c(1-p_c)}{n_c} + \frac{p_t(1-p_t)}{n_t}\right).$$

Binary example: A fixed sample test

A fixed sample test requires information

$$\begin{aligned}\mathcal{I}_{fix} &= \{\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\beta)\}^2 / \delta^2 \\ &= (\{\Phi^{-1}(0.975)+\Phi^{-1}(0.9)\}^2 / 0.05^2 \\ &= 4203.2.\end{aligned}$$

With equal allocation to the two treatments and $n_t = n_c = n$,

$$\mathcal{I} = (\mathsf{Var}(\widehat{\theta}))^{-1} = \left(\frac{p_c(1-p_c)}{n} + \frac{p_t(1-p_t)}{n}\right)^{-1}.$$

Calculating power under the alternative $p_c = 0.25$ and $p_t = 0.2$, we find a fixed sample size test requires

$$n = 1461$$

subjects per treatment arm.

NB This sample size depends on p_c and p_t, not just $\theta = p_c - p_t$.

Binary example: A group sequential design

Suppose investigators choose:

- A ρ-family, one-sided error spending test with $\rho = 3$ (in f and g),
- Type I error rate $\alpha = 0.025$, power 0.9 when $\theta = 0.05$,
- A total of 5 analyses, and a **binding** futility boundary.

This test has inflation factor $R = 1.049$, so the maximum information level is

$$\mathcal{I}_{\max} \;=\; 1.049 \times 4203.2 \;=\; 4409.2.$$

Since $\mathcal{I} \;=\; n\left\{p_c(1-p_c) + p_t(1-p_t)\right\}^{-1}$, this will require up to 1533 subjects per treatment when $p_c = 0.25$ and $p_t = 0.2$.

Using an error spending test in a maximum information design allows re-assessment of the sample size needed to reach $\mathcal{I}_{max}$.

Binary example: Applying the error spending test

At analysis k: Using current estimates $\hat{p}_c$ and $\hat{p}_t$, calculate

$$\widehat{\mathcal{I}}_k \;=\; \{\,\hat{p}_c(1-\hat{p}_c)/n_c + \hat{p}_t(1-\hat{p}_t)/n_t\,\}^{-1}$$

and

$$Z_k \;=\; \frac{\hat{p}_c - \hat{p}_t}{\sqrt{\{\,\hat{p}_c(1-\hat{p}_c)/n_c + \hat{p}_t(1-\hat{p}_t)/n_t\,\}}} \;=\; \widehat{\theta}_k\,\sqrt{\widehat{\mathcal{I}}_k}.$$

Compute boundary values a_k and b_k using error spending functions

$$f(\mathcal{I}/\mathcal{I}_{\max}) = 0.025\,\min\{1,\,(\mathcal{I}/\mathcal{I}_{\max})^3\},$$

$$g(\mathcal{I}/\mathcal{I}_{\max}) = 0.1\,\min\{1,\,(\mathcal{I}/\mathcal{I}_{\max})^3\}.$$

Apply the stopping rule

If $Z_k < a_k$: stop, accept H_0,

If $Z_k > b_k$: stop, reject H_0.

Binary example: Information monitoring

The re-admission rates used in sample size calculations, $p_c = 0.25$ and $p_t = 0.2$, may not hold in practice.

These rates can be re-estimated from observed data.

Information is related to sample size per treatment by

$$\mathcal{I} \;=\; n\left\{p_c(1-p_c) + p_t(1-p_t)\right\}^{-1} \;=\; n\,\gamma^{-1}, \quad \text{say.}$$

At an interim analysis, estimate $\gamma = p_c(1-p_c) + p_t(1-p_t)$ by

$$\widehat{\gamma} \;=\; \hat{p}_c\,(1-\hat{p}_c) + \hat{p}_t\,(1-\hat{p}_t).$$

Then, use this value to compute the target sample size per treatment group,

$$\widehat{n}_{\max} \;=\; \widehat{\gamma}\,\mathcal{I}_{\max}$$

and modify remaining group sizes to reach this target at the final planned analysis.

Binary example: Illustrative data

Analysis 1

Control treatment		Experimental treatment	
$n_c = 310,$	$y_c = 73$	$n_t = 306,$	$y_t = 70$
$\hat{p}_c = 0.236$	$(s.e.\ 0.024)$	$\hat{p}_t = 0.229$	$(s.e.\ 0.024)$
$\widehat{\theta}_1 = 0.007$	$(s.e.\ 0.034)$		
$Z_1 = 0.20,$	$\mathcal{I}_1 = 864$	$a_1 = -1.70,$	$b_1 = 3.56$

Analysis 2

Control treatment		Experimental treatment	
$n_c = 612,$	$y_c = 151$	$n_t = 602,$	$y_t = 141$
$\hat{p}_c = 0.247$	$(s.e.\ 0.017)$	$\hat{p}_t = 0.234$	$(s.e.\ 0.017)$
$\widehat{\theta}_2 = 0.013$	$(s.e.\ 0.024)$		
$Z_2 = 0.51,$	$\mathcal{I}_2 = 1662$	$a_2 = -0.54,$	$b_2 = 3.03$

Binary example: Illustrative data

Analysis 3

Control treatment		Experimental treatment	
$n_c = 915,$	$y_c = 238$	$n_t = 925,$	$y_t = 202$
$\hat{p}_c = 0.260$	$(s.e.\ 0.014)$	$\hat{p}_t = 0.218$	$(s.e.\ 0.014)$
$\widehat{\theta}_3 = 0.042$	$(s.e.\ 0.020)$		
$Z_3 = 2.10,$	$\mathcal{I}_3 = 2532$	$a_3 = 0.39,$	$b_3 = 2.65$

Analysis 4

Control treatment		Experimental treatment	
$n_c = 1225,$	$y_c = 324$	$n_t = 1222,$	$y_t = 268$
$\hat{p}_c = 0.264$	$(s.e.\ 0.013)$	$\hat{p}_t = 0.219$	$(s.e.\ 0.012)$
$\widehat{\theta}_4 = 0.045$	$(s.e.\ 0.017)$		
$Z_4 = 2.61,$	$\mathcal{I}_4 = 3345$	$a_4 = 1.12,$	$b_4 = 2.37$

— Stop, reject H_0 —

Binary example: Illustrative data

Summary of the application of a one-sided error spending test:

Analysis k	$\mathcal{I}_k$	*Boundary* a_k	b_k	$\hat{\theta}_k$	s.e. $(\hat{\theta}_k)$	Z_k
1	864	-1.70	3.56	0.007	0.034	0.20
2	1662	-0.54	3.03	0.013	0.024	0.51
3	2532	0.39	2.65	0.042	0.020	2.10
4	3345	1.12	2.37	0.045	0.017	2.61

The upper boundary is crossed at analysis 4 out of 5 and the null hypothesis H_0: $\theta \leq 0$ is rejected.

Had the trial continued past analysis 4, putting $\hat{p}_c = 0.264$ and $\hat{p}_t = 0.219$ in the formula for $\text{Var}(\widehat{\theta})$ would have led to a target of 1611 patients per treatment arm to achieve $\mathcal{I}_5 = 4409.2$.

East: Demonstration, Binary data

We wish to plan a two treatment comparison, testing for *superiority*, i.e., testing

$$H_0\text{: } \theta \leq 0 \quad \text{vs} \quad \theta > 0,$$

where $\theta = p_c - p_t$.

We shall consider how to:

- Design a one-sided, error spending test with

  Type I error probability, $\alpha = 0.025$,

  Power $1 - \beta = 0.9$ at $\theta = 0.05$,

- Apply the test with "information monitoring"

- Include a binding or non-binding futility boundary.

1.6 (c) An error spending test with survival data

Example: Oropharynx Clinical Trial Data

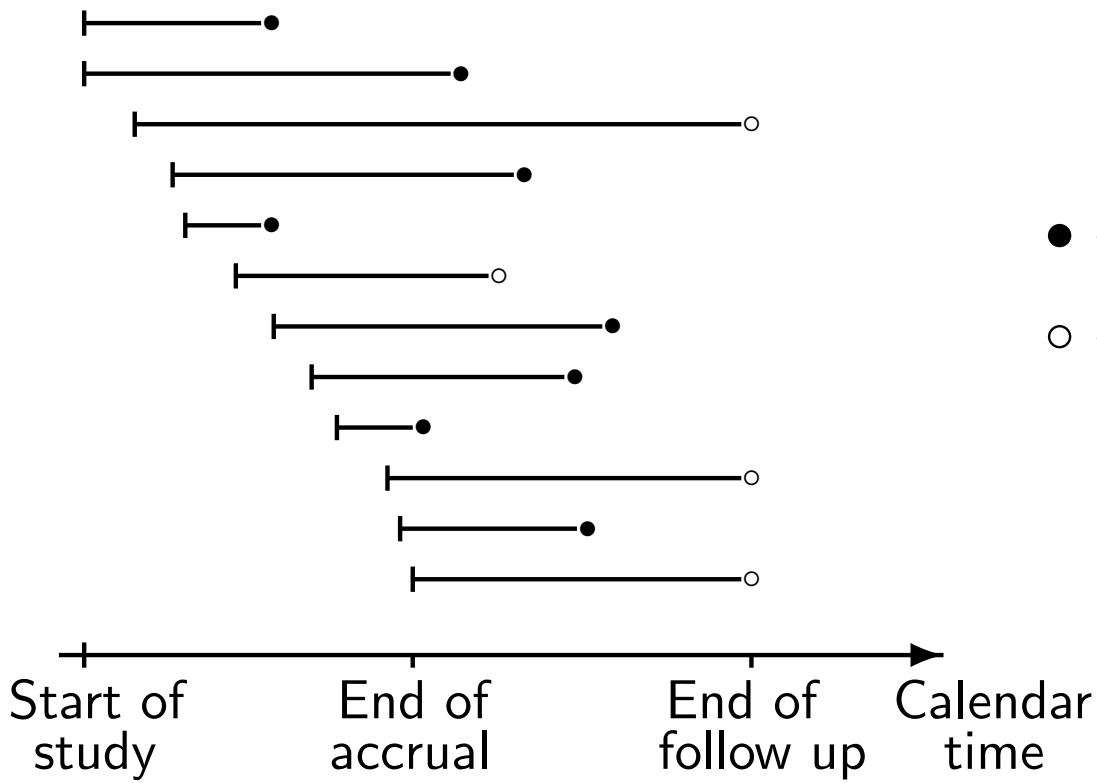
Survival of patients on experimental Treatment A and standard Treatment B.

Analysis k	Date	Number entered Trt A	Number entered Trt B	Number of deaths Trt A	Number of deaths Trt B
1	12/69	38	45	13	14
2	12/70	56	70	30	28
3	12/71	81	93	44	47
4	12/72	95	100	63	66
5	12/73	95	100	69	73

From Kalbfleisch & Prentice (2002) *The Statistical Analysis of Failure Time Data, 2nd edition*, Appendix A, Data Set II.

See also JT, Ch. 13.

Accrual and follow up in a survival study



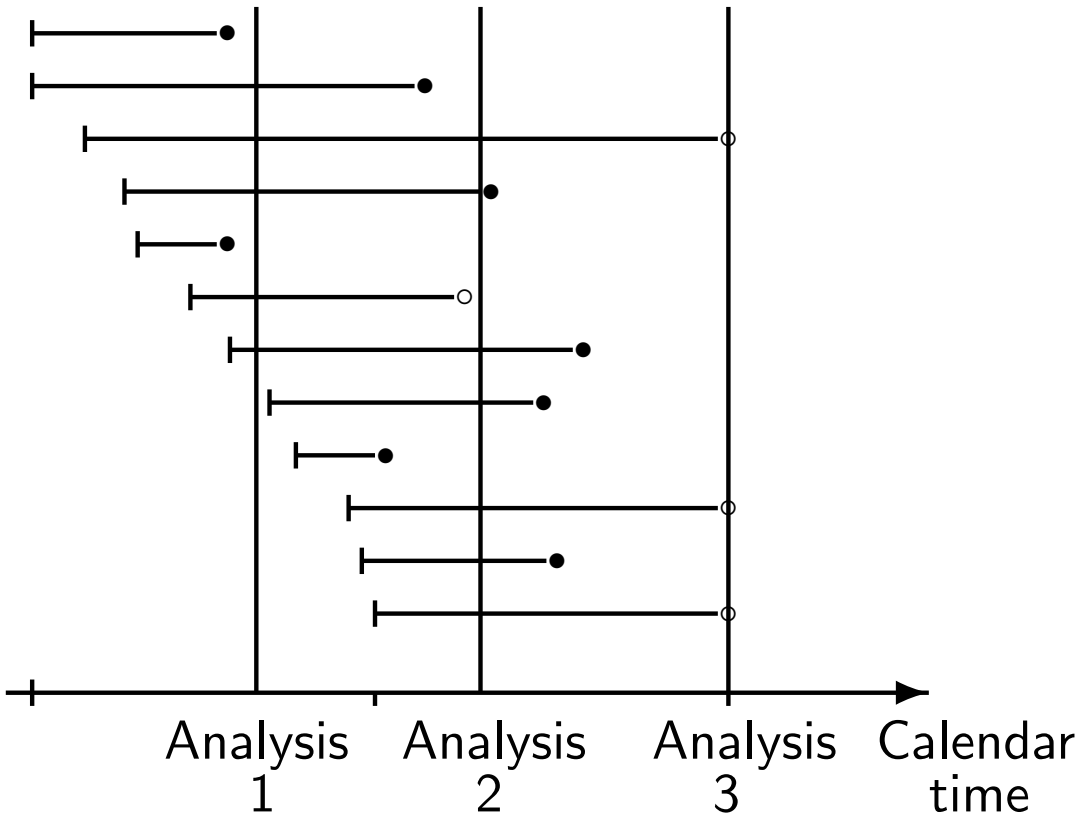
Key

- • death time observed
- ○ censored observation

Subjects are randomised to a treatment as they enter the study.

Survival is measured from entry to the study.

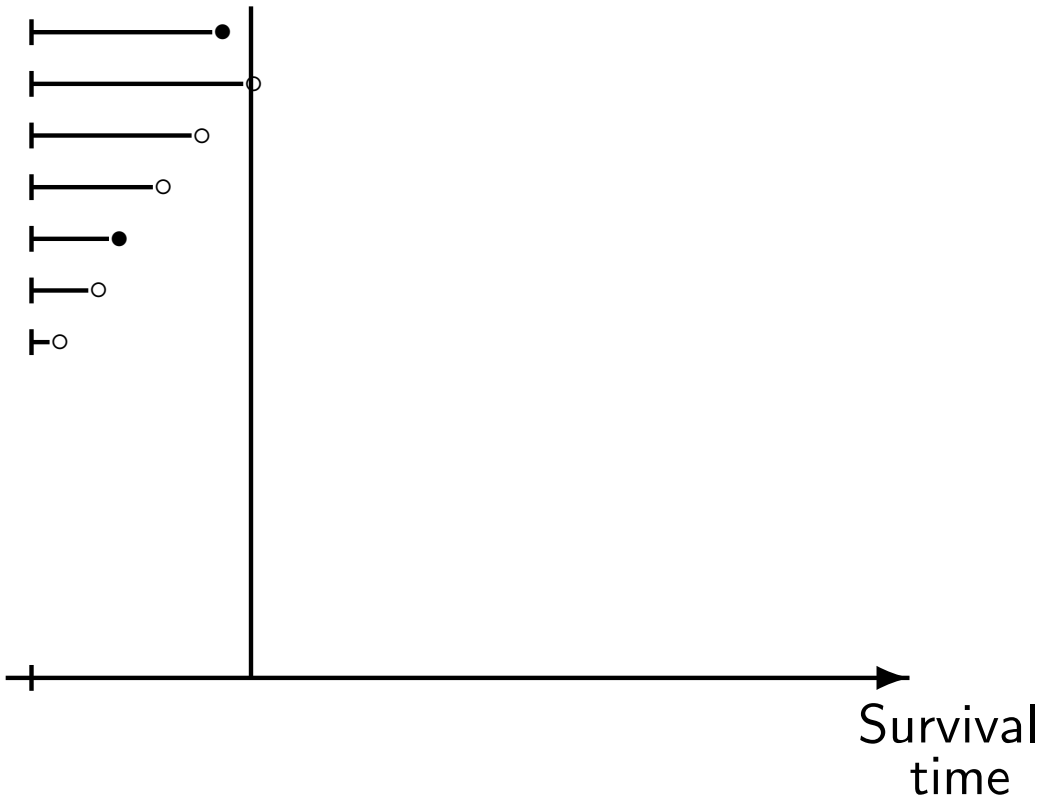
Interim analyses



At an interim analysis, subjects are censored if they are still alive.

Information on such patients continues to accrue at later analyses.

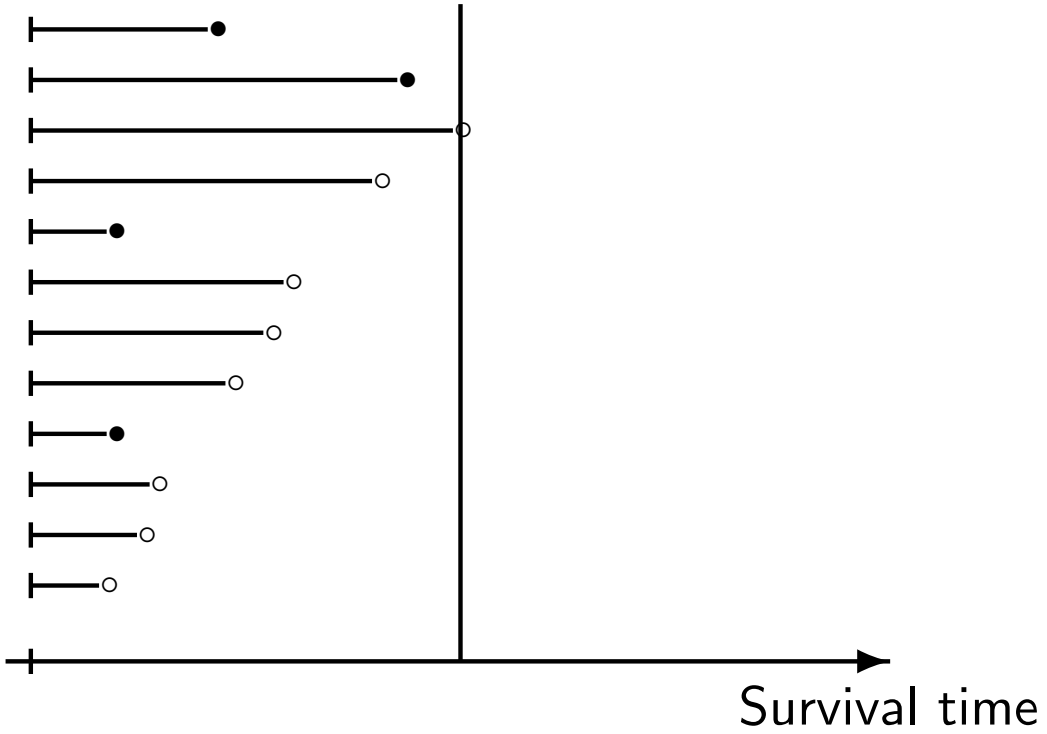
Interim analysis 1



We analyse data on survival from time of randomisation.

Survival times start at zero and "analysis time" censoring occurs for subjects surviving past this first analysis.

Interim analysis 2



At interim analysis 2, there is further follow-up of subjects who were censored at analysis 1.

In addition, there is initial information on the survival times of subjects entering the trial since analysis 1.

The logrank statistic

At stage k, the observed number of deaths is d_k.

Elapsed times between entry to the study and these deaths are

$$\tau_{1,k} < \tau_{2,k} < \ldots < \tau_{d_k,k} \quad \text{(assuming no ties)}.$$

Define variables at analysis k

$r_{iA,k}$ and $r_{iB,k}$	Numbers at risk on Trts A and B at $\tau_{i,k}-$
$r_{ik} = r_{iA,k} + r_{iB,k}$	Total number at risk at $\tau_{i,k}-$
O_k	Observed number of deaths on Trt B
$E_k = \sum_{i=1}^{d_k} r_{iB,k}/r_{ik}$	"Expected" number of deaths on Trt B
$V_k = \sum_{1}^{d_k} r_{iA,k} r_{iB,k}/r_{ik}^2$	"Variance" of O_k
$Z_k = (O_k - E_k)/\sqrt{V_k}$	Standardised logrank statistic

Canonical joint distribution of logrank-statistics

In the **Proportional Hazards Model:** We assume hazard rates h_A on Treatment A and h_B on Treatment B are related by

$$h_B(t) = \lambda\, h_A(t).$$

The log hazard ratio is $\theta = \ln(\lambda)$.

Then, with $\mathcal{I}_k = V_k$, we have approximately

$$Z_k \sim N(\theta\sqrt{\mathcal{I}_k},\, 1), \quad k = 1, \ldots, K,$$

$$\text{Cov}(Z_{k_1}, Z_{k_2}) = \sqrt{(\mathcal{I}_{k_1}/\mathcal{I}_{k_2})} \quad \text{for } k_1 < k_2.$$

In addition, $(Z_1, \ldots, Z_K)$ is approximately multivariate normal — so the statistics $Z_1, \ldots, Z_K$ follow the canonical joint distribution.

The kth score statistic is $S_k = Z_k\sqrt{\mathcal{I}_k}$, with variance $V_k = \mathcal{I}_k$, and the sequence $\{S_1, \ldots, S_K\}$ has uncorrelated increments.

Canonical joint distribution of estimates of the hazard ratio

Observed information: Recall that

$$\mathcal{I}_k \;=\; V_k \;=\; \sum_{i=1}^{d_k} \frac{r_{iA,k}\, r_{iB,k}}{(r_{iA,k} + r_{iB,k})^2}.$$

If equal numbers are randomised to Treatments A and B and $\lambda \approx 1$, we can expect $r_{iA,k} \approx r_{iB,k}$ for each k, and so

$$\mathcal{I}_k = V_k \approx d_k/4.$$

Estimating θ:

Since $Z_k \sim N(\theta\sqrt{\mathcal{I}_k},\, 1)$, we can estimate θ at analysis k by

$$\widehat{\theta}_k = \frac{Z_k}{\sqrt{\mathcal{I}_k}}.$$

It follows that

$$\widehat{\theta}_k \;\sim N(\theta,\, \mathcal{I}_k^{-1}) \quad \text{approximately.}$$

Design of the Oropharynx trial

Suppose we wish to create a one-sided test of H_0: $\theta \leq 0$ vs $\theta > 0$.

Note $\theta > 0 \Rightarrow \lambda > 1$, i.e., Treatment A is better.

We require:

Type I error probability $\alpha = 0.025$,

Power $1 - \beta = 0.8$ at $\theta = 0.5$, i.e., at $\lambda = 1.65$.

Information needed for a fixed sample study is

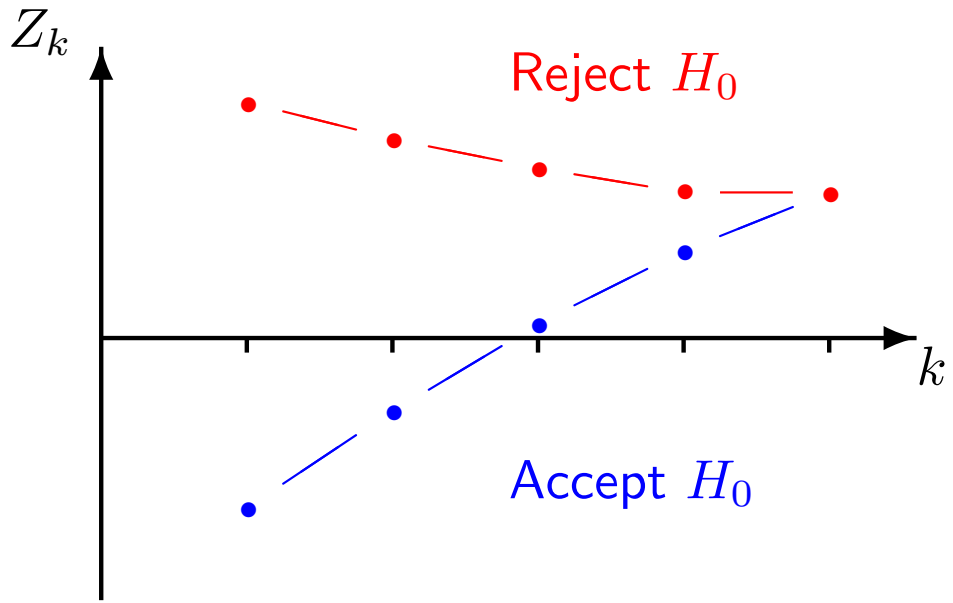
$$\mathcal{I}_{fix} = \frac{\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2}{0.5^2} = 31.40.$$

Under the approximation $\mathcal{I} \approx d/4$, the total number of failures to be observed is

$$d_f = 4\mathcal{I}_{fix} \approx 126.$$

Design of the Oropharynx trial

For a one-sided test with up to 5 analyses, we could try to use a standard design created for equally spaced information levels.



However, increments in information between analyses are unpredictable.

So, an error spending design is a natural choice.

A one-sided, error spending design

Specification:

One-sided test of H_0: $\theta \leq 0$ vs $\theta > 0$,

Type I error probability $\alpha = 0.025$,

Power $1 - \beta = 0.8$ at $\theta = \ln(\lambda) = 0.5$,

Binding futility boundary.

When designing, assume $K = 5$ equally spaced information levels.

Use a power-family test with $\rho = 2$ to spend error $\propto (\mathcal{I}/\mathcal{I}_{\max})^2$.

Information for a fixed sample test has to be inflated by $R = 1.098$.

So, we require $\mathcal{I}_{\max} = 1.098 \times 31.40 = 34.48$, which needs a total of $4 \times 34.48 \approx 138$ observed deaths.

A one-sided, error spending design

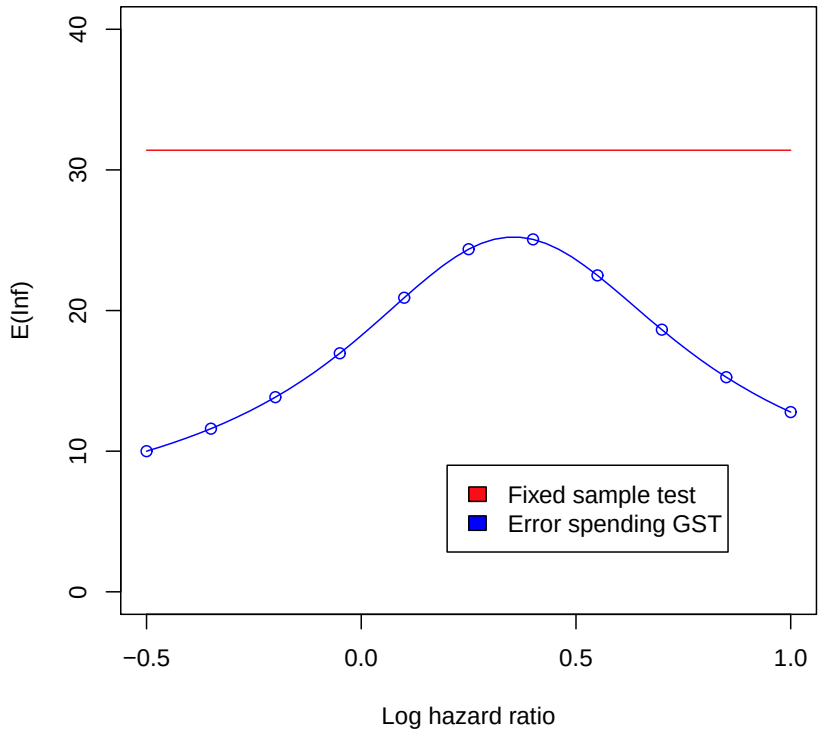
Suppose that, as assumed when planning the trial, information levels are equally spaced up to $\mathcal{I}_5 = \mathcal{I}_{\max} = 34.48$.

Then, we would have the following boundary values $(a_1, b_1), \ldots, (a_5, b_5)$ for the standardised logrank statistics $Z_1, \ldots, Z_5$.

k	$\mathcal{I}_k$	a_k	b_k
1	6.90	-1.10	3.09
2	13.79	-0.05	2.71
3	20.69	0.72	2.47
4	27.58	1.39	2.28
5	34.48	2.06	2.06

A one-sided, error spending design

If information levels $\mathcal{I}_k = (k/5)\,34.48,\ k = 1, \ldots, 5,$ are observed, the expected information on termination is the following function of the log hazard ratio, θ.



Note that the GST has uniformly lower $\mathbb{E}(\mathcal{I})$ than the fixed sample design's $\mathcal{I}_{fix} = 31.40$.

Summary data and critical values for the Oropharynx trial

In reality, we construct error spending boundaries using the *observed* information levels.

The design with a non-binding futility boundary has the following boundary values $(a_1, b_1), \ldots, (a_5, b_5)$ for the standardised logrank statistics $Z_1, \ldots, Z_5$.

Analysis k	*Number entered*	*Number of deaths*	$\mathcal{I}_k$	a_k	b_k	Z_k
1	83	27	5.43	−1.41	3.23	−1.04
2	126	58	12.58	−0.21	2.76	−1.00
3	174	91	21.11	0.78	2.44	−1.21
4	195	129	30.55	1.68	2.16	−0.73
5	195	142	33.28	2.14	2.14	−0.87

The trial would have terminated at analysis 2 to accept H_0.

An error spending test with a non-binding futility boundary

If a **non-binding** futility boundary is used, the required maximum information level is a little higher at 35.58.

Applying this design to the observed information levels gives:

Analysis k	*Number entered*	*Number of deaths*	$\mathcal{I}_k$	a_k	b_k	Z_k
1	83	27	5.43	-1.44	3.25	-1.04
2	126	58	12.58	-0.23	2.78	-1.00
3	174	91	21.11	0.75	2.46	-1.21
4	195	129	30.55	1.64	2.20	-0.73
5	195	142	33.28	2.09	2.09	-0.87

Again, the trial terminates at analysis 2 with acceptance of H_0.

Covariate adjustment in the Oropharynx trial

Covariate information was recorded for subjects: *Institution (6), Gender, Initial condition, T-staging, N-staging, Tumour site (3).*

A proportional hazards regression model includes

- Strata $l = 1, \ldots, 6$ for the six participating institutions,
- Treatment effect β_1,
- Coefficients $\beta_2, \ldots, \beta_5$ for Gender and the continuous variables Initial condition, T-staging and N-staging,
- Coefficients β_6 and β_7 for the categorical variable Tumour site.

Modelling the hazard rate for patient i as

$$h_{il}(t) = h_{0l}(t)\, e^{\{\beta_1 I(\text{Patient } i \text{ on Trt B}) + \Sigma_{j=2}^{7} x_{ij}\beta_j\}},$$

the objective is to test H_0: $\beta_1 \leq 0$ against $\beta_1 > 0$.

Covariate adjustment in the Oropharynx trial

Standard software for Cox regression can provide an estimate of the parameter vector, β, and its estimated variance.

We are interested in the treatment effect β_1.

At stage k we have

$$\widehat{\beta}_1^{(k)}$$

$$v_k = \widehat{\mathsf{Var}}\left(\widehat{\beta}_1^{(k)}\right)$$

$$\mathcal{I}_k = v_k^{-1}$$

$$Z_k = \widehat{\beta}_1^{(k)} / \sqrt{v_k}.$$

Theory tells us: The standardised statistics $Z_1, \ldots, Z_5$ have, approximately, the canonical joint distribution.

Covariate-adjusted analysis of the Oropharynx trial

Constructing the error spending test with a **non-binding** futility boundary gives critical values $(a_1, b_1), \ldots, (a_5, b_5)$ for $Z_1, \ldots, Z_5$.

k	$\mathcal{I}_k$	a_k	b_k	$\widehat{\beta}_1^{(k)}$	Z_k
1	4.11	-1.77	3.40	-0.79	-1.60
2	10.89	-0.47	2.87	-0.14	-0.45
3	19.23	0.55	2.52	-0.08	-0.33
4	28.10	1.41	2.27	0.04	0.20
5	30.96	2.27	2.27	0.01	0.04

Under this stopping rule, the study would have continued — just — at analysis 2 and stopped to accept H_0 at analysis 3.

Note that β_1 is the log hazard ratio after covariate adjustment. For $\beta_1 > 0$, we should expect $\beta_1 > \lambda$ where λ is the log hazard ratio in a model without covariates.

East: Demonstration, Survival data

We wish to plan a two treatment comparison, testing for *superiority*, i.e., testing

$$H_0\text{: } \theta \leq 0 \text{ vs } \theta > 0,$$

where λ is the hazard ratio between treatments and $\theta = \log(\lambda)$.

We shall consider how to:

- Design a one-sided, error spending test with

  Type I error probability, $\alpha = 0.025$,

  Power $1 - \beta = 0.8$ at $\lambda = 1.65$,

- Using a sequential logrank test or a sequential test based on parameter estimates in a Cox model,

- Include a binding or non-binding futility boundary.

Recapitulation: Group sequential tests (1)

- It is natural to monitor clinical trials with a view to possible early stopping.
- Distribution theory supports a general approach to design group sequential tests for a variety of response types.
- Numerical integration allows us to compute properties of group sequential designs precisely and set stopping boundaries and decision rules that control the type I error rate.
- Group sequential designs can be optimised for a given objective.
- Error spending designs offer efficient, flexible monitoring of a variety of response types, including survival data.

Part 2. Group sequential tests (2)

2.1 Group sequential testing for a delayed response

Reference: Hampson & Jennison (*JRSS B*, 2013), hereafter "HJ"

Group sequential designs are most often developed supposing observations will be recorded immediately after treatment.

Thus, if it is decided to stop a trial at an interim analysis, it is assumed the current observations will form the final set of data.

In practice, responses are observed some time after treatment.

Thus, when it is decided to stop a trial at an interim analysis, one should expect additional data from patients who have been treated but whose responses have not yet been observed.

We shall refer to such patients as "in the pipeline".

How should the additional data be analysed?

Examples of group sequential trials with delayed response

Example 1: HJ describe a study of a cholesterol lowering drug. The primary endpoint is reduction in cholesterol after 4 weeks.

A total of 96 patients are to be recruited at a rate of 4 patients per week. At each interim analysis we can expect 16 subjects to have been treated but not yet produced a response.

If the study is stopped at an interim analysis, investigators will still follow up the $\sim$16 pipeline subjects and observe their responses.

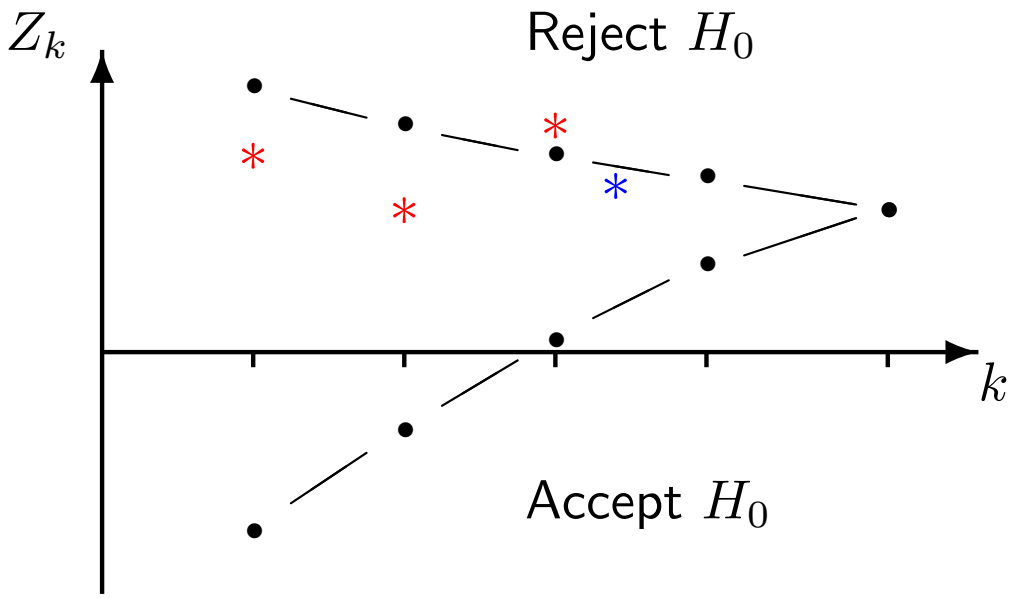
Example 2: Consider a clinical trial with a time-to-event endpoint.

Data are locked before each interim analysis. Time passes as data are cleaned, the DMC meets, and — at one analysis — the DMC recommends to the Steering Committee that the trial be stopped.

When stopping actually happens, more events will have occurred and other potential events will have been adjudicated.

The cholesterol reduction trial

Suppose a standard group sequential test (GST) is applied.



We observe $Z_3 = 2.4$, which exceeds the boundary value of 2.3.

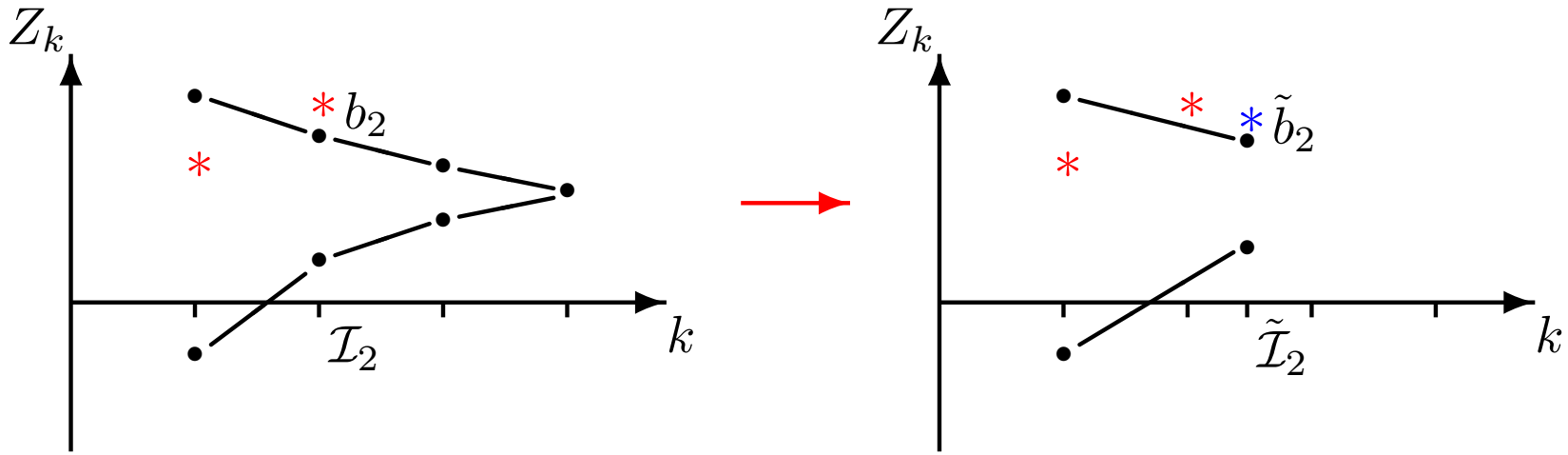
The trial stops but, with the pipeline data included, $Z = 2.1$.

Can the investigators claim significance at level α?

Incorporating delayed observations after a GST terminates

Whitehead (*Cont. Clin. Trials*, 1992) proposed the "deletion method".

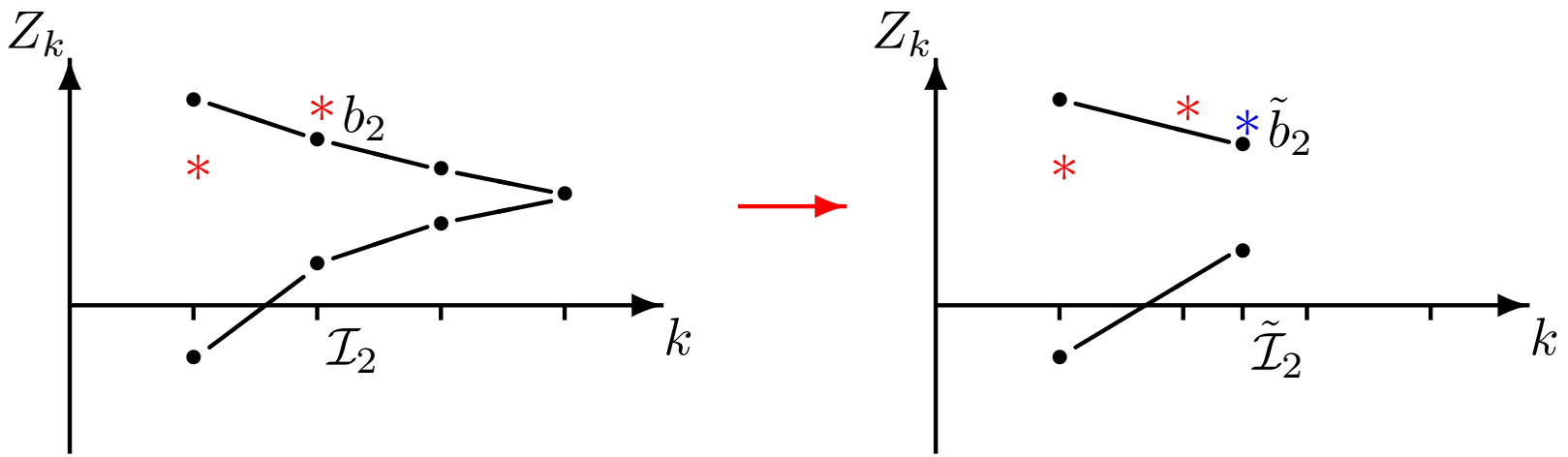
The analysis k at which termination occurs is deleted and one behaves as if analysis k had occurred with the information level $\tilde{\mathcal{I}}_k$ arising from the final set of responses.



A boundary value $\tilde{b}_k$ is computed and H_0 is rejected if, for the test statistic including pipeline data, $\tilde{Z}_k \geq \tilde{b}_k$.

Note: In order to reject H_0, the test statistics must first cross the upper boundary of the original group sequential design.

Incorporating delayed observations after a GST terminates



For H_0 to be rejected, the test statistics must first cross the upper boundary of the original group sequential design. Thus, this method protects the type I error rate conservatively.

Sorriyarachchi et al. (*Biometrics*, 2003) investigated the "deletion method" and several other proposals.

They found that tests using additional "pipeline" data often had lower power than simple GSTs which ignored these data — but extra information ought to help!

Incorporating delayed observations after a GST terminates

The method of Whitehead (1992) applies a GST as if response were immediate, then we try to accommodate additional pipeline data once this GST has terminated.

A more systematic approach is to recognise that there will be pipeline data when designing the trial.

Interestingly, T. W. Anderson (*JASA*, 1964) recognised this issue, well before the advent of modern group sequential methods.

The methods of Hampson & Jennison (*JRSS, B*, 2013) follow the same basic structure that was proposed by Anderson.

With delayed response data, a trial comes to an end in two stages:

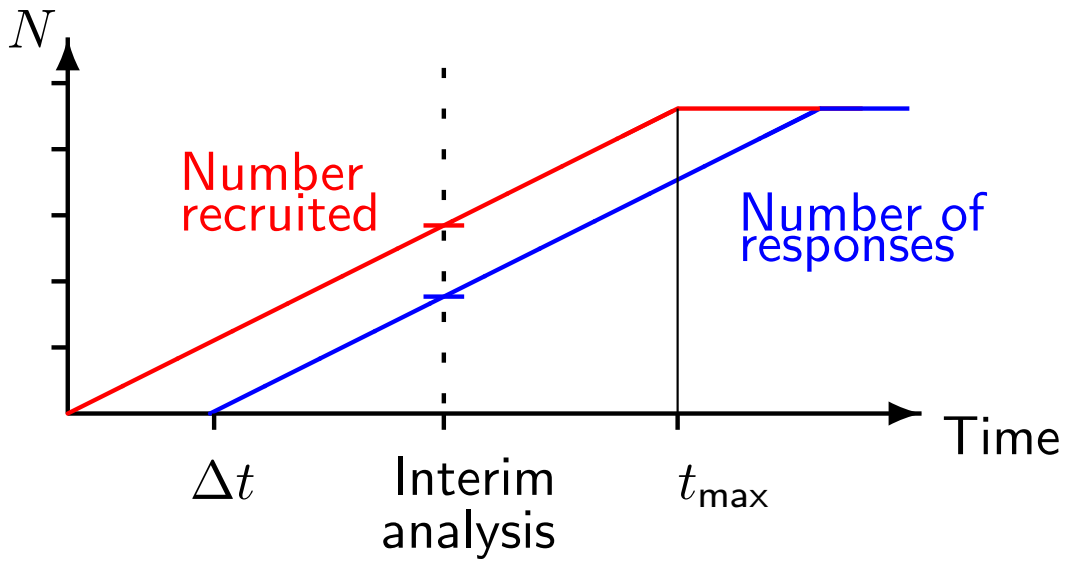
1. Stop recruitment of any more subjects,
2. After responses have been observed for all recruited subjects, make a decision to accept or reject H_0.

Defining a group sequential test with delayed responses

For now, we assume, as in Example 1:

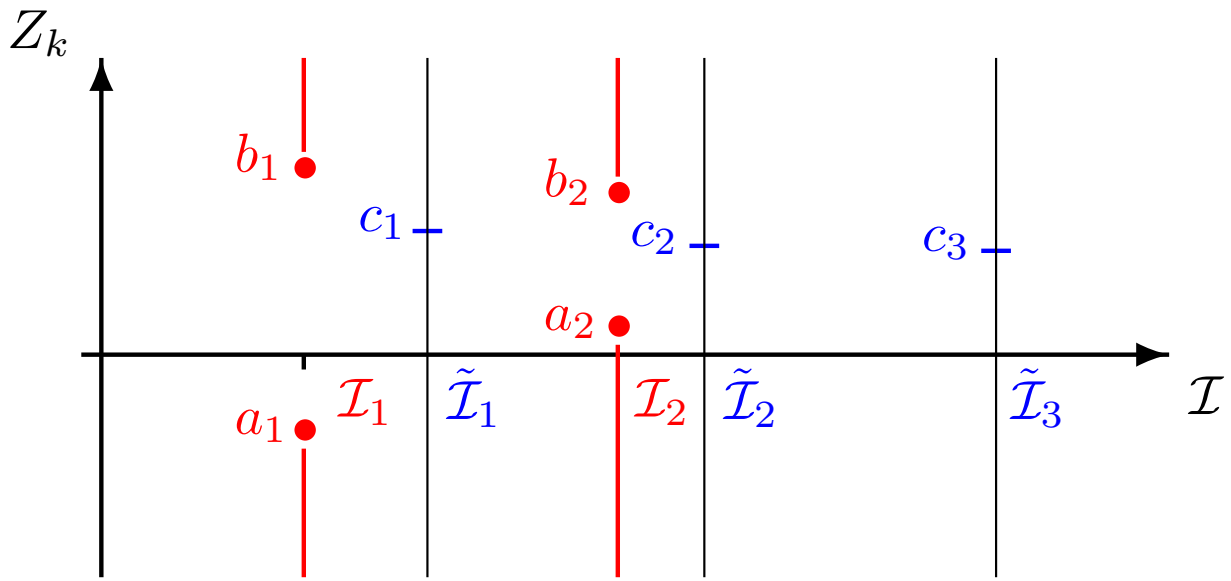
- The primary endpoint is measured a fixed time after treatment commences,
- The endpoint will be known (eventually) for all treated subjects,
- If recruitment is stopped, it cannot be re-started.

Consider a trial with responses observed time Δ_t after treatment.



Boundaries for a Delayed Response GST

At **interim** analysis k, observed information is $\mathcal{I}_k = \{\text{Var}(\hat{\theta}_k)\}^{-1}$.



If $Z_k > b_k$ or $Z_k < a_k$ at analysis k, we cease enrolment of patients and follow-up all recruited subjects to observe their responses.

At the subsequent decision analysis, information by $\tilde{\mathcal{I}}_k$ and the standardised test statistic by $\tilde{Z}_k > c_k$. We reject H_0 if $\tilde{Z}_k > c_k$.

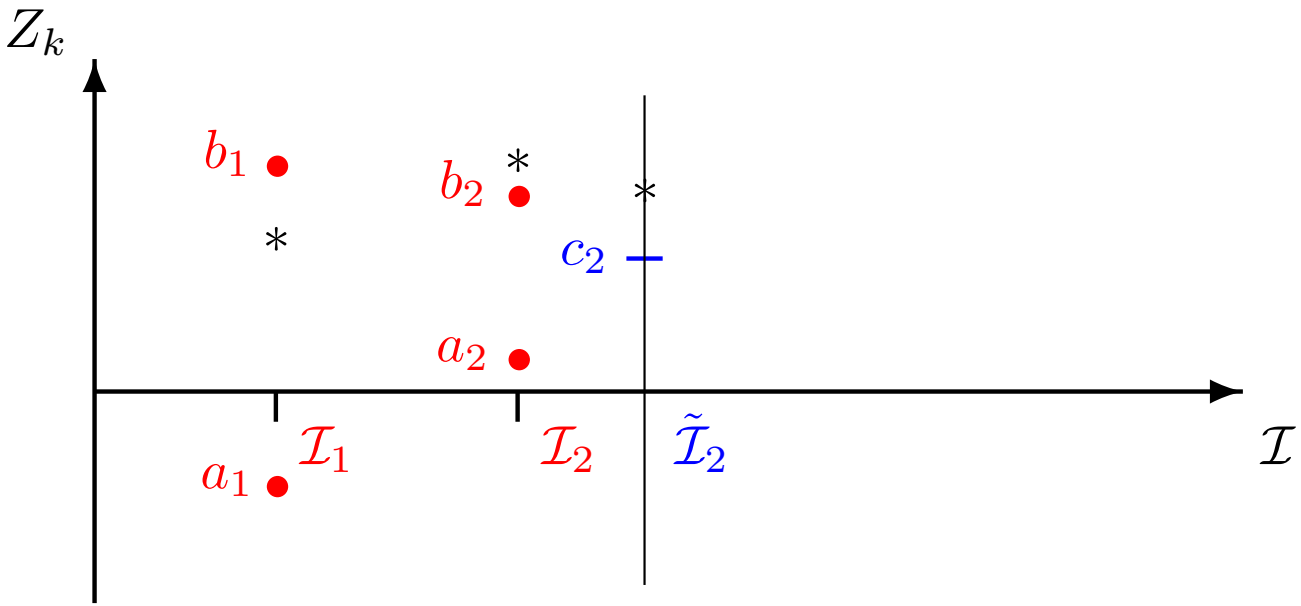
If we reach the **final** analysis K, we reject H_0 if $\tilde{Z}_K > c_K$.

Delayed Response GSTs

For a particular sequence of observed responses, we apply boundary points at a sequence of information levels of the form

$$\mathcal{I}_1, \ldots, \mathcal{I}_k, \tilde{\mathcal{I}}_k.$$

In the example below, recruitment ceases at the second analysis and the final decision is made with the additional "pipeline" data bringing the information up to $\tilde{\mathcal{I}}_2$.



Calculations for a Delayed Response GST

The type I error rate, power and expected sample size of a Delayed Response GST depend on the joint distributions of test statistic sequences:

$$\{Z_1, \ldots, Z_k, \tilde{Z}_k\}, \quad k = 1, \ldots, K-1,$$

and

$$\{Z_1, \ldots, Z_{K-1}, \tilde{Z}_K\}.$$

Each sequence is based on accumulating data sets.

Given $\{\mathcal{I}_1, \ldots, \mathcal{I}_k, \tilde{\mathcal{I}}_k\}$, the sequence $\{Z_1, \ldots, Z_k, \tilde{Z}_k\}$ follows the canonical distribution we saw earlier for the sequence of Z-statistics in a GST with immediate responses (JT, Ch. 11).

Thus, properties of Delayed Response GSTs can be calculated using the same numerical routines that were needed for standard group sequential designs.

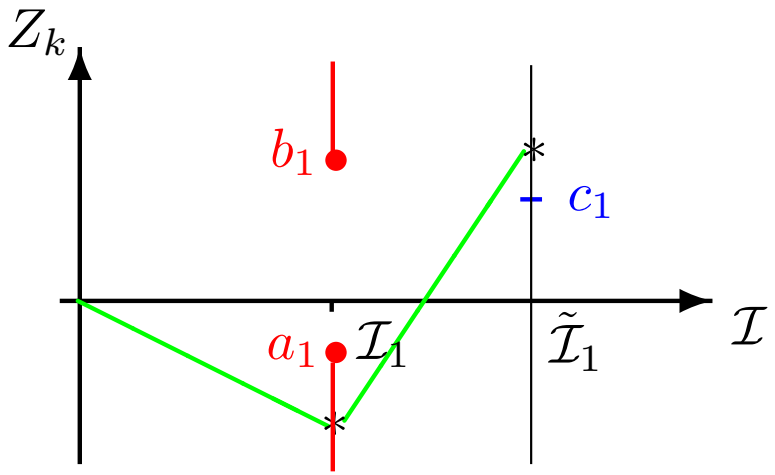
The value of information from pipeline subjects

When recruitment is terminated at interim analysis k with $Z_k > b_k$ or $Z_k < a_k$, current data suggest the likely final decision.

Pipeline data give more information to use in making this decision.

The pipeline data may produce a "reversal", with the final decision differing from that anticipated when recruitment was terminated.

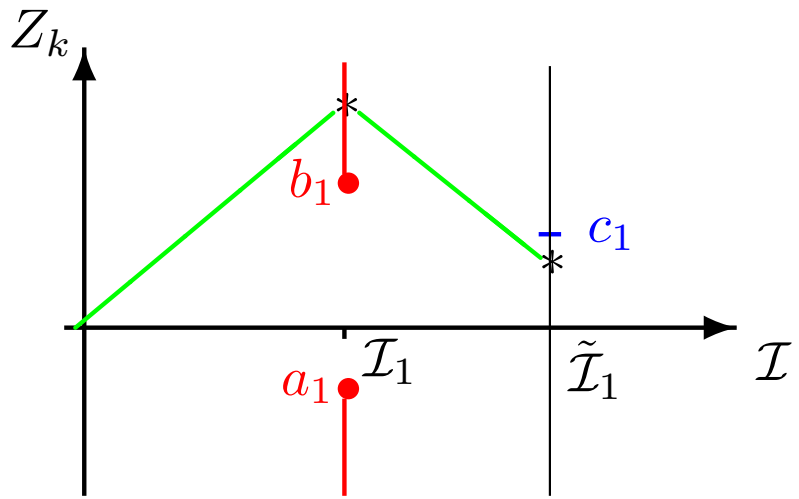
We could, for example, observe:



Here, accrual stops at analysis 1 because of unpromising results, but H_0 is rejected when the pipeline data are observed.

The value of information from pipeline subjects

Or, recruitment may cease with promising data only for H_0 to be accepted.



Note: There is no option of "banking" the evidence at analysis 1 — we assume all pipeline subjects will eventually be observed.

Decisions based on more data ought to be more accurate: perhaps these pipeline data have helped to avoid a false positive conclusion.

An optimised design will place boundary points to achieve high power for the permitted type I error rate, α.

Optimising a Delayed Response GST

We specify the type I error rate α and power $1-\beta$ at $\theta = \delta$.

We set the maximum sample size $n_{\max}$, number of stages K, and the analysis schedule.

Suppose there are $r\, n_{\max}$ pipeline subjects at each interim analysis.

Let N denote the total number of subjects recruited.

Objective:

Given α, β, δ, $n_{\max}$, K and r, find the Delayed Response GST minimising

$$F = \int \mathbb{E}_\theta(N)\, f(\theta)\, d\theta$$

where $f(\theta)$ is the density of a $N(\delta/2, (\delta/2)^2)$ distribution.

Other weighted combinations of $\mathbb{E}_\theta(N)$ can also be used.

Computing optimal Delayed Response GSTs

We follow the same approach as for optimising a GST with immediate response.

We create a Bayes sequential decision problem, placing a prior on θ and defining costs for sampling and for making incorrect decisions.

This problem can be solved rapidly by dynamic programming.

We then search for the combination of prior and costs such that the solution to the (unconstrained) Bayes decision problem has the specified frequentist error rates α at $\theta = 0$ and β at $\theta = \delta$.

The resulting design solves both the Bayes decision problem and the original frequentist problem.

Again, the Bayes decision problem is introduced as a computational device, but the derivation demonstrates the relationship between admissible frequentist designs and Bayes procedures.

An optimal design for the cholesterol treatment example

In the cholesterol treatment trial (Example 1), the primary endpoint is reduction in serum cholesterol after 4 weeks.

Responses are assumed normally distributed with variance $\sigma^2 = 2$.

The treatment effect θ is the difference in mean response between the new treatment and control.

An effect $\theta = 1$ is regarded as clinically significant.

It is required to test H_0: $\theta \leq 0$ against $\theta > 0$ with

Type I error rate $\alpha = 0.025$,

Power 0.9 at $\theta = 1$.

A fixed sample test needs $n_{fix} = 85$ subjects over the two treatments.

An optimal design for the cholesterol treatment example

We consider designs with a maximum sample size of 96.

We assume a recruitment rate of 4 per week:

Data start to accrue after 4 weeks,

Each interim analysis will have $4 \times 4 = 16$ pipeline subjects, so the "pipeline fraction" is $r = 16/96 = 0.17$.

Recruitment will close after 24 weeks.

Interim analyses are planned after $n_1 = 28$ and $n_2 = 54$ observed responses and the final decision is based on:

$\tilde{n}_1 = 44$ responses if recruitment stops at interim analysis 1,

$\tilde{n}_2 = 70$ responses if recruitment stops at interim analysis 2,

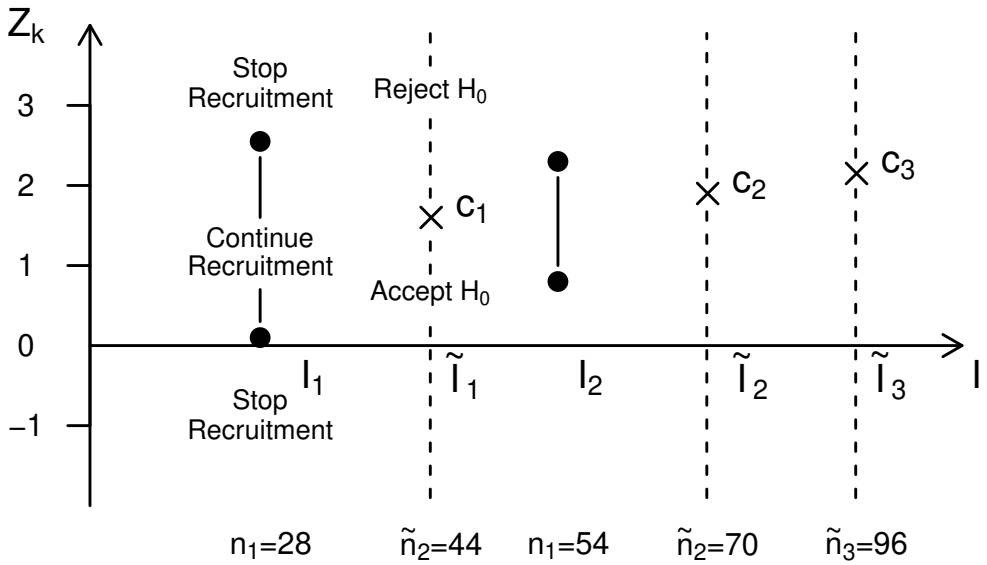
$\tilde{n}_3 = 96$ responses if there is no early stopping.

An optimal design for the cholesterol treatment example

The following Delayed Response GST minimises

$$F = \int \mathbb{E}_\theta(N)\, f(\theta)\, d\theta,$$

where $f(\theta)$ is the density of a $N(0.5, 0.5^2)$ distribution.



The values of c_1 and c_2 are less than 1.96. These can be raised to 1.96 with little change to the design's power curve.

2.2 Error spending Delayed Response GSTs

HJ show how to construct error spending Delayed Response GSTs.

Here, we present a variation on these methods which allows a non-binding futility boundary.

The test is defined through two error spending functions:

$f(\mathcal{I}/\mathcal{I}_{\max})$ for type I error probability,

$g(\mathcal{I}/\mathcal{I}_{\max})$ for type II error probability.

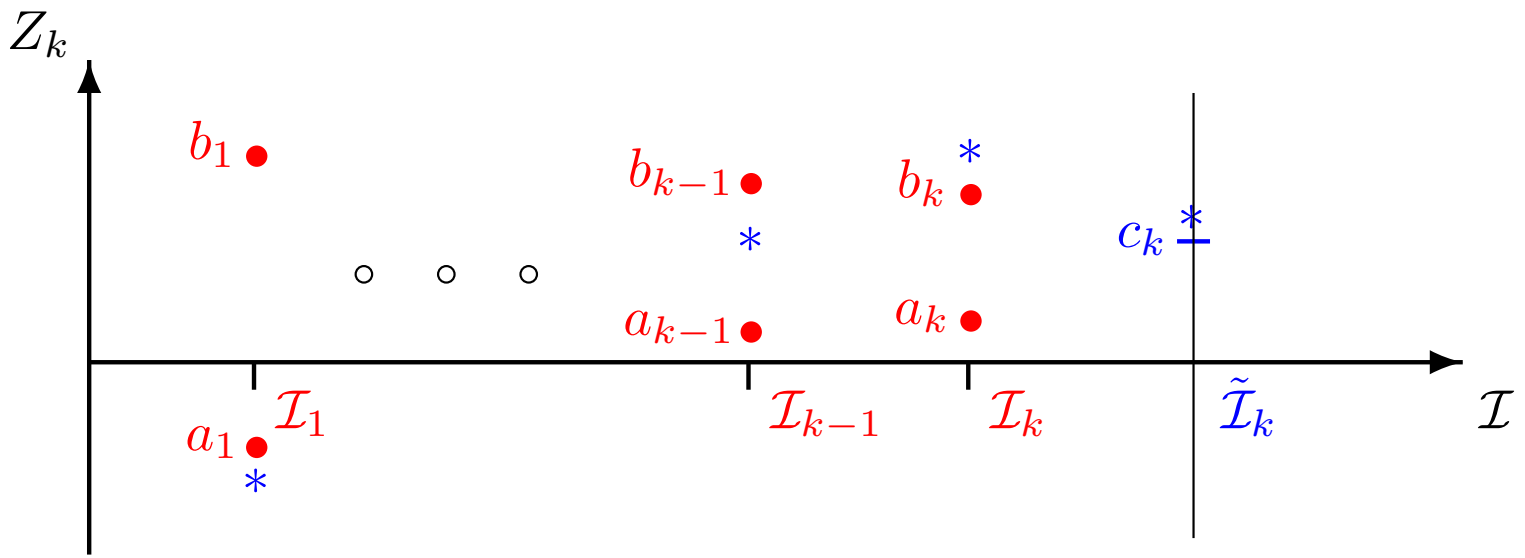
Recruitment stops when the target information $\mathcal{I}_{max}$ is reached (or will be reached with the responses from pipeline subjects).

After analysis k and its subsequent decision analysis:

The cumulative type I error will be exactly $f(\mathcal{I}_k/\mathcal{I}_{\max})$,

The cumulative type II error will be approximately $g(\mathcal{I}_k/\mathcal{I}_{\max})$ (depending on how accurately $\tilde{\mathcal{I}}_k$ can be predicted).

Error spending Delayed Response GSTs

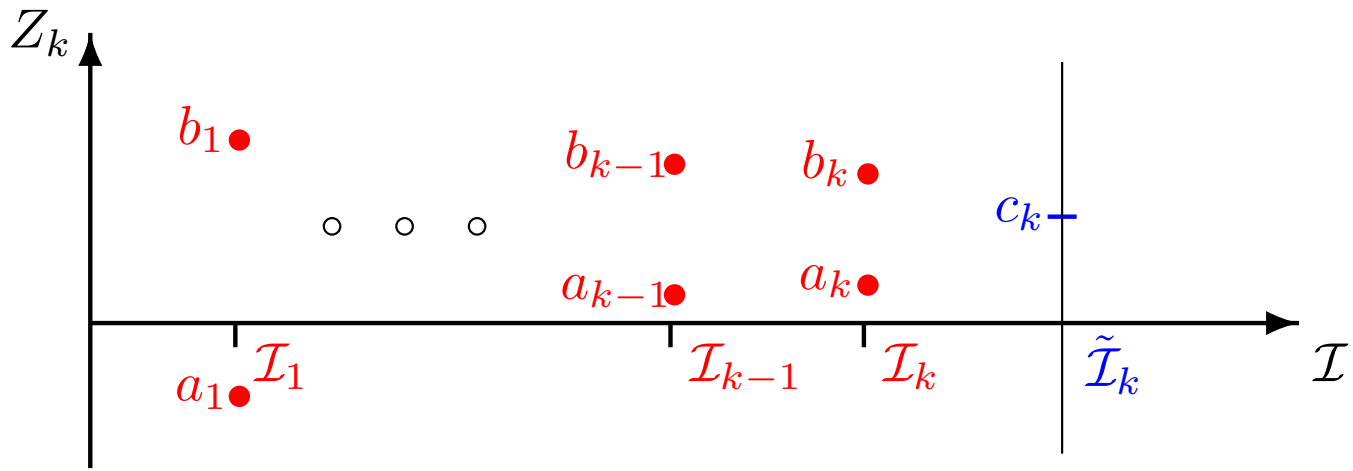


Non-binding futility boundary:

Type I error is calculated assuming recruitment still continues if $Z_k < a_k$ at interim analysis k and the futility boundary is crossed.

If recruitment is stopped when $Z_k < a_k$, a final decision to reject H_0 is not permitted, even if $\tilde{Z}_k > c_k$.

Computing an error spending Delayed Response GST



If we can predict $\tilde{\mathcal{I}}_k$ accurately, then a_k, b_k and c_k must satisfy

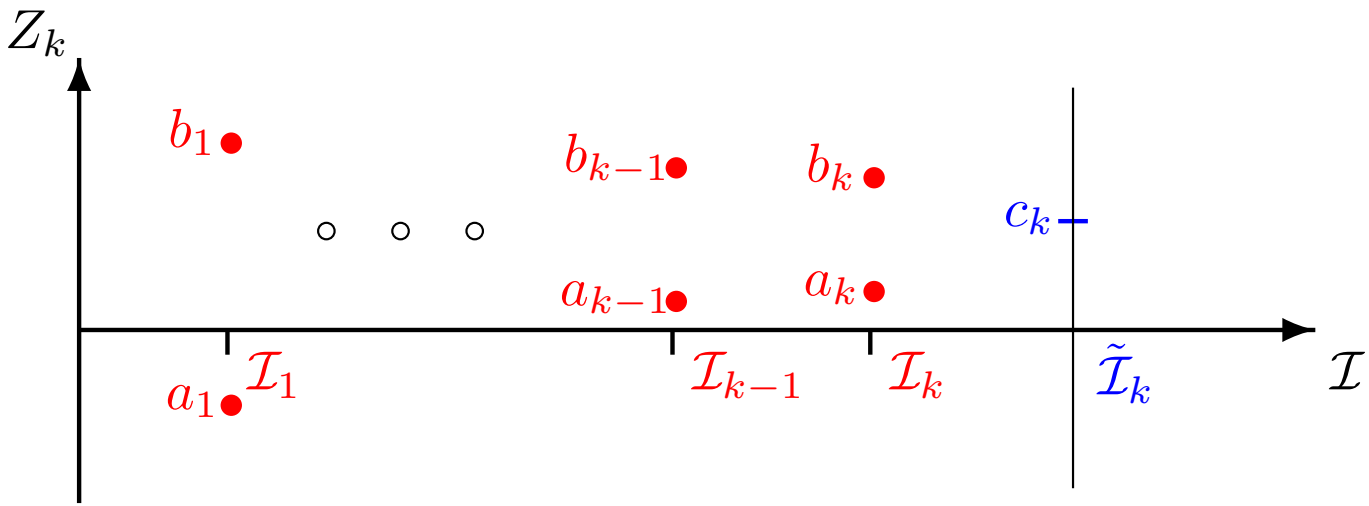
$$P_{\theta=0}\{Z_1 < b_1, \ldots, Z_{k-1} < b_{k-1}, Z_k > b_k, \tilde{Z}_k > c_k\}$$
$$= f(\mathcal{I}_k/\mathcal{I}_{\max}) - f(\mathcal{I}_{k-1}/\mathcal{I}_{\max}),$$

and

$$P_{\theta=\delta}\{a_1 < Z_1 < b_1, \ldots, a_{k-1} < Z_{k-1} < b_{k-1} \text{ and } [Z_k < a_k \text{ or}$$
$$(Z_k > b_k \text{ and } \tilde{Z}_k < c_k)\ ]\} = g(\mathcal{I}_k/\mathcal{I}_{\max}) - g(\mathcal{I}_{k-1}/\mathcal{I}_{\max}).$$

Note: We have two equations but three unknowns, a_k, b_k and c_k.

Computing an error spending Delayed Response GST



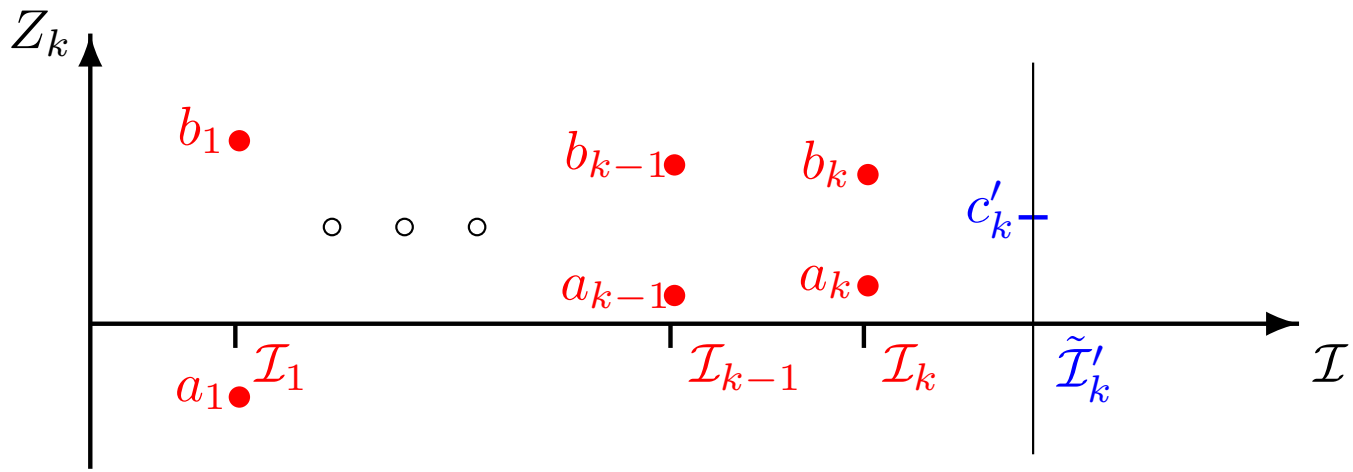
HJ noted their optimal Delayed Response GSTs with $\alpha = 0.025$ often had values of $c_1, \ldots, c_{K-1}$ less than $\Phi^{-1}(1-\alpha) = 1.96$.

For reasons of credibility, they suggested increasing the values of $c_1, \ldots, c_{K-1}$ to 1.96 — or set $c_1 = \cdots = c_{K-1} = 1.96$ before optimising over the remaining constants.

In an error spending design, we can set $c_k = \Phi^{-1}(1-\alpha) = 1.96$, then we have two equations to determine a_k and b_k.

Updating c_k on observing $\tilde{\mathcal{I}}_k$

The above boundary spends the required increments in type I and II error probability exactly — if the predicted $\tilde{\mathcal{I}}_k$ is actually observed.



If, in fact, the final information level is $\tilde{\mathcal{I}}'_k$, we find c'_k such that

$$P_{\theta=0}\{Z_1 < b_1, \ldots, Z_{k-1} < b_{k-1}, Z_k > b_k, \tilde{Z}_k > c'_k\}$$

$$= f(\mathcal{I}_k/\mathcal{I}_{\max}) - f(\mathcal{I}_{k-1}/\mathcal{I}_{\max})$$

and increase this to $c'_k = 1.96$ if the result is less than 1.96.

(This leads to $c'_k > 1.96$ if $\tilde{\mathcal{I}}'_k < \tilde{\mathcal{I}}_k$ and $c'_k = 1.96$ if $\tilde{\mathcal{I}}'_k > \tilde{\mathcal{I}}_k$.)

The ρ-family of error spending functions

HJ considered ρ-family error spending functions of the form

$$f(\mathcal{I}/\mathcal{I}_{\max}) = \alpha \min\{1,\ (\mathcal{I}/\mathcal{I}_{\max})^{\rho}\},$$

$$g(\mathcal{I}/\mathcal{I}_{\max}) = \beta \min\{1,\ (\mathcal{I}/\mathcal{I}_{\max})^{\rho}\}.$$

They found the resulting Delayed Response GSTs to have close to optimal efficiency for the objective function

$$F = \int \mathbb{E}_{\theta}(N)\, f(\theta)\, d\theta,$$

where $f(\theta)$ is the density of a $N(0.5, 0.5^2)$ distribution.

We shall use the functions f and g to define error spending Delayed Response GSTs with non-binding futility boundaries.

We consider designs for Example 1: the cholesterol treatment trial.

Example 1: A ρ-family error spending GST

Given α, β and δ, we can choose an error spending delayed response GST whose boundaries will converge at the final analysis if $\{\mathcal{I}_1, \tilde{\mathcal{I}}_1, \ldots, \mathcal{I}_{K-1}, \tilde{\mathcal{I}}_{K-1}, \tilde{\mathcal{I}}_K\}$ follow anticipated values.

In the cholesterol trial, the anticipated sample sizes

$$n_1 = 28, \quad \tilde{n}_1 = 44, \quad n_2 = 54, \quad \tilde{n}_2 = 72, \quad \tilde{n}_3 = 96$$

lead to

$$\mathcal{I}_1 = 3.5, \quad \tilde{\mathcal{I}}_1 = 5.5, \quad \mathcal{I}_2 = 6.75, \quad \tilde{\mathcal{I}}_2 = 8.75, \quad \tilde{n}_3 = 12.$$

With these information levels, the boundaries of a ρ-family error spending test with $\rho = 1.345$ will meet up at analysis 3.

In this case, the boundary values are

$$a_1 = -0.409, b_1 = 2.437, c_1 = 1.960;$$

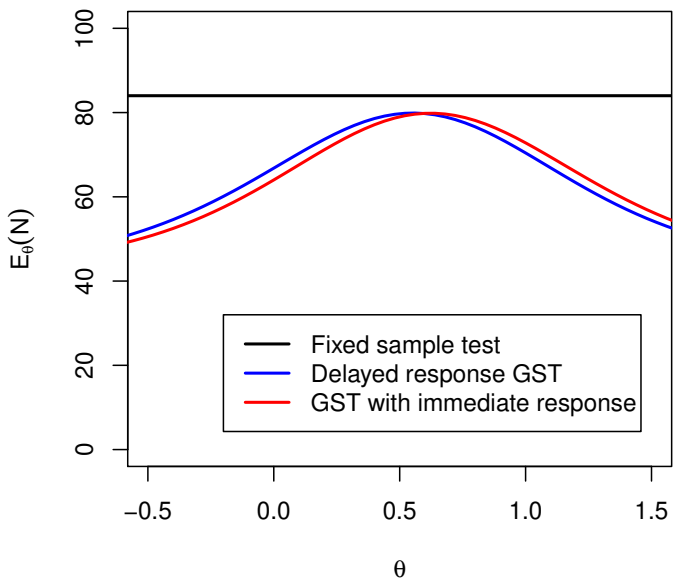
$$a_2 = 0.664, b_2 = 2.244, c_2 = 1.960;$$

$$c_3 = 2.069.$$

Example 1: A ρ-family error spending GST

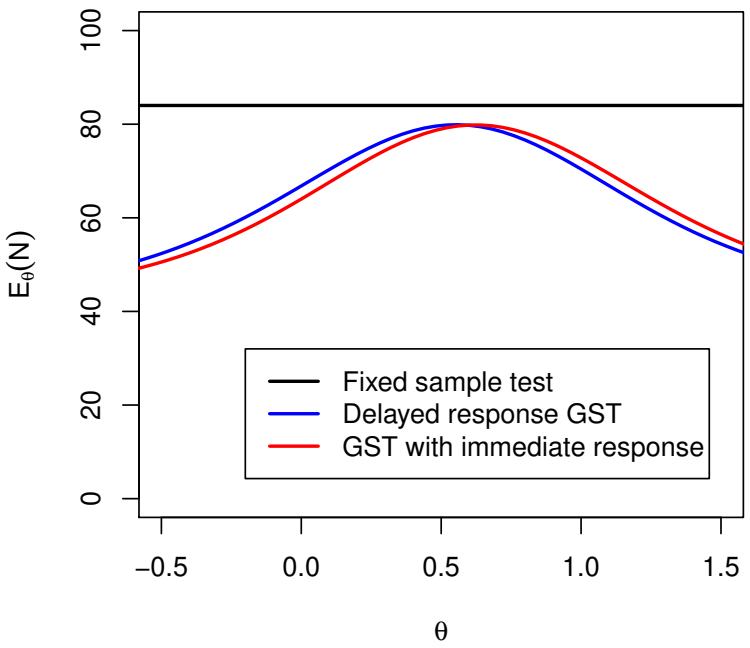
The figure shows $\mathbb{E}_\theta(N)$ for:

1. A fixed sample study design
2. Error spending delayed response GST ($\rho = 1.345$)
3. Error spending GST ignoring pipeline data ($\rho = 1.368$) but counting these subjects in $\mathbb{E}_\theta(N)$



Both GSTs have non-binding futility boundaries.

Example 1: A ρ-family error spending GST



Making use of the pipeline data leads to some efficiency gains for $\theta > 0.5$.

Importantly, the pipeline data do not have a detrimental effect.

In contrast, if we apply Whitehead's deletion method, starting from the ρ-family error spending GST for immediate response, power at $\theta = 1$ falls from 0.9 to 0.872. A 10% increase in overall sample size would be needed to this power.

Example 2: A study with a time-to-event endpoint

Suppose a study's endpoint is survival or progression free survival.

Events are likely to be recorded between the data set lock for an interim analysis and a decision to stop recruitment.

If events require adjudication, a further increase may follow.

The same approach can be taken as in Example 1 to create an error-spending Delayed Response GST.

Predicting $\tilde{\mathcal{I}}_k$ may be harder — but the methods can handle this.

Pipeline data may provide a substantial amount of additional information. Then, the guiding principles should be that:

If $\theta = 0$ using these data may help avoid a type I error;

If $\theta = \delta$ pipeline data are unlikely to "reverse a positive result".

Detailed calculations for Example 1 show this is possible!

Further development of Delayed Response GSTs

A variety of optimality criteria

HJ show how designs can be optimised for criteria involving both the number of subjects recruited and the time to a final decision.

The nature of a specific clinical trial will determine which approaches may be possible, depending on whether:

- All pipeline subjects must be followed to the response time;
- Investigators may decide not to wait to observe pipeline subjects;
- Data from (some) pipeline subjects will not be "valid" and cannot be used.

Discussants of the paper commented on the nature of "pipeline" data and HJ categorised possible cases in their response.

Further development of Delayed Response GSTs

Inference on termination

HJ define P-values and confidence intervals, with the usual frequentist properties, on termination of a Delayed Response GST.

These methods can also provide median unbiased point estimates.

Bias of maximum likelihood estimates can be corrected by applying Whitehead's (*Biometrika*, 1986) methods for standard GSTs.

Use of short-term endpoints

A delay in response reduces the benefits of sequential testing.

Pipeline patients contribute to sample size at an interim analysis but do not provide information to help decide whether to stop.

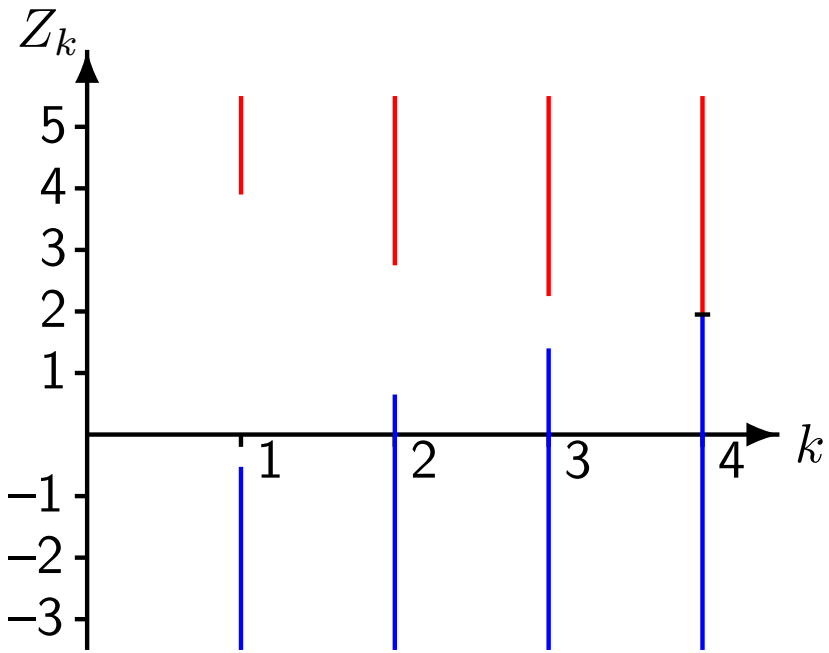
HJ show how a short term endpoint, correlated with the primary endpoint can be used to recover some of this efficiency loss.

2.3 Analysis on termination of a group sequential test

How should we analyse the data after a GST terminates?

Our sample space is all possible pairs (k, Z_k) on termination.

Sample space for a Pampallona & Tsiatis with $\Delta = 0$, $K = 4$ analyses, $\alpha = 0.025$ and power $1 - \beta = 0.8$ at $\theta = 1$:



Frequentist inference involves probabilities on this *sample space*.

The need for special methods

Suppose our 4-stage study with a Pampallona & Tsiatis boundary ends at stage 3 with $Z_3 = 2.6$.

It may be tempting to quote a 1-sided P-value of

$$P\{N(0,1) > 2.60\} = 0.0047.$$

But, using this definition, we would also get a P-value ≤ 0.0047 by

stopping at stage 1 with $Z_1 > 3.90$,

stopping at stage 2 with $Z_2 > 2.76$,

stopping at stage 3 with $Z_3 > 2.60$,

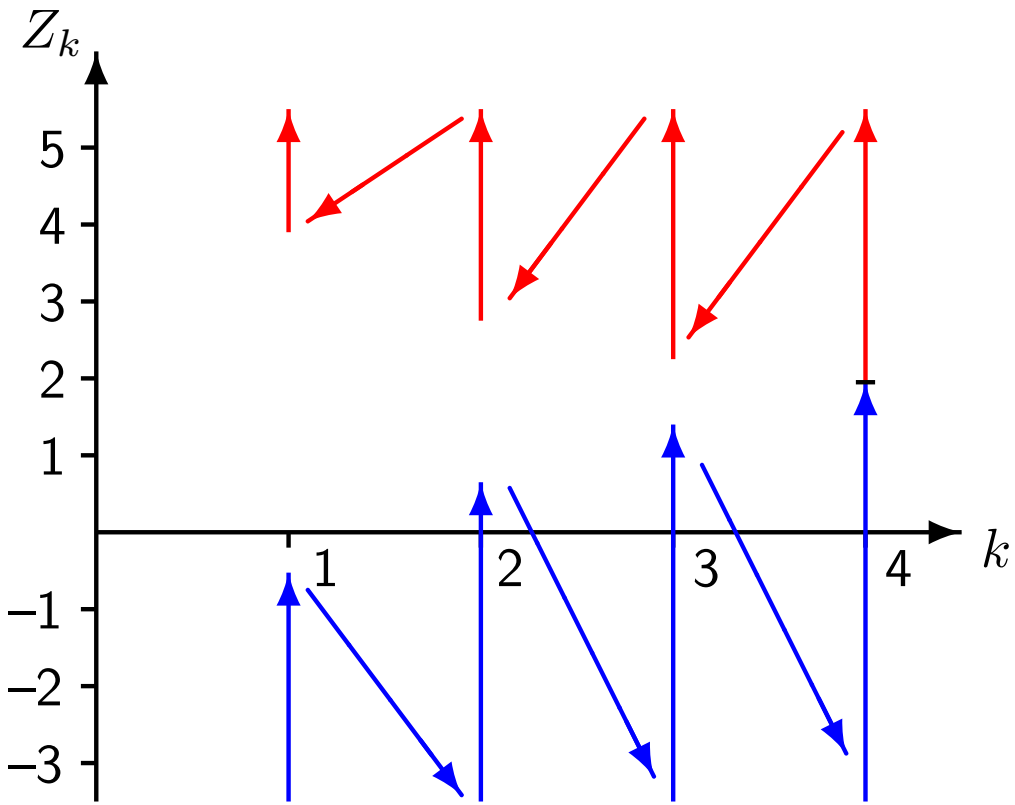
stopping at stage 4 with $Z_4 > 2.60$,

and the total probability under $\theta = 0$ of "$P \leq 0.0047$" is 0.0076.

So, this "P-value" does *not* have the null distribution $U(0,1)$.

Analysis on termination of a group sequential test (GST)

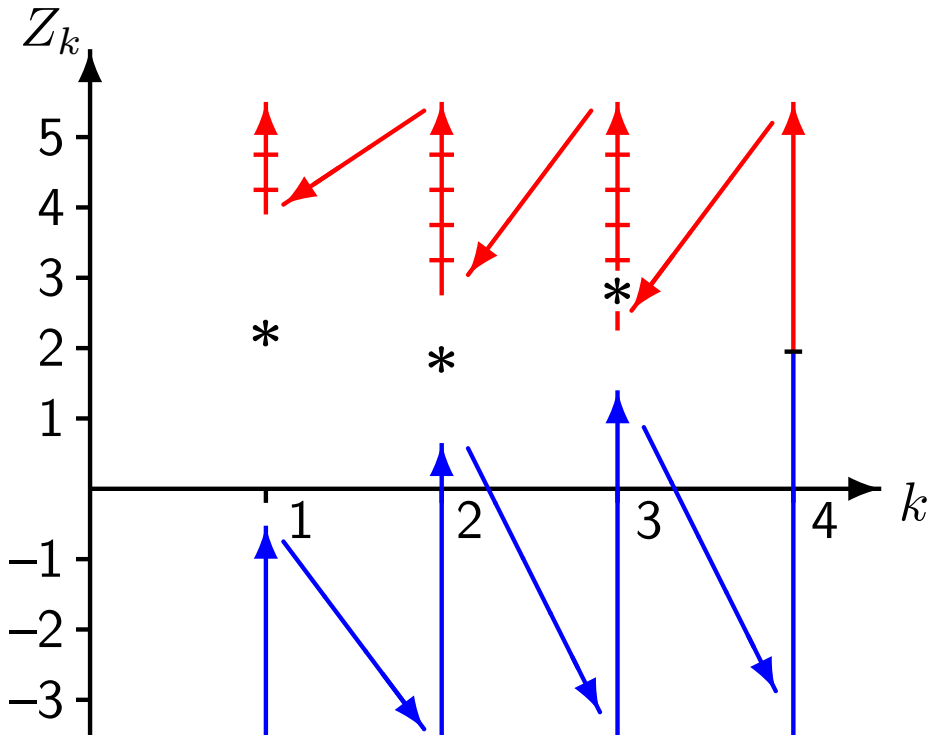
For proper frequentist inference, we first order the sample space.



Then, we define P-values and confidence intervals with respect to this ordering.

(i) A P-value on termination

The *P-value* for H_0: $\theta = 0$ is the probability under H_0 of seeing an outcome as extreme as that observed.



On stopping with $Z_3 = 2.60$, the 1-sided P-value for H_0: $\theta \leq 0$ is

$$P_{\theta=0}\{\text{Stop with } Z_1 \geq 3.90 \text{ or } Z_2 \geq 2.76 \text{ or } Z_3 \geq 2.60\} = 0.0063.$$

A P-value on termination

With the above definition, based on a specific ordering of the sample space:

The P-value has a $U(0,1)$ distribution under H_0.

If the group sequential test has one-sided type I error probability α, the P-value is $\leq \alpha$ precisely when the test stops with rejection of H_0,

i.e., in the part of the sample space coloured red in the previous slide.

The P-value will tend to take low values when θ is large and positive.

(ii) A confidence interval on termination

First, we specify an ordering of the GST's sample space.

Suppose the trial terminates at analysis k^* with $Z_{k^*} = Z^*$.

We define the $100(1-2\alpha)\%$ confidence interval for θ to be the

> set of values θ for which the observed (k^*, Z^*) is in the middle $(1-2\alpha)$ of the probability distribution under θ.

This is the interval (θ_1, θ_2) where

$$P_{\theta=\theta_1}\{\text{An outcome above } (k^*, Z^*)\} = \alpha$$

and

$$P_{\theta=\theta_2}\{\text{An outcome below } (k^*, Z^*)\} = \alpha.$$

There is a duality between this $100(1-2\alpha)\%$ confidence interval and the family of level 2α, two-sided tests of hypotheses $H\colon\ \theta = \tilde{\theta}$.

A confidence interval on termination

Example:

Suppose the trial stops at analysis 3 with $Z_3 = 2.6$.

Using our specified ordering, the 95% confidence interval for θ is

$$(0.22,\ 1.77)$$

In contrast:

The "naive" fixed sample CI would be $(0.25,\ 1.78)$.

However, it is not appropriate to use this fixed sample interval as this fails to take account of the sequential stopping rule.

Consequently, the coverage probability of this naive, fixed sample interval is *not* $1 - 2\alpha$.

Consistency of hypothesis testing and CI on termination

Suppose a group sequential trial is run to test H_0: $\theta \leq 0$ vs $\theta > 0$ with one-sided type I error probability α.

Then, a $1 - 2\alpha$, equal-tailed confidence interval on termination should lie completely above $\theta = 0$ if and only if H_0 is rejected.

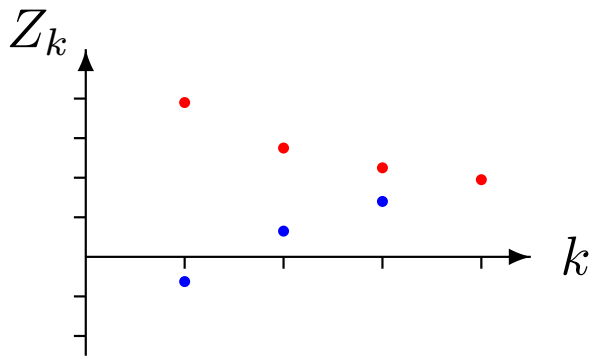
This happens automatically if outcomes for which we reject H_0 are at the top end of the sample space ordering — and any sensible ordering does this.

Why the naive approach does not work

A naive $1 - 2\alpha$ level CI on termination lies completely above $\theta = 0$ if an *unadjusted* α level, one-sided significance test rejects H_0.

Since there are multiple analyses, the probability of such an outcome is liable to be greater than the desired level α.

(iii) Estimating θ after a group sequential test



In a two-treatment comparison, the maximum likelihood estimate (MLE) of θ on termination of the trial at analysis k is

$$\widehat{\theta}_M \;=\; \bar{X}_{Ak} \;-\; \bar{X}_{Bk}.$$

For large, positive values of θ:

high values of $\widehat{\theta}$ lead to early stopping,

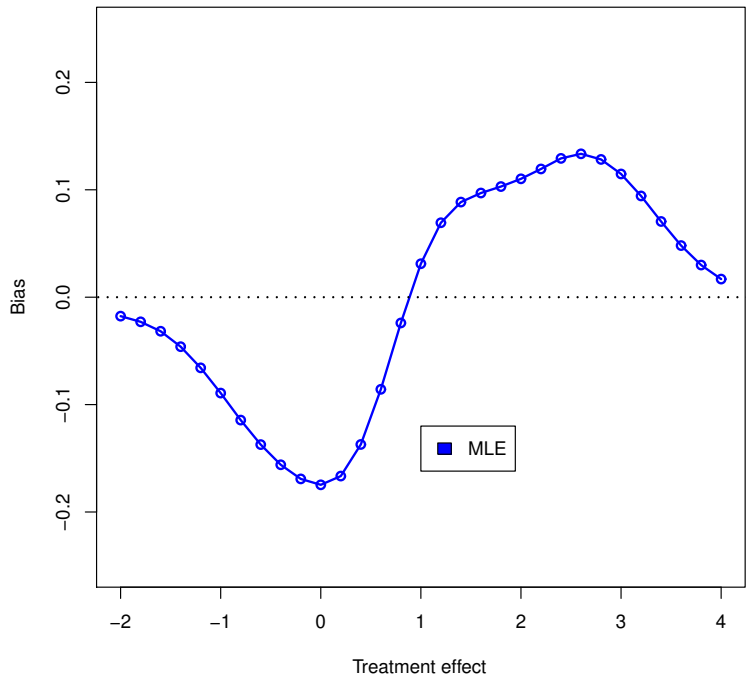
lower values of $\widehat{\theta}$ result in more observations, so $\widehat{\theta}$ can increase.

Thus, the MLE is biased with $E_\theta(\widehat{\theta}_M) > \theta$ for high values of θ.
Similarly, $E_\theta(\widehat{\theta}_M) < \theta$ for low values of θ.

Bias of the MLE of θ after a Pampallona & Tsiatis test

Consider the Pampallona & Tsiatis GST with $\Delta = 0$, $K = 4$ analyses, $\alpha = 0.025$ and power $1 - \beta = 0.8$ at $\theta = 1$

The bias of the MLE can be calculated as a function of the true effect size, θ.



The bias of the MLE is around 0.1 at values of θ just above 1.

Correcting the bias of the MLE

Denote the bias function of the MLE by

$$b(\theta) = E_\theta(\widehat{\theta}_M) - \theta.$$

Whitehead (*Biometrika*, 1986) suggested correcting the MLE by subtracting an estimate of its bias.

Although the true θ is unknown, the bias of the MLE can be estimated by $b(\widehat{\theta}_M)$.

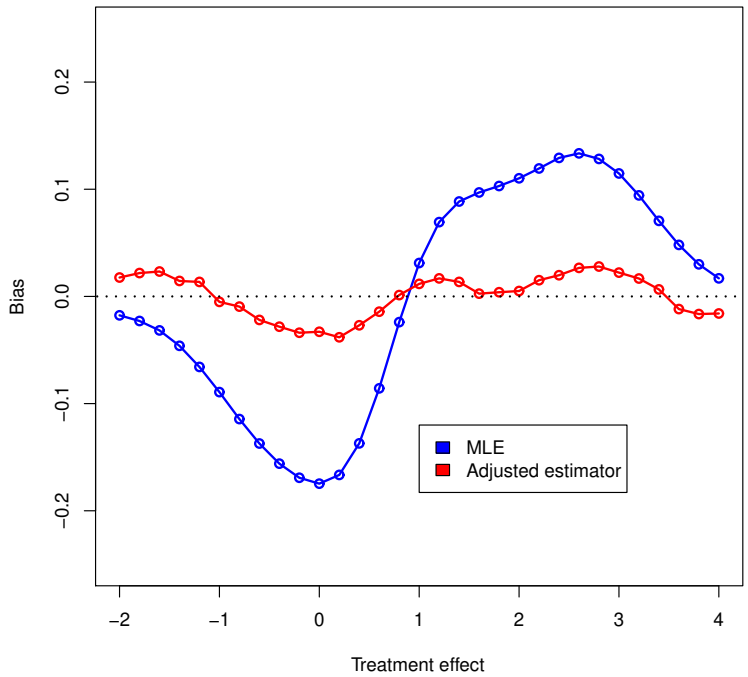
The adjusted estimator is then

$$\widehat{\theta}_{adj} = \widehat{\theta}_M - b(\widehat{\theta}_M).$$

Bias of the MLE of θ after a Pampallona & Tsiatis test

Simulation results show that Whitehead's adjusted estimator has much smaller bias than the MLE on which it is based.

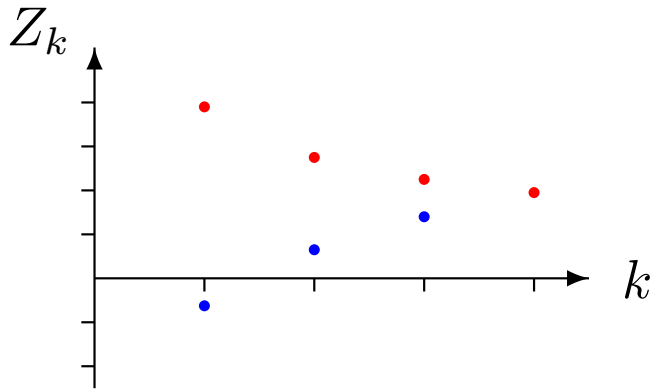
For our example:



The adjustment almost completely removes the bias in the MLE.

(iv) Estimation for a secondary endpoint after a GST

Stopping boundary for the primary endpoint



Denote the treatment effect on the primary endpoint by θ_1.

Suppose the trial stops and rejects H_0: $\theta_1 \leq 0$ in favour of $\theta_1 > 0$.

On stopping, data on a secondary endpoint are analysed to estimate the treatment effect, θ_2, on this endpoint.

For an individual, primary and secondary responses are correlated.

The group sequential design leads to bias in the MLE $\hat{\theta}_1$ — and the correlated responses imply that bias is passed on to the MLE $\hat{\theta}_2$.

Estimation for a secondary endpoint after a GST

Suppose an individual's responses are bivariate normal with correlation ρ.

For a patient on Treatment A,

$$\begin{aligned} &\text{Primary endpoint} && X_1 \sim N(\mu_{A1}, \sigma_1^2), \\ &\text{Secondary endpoint} && X_2 \sim N(\mu_{A2}, \sigma_2^2). \end{aligned}$$

Similarly, for a patient on Treatment B,

$$\begin{aligned} &\text{Primary endpoint} && X_1 \sim N(\mu_{B1}, \sigma_1^2), \\ &\text{Secondary endpoint} && X_2 \sim N(\mu_{B2}, \sigma_2^2). \end{aligned}$$

The primary treatment effect is

$$\theta_1 = \mu_{A1} - \mu_{B1}$$

and the secondary treatment effect is

$$\theta_2 = \mu_{A2} - \mu_{B2}.$$

Estimation for a secondary endpoint after a GST

Consider a group sequential design where the bias in the MLE $\hat{\theta}_1$ is

$$b_1(\boldsymbol{\theta}) \;=\; E_\theta(\hat{\theta}_1) - \theta_1$$

when the true treatment effects are $\boldsymbol{\theta} = (\theta_1, \theta_2)$.

Note that $E_\theta(\hat{\theta}_1)$ depends on θ_1 and not on θ_2.

Whitehead (*Biometrics*, 1986) shows that the MLE $\hat{\theta}_2$ has bias

$$b_2(\boldsymbol{\theta}) \;=\; E_\theta(\hat{\theta}_2) - \theta_2 \;=\; \rho \sqrt{(\sigma_2^2/\sigma_1^2)}\; b_1(\boldsymbol{\theta})$$

when the true treatment effects are $\boldsymbol{\theta} = (\theta_1, \theta_2)$.

Note that this bias depends on θ_1 — and not on θ_2.

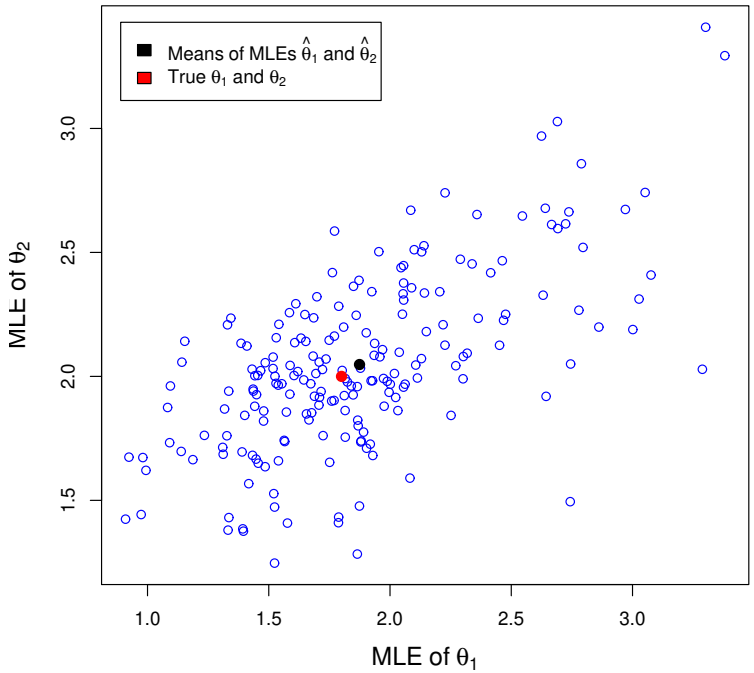
As for the primary endpoint, we can adjust the MLE, $\hat{\theta}_2$, by subtracting an estimate of its bias, $(\rho\,\sigma_2/\sigma_1)\,b_1(\hat{\boldsymbol{\theta}})$.

Estimation for a secondary endpoint: Example

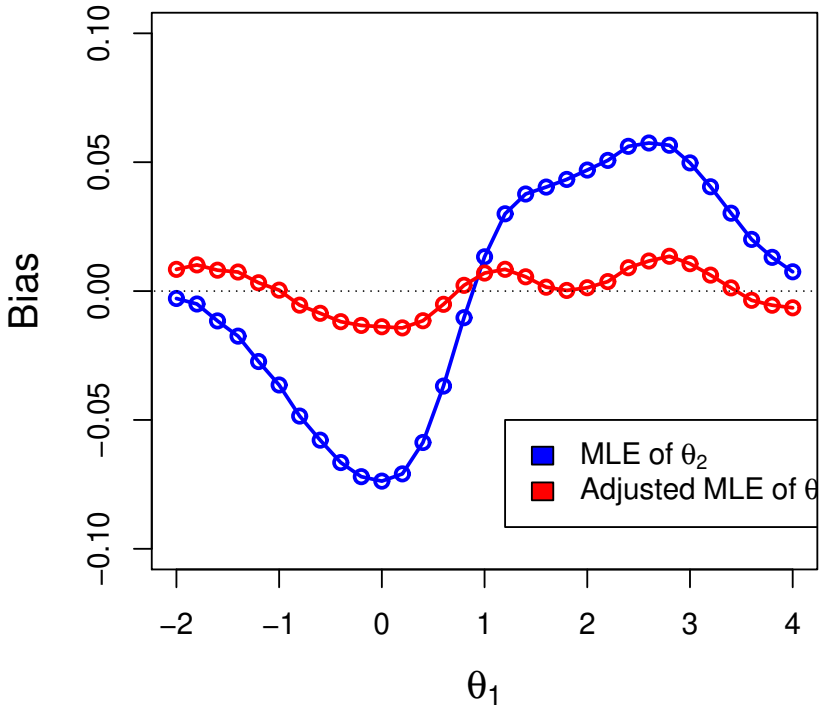
Suppose a trial tests its primary endpoint, using a Pampallona & Tsiatis GST with $\Delta = 0$, $\alpha = 0.025$ and power 0.8 at $\theta_1 = 1$.

Responses are bivariate normal, $\rho = 0.6$ and $\sigma_1^2/\sigma_2^2 = 2$.

The plot, for the case $\theta_1 = 1.8$ and $\theta_2 = 2$, shows the correlation between the MLEs, $\hat{\theta}_1$ and $\hat{\theta}_2$, on termination of the GST.



Estimation for a secondary endpoint: Example



We see that the bias in the MLE $\hat{\theta}_2$ is largely eliminated in the adjusted estimator

$$\hat{\theta}_2 - \rho \sqrt{\frac{\sigma_2^2}{\sigma_1^2}}\, b_1(\hat{\boldsymbol{\theta}}).$$

(v) Testing a secondary endpoint after a GST

In a trial of two treatments, A and B, a group sequential test is carried out on the primary endpoint, which has treatment effect θ_1.

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

The investigators wish to test whether Treatment A is also superior for a secondary endpoint, with treatment effect denoted by θ_2.

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

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

Is this approach to testing the two endpoints valid?

Testing a secondary endpoint: Example

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

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

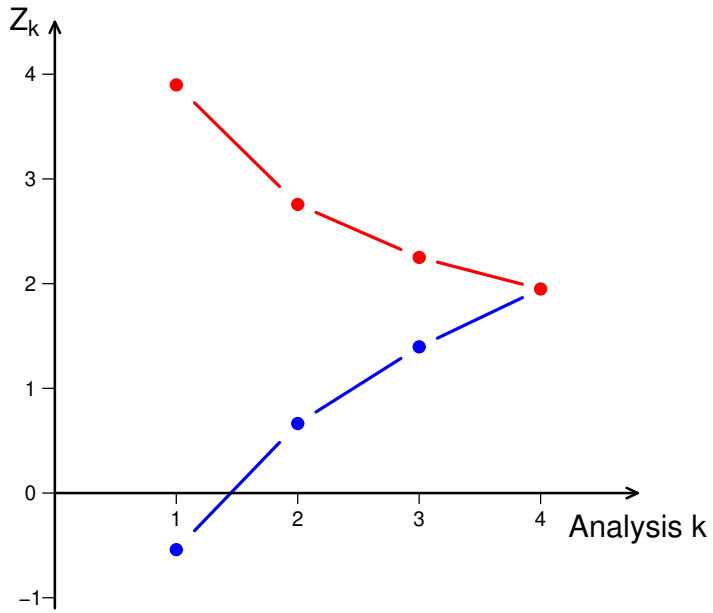
This test has upper boundary:

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

and lower boundary

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

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

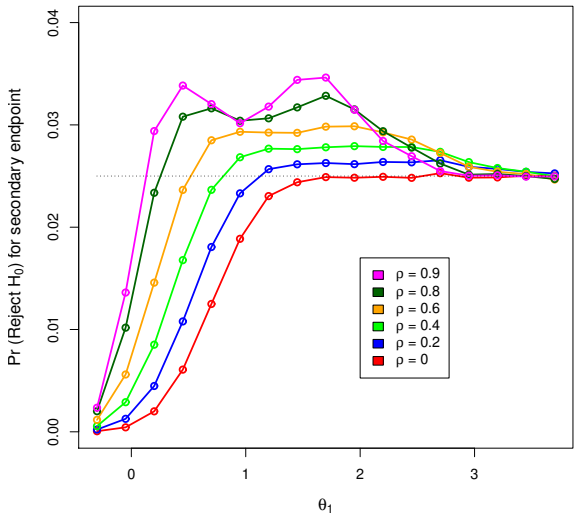


If the upper boundary is crossed, the secondary endpoint is tested in a level α, fixed sample size test, using current data.

Testing a secondary endpoint: Example

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

For modest values of ρ, the correlation between the two endpoints, the type I error rate for testing H_2 exceeds the nominal 0.025.



Hung, Wang and O'Neill (*J. Biopharm. Statist.*, 2007) noted that this approach to testing a secondary endpoint is not valid.

So, how should the secondary endpoint be tested?

To answer this, we need to consider **multiple testing procedures**.

Recapitulation: Group sequential tests

- It is natural to monitor clinical trials with a view to early stopping.
- Theory and computational methods support a variety of group sequential designs that control the type I error rate.
- These designs can be optimised for a given objective.
- Error spending designs offer efficient, flexible monitoring of a variety of response types, including survival data.
- Information monitoring facilitates changes to ensure the trial has enough subjects or events to achieve the desired statistical power.
- Group sequential tests can accommodate a delayed response.
- Inference on termination can provide P-values, confidence intervals and approximately unbiased point estimates.
- In order to conduct **multiple** hypothesis tests group sequentially, **we shall need additional methodology.**

Part 3. Multiple testing procedures

3.1. Introduction: The familywise error rate

3.2. Bonferroni tests

3.3. Recycling type I error probability

3.4. Example: Primary and secondary endpoints

3.5. Graphical representation of multiple testing procedures

3.6. Combining multiple testing and group sequential design

3.7. Example: Testing a secondary endpoint after a group sequential test for the primary endpoint

3.1 Multiple testing: The familywise error rate

We just saw an example with a primary and a secondary endpoint.

More generally, a clinical trial may involve

Co-primary endpoints

Positive outcomes required for at least one endpoint

Positive outcomes required on all endpoints

Secondary endpoints, tertiary endpoints, ...

The trial may have

Multiple treatments,

Pre-defined sub-populations of patients.

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

The familywise error rate

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

A testing procedure's **familywise error rate** under a set of values $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_h)$ is

$$P_{\boldsymbol{\theta}}\{\text{Reject } H_i \text{ for some } i \text{ with } \theta_i \leq 0\}$$

$$= P_{\boldsymbol{\theta}}\{\text{Reject at least one true } H_i\}.$$

The familywise error rate is controlled **strongly** at level α if this error rate is at most α for all possible combinations of θ_i values.

Then

$$P_{\boldsymbol{\theta}}\{\text{Reject any true } H_i\} \leq \alpha \quad \text{for all } \boldsymbol{\theta} = (\theta_1, \ldots, \theta_h).$$

3.2 Bonferroni adjustment (Carlo Bonferroni, 1892–1960)

Suppose we test h null hypotheses, each at significance level α/h.

If $\boldsymbol{\theta}$ is such that all h null hypotheses are true,

$$P_{\boldsymbol{\theta}}\{\text{Reject at least one of } H_1, \ldots, H_h\}$$

$$\leq P_{\boldsymbol{\theta}}\{\text{Reject } H_1\} + \ldots + P_{\boldsymbol{\theta}}\{\text{Reject } H_h\} \leq h\,\frac{\alpha}{h} = \alpha.$$

If $\boldsymbol{\theta}$ is such that only some of the h null hypotheses are true,

$$P_{\boldsymbol{\theta}}\{\text{Reject at least one true } H_i\} < \alpha.$$

So we have **strong control** of the **familywise error rate**.

We start by considering applications in fixed sample size study designs ...

Example: A Bonferroni test with co-primary endpoints

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

Endpoint 1, Core MACE (*Major Adverse Cardiac Event* — CV-related death, nonfatal stroke, or nonfatal MI)

Endpoint 2, Expanded MACE (Core MACE plus hospitalization for unstable angina or coronary revascularization).

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

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

An effect on Core MACE is declared if

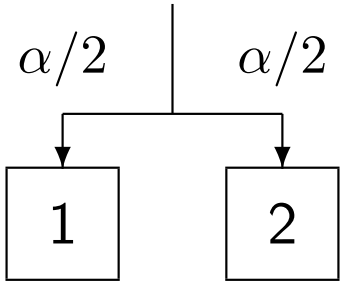
$$Z_1 > \Phi^{-1}(1 - \alpha/2) = 2.24,$$

An effect on Expanded MACE is declared if

$$Z_2 > \Phi^{-1}(1 - \alpha/2) = 2.24.$$

Example: Co-primary endpoints

This Bonferroni procedure can be represented graphically as:



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

Hence, familywise type I error is protected conservatively.

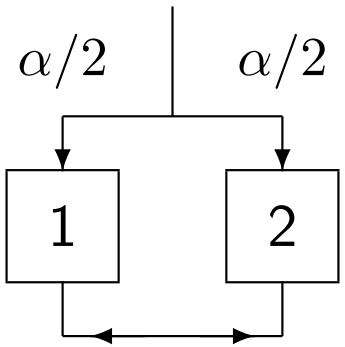
Also, if one hypothesis is false, the familywise type I error probability is at most $\alpha/2$.

Power when H_1 and H_2 are false can be increased by "recycling" type I error after one or other hypothesis is rejected.

3.3 Recycling type I error probability

The Holm procedure is a version of the Bonferroni procedure that "recycles" error probability after rejecting H_1 or H_2.

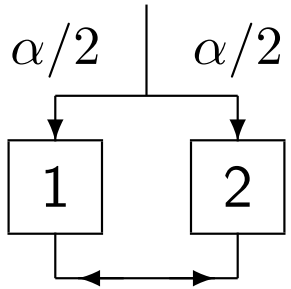
This method can be represented as:



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

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

Proof that FWER is protected in the Holm procedure



If H_1 and H_2 are both true,

$$\begin{aligned}
\text{FWER} &= P_{\boldsymbol{\theta}}\{\text{Reject } H_1 \text{ or } H_2\} \\
&\leq P_{\boldsymbol{\theta}}\{Z_1 > \Phi^{-1}(1-\alpha/2)\} + P_{\boldsymbol{\theta}}\{Z_2 > \Phi^{-1}(1-\alpha/2)\} \\
&\leq \alpha/2 + \alpha/2 = \alpha.
\end{aligned}$$

If H_1 is true and H_2 is false,

$$\text{FWER} = P_{\boldsymbol{\theta}}\{\text{Reject } H_1\} \leq P_{\boldsymbol{\theta}}\{Z_1 > \Phi^{-1}(1-\alpha)\} \leq \alpha.$$

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

H_1 and H_2 both false: A type I error cannot be made.

3.4 Example: Primary and secondary endpoints

A hierarchical testing or "gatekeeping" procedure

Consider a trial where

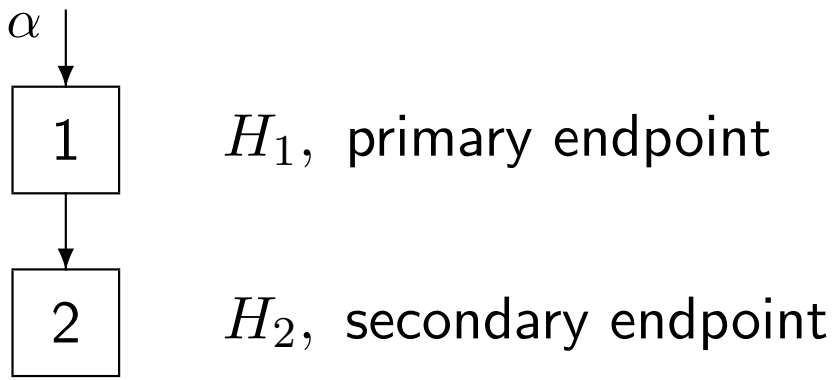
The null hypothesis H_1 concerns the primary endpoint,

The null hypothesis H_2 relates to a secondary endpoint,

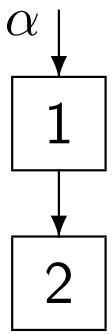
and H_2 will only be tested if H_1 has already been rejected.

First, we test H_1 at significance level α.

If H_1 is rejected, we continue and test H_2 at significance level α.



Proof: FWER is protected in the gatekeeping procedure



H_1, primary endpoint

H_2, secondary endpoint

Suppose H_1 is true.

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

$$\text{FWER} = P_{\boldsymbol{\theta}}\{\text{Reject } H_1\} = P_{\boldsymbol{\theta}}\{Z_1 > \Phi^{-1}(1-\alpha)\} \leq \alpha.$$

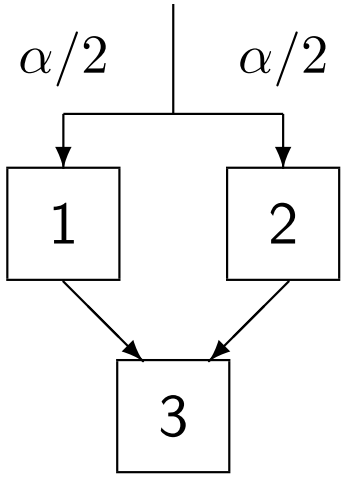
If H_1 is false and H_2 is true,

$$\begin{aligned} \text{FWER} &= P_{\boldsymbol{\theta}}\{\text{Reject } H_1 \text{ and then reject } H_2\} \\ &\leq P_{\boldsymbol{\theta}}\{Z_2 > \Phi^{-1}(1-\alpha)\} \leq \alpha. \end{aligned}$$

If H_1 and H_2 are both false, a type I error cannot be made.

Example: Testing co-primary and secondary endpoints

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



H_1, H_2: co-primary endpoints

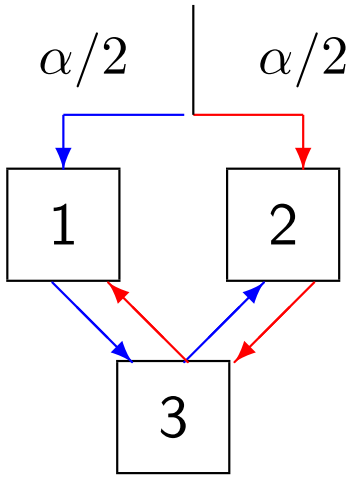
H_3: secondary endpoint

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

We can prove there is strong control of FWER at level α by considering all combinations of H_1, H_2 and H_3 being True or False.

Testing co-primary and secondary endpoints

We can add more "recycling" to the previous testing procedure.



H_1, H_2: co-primary endpoints

H_3: secondary endpoint

The additional lines in the graph indicate that

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

If $P_2 \leq \alpha/2$ and $P_3 \leq \alpha/2$, then H_1 is tested at level α.

Testing co-primary and secondary endpoints

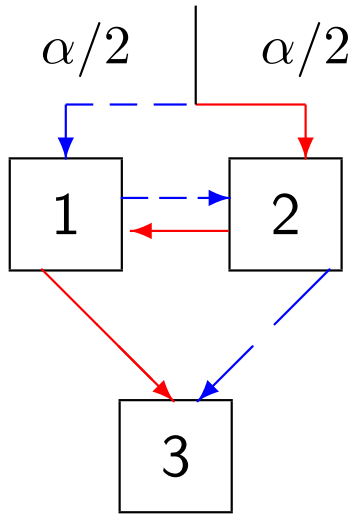
We may prefer to gain maximum power for tests of co-primary endpoints before testing a secondary endpoint.

To do this, we recycle type I error probability between H_1 and H_2 before allocating any error probability to H_3.

A graphical representation is:

H_1, H_2: co-primary endpoints

H_3: secondary endpoint



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

The other half is cycled through H_2, H_1 and on to H_3.

3.5 Graphical representation of multiple testing procedures

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

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

> *"A recycling framework for the construction of Bonferroni-based multiple tests" by Burman, Sonesson & Guilbaud,* Statistics in Medicine, *2009.*

> *"A graphical approach to sequentially rejective multiple test procedures" by Bretz, Maurer, Brannath & Posch,* Statistics in Medicine, *2009.*

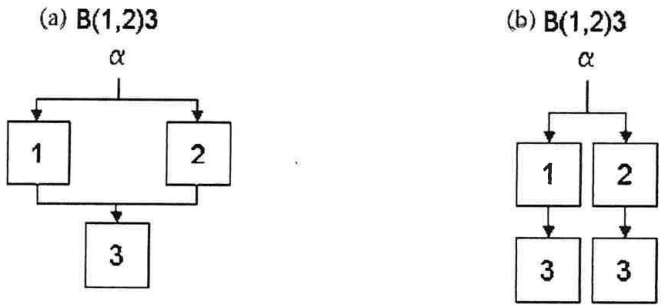
These procedures are closed testing procedures in which the tests of intersection hypotheses are weighted Bonferroni tests.

It is implicit in their method of construction that these procedures provide strong control of the FWER.

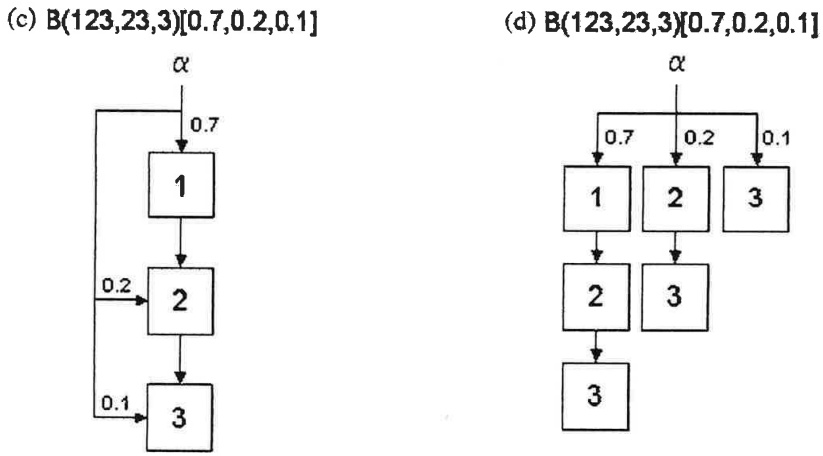
A figure from Burman et al. (2009)

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

(a) and (b) A parallel gatekeeping procedure



(c) and (d) A fallback procedure



A figure from Bretz et al. (2009)

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

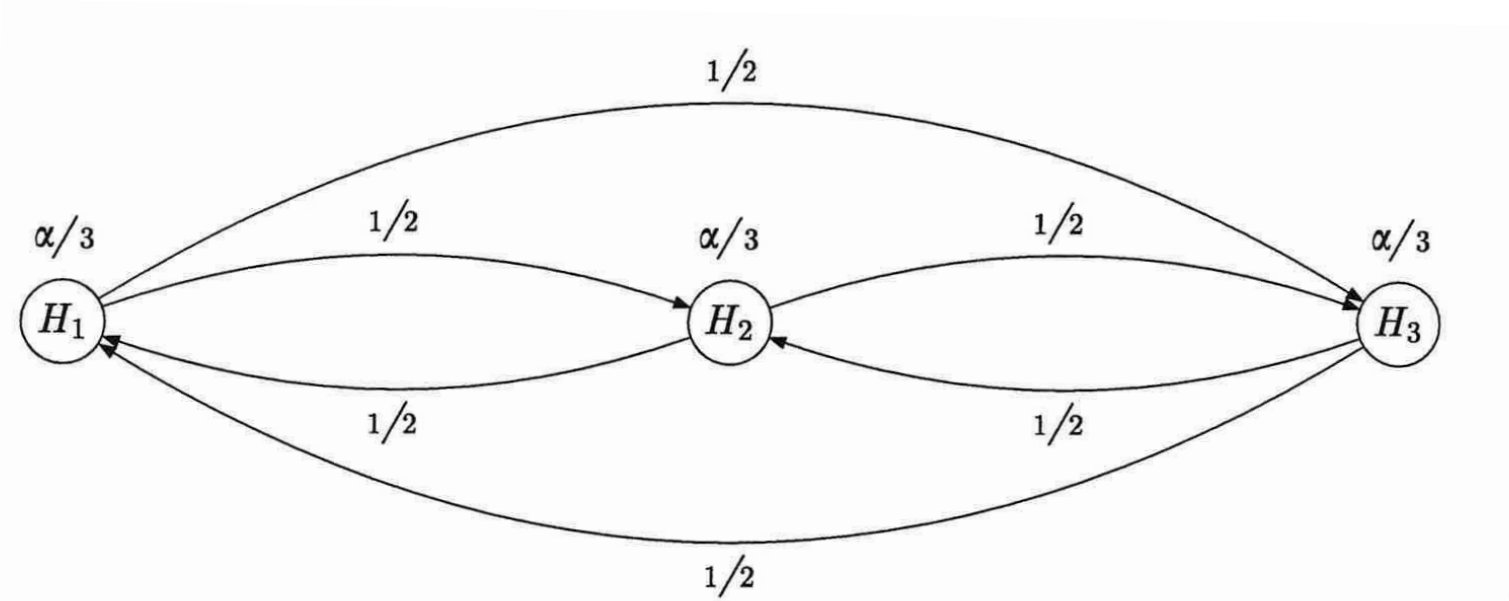


Figure 3. Graphical illustration of the Bonferroni–Holm procedure with $m=3$ hypotheses and initial allocation $\boldsymbol{\alpha}=(\alpha/3, \alpha/3, \alpha/3)$.

Question: How can we apply such a procedure in a group sequential trial?

3.6 Multiple testing within a group sequential design

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

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

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

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

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

Stop the study when key conclusions have been reached.

Combining multiple testing and group sequential design

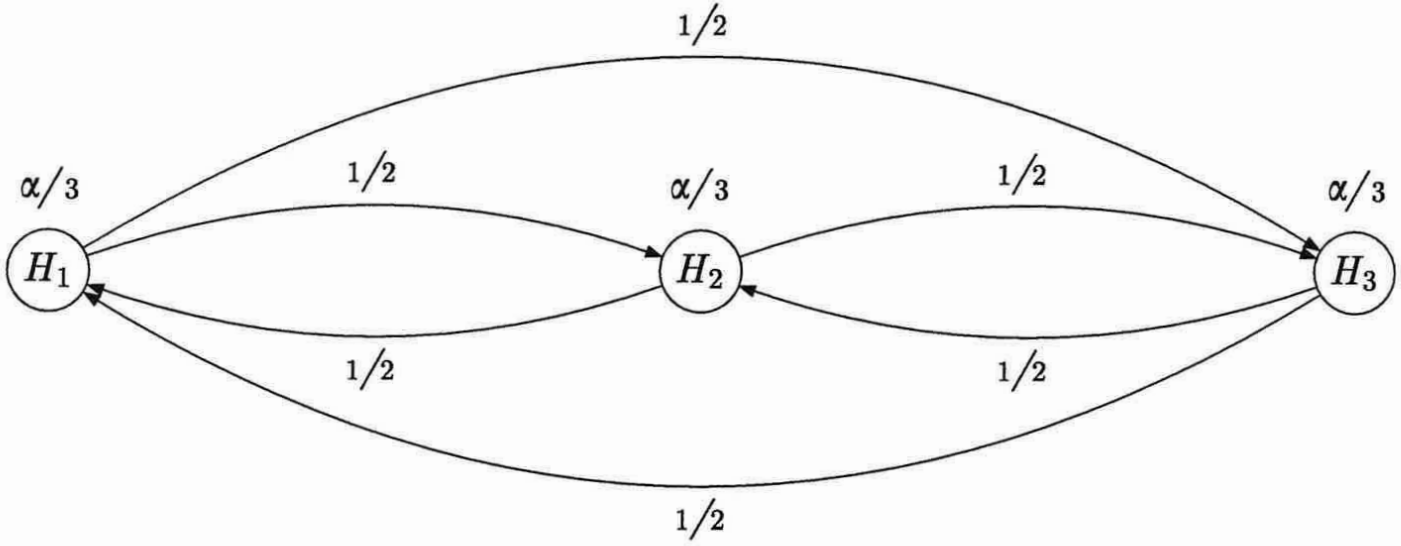


Figure 3. Graphical illustration of the Bonferroni–Holm procedure with $m = 3$ hypotheses and initial allocation $\boldsymbol{\alpha} = (\alpha/3, \alpha/3, \alpha/3)$.

For group sequential implementation of the above multiple testing procedure, we need

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

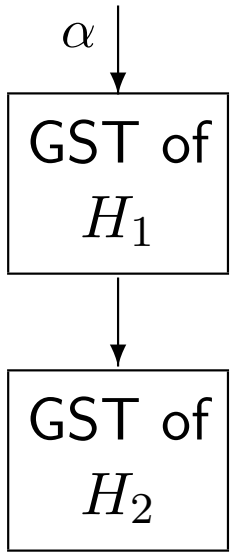
for each of the hypotheses, H_1, H_2 and H_3.

3.7 Testing a secondary endpoint after a sequential test

A correct gatekeeping procedure

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

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



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

The group sequential test of H_2 is used for the secondary analysis if and when H_1 is rejected

A correct gatekeeping procedure

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

The group sequential test of H_1 stops at analysis k to

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

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

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

$$\sum_{k=1}^{K} P_{\theta_1=0}\{Z_{1,1} \in (a_1, b_1), \ldots, Z_{1,k-1} \in (a_{k-1}, b_{k-1}), Z_{1,k} > b_k\} = \alpha.$$

Suppose this GST stops to reject H_1 at analysis k^* ...

A correct gatekeeping procedure

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

The level α group sequential test of H_2 rejects H_2 at analysis k if $Z_{2,k} \geq c_k$, where

$$\sum_{k=1}^{K} P_{\theta_2=0}\{Z_{2,1} < c_1, \ldots, Z_{2,k-1} < c_{k-1}, Z_{2,k} > c_k\} = \alpha.$$

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

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

A gatekeeping procedure *could* reject H_2 if

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

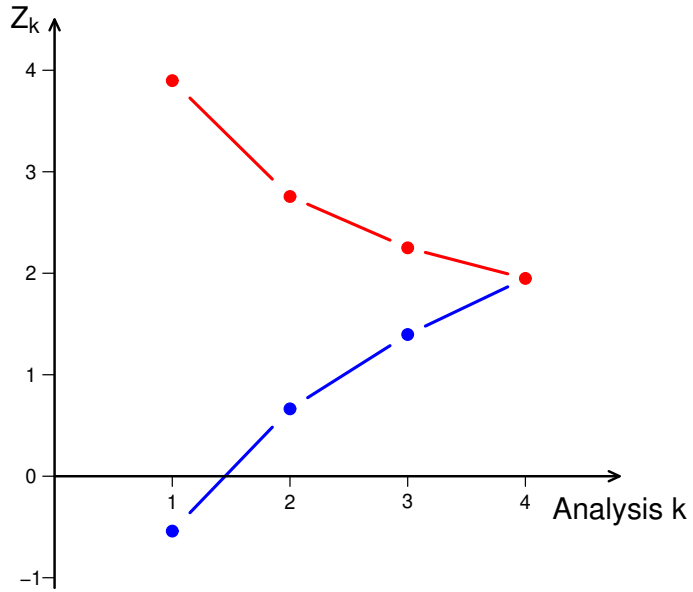
so the FWER is protected conservatively.

Example: Testing primary and secondary endpoints

In a trial comparing two treatments, denote the treatment effects on the primary and secondary endpoints by θ_1 and θ_2.

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

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



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

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

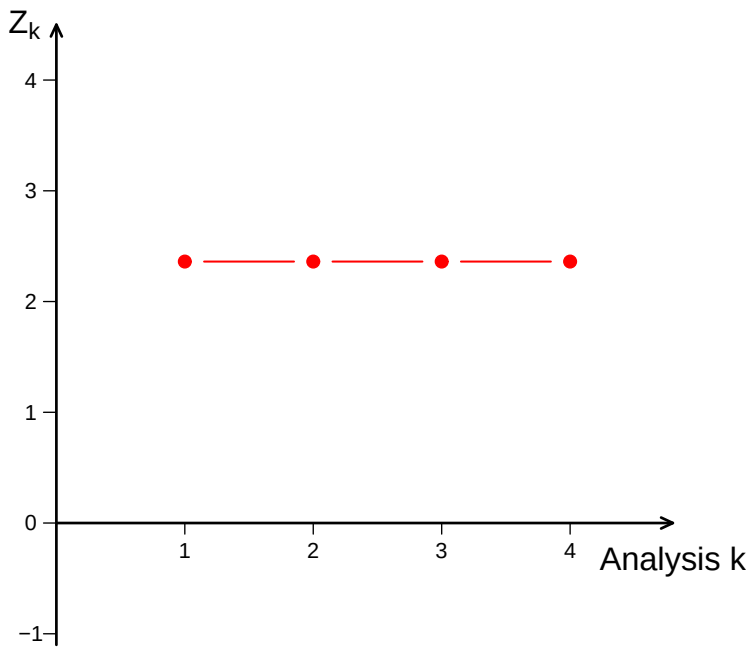
We consider two options for this test of H_2.

Example: Testing primary and secondary endpoints

Consider two options for the group sequential test of H_2.

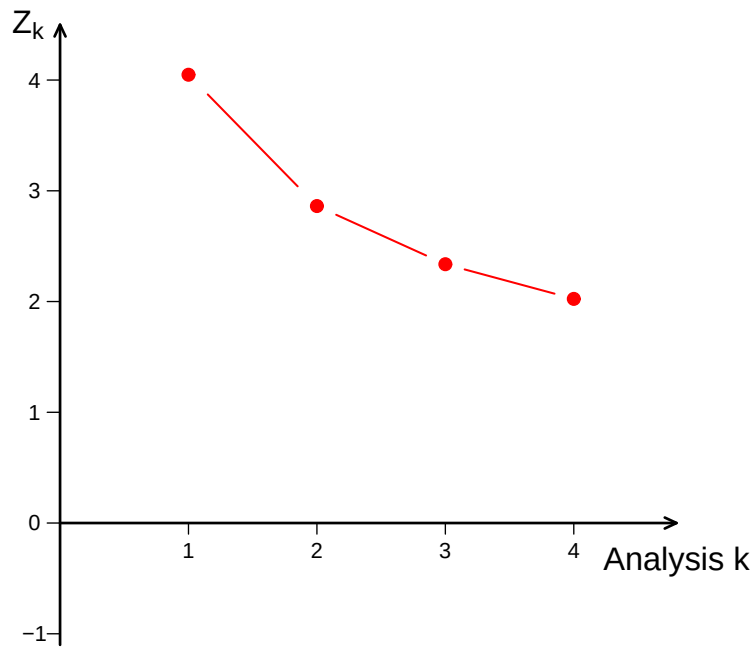
A: Pocock boundary for H_2

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



B: OBF boundary for H_2

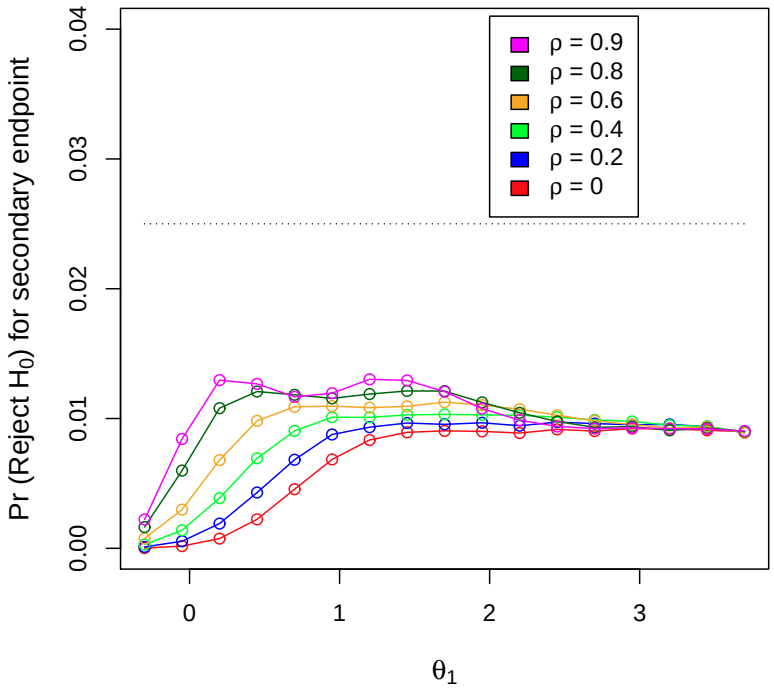
$$c_k = 2.024\sqrt{4/k}, \quad k = 1, \ldots, 4.$$



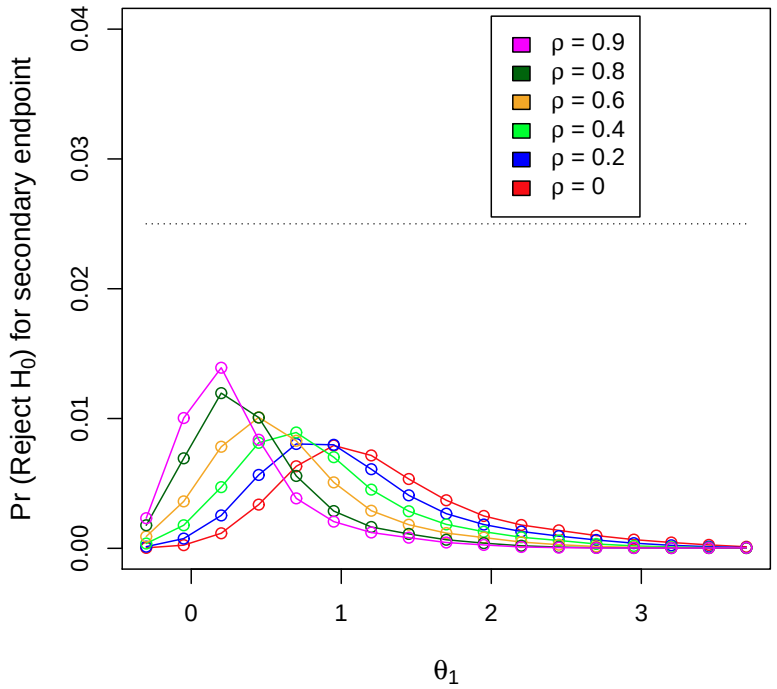
Note: The O'Brien & Fleming boundary requires a very high value of Z_{2,k^*} to reject H_2 if the GST of H_1 stops at the first analysis.

Type I error probability for testing H_2

A: Pocock boundary for H_2



B: OBF boundary for H_2

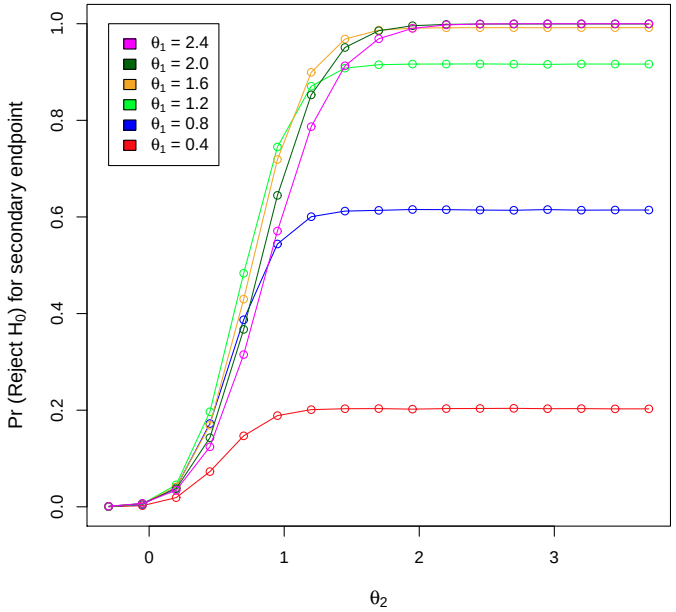


Type I error probabilities are calculated under $\theta_2 = 0$, but they also depend on θ_1 and the correlation, ρ, between the primary and secondary endpoints.

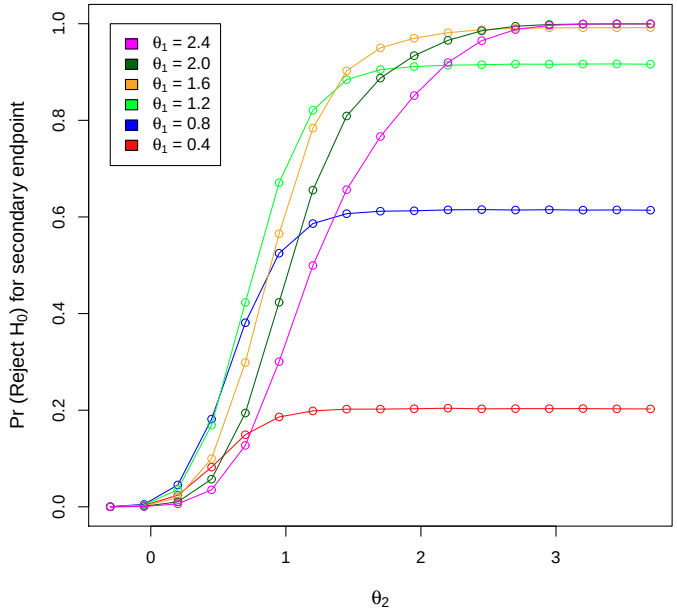
The OBF test of H_2 is particularly conservative when θ_1 is large.

Power for testing H_2, $\rho = 0.25$

A: Pocock boundary for H_2



B: OBF boundary for H_2



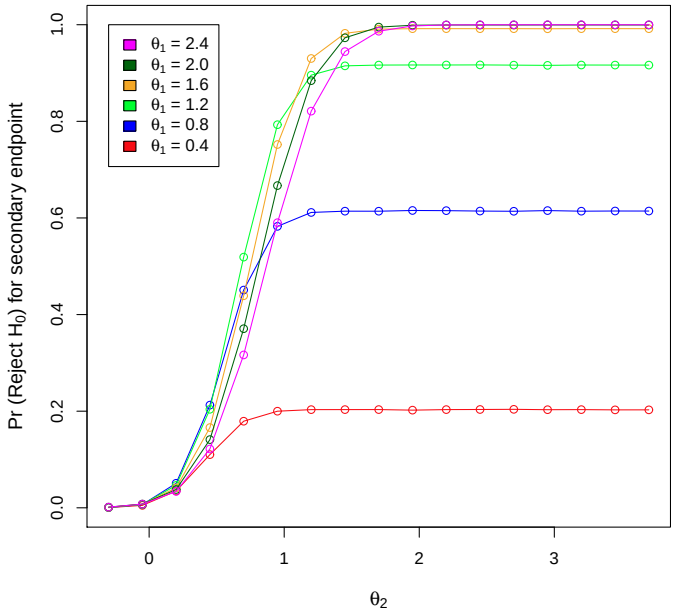
Results are shown for the case that the variance of the secondary response is 0.5 times that for the primary response.

Power is shown as a function of θ_2 for selected values of θ_1.

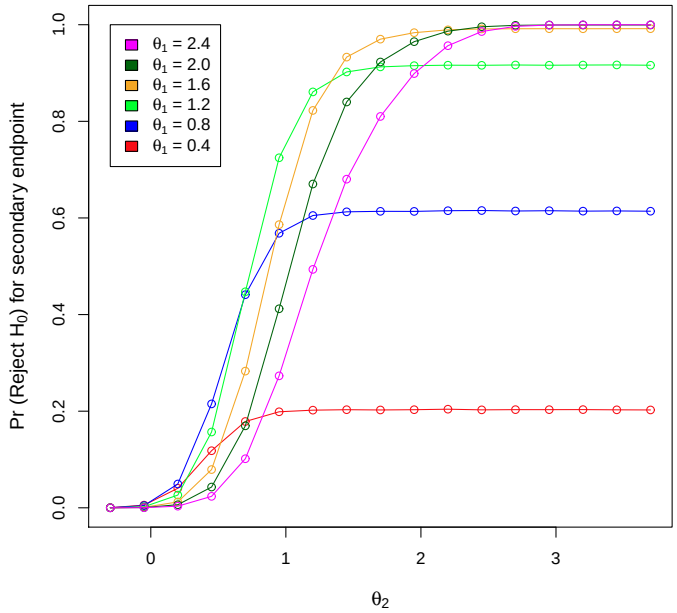
The Pocock boundary for H_2 deals better with the trial's uncertain termination time — which depends significantly on the value of θ_1.

Power for testing H_2, $\rho = 0.5$

A: Pocock boundary for H_2



B: OBF boundary for H_2



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

Power is shown as a function of θ_2 for selected values of θ_1.

Again, the Pocock boundary for H_2 deals better with the trial's uncertain termination time — which depends significantly on θ_1.

Testing a secondary endpoint: Further options

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

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

There are options for reducing conservatism and increasing power:

1. Reject H_2 if $Z_{2,k} \geq c_k$ for some $k < k^*$, even though $Z_{2,k^*} < c_{k^*}$.

   However, ignoring more recent data (and not using the sufficient statistic for θ_2) may detract from the credibility of this decision.

2. Continue the trial to see if $Z_{2,k} \geq c_k$ at a future analysis.

   However, if the primary endpoint is observed for future subjects, the positive result on the primary endpoint could be "lost".

Several authors have considered option (2), retaining a positive outcome for H_1, whatever the additional information about θ_1.

Testing a secondary endpoint: Further options

3. In some cases, the worst case scenario, in which a procedure's maximum FWER occurs, can be identified.

Then, the procedure may be calibrated so that the FWER is equal to the specified level α in this worst case scenario. See:

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

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

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

Tamhane, Gou, Jennison, Mehta & Curto (Biometrics, 2018) A gatekeeping procedure to test a primary and a secondary endpoint in a group sequential design with multiple interim looks.

Recapitulation: Multiple hypothesis procedures

- There are many multiple testing schemes to choose from.

  The most suitable choice will depend on the importance to investigators of rejecting each null hypothesis and the likelihood of each null hypothesis being true or false.

- Graphical representations (SiM papers, 2009) can help in selecting — and understanding — an appropriate multiple testing procedure.

- Methods are available to test multiple hypotheses in a group sequential design AND control the overall type I error probability.

- When testing multiple hypotheses in a group sequential setting, the key is to use GSTs as "testing rules" in the multiple testing scheme: if this is not done correctly, FWER may be inflated.

Part 4. Adaptive clinical trial designs (1)

4.1. Motivation for adaptive designs

4.2. Combination tests

4.3. Sample size re-estimation

4.4. Adaptive trials that test multiple hypotheses

4.5. Closed Testing Procedures

4.6. Using combination tests in a Closed Testing Procedures

4.1 Motivation for adaptation in clinical trials

Wall Street Journal, July 2006:

FDA Signals it's Open to Drug Trials that Shift Midcourse

Adaptive designs may allow trials to be adjusted:

- Route more patients to the treatment that seems to work best
- Drop treatments that don't seem to be effective
- Add more of the type of patients ... reacting best to a particular treatment
- Merge two different phases of drug development into one trial

With views from:

Bob O'Neill , FDA

Michael Krams, Wyeth

Paul Gallo, Novartis

Don Berry, M. D. Anderson Cancer Center

Tom Fleming, Univ. Washington

Bruce Turnbull, Cornell University

Adaptation in clinical trials

Adaptation to external factors

Changes in the clinical setting or economic background

Following withdrawal of a competing treatment, a smaller treatment effect is now of clinical interest.

An improved financial position means sponsors can invest more in this trial.

Adaptation to internal factors

Nuisance parameters affecting sample size

In-study estimates of sample variance indicate a greater sample size is needed to achieve the intended power.

Overall failure rates in a survival study are low: higher accrual and longer follow-up are required.

Adaptation in clinical trials

Adaptation to internal factors ...

Safety outcomes

Higher than expected toxicity implies dose should be reduced.

A lower rate of adverse events in the experimental treatment suggests it will suffice to demonstrate non-inferiority, rather than superiority.

Sub-group analyses

The new treatment benefits a particular sub-group: investigators wish to re-define the target population.

Change of endpoint

An alternative endpoint provides better discrimination between treatment groups: investigators wish to re-define the primary endpoint.

Adaptation in clinical trials

Adaptation to internal factors ...

Response on primary endpoint

Results on the new treatment are good and it is desirable to reach a conclusion as rapidly as possible.

Responses on the new treatment are not as good as anticipated: investigators wish to increase sample size to enhance power at lower effect sizes.

Response-dependent treatment allocation

Interim data suggest one treatment arm could be superior but results are not yet statistically significant. In order to improve treatment of patients in the trial, weight random allocation in favour of the currently superior treatment arm.

Adaptation in clinical trials

A trial with multiple treatments or dose levels

Eliminate weaker treatments as the study progresses.

Using a dose-response model, optimize treatment allocation to learn most efficiently about the best choice of dose level.

Seamless Phase IIb – Phase III trials

Select the best dose level in Phase IIb.

Proceed directly to Phase III and test the treatment at this dose level, eliminating "white space" between phases.

Combine Phase IIb and Phase III data in the final statistical analysis.

FDA Guidance for Industry, February 2010

Adaptive Design Clinical Trials for Drugs and Biologics

Section XI of the 2010 document stated that adaptations should be pre-specified and access to interim data should be strictly controlled.

> *" . . . there should be comprehensive and prospective, written standard operating procedures (SOPs) that define who will implement the interim analysis and adaptation plan, and all monitoring and related procedures for accomplishing the implementation, providing for the strict control of access to unblinded data (see the DMC guidance)*
>
> *. . . It is likely that the measures defined by the SOPs will be related to the type of adaptation and the potential for impairing study integrity."*

FDA Guidance for Industry, October 2018

Adaptive Design Clinical Trials for Drugs and Biologics

The updated guidance from the FDA sets out principles for the planning and conduct of adaptive designs under the headings

- Controlling the Chance of Erroneous Conclusions
- Estimating Treatment Effects
- Trial Planning
- Maintaining Trial Conduct and Integrity

The guidance is supportive of a variety of adaptive designs.

However, designs must guarantee control of type I error and provide valid additional inferences on termination.

Safeguards are required to ensure the process of adaptation and related flow of information do not undermine a trial's integrity.

Adaptive Designs for Medical Device Clinical Studies

The FDA's guidance for Studies of Medical Devices is similar in spirit to that for Adaptive Trials for Drugs and Biologics.

The phrase "adaptive design" includes group sequential designs with early stopping for efficacy or futility.

The document stresses the importance of:

Pre-planning and the use of pre-specified adaptation rules,

Protection of the type I error rate,

Firewalls to prevent information leakage,

Assessment of the benefits of adaptive design and weighing of these against more complex logistical requirements,

The benefits of adaptive designs as a "learning paradigm".

Protecting the type I error probability

The importance of the type I error rate is widely recognised:

- ICH E9 (p. 25)

  "The procedures selected should always ensure that the overall probability of type I error is controlled."

- PhRMA White paper (*J. Biopharmaceutical Statistics*, 2006)

  "The key issue in most contexts is preservation of the type I error rate."

- Pocock & Hughes (*Controlled Clinical Trials*, 1989)

  "Control of type I error is a vital aid to prevent a flood of false positives into the medical literature."

However, for a complex adaptive procedure, it may be difficult to demonstrate control of the type I error rate over the whole of a multi-dimensional null hypothesis.

4.2 Combination tests

Suppose we run a clinical trial adaptively in two stages:

Set the design of Stage 1, then conduct this part of the trial,

Analyse results from Stage 1,

Consider external information, if appropriate.

Set the design of Stage 2, informed by Stage 1 results and external information,

Conduct Stage 2,

Analyse the results from Stage 2.

How can we test a null hypothesis with proper protection of the type I error rate?

Combination tests

Before the trial commences, define the null hypothesis.

Let θ denote the treatment effect *vs* control for a specified form of the treatment, patient population and endpoint.

We test H_0: $\theta \leq 0$ against $\theta > 0$, with type I error rate α at $\theta = 0$.

Define one-sided P-values $P^{(1)}$ and $P^{(2)}$ from hypothesis tests of H_0 based on Stage 1 and Stage 2 data, respectively.

Under $\theta = 0$

$P^{(1)} \sim U(0,1)$.

Conditionally on all Stage 1 data and the Stage 2 design, $P^{(2)} \sim U(0,1)$.

Hence, if $\theta = 0$, $P^{(1)}$ and $P^{(2)}$ are independent $U(0,1)$ variates.

The inverse χ^2 combination test

Reference: Bauer & Köhne (*Biometrics*, 1994).

Initial design

Define H_0 and specify the **inverse χ^2 combination test.**

Design Stage 1, fixing the sample size and test statistic.

Stage 1

Observe the one-sided P-value, $P^{(1)}$, based on Stage 1 data.

Design Stage 2 in the light of Stage 1 data.

Stage 2

Observe the P-value, $P^{(2)}$, based on **only** Stage 2 data.

NB: Under $\theta = 0$, $P^{(1)} \sim U(0,1)$, $P^{(2)} \sim U(0,1)$, independent.

Bauer & Köhne's inverse χ^2 combination test

Bauer & Köhne's test rejects H_0 for low values of $P^{(1)}\,P^{(2)}$.

If $P \sim U(0,1)$, then

$$-\ln(P) \sim \text{Exp}\,(1) = \frac{1}{2}\,\chi^2_2.$$

Thus, under $\theta = 0$,

$$-\ln(P^{(1)}\,P^{(2)}) \sim \frac{1}{2}\,\chi^2_4.$$

Combining the two P-values in an overall test, we reject H_0 if

$$-\ln(P^{(1)}\,P^{(2)}) > \frac{1}{2}\,\chi^2_{4,\,1-\alpha}.$$

If $\theta < 0$, then $P^{(1)}$ and $P^{(2)}$ are stochastically larger than $U(0,1)$ random variables and the type I error rate is less than α.

This χ^2 test was originally proposed for combining results of several studies by R. A. Fisher in 1932.

The inverse normal combination test

Initial design

Specify the **inverse normal test** for null hypothesis H_0, with weights w_1 and w_2 where $w_1^2 + w_2^2 = 1$.

Design Stage 1, fixing sample size and test statistic.

Stage 1

Observe the one-sided P-value, $P^{(1)}$, based on Stage 1 data.

Compute $Z^{(1)} = \Phi^{-1}(1 - P^{(1)})$.

Design Stage 2 in the light of Stage 1 data.

Stage 2

Observe the P-value, $P^{(2)}$, based **only** on Stage 2 data.

Compute $Z^{(2)} = \Phi^{-1}(1 - P^{(2)})$.

NB Under $\theta = 0$, $Z^{(1)} \sim N(0,1)$, $Z^{(2)} \sim N(0,1)$, independent.

The inverse normal combination test

The combination test is based on the statistic $w_1 Z^{(1)} + w_2 Z^{(2)}$.

Under $\theta = 0$, $Z^{(1)}$ and $Z^{(2)}$ are independent $N(0,1)$ so, with $w_1^2 + w_2^2 = 1$,

$$w_1 Z^{(1)} + w_2 Z^{(2)} \sim N(0,1).$$

Hence, for an overall one-sided test with type I error rate α, we reject H_0 if

$$w_1 Z^{(1)} + w_2 Z^{(2)} > \Phi^{-1}(1-\alpha).$$

If $\theta < 0$, then $Z^{(1)}$ and $Z^{(2)}$ are stochastically smaller than $N(0,1)$ random variables and the type I error rate is less than α.

An attractive property

If w_1 and w_2 are proportional to the square roots of the Stage 1 and Stage 2 sample sizes then $w_1 Z^{(1)} + w_2 Z^{(2)}$ is the standard Z-statistic based on the data at the end of Stage 2.

A combination test with more than two groups

The combination test approach can be applied in a group sequential framework (Lehmacher & Wassmer, *Biometrics*, 1999).

For $k = 1, \ldots, K$, define $Z^{(k)}$ to be the standardised test statistic based on Stage k data alone, and define cumulative Z-statistics

$$Z_k \;=\; (w_1 Z^{(1)} + \ldots + w_k Z^{(k)}) \,/\, (w_1^2 + \ldots + w_k^2)^{1/2}.$$

Under $\theta = 0$, the sequence $\{Z_k\}$ follows the canonical joint distribution with

$$\text{Cov}(Z_{k_1}, Z_{k_2}) \;=\; \sqrt{\frac{w_1^2 + \ldots + w_{k_1}^2}{w_1^2 + \ldots + w_{k_2}^2}} \quad \text{for } \; k_1 < k_2.$$

At analysis k, we compare Z_k with critical values a_k and b_k that define a group sequential test with type I error probability α.

Then, under $\theta = 0$, the probability of rejecting H_0 is exactly α.

4.3 Sample size re-estimation

A combination test can be used to protect the type I error rate when a trial's sample size is changed.

Consider a two-treatment comparison in which observations on Treatments A and B, respectively, are distributed as

$$X_{Ai} \sim N(\mu_A, \sigma^2) \quad \text{and} \quad X_{Bi} \sim N(\mu_B, \sigma^2).$$

Objective

It is desired to test H_0: $\theta = \mu_A - \mu_B \leq 0$ against $\theta > 0$ with type I error rate α and power $1 - \beta$ at $\theta = \delta$.

In the case of known variance, the sample size formula

$$n = \frac{\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2 \, 2\,\sigma^2}{\delta^2} \tag{1}$$

gives the required value of n, the sample size per treatment.

However, in practice, only an estimate of σ^2 is usually available.

Sample size re-estimation for a response variance

We can follow an adaptive approach, using an estimate of σ^2 from early trial data to modify the initial choice of sample size.

Initial design

Specify a two-stage adaptive design, using the inverse χ^2 combination rule to test H_0: $\theta \leq 0$ against $\theta > 0$.

Use an initial estimate σ_0^2 in the sample size formula (1) to obtain a sample size of n_0 per treatment.

Stage 1

Conduct Stage 1 with $n_1 = n_0/2$ subjects per treatment.

Observe estimates $\widehat{\theta}_1$, $\hat{\sigma}_1^2$ and the t-statistic t_1 for testing H_0.

Convert t_1 to a one-sided P-value, $P^{(1)} = P_{\theta=0}\{T_{2n_1-2} > t_1\}$.

Sample size re-estimation for a response variance

After Stage 1

Calculate a new Stage 2 sample size of n_2 per treatment arm.

Here, n_2 may be obtained simply by using the new variance estimate $\hat{\sigma}_1^2$ in the original sample size formula.

Or, n_2 might be chosen to give conditional power $1-\beta$ given $P^{(1)}$, assuming $\theta = \widehat{\theta}_1$ and $\sigma^2 = \hat{\sigma}_1^2$.

Stage 2

Calculate the t-statistic t_2 for testing H_0 based on Stage 2 data alone, and obtain the P-value $P^{(2)} = P_{\theta=0}\{T_{2n_2-2} > t_2\}$.

The inverse χ^2 combination test, which rejects H_0 if

$$-\ln(P^{(1)}\,P^{(2)}) > \frac{1}{2}\,\chi^2_{4,\,1-\alpha}$$

has type I error rate exactly equal to α.

Sample size re-estimation for a response variance

The above approach adds to the variety of methods for dealing with an unknown parameter, ϕ, that affects sample size.

Internal pilot:

Wittes & Brittain (*Statistics in Medicine*, 1990) proposed a simple "plug in" of the current estimate $\hat{\phi}$ to update sample size. Bias in the final $\hat{\phi}$ tends to cause a small inflation of the type I error rate.

Unblinded variance estimation:

Friede & Miller (*Applied Statistics*, 2012) show that, for normally distributed data, sample size modification based on an unblinded estimate of σ^2 leads to almost zero type I error rate inflation.

Information monitoring:

Mehta & Tsiatis (*Drug Information Journal*, 2001) "plug in" estimated information in an error spending group sequential design. Typically, this leads to a small inflation of the type I error rate.

Sample size re-estimation for σ^2 in a GST

A Lehmacher-Wassmer K-stage group sequential design:

Weights $w_1, \ldots, w_K$ are specified, where $w_1^2 + \ldots + w_K^2 = 1$.

Let t_k be the t-statistic, on ν_k degrees of freedom, testing H_0: $\theta \leq 0$ vs $\theta > 0$ based on responses in group k alone.

We compute the P-value $P^{(k)} = P_{\theta=0}\{T_{\nu_k} > t_k\}$ and define

$$Z^{(k)} \;=\; \Phi^{-1}(1 - P^{(k)}).$$

At analysis k, we compare the cumulative statistic

$$Z_k \;=\; (w_1 Z^{(1)} + \ldots + w_k Z^{(k)}) \,/\, (w_1^2 + \ldots + w_k^2)^{1/2},$$

with pre-specified critical values a_k and b_k that define a group sequential test with type I error probability α.

Under $\theta = 0$, $P^{(k)} \sim U(0, 1)$ and $Z^{(k)} \sim N(0,\, 1)$, so this group sequential combination test controls the type I error rate exactly.

An example of sample size modification

Consider a clinical trial studying treatment of heart failure.

Failure rates on control and experimental treatments are p_c and p_t.

From historical data, $p_c \approx 0.25$ and a reduction of 0.05 is desired.

Writing $\theta = p_c - p_t$, it is desired to test H_0: $\theta \leq 0$ against $\theta > 0$ with type I error rate $\alpha = 0.025$ and power $1 - \beta = 0.9$ if $\theta = 0.05$.

Initial design

A Bauer & Köhne two-stage design is specified.

For the above type I error rate and power, assuming $p_c = 0.25$, a fixed sample test needs 1461 subjects per treatment arm.

Stage 1 is planned with 730 subjects per treatment, with a view to re-assessing requirements for the remainder of the study in the light of their responses.

Example: Sample size modification

Stage 1 with 730 subjects per treatment

We observe $\hat{p}_c^{(1)} = 0.253$ and $\hat{p}_t^{(1)} = 0.219$, so $\widehat{\theta}_1 = 0.034$ with standard error 0.0222.

A test of H_0: $\theta \leq 0$ has $Z^{(1)} = 0.034/0.0222 = 1.53$ and P-value

$$P^{(1)} = 1 - \Phi(1.53) = 0.0629.$$

The final test will reject H_0 if $-\ln(P^{(1)}P^{(2)}) > 0.5\chi^2_{4,0.975} = 5.57$.

Since $-\ln(0.0629) = 2.77$, results thus far are promising — but a positive outcome is by no means certain.

It is learnt that the trial of a competing treatment is unsuccessful.

It is decided to increase the second stage sample size to give a higher probability of a positive outcome under the original alternative, $\theta = 0.05$ — and under smaller effect sizes.

Example: Sample size modification

Planning Stage 2 sample size

p_c	p_t	θ	Stage 2 sample size	Conditional power
0.25	0.22	0.03	750	0.43
			1000	0.51
			1250	0.59
0.25	0.21	0.04	750	0.62
			1000	0.72
			1250	0.80
0.25	0.20	0.05	750	0.78
			1000	0.87
			1250	0.93

Example: Sample size modification

Stage 2

Suppose we decide on 1000 patients per treatment arm in Stage 2.

Stage 2 data (alone) yield $\hat{p}_c^{(2)} = 0.251$ and $\hat{p}_t^{(2)} = 0.221$.

Hence, the Stage 2 data give $\widehat{\theta}_2 = 0.030$ with standard error 0.019.

The test of H_0: $\theta \leq 0$ based on Stage 2 data alone has $Z = 0.030/0.019 = 1.58$ and the P-value is

$$P^{(2)} = 1 - \Phi(1.58) = 0.0570.$$

In the overall test,

$$-\ln(P^{(1)}\, P^{(2)}) = -\ln(0.0629) - \ln(0.0570) = 2.77 + 2.87 = 5.64.$$

This is greater than $\frac{1}{2}\,\chi^2_{4,\,0.975} = 5.57$, so H_0 is rejected and it is concluded that the new treatment has a lower failure rate.

Sample size re-estimation in response to $\widehat{\theta}$

In the early 2000s, the possibility of adaptive design prompted interest in procedures that increase sample size in response to a low interim estimate of the treatment effect — while protecting the type I error probability.

The objective here is to increase power, recognising that the effect size used in the original power calculation was over-optimistic.

The resulting procedures have an overall maximum possible sample size but, depending on the observed data, the actual sample size can be smaller than this — just like a GST.

JT have compared the properties of adaptive designs and GSTs (*Statistics in Medicine*, 2003 and 2006, *Biometrika*, 2006).

They conclude that familiar group sequential designs can provide almost all the available efficiency gains — whereas some proposals for adaptive designs can be quite inefficient.

Sample size re-estimation in response to $\widehat{\theta}$

Mehta & Pocock (MP, *Statistics in Medicine*, 2011) proposed their "promising zone" trial design, which has the option of adding subjects to increase the (conditional) power for a range of $\widehat{\theta}$ values.

MP describe a trial in which the response is measured 26 weeks after the start of treatment and a large proportion of the total sample will be treated but not yet observed at the interim analysis.

The presence of such "pipeline" subjects imposes a minimum for the total sample size in a design with sample size re-estimation.

Most group sequential tests are designed for the case of an immediate response, with just a few suggestions made for incorporating data from pipeline subjects after a trial is stopped.

However, the Delayed Response GSTs proposed by Hampson & Jennison (*JRSS, B*, 2013) provide a way to handle pipeline data.

Sample size re-estimation in response to $\widehat{\theta}$

JT (*Statistics in Medicine*, 2015) discuss Mehta & Pocock's designs and suggest ways to improve their efficiency.

JT define sample size rules that optimize certain sample size and power criteria — achieving the same overall power as MP designs with lower expected sample size.

The designs proposed by JT have smaller increases in sample size occurring over a wider range than MP's original "promising zone".

JT's developments within the MP framework lead, ultimately, to the adaptive version of Hampson and Jennison's Delayed Response GSTs — so we see a convergence of the two approaches.

If fixed group sizes are desired, one can use Hampson and Jennison's non-adaptive delayed response GSTs and their efficiency is still close to the best fully adaptive designs.

4.4 Adaptive trial designs to test multiple hypotheses

There may be changes to key elements during the course of a trial:

Change of treatment definition,

Change of endpoint,

Switching from a test of superiority to a test of non-inferiority.

Or, a trial may proceed in stages with design choices being made at interim analyses

Seamless Phase 2/Phase 3 trial: *Treatment selection, followed by a confirmatory stage*

Enrichment trial: *Possible restriction to patients in a pre-specified sub-group*

Multi-arm multi-stage trial: *Several treatments compared to a control with elimination of poorly performing treatments.*

Adaptive trial designs to test multiple hypotheses

In some of the preceding examples, several null hypotheses may be tested at the end of the trial.

In other cases, a single null hypothesis will be tested but the choice of this hypothesis is data dependent — so care is needed to avoid introducing a **selection bias.**

We shall apply multiple testing methods that guarantee protection of the overall type I error rate.

These multiple testing methods will be used in conjunction with combination tests that merge the two sets of data either side of the interim analysis at which an adaptive decision is taken.

Selected references:

Bretz, Schmidli et al. (*Biometrical Journal*, 2006),

Schmidli, Bretz et al. (*Biometrical Journal*, 2006),

Jennison & Turnbull (*J. Biopharmaceutical Statistics*, 2007).

Procedures for testing multiple hypotheses

The familywise error rate

Suppose we have h null hypotheses, H_i: $\theta_i \leq 0$ for $i = 1, \ldots, h$.

A procedure's **familywise error rate** when $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_h)$ is

$$P_{\boldsymbol{\theta}}\{\text{Reject } H_i \text{ for some } i \text{ with } \theta_i \leq 0\}.$$

The familywise error rate is controlled **strongly** at level α if this error rate is at most α for all possible combinations of θ_i values.

Then

$$P_{\boldsymbol{\theta}}\{\text{Reject any true } H_i\} \leq \alpha \quad \text{for all } \boldsymbol{\theta} = (\theta_1, \ldots, \theta_h).$$

Using such a procedure, the probability of choosing to focus on a parameter θ_{i^*} and then falsely claiming significance for the associated null hypothesis H_{i^*} is at most α.

4.5 Closed testing procedures

Marcus et al. (*Biometrika*, 1976) introduced a **closed testing procedure** which provides strong control of FWER by combining level α tests of each H_i and of intersections of these hypotheses.

Suppose we have null hypotheses H_i, $i = 1, \ldots, h$.

For each subset I of $\{1, \ldots, h\}$, define the intersection hypothesis

$$H_I = \cap_{i \in I} H_i.$$

Construct a level α test of each intersection hypothesis H_I, i.e., a test which rejects H_I with probability at most α whenever all hypotheses specified in H_I are true.

Closed testing procedure

The simple hypothesis H_j: $\theta_j \leq 0$ is rejected overall if, and only if, H_I is rejected for every set I containing index j.

Closed testing procedures

Proof of strong control of familywise error rate

In the closed testing procedure, overall rejection of the simple hypothesis H_j can only occur if H_I is rejected for every set I containing index j.

Let $\tilde{I}$ be the set of indices of all true hypotheses H_i.

Since $H_{\tilde{I}}$ is true, $P\{\text{Reject } H_{\tilde{I}}\} = \alpha$.

For a familywise error to be committed, $H_{\tilde{I}}$ must be rejected.

Hence, the probability of a familywise error is no greater than α.

Testing an intersection hypothesis

Suppose the intersection hypothesis $H_I = \cap_{i \in I} H_i$ is the intersection of m simple hypotheses.

For each $i \in I$, let P_i be the 1-sided P-value for testing H_i.

Denote the ordered values of the P_i by $P_{[1]} \leq P_{[2]} \leq \ldots \leq P_{[m]}$.

There are several ways to test an intersection hypothesis.

Bonferroni adjustment

The overall P-value for testing H_I is $P_I = m\, P_{[1]}$.

Simes' method (Biometrika, 1986)

The overall P-value for H_I is

$$P_I = \min_{k=1,\ldots,m} (m\, P_{[k]}/k).$$

Bonferroni and Simes' methods

The Bonferroni adjustment is simple, but conservative.

In the definition of Simes' P-value,

$$P_I = \min_{k=1,\ldots,m} (m\, P_{[k]}/k),$$

the term for $k = 1$ is $mP_{[1]}$, i.e., the Bonferroni adjusted P-value.

Other low P-values can reduce the overall result, e.g., if $P_{[2]}$ is only a little higher than $P_{[1]}$ so $P_{[2]}/2 < P_{[1]}$, then this will reduce P_I.

The Simes method is valid — and still slightly conservative — when the P_i are independent or positively dependent.

Such positive dependence arises in a comparison of m treatments with a common control or in tests for a treatment effect in overlapping sub-populations.

Dunnett's method (JASA, 1955)

Suppose m treatments are compared with a control, responses are normal with known variance, and sample sizes on each treatment and the control are equal.

Each null hypothesis H_i says treatment i is no better than control.

We are to test the intersection hypothesis $H_I = \cap_{i \in I} H_i$.

Denote the Z-statistic arising from the test of H_i by Z_i.

When each treatment effect for an $H_i \in H_I$ is zero,

$$Z_i \sim N(0,1), \quad i \in I, \quad \text{Cov}(Z_i, Z_{i'}) = 0.5, \quad i \neq i'.$$

The P-value for testing H_I using Dunnett's test is

$$P\{\max_{i \in I} Z_i > z^*\},$$

where z^* is the observed value of $\max_{i \in I} Z_i$, and the probability is under the above multivariate normal distribution for $\{Z_i,\, i \in I\}$.

4.6 Using combination tests in a closed testing procedure

Suppose we have null hypotheses H_i, $i = 1, \ldots, h$, and wish to test these in a closed testing procedure with FWER α.

We need to define a level α test for each intersection hypothesis

$$H_I = \cap_{i \in I} H_i$$

— a simple hypothesis H_j is a special case where $I = \{j\}$.

In an adaptive trial with two stages, we can test each H_I by applying a combination test across the stages.

Each stage provides a P-value for H_I and we combine these by a pre-specified method, e.g., "an inverse normal combination test with equal weights".

The P-value for the second stage must be defined before that stage commences, but it may depend on first stage responses.

Adaptive trial designs with multiple hypotheses

The approach of using combination tests within a closed testing procedure is widely applicable.

It provides a way to run trials with the different types of adaptation listed earlier:

Change of treatment definition,

Change of endpoint,

Testing either superiority or non-inferiority,

Seamless Phase 2/Phase 3 trials,

Enrichment trials,

Multi-arm multi-stage trial.

We shall conclude by studying an example of a Phase 3 trial incorporating adaptive treatment selection.

Part 5. Adaptive clinical trial designs (2)

5.1. Enrichment designs

Combination tests and Closed Testing Procedures

Examples

Assessing the benefits of enrichment designs

5.2. Seamless Phase 2/3 designs

Combination tests and Closed Testing Procedures

Examples

Efficiency gains from combining Phases 2 and 3

5.1 Enrichment designs

Consider a new treatment which is expected to be particularly effective in an identified sub-group of the target population.

An "enrichment design" aims to identify differential treatment effects in patient subgroups and adapt the trial's focus to patients for whom the benefit is greatest.

Results may support a licence for the full population or just the sub-population.

"Adaptive signature designs" (Friedlin & Simon, *Clin. Cancer Res.*, 2005) aim to characterise a sub-population and demonstrate a treatment effect in this sub-population in a **single** trial.

However, it is more common to commence a Phase III trial with a clearly defined sub-population and with validated assays to determine whether a patient is a member of this sub-population.

Enrichment designs: A single sub-population

We shall consider clinical trials with enrichment in which we:

Start by comparing the new treatment against control in the full population.

Examine responses at an interim stage.

If there is no evidence of treatment effect, stop for futility.

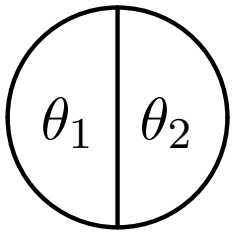
If the new treatment appears effective in the full population, continue as before.

If the new treatment appears to benefit just the subgroup, recruit only from the subgroup and increase the numbers in this subgroup.

On conclusion of the trial, test null hypotheses concerning the full population or the sub-population, as appropriate.

Enrichment designs: A single sub-population

Sub-population
(proportion λ_1)



Rest of the population
(proportion λ_2)

Denote the treatment effect:

In the sub-population by θ_1,

In the complement of the sub-population by θ_2,

Aggregated over the whole population by $\theta_3 = \lambda_1\theta_1 + \lambda_2\theta_2$.

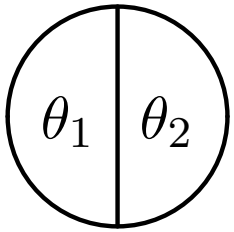
The null hypothesis for the sub-population is H_1: $\theta_1 \leq 0$.

The null hypothesis for the full target population is H_3: $\theta_3 \leq 0$.

All tests are against one-sided alternatives, $\theta_1 > 0$ or $\theta_3 > 0$.

Enrichment designs: A single sub-population

H_1: No effect in the sub-population



H_3: No overall effect in the full population

Stage 1

P-value for H_1 from the sub-population data only is $P_1^{(1)}$.

P-value for H_3 from the full data is $P_3^{(1)}$.

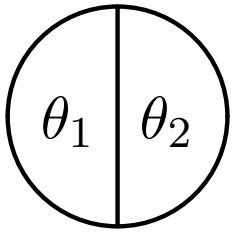
We use Simes' method to obtain a P-value from Stage 1 data for $H_{13} = H_1 \cap H_3$

$$P_{13}^{(1)} = \min\{2\min(P_1^{(1)}, P_3^{(1)}), \max(P_1^{(1)}, P_3^{(1)})\}.$$

(Simes' method is conservative here since $P_1^{(1)}$ and $P_3^{(1)}$ are positively correlated through their overlapping sets of subjects.)

Enrichment designs: A single sub-population

H_1: No effect in the sub-population



H_3: No overall effect in the full population

Case A: Stage 2 continuing with the full population

P-value for H_1 from the sub-population data only is $P_1^{(2)}$.

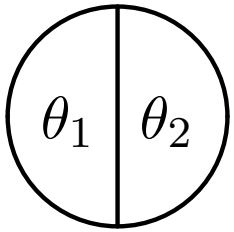
P-value for H_3 from the full Stage 2 data is $P_3^{(2)}$.

Simes' method gives the P-value for H_{13} from the Stage 2 data

$$P_{13}^{(2)} = \min\{2\min(P_1^{(2)}, P_3^{(2)}), \max(P_1^{(2)}, P_3^{(2)})\}.$$

Enrichment designs: A single sub-population

H_1: No effect in the sub-population



H_3: No overall effect in the full population

Case B: Stage 2 restricted to the sub-population

P-value for H_1 from the sub-population data is $P_1^{(2)}$.

No P-value is available for testing H_3 in Stage 2

P-value for H_{13} from Stage 2 data is simply $P_{13}^{(2)} = P_1^{(2)}$.

Enrichment designs: Combining stages

A rule must be pre-specified for combining P-values from the two stages to give tests of H_1, H_3 and H_{13}.

If a certain null hypothesis has P-values $p^{(1)}$ and $p^{(2)}$ in Stages 1 and 2, a weighted inverse normal rule gives

$$Z(p^{(1)}, p^{(2)}) = w_1 \Phi^{-1}(1 - p^{(1)}) + w_2 \Phi^{-1}(1 - p^{(2)}).$$

Here, the weights w_1 and w_2 should reflect the relative sample sizes planned for Stage 1 and Stage 2 and satisfy $w_1^2 + w_2^2 = 1$.

Then, $Z \sim N(0, 1)$ when $P^{(1)}$ and $P^{(2)}$ are independent $U(0, 1)$, and the null hypothesis is rejected at level α if $Z > \Phi^{-1}(1 - \alpha)$.

Global tests of H_1 and H_3 with familywise error rate α are formed from "local" level-α tests of H_1, H_3 and H_{13}, each combining P-values from the two stages.

Enrichment designs: Global tests

Case A: Stage 2 continuing with the full population

Global test of H_1 (sub-population effect)

H_1 is rejected at global level α if both H_1 and H_{13} are rejected in combination tests for Stage 1 and Stage 2 data.

NB This is a stronger requirement than rejection of just the elementary hypothesis H_1 due to testing multiple hypotheses.

Global test of H_3 (full population effect)

H_3 is rejected at global level α if both H_3 and H_{13} are rejected (a stronger requirement than rejection of just H_3).

When continuing with the full population, it is possible to reject H_1 but not H_3, finding a treatment effect in the sub-population only.

Enrichment designs: Global tests

Case B: Stage 2 restricted to the sub-population

Global test of H_1 (sub-population effect)

H_1 is rejected at global level α if both H_1 and H_{13} are rejected in combination tests for Stage 1 and Stage 2 data.

No Stage 2 P-value is available for H_3 and an overall test of H_3 is not possible.

This is in keeping with the decision to enrich and focus on just the sub-population.

Example 1 (JT, *J. Biopharmaceutical Statistics*, 2007)

A trial is conducted to compare a new treatment against control in a general population (null hypothesis H_3).

The new treatment is expected to be particularly effective in men over 50 years old (null hypothesis H_1).

In testing H_1 and H_3, the familywise type I error rate is set at $\alpha = 0.025$.

After Stage 1 of the trial, a decision will be made whether to enrich and recruit only older men in the remainder of the study.

A Closed Testing Procedure will be followed, using Simes' test for the intersection hypothesis.

For each hypothesis and the intersection hypothesis H_{13}, Stage 1 and Stage 2 data will be combined in an inverse normal test with weights $w_1 = w_2 = 1/\sqrt{2}$.

Example 1, continued

Stage 1 results

Older men: $P_1^{(1)} = 0.02$

Full population: $P_3^{(1)} = 0.20$

Using Simes' method, the P-value for testing $H_{13} = H_1 \cap H_3$ from Stage 1 data is

$$P_{13}^{(1)} = \min\{2\min(P_1^{(1)}, P_3^{(1)}), \max(P_1^{(1)}, P_3^{(1)})\} = 0.04.$$

We decide not to enrich and recruitment continues from the full population in Stage 2.

Example 1, continued

Stage 2 results

Older men: $P_1^{(2)} = 0.03$

Full population: $P_3^{(2)} = 0.10$

By Simes' method: $P_{13}^{(2)} = 2 \times 0.03 = 0.06.$

Combining results from the two stages

$$Z_1 = \sqrt{0.5}\,\Phi^{-1}(1-0.02) + \sqrt{0.5}\,\Phi^{-1}(1-0.03) = 2.78,$$

$$Z_3 = \sqrt{0.5}\,\Phi^{-1}(1-0.20) + \sqrt{0.5}\,\Phi^{-1}(1-0.10) = 1.50,$$

$$Z_{13} = \sqrt{0.5}\,\Phi^{-1}(1-0.04) + \sqrt{0.5}\,\Phi^{-1}(1-0.06) = 2.34,$$

$$P_1 = 1-\Phi(2.78) = 0.003,$$

$$P_3 = 1-\Phi(1.50) = 0.067,$$

$$P_{13} = 1-\Phi(2.34) = 0.010.$$

Example 1, continued

We have:

$$P_1 = 0.003, \quad P_3 = 0.0673, \quad P_{13} = 0.010.$$

Thus, H_1 and H_{13} are rejected in **local** tests at level $\alpha = 0.025$.

Rejection of H_1 and H_{13} in local level-α tests implies $\boldsymbol{H_1}$ **is rejected globally** — in a Closed Testing Procedure with familywise error probability 0.025.

The adjusted P-value for H_1 is $\tilde{P}_1 = \max\{P_1, P_{13}\} = 0.010$.

The local test of H_3 has P-value 0.067. Since this is greater than $\alpha = 0.025$, $\boldsymbol{H_3}$ **is not rejected globally**.

Example 1b

Consider a realisation of the trial where Stage 1 data are as before, but it is decided to restrict recruitment to "Older men" in Stage 2.

Stage 1 results (as before)

Older men: $P_1^{(1)} = 0.02$

Full population: $P_3^{(1)} = 0.20$

By Simes' method: $P_{13}^{(1)} = 2 \times 0.02 = 0.04.$

Stage 2 results

Older men: $P_1^{(2)} = 0.03$

Full population: $P_3^{(2)}$ is undefined

By Simes' method: $P_{13}^{(2)} = P_1^{(2)} = 0.03$

Example 1b, continued

Combining results from the two stages

$$Z_1 = \sqrt{0.5}\,\Phi^{-1}(1-0.02) + \sqrt{0.5}\,\Phi^{-1}(1-0.03) = 2.78,$$
$$Z_{13} = \sqrt{0.5}\,\Phi^{-1}(1-0.04) + \sqrt{0.5}\,\Phi^{-1}(1-0.03) = 2.34,$$

$$P_1 = 1-\Phi(2.78) = 0.003,$$
$$P_{13} = 1-\Phi(2.57) = 0.005.$$

$\boldsymbol{H_1}$ **is rejected overall** since both H_1 and H_{13} are rejected in local tests at $\alpha = 0.025$.

The adjusted P-value for H_1 is $\tilde{P}_1 = \max\{P_1, P_{13}\} = 0.005$.

As there is no Stage 2 P-value $P_3^{(2)}$, we do not have a combination test for H_3. Hence, there is no global test of H_3.

Enrichment designs: More sub-populations

The preceding method extends to 2-stage trials with more designated sub-populations. To do this, we must specify:

- The full population and all sub-populations to be considered,
- The rule for testing an intersection hypothesis with data from a single stage of the trial,
- The rule for combining results from two stages of the trial to give an overall hypothesis test.

The study commences with recruitment from the full population.

At the assigned re-design point, interim data are studied and a decision is taken:

- To continue recruiting from the full population, or to restrict recruitment to specific sub-populations.
- To select a suitable sample for Stage 2 size in either case.

Example 2: (JT, *J. Biopharm. Statist.*, 2007)

In our example, suppose there are specified sub-populations:

1. The entire population
2. Men only
3. Men over 50
4. Men who are smokers

Let $\theta_1, \ldots, \theta_4$ denote the treatment effects within the four specified sub-populations.

The elementary null hypotheses are H_i: $\theta_i \leq 0$ for $i = 1, \ldots, 4$.

In order to protect the familywise error rate at level α, we test these hypotheses using a Closed Testing Procedure.

For each individual hypothesis and intersection hypothesis, we combine P-values from the two stages by the inverse normal rule.

Example 2: Several sub-populations

We have elementary null hypotheses H_i: $\theta_i \leq 0$ for $i = 1, \ldots, 4$.

Within a stage, each intersection hypothesis will be tested by combining P-values for individual hypotheses using Simes' method.

In Stage 1

We test each H_i against $\theta_i > 0$ using the estimate $\widehat{\theta}_i$ from subjects in the relevant sub-population, giving P-value $P_i^{(1)}$.

In Stage 2

It may only be possible to test some of the H_i using Stage 2 data.

For example, if recruitment is restricted to "men only", we can test H_2, H_3 and H_4 but not H_1, since θ_1 is a weighted average of effects on both men and women.

Thus, we obtain Stage 2 P-values $P_2^{(2)}$, $P_3^{(2)}$ and $P_4^{(2)}$ for hypotheses H_2, H_3 and H_4 but we have no $P_1^{(2)}$ for testing H_1.

Example 2: Closed Testing Procedure

Formally:

In order to reject H_{i^*}: $\theta_{i^*} \leq 0$, we need to reject each intersection hypothesis H_I with $i^* \in I$ at level α, based on combined Stage 1 and Stage 2 data.

Here, $H_I = \cap_{i \in I} H_i$ states that $\theta_i \leq 0$ for all $i \in I$.

Intuitively:

Some sub-populations will have better than average results due to random variation in patient responses.

In order to avoid an inflated type I error rate, we must allow for the multiplicity of hypotheses being tested.

If $\theta_i = 0$ for $i = 1, \ldots, 4$, the sub-population with the most favorable results should be viewed as the *best out of four comparisons of an ineffective treatment with the control*, rather than typical results for a single, pre-specified comparison.

Example 2: Closed Testing Procedure

Testing an intersection hypothesis H_I: $\theta_i \leq 0$ for all $i \in I$

(a) We need to test the intersection hypothesis in each stage.

(b) We need to combine data from two stages.

Task (b):

A weighted inverse normal combination test is specified.

Letting $P_I^{(1)}$ and $P_I^{(2)}$ denote P-values for testing H_I from Stage 1 and 2 data respectively, we calculate

$$Z(P_I^{(1)}, P_I^{(2)}) = w_1 \Phi^{-1}(1 - P_I^{(1)}) + w_2 \Phi^{-1}(1 - P_I^{(2)}),$$

using the pre-specified w_1 and w_2.

Then, we reject H_I if

$$Z(P_I^{(1)}, P_I^{(2)}) > \Phi^{-1}(1 - \alpha).$$

Task (a): Testing an intersection hypothesis in each stage

Suppose m hypotheses are involved in $H_I = \cap_{i \in I} H_i$.

(a) Testing H_I in Stage 1

We calculate a P-value $P_i^{(1)}$ from Stage 1 data for each $H_i \in H_I$ and apply Simes' method to test the intersection hypothesis H_I.

With $P_{[i]}^{(1)}$, $i = 1, \ldots, m$, denoting the m P-values in increasing order, the P-value for testing H_I is

$$P_I^{(1)} = \min_{k=1,\ldots,m} (m\, P_{[k]}^{(1)}/k).$$

(b) Testing H_I in Stage 2

Applying the same procedure to the Stage 2 P-values $P_i^{(2)}$ gives the Stage 2 P-value $P_I^{(2)}$ for H_I.

If some sub-populations are dropped in Stage 2, consider only those H_i for which a P-value $P_i^{(2)}$ is available and reduce m accordingly.

Testing an intersection hypothesis in each stage

Testing H_I in Stage 1

Suppose

$$P_1^{(1)} = 0.2, \quad P_2^{(1)} = 0.04, \quad P_3^{(1)} = 0.05, \quad P_4^{(1)} = 0.03.$$

In the global test of H_{i^*}, we consider all sets I containing i^*.

So, for $i^* = 4$, we need P-values for H_I with

$$I = \{4\}, \quad I = \{1,4\}, \quad I = \{2,4\}, \quad I = \{3,4\}, \quad I = \{1,2,4\},$$
$$I = \{1,3,4\}, \quad I = \{2,3,4\}, \quad I = \{1,2,3,4\}.$$

For the case $I = \{1,3,4\}$, the ordered P-values are

$$P_{[1]}^{(1)} = 0.03, \quad P_{[2]}^{(1)} = 0.05, \quad P_{[3]}^{(1)} = 0.2$$

and Simes' test gives

$$P_I^{(1)} = \min_{k=1,\ldots,3} (3\, P_{[k]}^{(1)}/k) = 3 \times 0.05/2 = 0.075.$$

Testing an intersection hypothesis in each stage

Testing H_I in Stage 2

In Stage 2, we have P-values, $P_i^{(2)}$, for *some* of the H_i: $\theta_i \leq 0$, depending on the section of the population from which recruitment took place in this stage.

Let $\tilde{I}$ be the set of indices $i \in I$ for which we have a P-value $P_i^{(2)}$ and suppose there are $\widetilde{m}$ such indices.

We can apply Simes' method on the reduced set $\tilde{I}$, as long as it is non-empty, yielding the P-value for testing H_I

$$P_I^{(2)} = \min_{k=1,\ldots,\widetilde{m}} (\widetilde{m}\, P_{[k]}^{(2)}/k),$$

where $P_{[k]}^{(2)}$, $k = 1, \ldots, \widetilde{m}$, are the $\widetilde{m}$ available P-values in increasing order.

Note that if the set $\tilde{I}$ is empty, its P-value will not be needed.

Restrictions on testing sub-populations in Stage 2

If recruitment continues from the full population in Stage 2, a P-value can be calculated for each sub-population, so all combination tests are feasible.

If recruitment is restricted, elementary tests are only possible for sub-populations contained completely in the new recruitment pool.

If, in our example, recruitment is restricted to "Men only", the sub-populations "Men over 50" and "Men who are smokers" are still sampled fully so we can test H_3 and H_4 as well as H_2.

However, we cannot test H_1 since θ_1 is a weighted average of effects on both men and women and Stage 2 provides no information about women.

Consequently, we can test H_2, H_3 and H_4 at the global level.

As an example, in testing H_2, the relevant sets I all contain $i = 2$, so there is at least one element $(i = 2)$ in the reduced set $\tilde{I}$.

Example 2: (JT, *J. Biopharm. Statist.*, 2007)

Suppose we observe

Stage 1 results

Full population:	$P_1^{(1)} = 0.20$
All men:	$P_2^{(1)} = 0.10$
Men over 50 years:	$P_3^{(1)} = 0.03$
Men who smoke:	$P_4^{(1)} = 0.03.$

After restricting recruitment in Stage 2 to "Men only", we obtain

Stage 2 results

All men:	$P_2^{(2)} = 0.11$
Men over 50 years:	$P_3^{(2)} = 0.08$
Men who smoke:	$P_4^{(2)} = 0.04.$

Example 2, continued

Suppose, at the end of the trial, we have the stage-wise P-values $P_i^{(1)}$ and $P_i^{(2)}$ for individual hypotheses H_i, $i = 1, \ldots, 4$:

		Stage 1	Stage 2
H_1:	Entire population	0.20	—
H_2:	All men	0.10	0.11
H_3:	Men over 50 years	0.03	0.08
H_4:	Men who smoke	0.03	0.03

We wish to combine these P-values to test H_2, H_3 and H_4, while protecting the familywise error rate at level $\alpha = 0.025$.

We first calculate stage-wise P-values, $P_I^{(1)}$ and $P_I^{(2)}$ for each intersection hypothesis H_I.

We then apply a weighted inverse normal combination test to each pair $P_I^{(1)}$ and $P_I^{(2)}$ to give an overall P-value for H_I.

Finally, we apply the Closed Testing Procedure to these P-values.

Example 2: P-values for intersection hypotheses

	P-values		Combined	
	Stage 1	Stage 2	statistics	
H_I	$P_I^{(1)}$	$P_I^{(2)}$	Z_I	P_I
$H_{\{1\}}$	0.20	—	—	—
$H_{\{2\}}$	0.10	0.11	1.77	0.038
$H_{\{3\}}$	0.03	0.08	2.32	0.010
$H_{\{4\}}$	0.03	0.03	2.66	0.004
$H_{\{1,2\}}$	0.20	0.11*	1.46	0.072
$H_{\{1,3\}}$	0.06	0.08*	2.09	0.018
$H_{\{1,4\}}$	0.06	0.03*	2.43	0.008
$H_{\{2,3\}}$	0.06	0.11	1.97	0.025
$H_{\{2,4\}}$	0.06	0.06	2.20	0.014
$H_{\{3,4\}}$	0.03	0.06	2.43	0.008
$H_{\{1,2,3\}}$	0.09	0.11*	1.82	0.035
$H_{\{1,2,4\}}$	0.09	0.06*	2.05	0.020
$H_{\{1,3,4\}}$	0.045	0.06*	2.30	0.011
$H_{\{2,3,4\}}$	0.045	0.09	2.15	0.016
$H_{\{1,2,3,4\}}$	0.06	0.09*	2.05	0.020

* Stage 2 P-value $P^{(2)}_{\{1\}\cup I}$ is set equal to $P_I^{(2)}$ for $I \subseteq \{2,3,4\}$.

Example 2, continued

In order to reject H_i at global level $\alpha = 0.025$, each intersection hypothesis H_I with $i \in I$ must be rejected at this level.

The results table shows we can reject H_4 at global significance level $\alpha = 0.025$.

The other individual hypotheses are not rejected globally.

The adjusted P-value $\widetilde{P}_i$ for testing H_i is the maximum combined P-value P_I over all H_I with $i \in I$.

H_2: All men

$$\widetilde{P}_2 = \max\{P_2, P_{12}, P_{23}, P_{24}, P_{123}, P_{124}, P_{234}, P_{1234}\} = 0.072$$

H_3: Men over 50 years

$$\widetilde{P}_3 = \max\{P_3, \ldots, P_{1234}\} = 0.035$$

H_4: Men who smoke

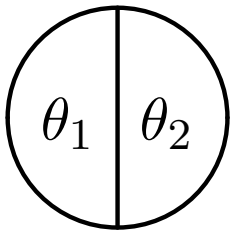
$$\widetilde{P}_4 = \max\{P_4, \ldots, P_{1234}\} = 0.020$$

So, we can reject H_4 and quote the adjusted P-value $\widetilde{P}_4 = 0.020$.

Assessing the benefits of an enrichment design

Consider a trial to investigate whether a new treatment is beneficial to the full population or, possibly, just a sub-population.

Sub-population (proportion λ_1)



Rest of the population (proportion λ_2)

The treatment effect is

θ_1 in the sub-population,

θ_2 in the complement of the sub-population,

θ_3 aggregated over the whole population by.

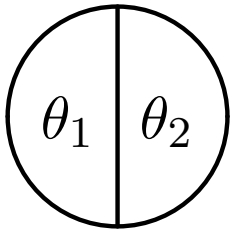
We may test

H_1: $\theta_1 \leq 0$ vs $\theta_1 > 0$ and/or

H_3: $\theta_1 \leq 0$ vs $\theta_1 > 0$.

Assessing the benefits of an enrichment design

Sub-population (proportion λ_1)



Rest of the population (proportion λ_2)

First, consider a design testing **just for a whole population effect**,

$$\theta_3 = \lambda_1\theta_1 + \lambda_2\theta_2.$$

The design has two analyses and one-sided type I error rate 0.025.

Sample size is set to achieve power 0.9 at $\theta_3 = 20$.

Data in each stage are summarised by a Z-value:

	Stage 1	*Stage 2*	*Overall*
H_3: $\theta_3 \leq 0$	$Z_3^{(1)}$	$Z_3^{(2)}$	$Z_3 = \frac{1}{\sqrt{2}} Z_3^{(1)} + \frac{1}{\sqrt{2}} Z_3^{(2)}$

Assessing the benefits of an enrichment design

A two stage design testing for a whole population effect θ_3 has

	Stage 1	*Stage 2*	*Overall*
$H_3\colon\ \theta_3 \leq 0$	$Z_3^{(1)}$	$Z_3^{(2)}$	$Z_3 = \frac{1}{\sqrt{2}} Z_3^{(1)} + \frac{1}{\sqrt{2}} Z_3^{(2)}$

Decision rules:

If $Z_3^{(1)} < 0$ Stop at Stage 1, Accept H_3

If $Z_3^{(1)} \geq 0$ Continue to Stage 2, then

- If $Z_3 < 1.96$ Accept H_3
- If $Z_3 \geq 1.96$ Reject H_3

Since the futility boundary can be regarded as non-binding, this design is slightly conservative.

Assessing the benefits of an enrichment design

Assume the sub-population comprises half the total population, so $\lambda_1 = \lambda_2 = 0.5$.

Properties of the design for the whole population effect, θ_3:

θ_1	θ_2	θ_3	*Power to reject* H_3: $\theta_3 \leq 0$
20	20	20	0.90
10	10	10	0.37
20	0	10	0.37

It is feasible to identify at Stage 1 that θ_1 may be quite high while a smaller value of θ_2 means that θ_3 is low.

In such a situation, we may choose to switch resources to test for a treatment effect in only the sub-population.

Assessing the benefits of an enrichment design

We wish to be able to consider two null hypotheses:

H_3: $\theta_3 \leq 0$ Treatment is not effective in the whole population,

H_1: $\theta_1 \leq 0$ Treatment is not effective in the sub-population.

Since $\theta_3 = 0.5\,\theta_1 + 0.5\,\theta_2$, either H_1 or H_3 may be true on its own.

In applying a Closed Testing Procedure, we also test the intersection hypothesis

$$H_{13}\text{: } \theta_1 \leq 0 \text{ and } \theta_3 \leq 0.$$

In order to reject H_1 globally in a Closed Testing Procedure with family-wise type I error rate α, we must reject both H_1 and H_{13} in local, level α tests.

Similarly, we reject H_3 overall if both H_3 and H_{13} are rejected in local, level α tests.

An adaptive enrichment design

At Stage 1, if $\widehat{\theta}_3^{(1)} < 0$, stop to accept H_3: $\theta_3 \leq 0$.

If $\widehat{\theta}_3^{(1)} > 0$ and the trial continues:

If $\widehat{\theta}_2^{(1)} < 0$ and $\widehat{\theta}_1^{(1)} > \widehat{\theta}_2^{(1)} + 8$,	Restrict to sub-population 1 and test H_1 only.
Otherwise,	Continue with full population and test H_1 and H_3.

Stage 2 sample size is the same in both cases, with numbers recruited from the sub-population increasing under enrichment.

Final decisions for global rejection of H_1 and H_3 follow the Closed Testing Procedure based on local, level α tests of H_1, H_3 and H_{13}.

The above rule for deciding whether to enrich is rather arbitrary: for a systematic approach to optimising the enrichment rule see Thomas Burnett's Bath PhD thesis (2017).

Local, level α tests for H_1, H_3 and H_{13}

We test each null hypothesis, H_1, H_3 and H_{13}, in a 2-stage group sequential test.

Let H_i be one of these hypotheses and $Z_i^{(1)}$ and $Z_i^{(2)}$ the Z-statistics for H_i based on data collected in Stages 1 and 2.

The 2-stage group sequential test rejects H_i after Stage 2 if

$$\frac{1}{\sqrt{2}} Z_i^{(1)} + \frac{1}{\sqrt{2}} Z_i^{(2)} \geq \Phi^{-1}(1-\alpha) = 1.96.$$

Defining $Z_i^{(1)}$ and $Z_i^{(2)}$

For H_1, we define $Z_1^{(1)}$ and $Z_1^{(2)}$ from estimates $\widehat{\theta}_1^{(1)}$ and $\widehat{\theta}_1^{(2)}$ of θ_1 based on data collected Stages 1 and 2, respectively.

For H_3, we define $Z_3^{(1)}$ and $Z_3^{(2)}$ from estimates $\widehat{\theta}_3^{(1)}$ and $\widehat{\theta}_3^{(2)}$ of θ_3 based on the data from each stage.

It remains to define $Z_{13}^{(1)}$ and $Z_{13}^{(2)}$ for testing H_{13}.

Local, level α tests for H_1, H_3 and H_{13}

Testing H_{13} in Stage 1

To maximise the probability of rejecting H_3 globally, we might set

$$Z_{13}^{(1)} = Z_3^{(1)}.$$

Alternatively, if our priority is to demonstrate an effect in the sub-population if only this effect is present, we may set

$$Z_{13}^{(1)} = Z_1^{(1)}.$$

A compromise is to combine these statistics as

$$\overline{Z}_{13}^{(1)} = (Z_3^{(1)} + Z_1^{(1)})/\sqrt{(2+\sqrt{2})}.$$

Here, the factor $1/\sqrt{(2+\sqrt{2})}$ makes the variance of $\overline{Z}_{13}^{(1)}$ equal to 1 — avoiding the conservatism of Simes' test.

Local, level α tests for H_1, H_3 and H_{13}

Testing H_{13} in Stage 2

If recruitment from the full population continues in Stage 2, we shall set

$$Z_{13}^{(2)} = Z_3^{(2)}.$$

If "enrichment" occurs and recruitment is restricted to the sub-population in Stage 2, we set

$$Z_{13}^{(2)} = Z_1^{(2)}.$$

Summary of Z-statistics for testing H_1, H_3 and H_{13}

When continuing with the full population:

	Stage 1	*Stage 2*
H_1	$Z_1^{(1)}$	$Z_1^{(2)}$
H_3	$Z_3^{(1)}$	$Z_3^{(2)}$
H_{13}	$\overline{Z}_{13}^{(1)}$	$Z_3^{(2)}$

When switching to the sub-population:

	Stage 1	*Stage 2*
H_1	$Z_1^{(1)}$	$Z_1^{(2)}$
H_{13}	$\overline{Z}_{13}^{(1)}$	$Z_1^{(2)}$

Note the common test statistic for H_{13} in Stage 1 — which must be pre-specified.

Results: Power of non-adaptive and adaptive designs

				Non-adapt$^{\mathbf{v}}$	**Adaptive**		
	θ_1	θ_2	θ_3	*Full pop^n*	*Sub-pop^n only*	*Full pop^n*	*Total*
1.	30	0	15	**0.68**	0.47	0.41	**0.88**
2.	20	0	10	**0.37**	0.33	0.25	**0.58**
3.	20	20	20	**0.90**	0.04	0.83	**0.87**
4.	20	10	15	**0.68**	0.15	0.57	**0.72**

Cases 1 & 2: Testing focuses (correctly) on H_1, but it is still possible to find an effect (wrongly) for the full population.

Overall power is increased.

Case 3: Restricting to the sub-population reduces power for finding an effect in the full population.

Case 4: Adaptation improves overall power a little.

The benefits of an enrichment design

In creating an enrichment design, we can favour specific goals when defining:

The rule for switching to the sub-population,

The test of the intersection hypothesis.

However, we cannot eliminate the probability of making an error in these decisions.

This is to be expected. In our example, the standard error of the interim estimates $\widehat{\theta}_1^{(1)}$ and $\widehat{\theta}_2^{(1)}$ is 12.3 — much higher than the differences between θ_1 and θ_2 that interest us.

Although restricting attention to a sub-population can help improve power, higher overall sample size is needed if we require accurate inference for each sub-population.

Conclusions: Enrichment designs

- With the advent of biomarkers and personalised medicine, there is a desire for methods that allow adaptation to focus on a subgroup of patients during a clinical trial.
- Such methods are feasible using combination tests and a closed testing procedure.
- Restricting attention to a sub-population can be effective in improving power.
- However, higher overall sample size is needed for more accurate sub-population inference.
- There is continuing research on more complex trial scenarios and inference on termination.

5.2 Seamless Phase 2/3 designs

Phase 2b

Several dose levels (or other variants) of a treatment and a control are compared to select a dose and to confirm that this treatment offers an improvement on the control.

Phase 3

This confirmatory study aims to demonstrate superiority against control of the treatment selected in Phase 2b.

Stages:

Write Phase 2b protocol, seek ethical and regulatory approval, (FDA, IRBs, . . .)

Run Phase 2b, analyse data, reach conclusions.

Write Phase 3 protocol, seek ethical and regulatory approval, (FDA, IRBs, . . .)

Run Phase 3, analyse data, reach final conclusion.

Planning the Phase 3 trial after Phase 2 is complete

Planning the Phase 3 trial after Phase 2 allows investigators to make use of information gained in Phase 2.

They may decide to modify:

Treatment definition,

Target population,

Primary endpoint,

Sample size.

Positive results in Phase 2 will help recruitment for participation in Phase 3.

But, planning and gaining approval for the Phase 3 trial can be time-consuming.

If the final outcome is likely to be positive, the sooner this conclusion can be reached, the better.

Inferentially seamless Phase 2/3 trials

Requirements

A single protocol for the combined Phase 2b and Phase 3 trials.

Rules for a committee managing the trials to follow as they:

Decide whether to proceed to Phase 3,

Select the dose level for Phase 3 considering efficacy and safety,

Use information from Phase 2, e.g., estimated treatment effect or response variance to set the Phase 3 sample size.

NB Expect everyone else will be blinded to the Phase 2 results.

Potential benefits

Eliminating the "white space" between phases,

Efficiency gain from using Phase 2 data in the final analysis.

A Closed Testing Procedure with combination tests

Reference: Bretz, Schmidli et al. (*Biometrical Journal*, 2006).

Suppose we test k dose levels plus the control in Phase 2.

Let θ_i, $i = 1, \ldots, k$, denote the effect size of dose level i *vs* the control treatment and define $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.

We consider the null hypotheses H_i: $\theta_i \leq 0$, for $i = 1, \ldots, k$.

Procedure

After Phase 2, select dose level i^* to advance to Phase 3.

Test H_{i^*}: $\theta_{i^*} \leq 0$ at significance level α, with allowance for selecting H_{i^*} from k null hypotheses based on Phase 2 results.

Formally

We shall define a procedure controlling the **familywise error rate**.

Then, for all possible vectors of treatment effects $\boldsymbol{\theta}$

$$P_{\boldsymbol{\theta}}\{\text{Reject any true } H_i\} \leq \alpha.$$

A Closed Testing Procedure

We have data:

Phase 2

Treatment effect estimates $\widehat{\theta}_i^{(1)}$, $i = 1, \ldots, k$,

Phase 3

Treatment effect estimate $\widehat{\theta}_{i^*}^{(2)}$.

Here, dose i^* is selected based on a high observed treatment effect in Phase 2, $\widehat{\theta}_{i^*}^{(1)}$, and good safety outcomes.

Analysing these data:

In order to reject H_{i^*}: $\theta_{i^*} \leq 0$, we need to reject each intersection hypothesis H_I with $i^* \in I$ at level α.

Here, $H_I = \cap_{i \in I} H_i$ states that $\theta_i \leq 0$ for all $i \in I$.

A Closed Testing Procedure

Formally:

In order to reject H_{i^*}: $\theta_{i^*} \leq 0$ globally, we need to reject each intersection hypothesis H_I with $i^* \in I$ in a local, level α test, based on the combined data from Phases 2 and 3.

Intuitively:

Dose i^* is chosen because of good Phase 2 results.

We must adjust for this choice in order to avoid **selection bias** when adding the Phase 2 data on dose level i^* to Phase 3 data in the final analysis .

Under the global null hypothesis $\boldsymbol{\theta} = (0, \ldots, 0)$, we should view $\widehat{\theta}_{i^*}^{(1)}$ as *the highest estimated effect from k comparisons of an ineffective dose with the control,* rather than typical results for a single, pre-specified dose.

A Closed Testing Procedure

Testing an intersection hypothesis H_I: $\theta_i \leq 0$ for all $i \in I$

(a) We need to test the intersection hypothesis in each stage.

(b) We need to combine data from two stages.

Task (b):

Denote the P-value for testing H_I in Phase 2 by $P_I^{(1)}$.

Denote the P-value for testing H_I in Phase 3 by $P_I^{(2)}$.

An inverse χ^2 combination test rejects H_I if at level α

$$-\log(P_I^{(1)}\, P_I^{(2)}) > \frac{1}{2}\,\chi^2_{4,\,1-\alpha}.$$

An inverse normal combination test rejects H_I if

$$w_1\Phi^{-1}(1-P_I^{(1)}) + w_2\Phi^{-1}(1-P_I^{(2)}) > \Phi^{-1}(1-\alpha).$$

where w_1 and w_2 are pre-specified and $w_1^2 + w_2^2 = 1$.

Task (a): Testing an intersection hypothesis in each stage

Testing H_I is most complex in Phase 2

For each $i \in I$, let $P_i^{(1)}$ denote the 1-sided P-value for testing H_i: $\theta_i \leq 0$ against $\theta_i > 0$.

Suppose H_I is the intersection of m simple hypotheses and denote their P-values in increasing order by $P_{[j]}^{(1)}$, $j = 1, \ldots, m$.

Bonferroni adjustment:

The P-value for testing H_I is

$$P_I^{(1)} = m\, P_{[1]}^{(1)}.$$

Simes' method (Biometrika, 1986):

The P-value for H_I is

$$P_I^{(1)} = \min_{j=1,\ldots,m} (m\, P_{[j]}^{(1)}/j).$$

Dunnett's method (JASA, 1955)

Suppose $H_I = \cap_{i \in I} H_i$ is the intersection of m simple hypotheses.

Each H_i concerns the comparison of one dose against the control.

Thus, we have the situation discussed in Section 4.5 where m treatments are compared with a control, responses are normal, and sample sizes on each treatment and the control are equal.

We can, therefore, use Dunnett's method.

Denote the Z-statistic arising from the test of H_i by Z_i. Let $Z^* = \max_{i \in I} Z_i$ and suppose the value attained by Z^* is z^*.

The P-value for testing H_I using Dunnett's test is

$$P\{\max_{i \in I} Z_i > z^*\}$$

when $\theta_i = 0$ for all $i \in I$, so the Z_i are multivariate normal, $Z_i \sim N(0,1),\ i = 1, \ldots, m$, and $\text{Cov}(Z_i, Z_{i'}) = 0.5$ for $i \neq i'$.

Testing an intersection hypothesis in each stage

Testing H_I in Phase 3

Having selected dose i^*, we need to reject each H_I with $i^* \in I$ in order to reject H_{i^*}: $\theta_{i^*} \leq 0$ overall.

Only dose level i^* and the control are studied in Phase 3.

Thus, a test of H_I with $i^* \in I$ using Phase 3 data must be based on $\widehat{\theta}_{i^*}^{(2)}$ — and there is just one such test.

Hence, all H_I of interest have the common P-value $P_I^{(2)} = P_{i^*}^{(2)}$.

Combining P-values from Phases 2 and 3

Having selected dose i^*, we consider each H_I with $i^* \in I$.

In testing H_I, we combine the Phase 2 P-value $P_I^{(1)}$ with the Phase 3 P-value $P_I^{(2)} = P_{i^*}^{(2)}$.

So, the inverse χ^2 combination test rejects H_I for high values of

$$-\log(P_I^{(1)}\, P_{i^*}^{(2)}),$$

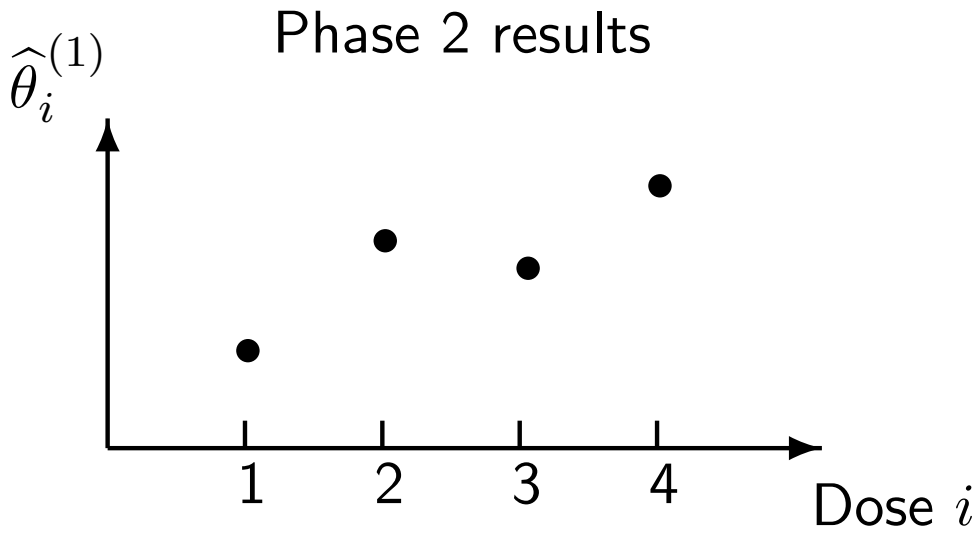
while the inverse normal combination test rejects H_I for high values of

$$w_1\Phi^{-1}(1-P_I^{(1)}) + w_2\Phi^{-1}(1-P_{i^*}^{(2)}).$$

Hence, rejection of H_{i^*} depends on the highest value of $P_I^{(1)}$ appearing in these formulae and the key statistic from Phase 2 is

$$\max_I P_I^{(1)} \text{ over sets } I \text{ containing } i^*.$$

An example using Simes' test



We select $i^* = 4$ for Phase 3.

Suppose $P_1^{(1)} = 0.2,\ P_2^{(1)} = 0.04,\ P_3^{(1)} = 0.05,\ P_4^{(1)} = 0.03.$

We need to find $\max_I P_I^{(1)}$ over sets I containing $i^* = 4$.

First consider single element sets I containing $i^* = 4$

There is just one set, $I = \{4\}$, with P-value $P_I^{(1)} = P_4^{(1)} = 0.03.$

An example using Simes' test, continued

In Phase 2: $P_1^{(1)} = 0.2,\ P_2^{(1)} = 0.04,\ P_3^{(1)} = 0.05,\ P_4^{(1)} = 0.03.$

Two-element sets I containing $i^* = 4$

The largest $P_I^{(1)}$ comes from $I = \{1, 4\}$.

The ordered P-values are $P_{[1]}^{(1)} = 0.03$ and $P_{[2]}^{(1)} = 0.2$, so

$$P_I^{(1)} = \min_{j=1,2} (2\, P_{[j]}^{(1)}/j) = 2 \times 0.03 = 0.06.$$

Three-element sets I containing $i^* = 4$

The largest $P_I^{(1)}$ comes from $I = \{1, 3, 4\}$.

The ordered P-values are $P_{[1]}^{(1)} = 0.03$, $P_{[2]}^{(1)} = 0.05$, and $P_{[3]}^{(1)} = 0.2$, so

$$P_I^{(1)} = \min_{j=1,\,2,\,3} (2\, P_{[j]}^{(1)}/j) = 3 \times 0.05/2 = 0.075.$$

An example using Simes' test, continued

In Phase 2: $P_1^{(1)} = 0.2,\ P_2^{(1)} = 0.04,\ P_3^{(1)} = 0.05,\ P_4^{(1)} = 0.03.$

Four-element sets I containing $i^* = 4$

There is just one four-element set, $I = \{1, 2, 3, 4\}$.

The ordered P-values for are $P_{[1]}^{(1)} = 0.03,\ P_{[2]}^{(1)} = 0.04,$ $P_{[3]}^{(1)} = 0.05$ and $P_{[4]}^{(1)} = 0.2$,so

$$P_I^{(1)} \ = \ \min_{j=1,\ldots,4} (4\, P_{[j]}^{(1)}/j) \ = \ 4 \times 0.05/3 \ = \ 0.067.$$

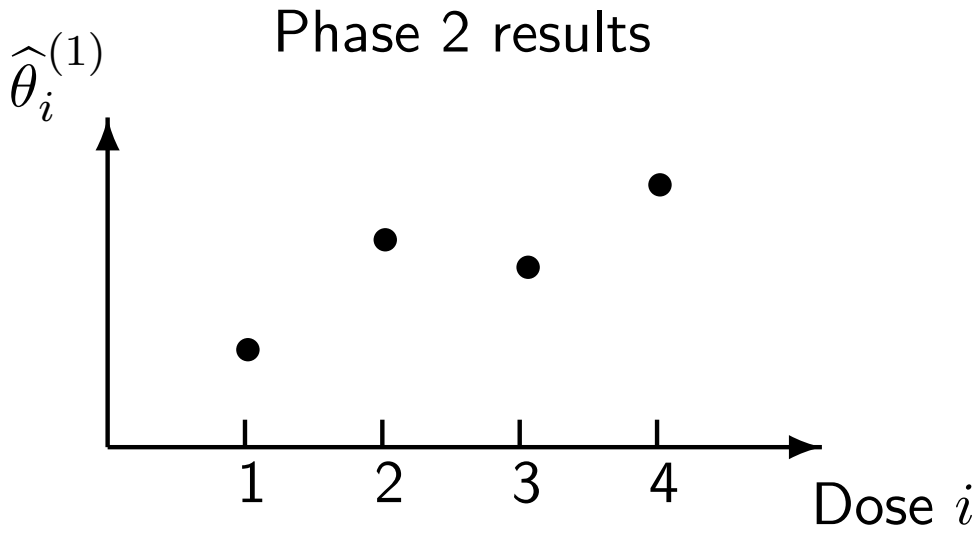
Conclusion:

The maximum value of $P_I^{(1)}$ over all sets I containing $i^* = 4$ comes from $I = \{1, 3, 4\}$, for which

$$P_I^{(1)} \ = \ 0.075.$$

This value will be carried forward to be combined with $P_4^{(2)}$.

An example using Dunnett's test



We select $i^* = 4$ for Phase 3.

$$P_1^{(1)} = 0.2, \quad P_2^{(1)} = 0.04, \quad P_3^{(1)} = 0.05, \quad P_4^{(1)} = 0.03.$$

The Dunnett P-value for intersection hypothesis H_I depends on the smallest $P_i^{(1)}$ for $i \in I$ and it increases with the number of elements of I.

For sets I containing $i^* = 4$, the smallest P-value is $P_4^{(1)} = 0.03$ and the largest such set is $I = \{1, 2, 3, 4\}$.

An example using Dunnett's test, continued

We need to find the Dunnett P-value for the intersection hypothesis H_I with $I = \{1, 2, 3, 4\}$.

The smallest Phase 2 P-value for a dose i with $i \in I$ is $P_4^{(1)} = 0.03$ and this has Z-value

$$P_4^{(1)} \;=\; \Phi^{-1}(1 - 0.03) \;=\; 1.88.$$

Thus, the Dunnett P-value for $I = \{1, 2, 3, 4\}$ is

$$P_I^{(1)} \;=\; P\{\max_{i=1,\ldots,4}(Z_i) > 1.881\} \;= 0.0918$$

— which is calculated assuming the Z_i are multivariate normal, $Z_i \sim N(0, 1),\ i = 1, \ldots, 4$, and $\text{Cov}(Z_i, Z_{i'}) = 0.5$ for $i \neq i'$.

This value will be carried forward to be combined with $P_4^{(2)}$.

Summary: Combining Phase 2 and Phase 3 data

In Phase 2: Select dose i^* and carry forward

$$P_1^* = \max_{I:\, i^* \in I} P_I^{(1)}.$$

In Phase 3: Test dose i^* against control and find $P_{i^*}^{(2)}$.

Take $P_{i^*}^{(2)}$ as the Phase 3 P-value for all tests of intersection hypotheses involving dose i^*.

Overall: Apply a combination test to P_1^* and $P_{i^*}^{(2)}$ to see if the Closed Testing Procedure rejects $H_{i^*}\colon\ \theta_{i^*} \leq 0$.

Flexibility: We can select dose i^* for efficacy, safety, or other factors — it is not necessarily the dose with maximum $\widehat{\theta}_i^{(1)}$.

Efficiency: The use of Phase 2 data in the final analysis should increase power, or reduce the sample size needed for a given power.

Method of Thall, Simon & Ellenberg (*Biometrika*, 1988)

Thall, Simon & Ellenberg (TSE) proposed the 2-stage design:

Phase 2

Take m_1 observations on each dose $i = 1, \ldots, k$ and control,

Identify $\widehat{\theta}_{i^*}^{(1)}$, the maximum of the effect estimates $\widehat{\theta}_i^{(1)}$,

If $\widehat{\theta}_{i^*}^{(1)} < C_1$, stop and accept H_0: $\theta_1 \leq 0, \ldots, \theta_k \leq 0$,

If $\widehat{\theta}_{i^*}^{(1)} \geq C_1$, select dose i^* and proceed to Phase 3.

Phase 3

Take m_2 observations on dose i^* and the control,

Combine data in $T_{i^*} = (m_1 \widehat{\theta}_{i^*}^{(1)} + m_2 \widehat{\theta}_{i^*}^{(2)})/(m_1 + m_2)$,

If $T_{i^*} < C_2$, accept H_0,

if $T_{i^*} \geq C_2$, reject H_0 and conclude $\theta_{i^*} > 0$.

The method of Thall, Simon & Ellenberg

In TSE's method, the statistic $T_{i^*} = (m_1\widehat{\theta}_{i^*}^{(1)} + m_2\widehat{\theta}_{i^*}^{(2)})/(m_1 + m_2)$ is the simple estimate of θ_{i^*} from pooled Phase 2 and 3 data.

The values of m_1, m_2, C_1 and C_2 can be chosen to satisfy type I error and power requirements.

Type I error

Treatment i^* is said to be "chosen" if

Treatment i^* is selected at the end of Phase 2, and

H_0 is rejected in favour of $\theta_{i^*} > 0$ in the final analysis.

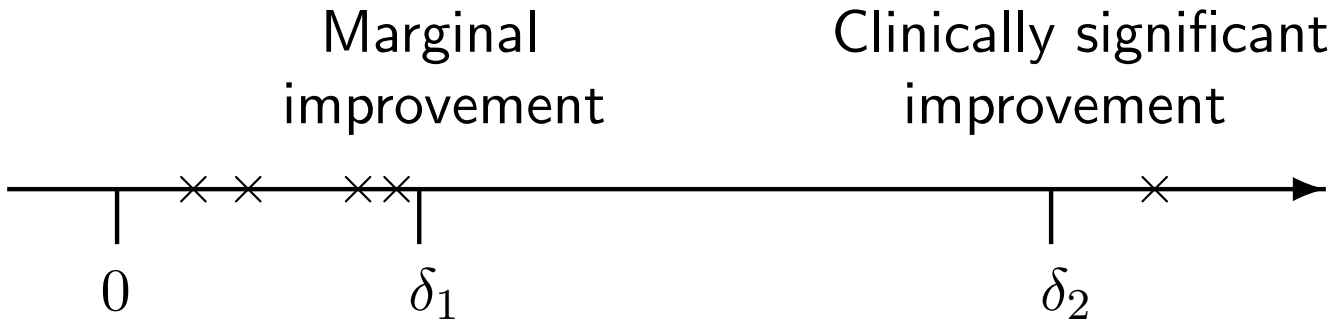
TSE define the type I error rate as the maximum value of

$$P\{\text{Any experimental treatment is "chosen"}\}$$

for H_0: $\theta_1 \leq 0, \ldots, \theta_k \leq 0$, which occurs when $\boldsymbol{\theta} = (0, \ldots, 0)$.

The method of Thall, Simon & Ellenberg

Power



Any dose with $\theta_i \geq \delta_2$ is said to be "acceptable".

TSE consider cases of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ where:

At least one dose is acceptable,

No θ_i lies in the interval (δ_1, δ_2).

They define the power function as

$$P_{\boldsymbol{\theta}}\{\text{An acceptable dose is selected and the corresponding } H_i \text{ is rejected}\}.$$

The method of Thall, Simon & Ellenberg

Power

TSE show that power is minimized in the cases described above under the *least favourable configuration* $\boldsymbol{\theta}^*$, where

$$\theta_1 = \ldots = \theta_{k-1} = \delta_1 \quad \text{and} \quad \theta_k = \delta_2.$$

Numerical integration enables parameters m_1, m_2, C_1 and C_2 to be found with type I error probability α when $\boldsymbol{\theta} = (0, \ldots, 0)$ and power $1 - \beta$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

Optimisation

There are two degrees of freedom left when choosing the four constants m_1, m_2, C_1 and C_2 to satisfy two constraints.

TSE find the design that minimises expected sample size averaged over the two cases $\boldsymbol{\theta} = (0, \ldots, 0)$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

The method of Thall, Simon & Ellenberg

The TSE method has familywise error rate α

TSE set their type I error probability α when $\boldsymbol{\theta} = (0, \ldots, 0)$.

For "strong control" of the familywise error rate we must bound the maximum probability of "choosing" a treatment with $\theta_i \leq 0$ for all vectors $\boldsymbol{\theta}$.

JT (*J. Biopharmaceutical Statistics*, 2007) show the familywise error rate is maximised when $\boldsymbol{\theta} = (0, \ldots, 0)$ and, hence, the familywise error rate is protected at level α.

Adding flexibility

JT express the TSE method as a Closed Testing Procedure.

They define the necessary tests of intersection hypotheses and explain how to adapt these when a treatment i^* is selected with

$$\widehat{\theta}_{i^*}^{(1)} \leq \max_{i=1,\ldots,k} \widehat{\theta}_i^{(1)}.$$

The type I error rate is then met conservatively.

Example JT (*J. Biopharm. Statistics*, 2007)

Disease: Asthma,

Endpoint: Asthma quality of life score (AQLS) at 6 weeks,

Four treatment groups (doses) and a control group,

Responses are assumed to be normal with variance 12.5.

Phase 2:

100 observations per group,

If all $\widehat{\theta}_i^{(1)} < 0$, $i = 1, \ldots, 4$, the trial is halted as futile,

Otherwise, take the dose with the highest $\widehat{\theta}_i^{(1)}$ to Phase 3.

Phase 3:

500 observations on the control and the dose selected in Phase 2.

Example

You are a new statistician just hired by a small pharma company.

You are presented with data from a Phase 2 trial and a Phase 3 trial for asthma treatments, conducted as described above.

You are to analyse these data and, if results are positive, present these in an NDA submission.

(The statistician who designed the trials has just moved to a competitor.)

How do you proceed?

Example: The results

Phase 2 results

	Control	Dose 1	Dose 2	Dose 3	Dose 4
n	100	100	100	100	100
P (1-sided)		0.20	0.04	0.05	0.03
Z		0.84	1.75	1.64	1.88

Dose $i^* = 4$ was selected to go forward to Phase 3.

Phase 3 results

	Control	Dose 4
n	500	500
P (1-sided)		0.04
Z		1.75

Can dose 4 be recommended at significance level $\alpha = 0.025$?

Example: What did the protocol or SAP specify?

You find one of the following:

1. **Conventional**

   Separate Phase 2 and Phase 3 trials were conducted with the final decision to be made using only the Phase 3 data.

2. **Bauer & Köhne**

   Combination of Phase 2 and Phase 3 results using

   (a) inverse χ^2 or (b) weighted inverse normal combination test,

   (i) Simes' test or (ii) Dunnett's test for intersection hypotheses.

3. **The method of Thall, Simon & Ellenberg**

   The final test is based on the estimate of θ_{i^*} from pooled Phase 2 and Phase 3 data — but the critical value accounts for the selection of dose i^*.

Results: Conventional protocol

The Phase 3 P-value of 0.04 is greater than 0.025,

The null hypothesis is not rejected,

The result of the trial is negative.

Results: Bauer & Köhne

Stage 1:

In testing intersection hypotheses in Stage 1, we find P_1^*, the largest P-value $P_I^{(1)}$ for a set I containing the index of the selected dose, $i^* = 4$.

(i) Simes' method

$$P_1 = \max_{I:\, 4 \in I} \{P_I^{(1)}\} = 0.075.$$

The maximum comes from $I = \{1, 3, 4\}$ and equals 0.075, as explained earlier.

(ii) Dunnett's method

$$P_1 = \max_{I:\, 4 \in I} \{P_I^{(1)}\} = 0.0918.$$

The maximum $P_I^{(1)}$ comes from $I = \{1, 2, 3, 4\}$.

Results: Bauer & Köhne

Stage 2:

The intersection hypotheses of interest are those containing $i^* = 4$.

All of these yield the same P-value, $P_I^{(2)} = 0.04$.

(a) Combining P_1^* and $P_I^{(2)}$ by the inverse χ^2 method

Simes' test: With $P_1^* = 0.075$, the inverse χ^2 statistic is

$$-2\log(P_1^* P_I^{(2)}) = -2\log(0.075 \times 0.04) = 11.62$$

with significance level $P\{\chi_4^2 > 11.6\} = 0.0204$.

Dunnett's test: With $P_1^* = 0.0918$, the inverse χ^2 statistic is

$$-2\log(P_1^* P_I^{(2)}) = -2\log(0.0918 \times 0.04) = 11.21$$

with significance level $P\{\chi_4^2 > 11.21\} = 0.0243$.

So, both global tests find dose 4 effective at level $\alpha = 0.025$.

Results: Bauer & Köhne

(b) Combining P_1^* and $P_I^{(2)}$ by the inverse normal method

Weights for each stage are proportional to the square root of sample size.

Simes' test: With $P_1^* = 0.075$ and $P_I^{(2)} = 0.04$, the inverse normal statistic for testing an H_I with $i^* = 4 \in I$ is

$$\sqrt{100/600}\,\Phi^{-1}(1-0.075) + \sqrt{500/600}\,\Phi^{-1}(1-0.04) = 2.186$$

with significance level $1 - \Phi(2.186) = 0.0144$.

Dunnett's test: With $P_1^* = 0.0918$ and $P_I^{(2)} = 0.04$, the inverse normal statistic is

$$\sqrt{100/600}\,\Phi^{-1}(1-0.0918) + \sqrt{500/600}\,\Phi^{-1}(1-0.04) = 2.141$$

with significance level $1 - \Phi(2.141) = 0.0161$.

Again, both global tests find dose 4 effective at level $\alpha = 0.025$.

Results: Thall, Simon & Ellenberg

Stage 1

$$Z_{i^*}^{(1)} = 1.88, \quad \widehat{\theta}_{i^*}^{(1)} = Z_{i^*}^{(1)} \sqrt{2\sigma^2/m_1} = 0.940.$$

Stage 2

$$Z_{i^*}^{(2)} = 1.75, \quad \widehat{\theta}_{i^*}^{(2)} = Z_{i^*}^{(2)} \sqrt{2\sigma^2/m_2} = 0.391.$$

Combining

The TSE statistic is

$$T_{i^*} = \frac{100\,\widehat{\theta}_{i^*}^{(1)} + 500\,\widehat{\theta}_{i^*}^{(2)}}{600} = 0.483,$$

or, equivalently,

$$Z_{i^*} = \sqrt{\frac{100}{600}} \cdot Z_{1,i^*} + \sqrt{\frac{500}{600}} \cdot Z_2 = 2.365.$$

Results: Thall, Simon & Ellenberg

The TSE statistic is $T_{i^*} = 0.483$ or, equivalently, $Z_{i^*} = 2.365$.

The critical value C_2 for T_{i^*} is determined by the requirement

$$P\{\widehat{\theta}_{i^*}^{(1)} > C_1 = 0 \text{ and } T_{i^*} > C_2\} = \alpha = 0.025$$

when $\theta_1 = \ldots = \theta_k = 0$, which gives $C_2 = 0.449$.

The corresponding critical value for Z_{i^*} is 2.20.

Since $T_{i^*} > 0.449$ (and $Z_{i^*} > 2.20$), the null hypothesis $H_4 = H_{i^*}$ is rejected by the overall test with familywise type I error 0.025.

Thus, the trial has a positive outcome and a recommendation can be made in support of dose 4.

Efficiency gains from combining Phase 2 and Phase 3 data

Inferentially seamless Phase 2/3 designs carry a high organisational cost, so it is important they should provide substantial benefits.

It is therefore of interest to compare:

Separate Phase 2 and Phase 3 trials

versus

Seamless designs with Phase 2 data used at the end of Phase 3.

Since the vector of treatment effects is k-dimensional, there are many scenarios to consider.

There are also many options for conducting the final analysis.

Efficiency gains from combining Phase 2 and Phase 3 data

Jennison & Turnbull (*J. Biopharm. Statistics*, 2007) investigated the power of different testing procedures for the Asthma trial.

They considered situations when the vector of treatment effects is of the form $\boldsymbol{\theta} = (0, 0, 0, \delta)$.

We shall extend their results to include:

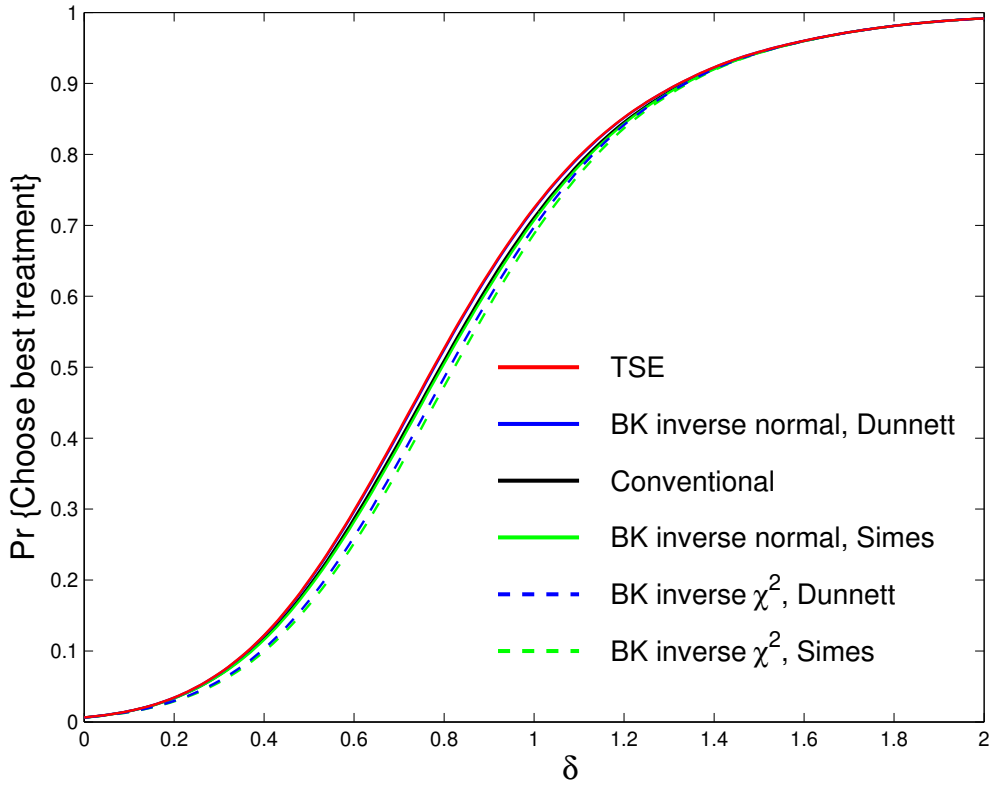
The "conventional" method, using only Phase 3 data in the final hypothesis test,

Inverse normal and inverse χ^2 combination tests paired with Simes and Dunnett rules for testing intersection hypotheses,

TSE: The method proposed by Thall, Simon & Ellenberg (*Biometrika*, 1988).

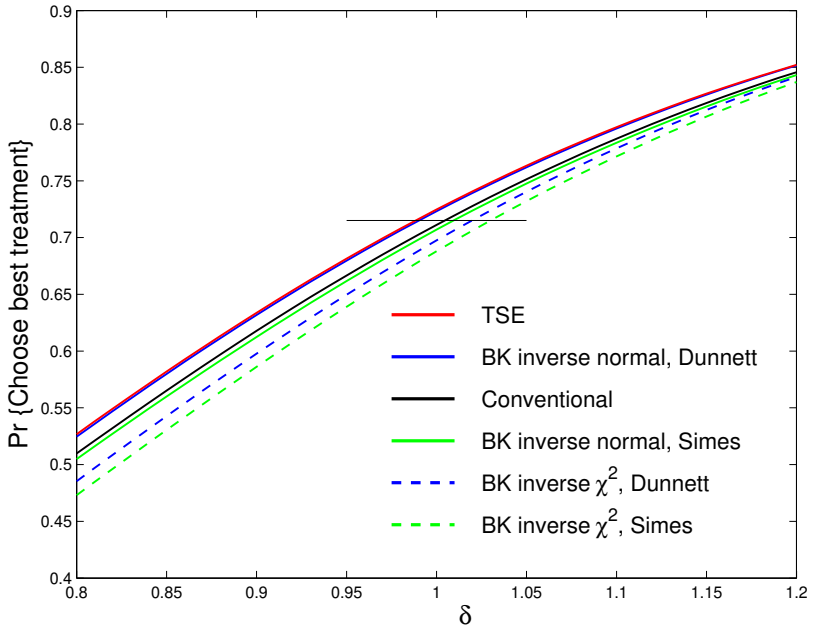
Power functions of six selection and testing procedures

Power of six 2-stage procedures when $\boldsymbol{\theta} = (0, 0, 0, \delta)$,
i.e., three doses are ineffective and the other has effect size δ:



The values of m_1 and m_2 and the stopping rule are the same for all six methods, so each procedure has the sample size distribution.

Power functions of six selection and testing procedures



The TSE procedure has slightly higher power than the inverse normal combination test using a Dunnett rule.

The Conventional procedure — with no data combination — is a little worse than the TSE test, and superior to three versions of the BK combination test !!

Differences in power equate to sample size differences of 4% to 8%.

Discussion of the six power functions

Variations in performance between the procedures are surprising.

While one BK method improves on the Conventional design, the others fail to gain any advantage from using Phase 2 data in the final analysis — but given the practical difficulties involved in a seamless Phase 2/3 design, a positive benefit is essential.

The **TSE** method and **BK with a weighted inverse normal combination test and Dunnett test for intersection hypotheses** perform well for vectors $\boldsymbol{\theta} = (0, 0, 0, \delta)$.

Are the same methods the best choices for other configurations of treatment means?

Is it possible to quantify the value of including Phase 2 data in the final analysis, e.g., in terms of an equivalent number of Phase 3 observations?

Assessing the benefits of Seamless Phase 2/3 designs

Reference: Hampson & Jennison, *Statistics in Medicine, 2015*, hereafter "H&J".

Hampson & Jennison create a decision theoretic formulation of the testing problem at the end of a Phase 2/3 trial and search for optimal decision rules in a variety of scenarios.

They conclude that good all round performance is achieved by

> The TSE method,
>
> An inverse normal combination test paired with Dunnett's test for intersection hypotheses.

Using Simes' rule, rather that the Dunnett test, can lead to inefficiency when some doses have low effect sizes.

Some comments on Simes' method

The Simes P-value for intersection hypothesis H_I

$$P_I^{(1)} = \min_{j=1,\ldots,m} (m\, P_{[j]}^{(1)}/j).$$

starts with $m\, P_{[1]}^{(1)}$ from dose i^* with the largest $\widehat{\theta}_i^{(1)}$, but may then "borrow strength" from other doses.

H&J show that, because of the correlation between the $\widehat{\theta}_i^{(1)}$s, it is optimal to place negative weights on the $\widehat{\theta}_i^{(1)}$, $i \neq i^*$. when testing $\boldsymbol{\theta} = (0, 0, 0, 0)$ vs $\boldsymbol{\theta} = (0, 0, 0, \delta)$.

H&J also show the TSE test is exactly optimal if the vector of treatment effects is of the form $\boldsymbol{\theta} = (\delta/2, \delta/2, \delta/2, \delta)$.

Since the TSE test does not involve $\widehat{\theta}_i^{(1)}$, $i \neq i^*$, we see that for "borrowing strength" to help, the second highest treatment effect must be at least half the maximum effect.

Value of Phase 2 data in an inferentially seamless design

H&J compare the power of efficient seamless designs with that of a "conventional" design with a higher Phase 3 sample size.

Let m_1 be the number of subjects treated at each dose in Phase 2.

Let γ be such that, in order to attain the power of the seamless design, a conventional design needs $\gamma\, m_1$ more observations on each of the selected treatment arm and the control arm.

A rule of thumb

In a variety of examples, H&J found γ to lie between 0.2 and 1.0.

In the most plausible scenarios, $\gamma \approx 0.5$, so the Phase 2 data on dose i^* and the control are worth about half their face value.

This advantage could justify the extra effort in running a seamless Phase 2/3 trial — particularly when there are low numbers of patients with the indication in question.

Further topics for Phase 2/3 trials

In variations and extensions of the preceding methods:

- More than one treatment may be carried forward to Phase 3,
- Sequential monitoring could result in early elimination of inferior treatments in Phase 2, or an early Phase 3 decision,
- In Phase 2, one may select and fit a dose-response model,
- An over-arching approach may be followed to optimise the Phase 2 and Phase 3 trials together.

See:

- TSE and H&J for the optimal division of sample size between Phases 2 and 3.
- Antonijevic et al., Ch. 6 of *Optimization of Pharmaceutical R&D Programs and Portfolios* (2015).
- Robbie Peck, University of Bath PhD Thesis (to appear).

Conclusions: Seamless Phase 2/3 designs

- The principles of Closed Testing Procedures and Combination Tests can be applied to create Seamless Phase 2/3 designs that protect the familywise error rate.
- Investigations have shown that using a Dunnett test for intersection hypotheses along with a weighted inverse normal combination test produces designs with robust efficiency across a variety of forms of the treatment effect vector.
- There are tangible benefits from including Phase 2 data in the final analysis: typically, the Phase 2 data on the selected dose and the control are worth about half their face value.
- Within this general framework, there is scope for further development and optimisation of Phase 2/3 designs.

Part 6. Multi-armed group sequential trials

6.1. Multi-armed multi-stage (MAMS) designs

Overall plan of MAMS trials

Closed Testing Procedures using Lehmacher-Wassmer multi-stage combination tests

An illustrative example

6.2. A survival trial with treatment selection

Overall plan of the trial

Avoiding error rate inflation in an adaptive trial with survival data

Choosing an adaptive design and assessing its benefits

6.1. Multi-armed multi-stage (MAMS) designs

One may wish to test several new treatments in a Phase 3 trial.

Perhaps there are variants of a new drug treatment to compare:

Should the drug be administered orally or subcutaneously?

Is the high dose necessary for efficacy?

Should the new drug be used with drug A, B or C in a combination therapy?

We shall consider a study in which Treatments 1 to J are compared with a control.

Denote the treatment effects by θ_j, $j = 1, \ldots, J$.

Then, we wish to test the null hypotheses H_j: $\theta \leq 0$ against one-sided alternatives $\theta_j > 0$.

We assume the familywise type I error rate must be at most α.

Multi-armed multi-stage (MAMS) designs

With treatment effects θ_j, $j = 1, \ldots, J$, let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$.

We assume familywise type error rate is to be protected at level α, so for all $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$

$$P_{\boldsymbol{\theta}}\{\text{Reject } H_j \text{ for some } j \text{ with } \theta_j \leq 0\} \leq \alpha.$$

(Some forms of multi-arm trial may not require this.)

We conduct the trial in stages, indexed by $k = 1$ to K.

At each analysis, we may drop poorly performing treatments or reach a positive conclusion about a treatment that performs well.

Having rejected H_j: $\theta_j < 0$ for one treatment, we could stop the whole trial and recommend Treatment j.

However, it may be of interest to see if other treatments are superior to the control, especially if they have a lower dose or comprise a more easily tolerated combination therapy.

Multi-armed multi-stage (MAMS) designs

Before the trial commences, the sponsor will explain the goals of the trial to the Data Monitoring Committee (DMC) and set out guidelines for:

When to drop a poorly performing treatment,

When to declare a positive result for an effective treatment,

When to terminate the whole study.

At its meetings, the DMC will consider all available information on efficacy and safety endpoints.

While clear guidelines are to be encouraged, it may be unrealistic to suppose formal stopping rules can anticipate every eventuality.

It is likely that the DMC will need to exercise judgement in applying the guidelines.

Therefore, we shall consider flexible, adaptive procedures.

Closed Testing Procedures for MAMS trials

The data

In each Stage $k = 1, \ldots, K$, denote the estimated effect of Treatment j vs control, based only on Stage k data, by $\widehat{\theta}_j^{(k)}$.

Suppose $\widehat{\theta}_j^{(k)}$ has variance $V_j^{(k)}$, then the associated Z-statistic is

$$Z_j^{(k)} = \widehat{\theta}_j^{(k)} / \sqrt{\{V_j^{(k)}\}}.$$

We shall apply a Closed Testing Procedure to deal with the multiple hypotheses $H_1, \ldots, H_J$.

This requires a test for each intersection hypothesis H_I where $I \subseteq \{1, \ldots, J\}$ — and these tests are to be conducted group sequentially.

We shall apply a Lehmacher-Wassmer multi-stage combination test to each H_I.

A Lehmacher-Wassmer K-stage test of H_I

We use Lehmacher-Wassmer multi-stage combination tests to

(i) Combine data across stages,

(ii) Allow early stopping for negative or positive outcomes.

We assume Stage k data yield the statistic $Z_I^{(k)}$ where, under H_I, $Z_I^{(k)} \sim N(0,1)$ or $Z_I^{(k)}$ is stochastically smaller than this.

We first specify weights $w_1, \ldots, w_K$.

Then at analysis k, we form the cumulative statistic

$$Z_{I,k} \;=\; (w_1 Z_I^{(1)} + \ldots + w_k Z_I^{(k)}) \,/\, (w_1^2 + \ldots + w_k^2)^{1/2}.$$

Under H_I, the sequence $\{Z_{I,k}\}$ has the "canonical joint distribution" with $\theta = 0$ and information levels $\mathcal{I}_k \propto w_k$, or $\{Z_{I,k}\}$ is stochastically smaller than this.

A Lehmacher-Wassmer K-stage test of H_I

Under H_I, the sequence $\{Z_{I,k}\}$ has the canonical distribution:

$(Z_{I,1}, \ldots, Z_{I,K})$ is multivariate normal,

$Z_{I,k} \sim N(0,1), \quad k = 1, \ldots, K,$

$$\text{Cov}(Z_{I,k_1}, Z_{I,k_2}) = \sqrt{(w_1^2 + \ldots + w_{k_1}^2)/(w_1^2 + \ldots + w_{k_2}^2)}, \; k_1 < k_2$$

or $\{Z_{I,k}\}$ is stochastically smaller than this.

Let b_k, $k = 1, \ldots, K$, be the upper boundary points of a GST with type I error rate α when the cumulative Z-statistics have the above distribution.

Then, the Lehmacher-Wassmer test rejects H_I at stage k if

$$Z_{I,k} > b_k.$$

One can add a non-binding futility boundary to this GST but, for simplicity, we shall not pursue this option here.

A Lehmacher-Wassmer K-stage test of H_I

Example

Consider a 4-stage multi-arm trial.

Suppose we anticipate the same group sizes per treatment arm in each stage, and so define equal weights

$$w_1 = w_2 = w_3 = w_4 = 1.$$

If treatments are dropped and group sizes per treatment arm increase in later stages, higher weights in these stages would be appropriate. However, the **weights must be fixed before the trial commences.**

We shall use a ρ-family error spending test with $\rho = 2$, type I error rate α, with no futility boundary.

This has upper boundary points

$$b_1 = 2.96, \quad b_2 = 2.56, \quad b_3 = 2.30, \quad b_4 = 2.09.$$

Testing an intersection hypothesis H_I in Stage k

For each Treatment j, we have the Z-statistic based on Stage k data only,

$$Z_j^{(k)} = \widehat{\theta}_j^{(k)} / \sqrt{\{V_j^{(k)}\}}.$$

Suppose the set I has m elements and let z^* be the maximum of the observed values of the $Z_j^{(k)}$, $j \in I$.

The P-value for testing H_I using Dunnett's test is

$$P_I^{(k)} = P\{\max_{j \in I} Z_j > z^*\},$$

when $(Z_1, \ldots, Z_m)$ is multivariate normal with each $Z_j \sim N(0,1)$ and $\text{Cov}(Z_j, Z_{j'}) = 0.5$, $j \neq j'$.

The associated Z-statistic is

$$Z_I^{(k)} = \Phi^{-1}(1 - P_I^{(k)})$$

and, by construction, $Z_I^{(k)} \sim N(0,1)$ if $\theta_j = 0$ for all $j \in I$.

Example: A MAMS trial design

Suppose 3 treatments, low, medium and high doses of a new drug, are to be compared against a control in a 4-stage trial.

We specify:

A Closed Testing Procedure,

Dunnett's method to be used to create stage-wise Z-values for intersection hypotheses,

Lehmacher-Wassmer, 4-stage combination tests for each H_I based on ρ-family error spending tests with

$\rho = 2$, $\alpha = 0.025$, no futility boundary.

The null hypothesis H_j: $\theta_j \leq 0$ can be rejected globally if the Lehmacher-Wassmer tests reject each H_I with $j \in I$.

Each treatment may be discontinued at any point for positive or negative reasons.

Example: Stage 1 data

Suppose the first stage produces Z-statistics $Z_1^{(1)}$, $Z_2^{(1)}$, and $Z_3^{(1)}$ for the three treatments, as shown below.

Treatment j	$Z_j^{(1)}$
1	1.26
2	1.84
3	2.76

We shall apply Dunnett's rule to find the Z-value $Z_I^{(1)}$ for each intersection hypothesis H_I.

At Stage 1, the $Z_I^{(1)}$ are also the cumulative Z-values, $Z_{1,I}$, that appear in the Lehmacher-Wassmer test.

Since the Lehmacher-Wassmer testing boundary has

$$b_1 = 2.96, \quad b_2 = 2.56, \quad b_3 = 2.30, \quad b_4 = 2.09,$$

we shall need to see $Z_{1,I} = Z_I^{(1)} \geq 2.96$ to reject H_I at this stage.

Example: Stage 1 data

Applying Dunnett's rule, Z-values for intersection hypotheses are

Hypothesis H_I	$Z_I^{(1)}$
$H_{\{1\}}$	1.26
$H_{\{2\}}$	1.84
$H_{\{3\}}$	2.76
$H_{\{1,2\}}$	1.56
$H_{\{1,3\}}$	2.54
$H_{\{2,3\}}$	2.54
$H_{\{1,2,3\}}$	2.41

As already noted, the $Z_I^{(1)}$ are also the cumulative Z-values, $Z_{1,I}$, that appear in the Lehmacher-Wassmer test.

As each $Z_{1,I} = Z_I^{(1)} < b_1 = 2.96$, no hypotheses are rejected here.

We suppose the trial continues with all 3 treatments still active.

Example: Stage 2 data

Results in Stage 2 (only) produce the Z-statistics $Z_1^{(2)}$, $Z_2^{(2)}$ and $Z_3^{(2)}$ shown below.

Treatment j	$Z_j^{(2)}$
1	-0.45
2	2.21
3	0.71

From these, we compute the Dunnett Z-values, $Z_I^{(2)}$, for each intersection hypothesis H_I.

Then, to apply the Lehmacher-Wassmer test, we calculate the cumulative Z-value for each H_I

$$Z_{I,2} = \frac{Z_I^{(1)} + Z_I^{(2)}}{\sqrt{2}}.$$

Example: Results after Stages 1 and 2

H_I	$Z_I^{(1)} = Z_{1,I}$	$Z_I^{(2)}$	$Z_{2,I}$
$H_{\{1\}}$	1.26	-0.45	0.57
$H_{\{2\}}$	1.84	2.21	2.86
$H_{\{3\}}$	2.76	0.71	2.45
$H_{\{1,2\}}$	1.56	1.96	2.49
$H_{\{1,3\}}$	2.54	0.34	2.04
$H_{\{2,3\}}$	2.54	1.96	3.18
$H_{\{1,2,3\}}$	2.41	1.81	2.98

The Lehmacher-Wassmer tests reject intersection hypotheses $H_{\{2\}}$, $H_{\{2,3\}}$ and $H_{\{1,2,3\}}$ since they have $Z_{2,I} > b_2 = 2.56$.

However, $H_{\{1,2\}}$ is not rejected so the Closed Testing Procedure does not allow global rejection of H_2.

Suppose the high dose Treatment 3 is dropped for safety reasons, so the trial continues with Treatments 1 and 2 and the control.

Example: Stage 3 data

Results in Stage 3 (only) produce Z-statistics $Z_1^{(3)}$ and $Z_2^{(3)}$

Treatment j	$Z_j^{(3)}$
1	0.90
2	1.41
3	—

In computing the Dunnett Z-value for an intersection hypothesis H_I with $3 \in I$, we set $Z_I^{(3)}$ equal to $Z_{I'}^{(3)}$ where $I' = I \setminus \{3\}$.

This cannot be done for $I = \{3\}$ — but that is not a problem as we are no longer interested in the global test of H_3.

The cumulative Z-value for each H_I after the first 3 stages is

$$Z_{I,3} = \frac{Z_I^{(1)} + Z_I^{(2)} + Z_I^{(3)}}{\sqrt{3}}.$$

Example: Results after Stages 1, 2 and 3

H_I	$Z_I^{(1)} = Z_{1,I}$	$Z_I^{(2)}$	$Z_{2,I}$	$Z_I^{(3)}$	$Z_{3,I}$
$H_{\{1\}}$	1.26	-0.45	0.57	0.90	0.99
$H_{\{2\}}$	1.84	2.21	2.86	1.41	3.15
$H_{\{3\}}$	2.76	0.71	2.45	—	—
$H_{\{1,2\}}$	1.56	1.96	2.49	1.10	2.67
$H_{\{1,3\}}$	2.54	0.34	2.04	0.90	2.19
$H_{\{2,3\}}$	2.54	1.96	3.18	1.41	3.41
$H_{\{1,2,3\}}$	2.41	1.81	2.98	1.10	3.07

The Lehmacher-Wassmer tests reject intersection hypotheses $H_{\{2\}}$, $H_{\{1,2\}}$, $H_{\{2,3\}}$ and $H_{\{1,2,3\}}$ since they have $Z_{3,I} > b_3 = 2.30$.

Thus, H_2 can be rejected globally and Treatment 2 declared superior to the control.

Suppose Treatment 2 is discontinued at this point and the trial continues with the low dose Treatment 1 and the control.

Example: Stage 4 data

Results in Stage 4 (only) produce the single Z-statistic $Z_1^{(4)}$.

Treatment j	$Z_j^{(4)}$
1	2.07
2	—
3	—

We shall use $Z_1^{(4)}$ to create Z-statistics for intersection hypotheses H_I involving Treatment 1.

We can then conduct the final analysis of the Lehmacher-Wassmer tests of these hypotheses using the test statistics

$$Z_{I,4} = \frac{Z_I^{(1)} + Z_I^{(2)} + Z_I^{(3)} + Z_I^{(4)}}{2}.$$

Example: Results after Stages 1, 2, 3 and 4

H_I	$Z_I^{(1)}$	$Z_I^{(2)}$	$Z_{2,I}$	$Z_I^{(3)}$	$Z_{3,I}$	$Z_I^{(4)}$	$Z_{4,I}$
$H_{\{1\}}$	1.26	-0.45	0.57	0.90	0.99	2.07	1.89
$H_{\{2\}}$	1.84	2.21	2.86	1.41	3.15	—	—
$H_{\{3\}}$	2.76	0.71	2.45	—	—	—	—
$H_{\{1,2\}}$	1.56	1.96	2.49	1.10	2.67	2.07	3.35
$H_{\{1,3\}}$	2.54	0.34	2.04	0.90	2.19	2.07	2.93
$H_{\{2,3\}}$	2.54	1.96	3.18	1.41	3.41	—	—
$H_{\{1,2,3\}}$	2.41	1.81	2.98	1.10	3.07	2.07	3.69

The Lehmacher-Wassmer tests reject intersection hypotheses $H_{\{1,2\}}$, $H_{\{1,3\}}$ and $H_{\{1,2,3\}}$ since they have $Z_{4,I} > b_4 = 2.09$.

However, $H_{\{1\}}$ is not rejected, so the Closed Testing Procedure does not allow global rejection of H_1.

Example: Conclusions

The conclusions from the study are that:

The medium dose, Treatment 2, was shown to be superior to the control in a testing procedure with familywise type I error rate $\alpha = 0.025$,

The low dose, Treatment 1, was not found to be superior to the control,

The high dose, Treatment 3, was found to have safety problems and was dropped half way through the study.

6.2 A survival trial with treatment selection

Consider a Phase 3 trial of cancer treatments comparing

Experimental Treatment 1: Intensive dosing

Experimental Treatment 2: Slower dosing

Control treatment

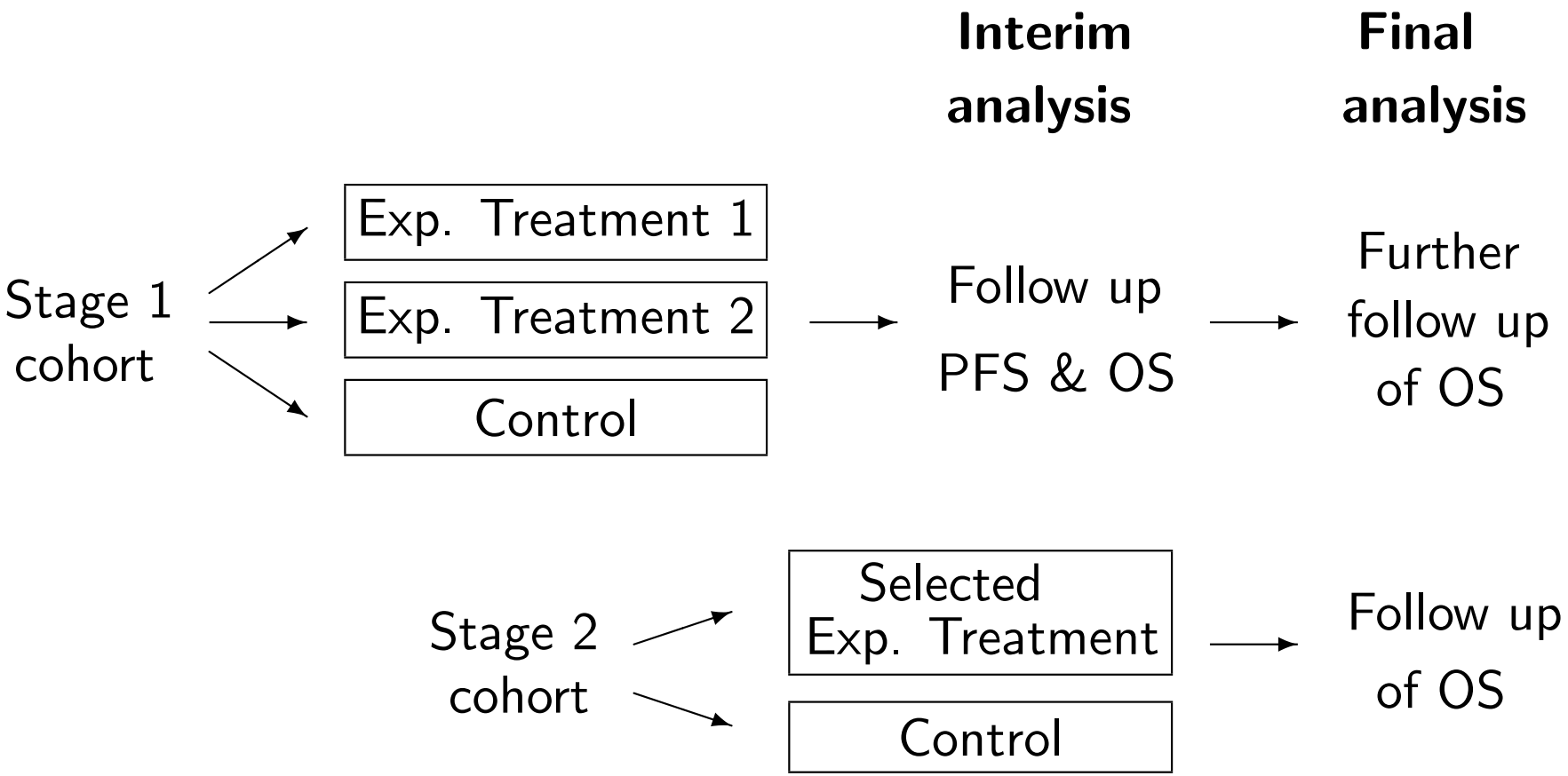
The primary endpoint is Overall Survival (OS).

At an interim analysis, information on OS, Progression Free Survival (PFS), PK measurements and safety will be used to choose between the two experimental treatments.

Note that PFS is useful here as it is more rapidly observed.

After the interim analysis, patients will only be recruited to the selected treatment and the control.

Overall plan of the trial



At the final analysis, we test the null hypothesis that OS on the selected treatment is no better than OS on the control treatment.

Protecting the type I error rate

We shall assume a proportional hazards model for OS with

$$\lambda_1 = \text{Hazard ratio, Control } \textit{vs} \text{ Exp Treatment 1}$$

$$\lambda_2 = \text{Hazard ratio, Control } \textit{vs} \text{ Exp Treatment 2}$$

$$\theta_1 = \log(\lambda_1), \quad \theta_2 = \log(\lambda_2).$$

We test null hypotheses

$$H_1\text{: } \theta_1 \leq 0 \text{ vs } \theta_1 > 0 \quad \textit{(Exp Treatment 1 superior to control)},$$

$$H_2\text{: } \theta_2 \leq 0 \text{ vs } \theta_2 > 0 \quad \textit{(Exp Treatment 2 superior to control)}.$$

In order to control the "familywise error rate", we require

$$P_{(\theta_1,\theta_2)}\{\text{Reject any true null hypothesis}\} \leq \alpha$$

for all (θ_1, θ_2).

A closed testing procedure

Define level α tests of

$$H_1\text{:}\ \ \theta_1 \leq 0,$$

$$H_2\text{:}\ \ \theta_2 \leq 0$$

and a level α test of the intersection hypothesis

$$H_{12} \ = \ H_1 \cap H_2\text{:}\ \ \ \theta_1 \leq 0 \ \text{ and } \ \theta_2 \leq 0.$$

Then:

Reject H_1 **overall** *if the above tests reject* H_1 *and* H_{12},

Reject H_2 **overall** *if the above tests reject* H_2 *and* H_{12}.

The requirement to reject H_{12} compensates for testing multiple hypotheses and the "selection bias" in choosing the treatment to focus on in Stage 2.

Combining data across stages

Consider testing a generic null hypothesis H_0: $\theta \leq 0$ against $\theta > 0$.

Suppose Stage 1 data produce $Z^{(1)}$ where

$$Z^{(1)} \sim N(0,1) \quad \text{if } \theta = 0.$$

On adaptation, Stage 2 data yield $Z^{(2)}$ *conditionally* distributed as

$$Z^{(2)} \sim N(0,1) \quad \text{if } \theta = 0,$$

while $Z^{(1)}$ and $Z^{(2)}$ are stochastically smaller if $\theta < 0$.

Weighted inverse normal Combination Test

With pre-specified weights w_1 and w_2 satisfying $w_1^2 + w_2^2 = 1$,

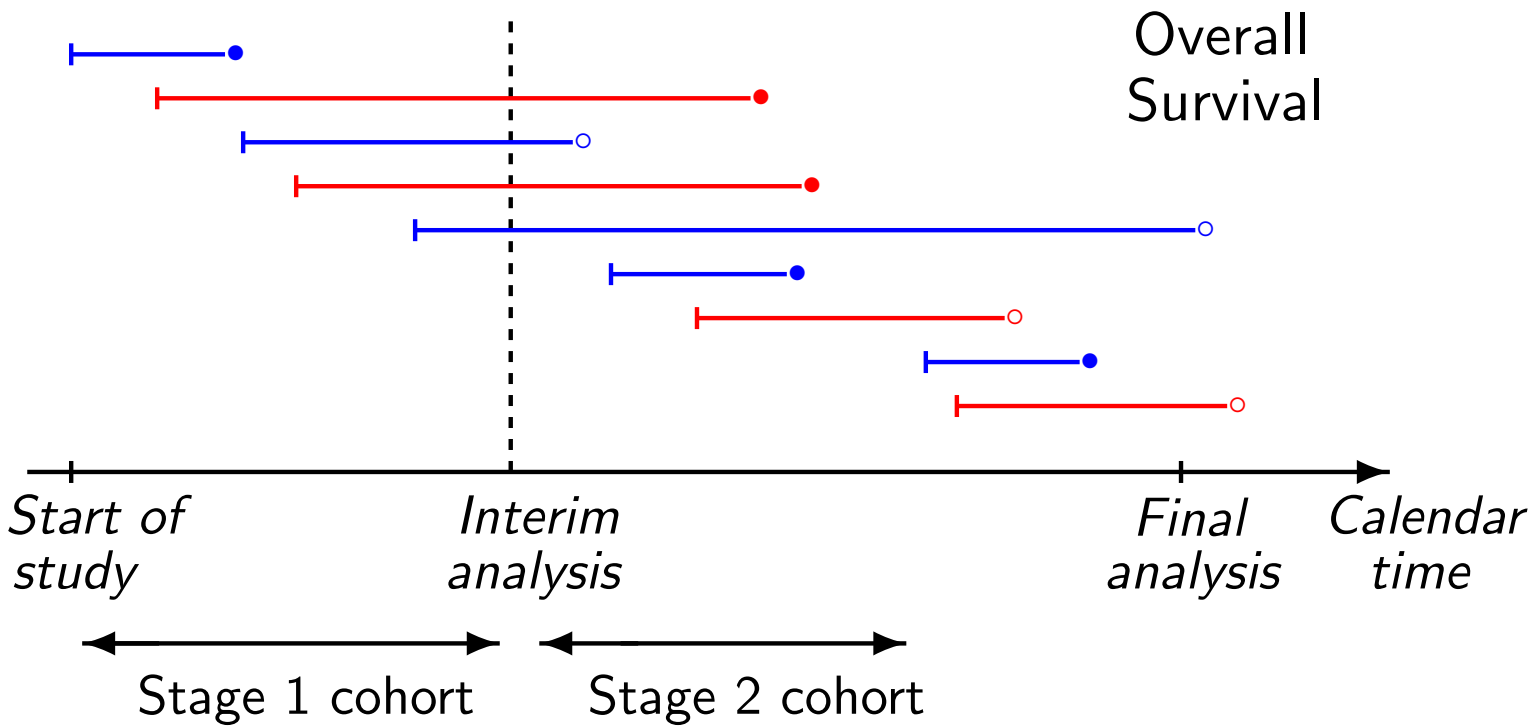
$$Z \;=\; w_1\, Z^{(1)} + w_2\, Z^{(2)} \;\sim\; N(0,1) \quad \text{if } \theta = 0,$$

and Z is stochastically smaller than $N(0,1)$ if $\theta < 0$.

So, for a level α test, we reject H_0 if $Z > \Phi^{-1}(1-\alpha)$.

Applying a combination test to survival data

For now, consider Experimental Treatment 1 vs Control.



Key:
- Subjects randomised to Exp Treatment 1
- Subjects randomised to Control
- ● Death observed
- ○ Censored observation

Properties of log-rank tests

Comparing Experimental Treatment 1 vs Control, define

S_1 = Unstandardised log-rank statistic at interim analysis,

$\mathcal{I}_1$ = Information for θ_1 at interim analysis $\approx$ (Number of deaths)/4

S_2 = Unstandardised log-rank statistic at final analysis,

$\mathcal{I}_2$ = Information for θ_1 at final analysis $\approx$ (Number of deaths)/4

Here, "Number of deaths" refers to the total number of deaths on Experimental Treatment 1 and Control arms only.

Then, approximately,

$$S_1 \sim N(\mathcal{I}_1 \theta_1, \mathcal{I}_1),$$

$$S_2 - S_1 \sim N(\{\mathcal{I}_2 - \mathcal{I}_1\} \theta_1, \{\mathcal{I}_2 - \mathcal{I}_1\})$$

and S_1 and $(S_2 - S_1)$ are **independent** (independent increments).

Reference: Tsiatis (*Biometrika*, 1981).

A combination test for survival data

We create Z statistics

Based on data at the interim analysis:

$$Z^{(1)} = \frac{S_1}{\sqrt{\mathcal{I}_1}},$$

Based on data accrued **between** the interim and final analyses:

$$Z^{(2)} = \frac{S_2 - S_1}{\sqrt{\mathcal{I}_2 - \mathcal{I}_1}}.$$

If $\theta_1 = 0$, $Z^{(1)} \sim N(0,1)$ and $Z^{(2)} \sim N(0,1)$ are independent.

If $\theta_1 < 0$, $Z^{(1)}$ and $Z^{(2)}$ are stochastically smaller than this.

So, we can use $Z = w_1 Z^{(1)} + w_2 Z^{(2)}$ in an inverse normal combination test of H_1: $\theta_1 \leq 0$.

A combination test for survival data: Caution!

The above distribution theory for logrank statistics of a single comparison requires

$$Z^{(2)} = \frac{S_2 - S_1}{\sqrt{\mathcal{I}_2 - \mathcal{I}_1}} \sim N(0,1) \quad \text{under } \theta_1 = 0,$$

regardless of decisions taken at the interim analysis.

Bauer & Posch (*Statistics in Medicine*, 2004) note this implies that the conduct of the second part of the trial should not depend on the prognosis of Stage 1 patients at the interim analysis.

Suppose prognoses are better for patients on Exp Treatment 1 than for those on Control, and the Stage 2 cohort size is reduced while follow up of Stage 1 patients is extended: then, the distribution of $Z^{(2)}$ could be biased upwards.

Our example has another potential source of bias, depending on how the Stage 2 statistic for testing H_{12} is defined.

Analysing an adaptive survival trial

In applying a Closed Testing Procedure, we require level α tests of

H_1: $\theta_1 \leq 0$,

H_2: $\theta_2 \leq 0$,

H_{12}: $\theta_1 \leq 0$ and $\theta_2 \leq 0$.

Combination tests for these hypotheses are formed from:

	Stage 1 data	*Stage 2 data*
H_1	$Z_1^{(1)}$	$Z_1^{(2)}$
H_2	$Z_2^{(1)}$	$Z_2^{(2)}$
H_{12}	$Z_{12}^{(1)}$	$Z_{12}^{(2)}$

The question is how should we define $Z_1^{(1)}$, $Z_1^{(2)}$, etc?

Analysing an adaptive survival trial

A natural choice is to:

Base $Z_1^{(1)}$, $Z_2^{(1)}$ and $Z_{12}^{(1)}$ on data at the interim analysis,

Base $Z_1^{(2)}$, $Z_2^{(2)}$ and $Z_{12}^{(2)}$ on the additional information accruing between interim and final analyses.

We could take $Z_1^{(1)}$ and $Z_2^{(1)}$ to be standardised log-rank statistics, and $Z_1^{(2)}$ and $Z_2^{(2)}$ standardised increments between analyses.

For intersection hypotheses: $Z_{12}^{(1)}$ is formed from $Z_1^{(1)}$ and $Z_2^{(1)}$, **while $\boldsymbol{Z_{12}^{(2)} = Z_j^{(2)}}$, where $\boldsymbol{j}$ is the selected treatment.**

However, treatment j is selected because it has better PFS outcomes at the interim analyses, so it is likely that future OS for these patients will also be better.

This approach would lead to a bias in the null distribution of $Z_{12}^{(2)}$.

The method of Jenkins, Stone & Jennison (2011)

If we base a combination test on the two parts of the data accrued before and after the interim analysis, bias can result:

	$Z^{(1)}$	$Z^{(2)}$
Stage 1 cohort	Overall survival (during Stage 1)	**Overall survival (during Stage 2)**
Stage 2 cohort		Overall survival (during Stage 2)

Instead, we divide the data into the parts from the two cohorts:

Stage 1 cohort	Overall survival (during Stage 1)	Overall survival (during Stage 2)	$Z^{(1)}$
Stage 2 cohort		Overall survival (during Stage 2)	$Z^{(2)}$

Partitioning data for a combination test

To avoid bias: All patients in the Stage 1 cohort are followed for overall survival up to a fixed time, shortly before the final analysis.

"Stage 1" statistics are based on Stage 1 cohort's **final** OS data

$Z_1^{(1)}$ from log-rank test of Exp Tr 1 vs Control

$Z_2^{(1)}$ from log-rank test of Exp Tr 2 vs Control

$Z_{12}^{(1)}$ from pooled log-rank test, or a Simes or Dunnett test.

"Stage 2" statistics are based on OS data for the Stage 2 cohort

If Exp Treatment 1 is selected:

$Z_1^{(2)}$ from log-rank test of Exp Tr 1 vs Control, $Z_{12}^{(2)} = Z_1^{(2)}$

If Exp Treatment 2 is selected:

$Z_2^{(2)}$ from log-rank test of Exp Tr 2 vs Control, $Z_{12}^{(2)} = Z_2^{(2)}$.

Partitioning data for a combination test

Discussion

Jenkins, Stone & Jennison (2011) introduced the proposed method in a design where a choice is made between testing for an effect in the full population or a sub-population.

They stipulated that the amount of follow up for the Stage 1 cohort should be fixed at the outset to avoid any risk of inflating the type I error rate.

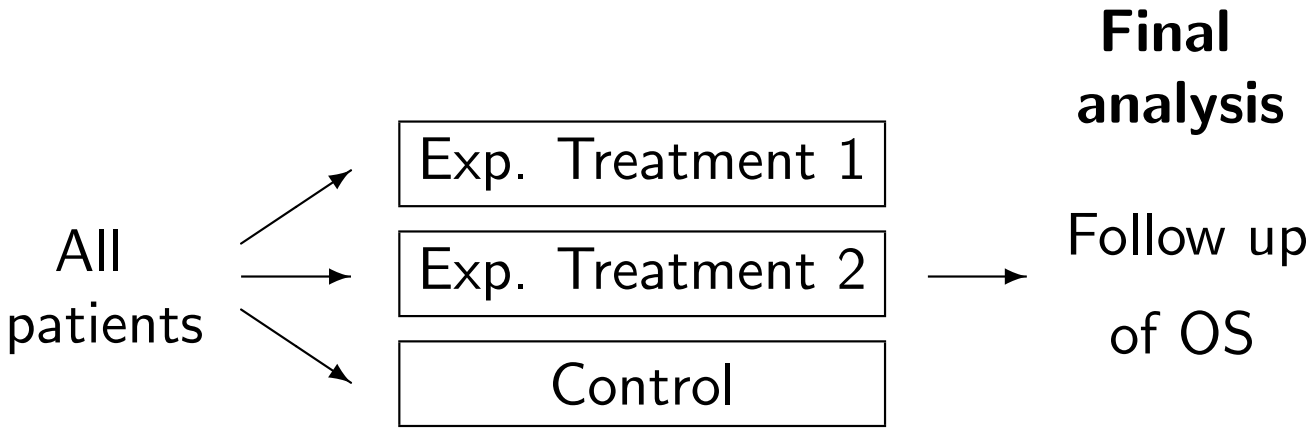
Some adaptive designs allow an early decision based on summaries of "Stage 1" data at an interim analysis.

In our three-treatment design, the statistics $Z_1^{(1)}$, $Z_2^{(1)}$ and $Z_{12}^{(1)}$ are not known at the time of the interim analysis, so we cannot define a formal stopping rule.

However, with only a little OS data available at the interim analysis, this is not a serious limitation.

Choosing an adaptive design and assessing its benefits

We compare the adaptive design with a non-adaptive trial in which randomisation is to both experimental treatments and control *throughout* the trial:



A closed testing procedure is used to control familywise error rate.

When the total numbers of patients and lengths of follow-up are the same in adaptive and non-adaptive designs,

Does the adaptive design provide higher power?

Are there other advantages?

Assessing the adaptive design: Model assumptions

Overall Survival

	Log hazard ratio
Exp Treatment 1 vs control	θ_1
Exp Treatment 2 vs control	θ_2

Logrank statistics are correlated due to the common control arm.

Progression Free Survival

	Log hazard ratio
Exp Treatment 1 vs control	ψ_1
Exp Treatment 2 vs control	ψ_2

Denote correlation between logrank statistics for OS and PFS by ρ.

In fact, hazard rates cannot be proportional for both endpoints.

However, it is the implications for the joint distribution of logrank statistics that matter, and it is convenient to describe these as if from two proportional hazards models.

Assessing the adaptive design: Model assumptions

Log hazard ratios for OS: $\theta_1,\ \theta_2$.

Log hazard ratios for PFS: $\psi_1,\ \psi_2$.

We suppose logrank statistics are distributed as if

$\psi_1 = \gamma \times \theta_1$ and $\psi_2 = \gamma \times \theta_2$

Final number of OS events for Stage 1 cohort $= 300$ (over 3 treatment arms)

Number of OS events for Stage 2 cohort $= 300$ (over 2 or 3 treatment arms)

Number of PFS events at interim analysis $= \lambda \times 300$.

When the log hazard ratio is θ, the standardised logrank statistic based on d observed events is, approximately, $N(\theta\sqrt{d/4}, 1)$.

Testing the intersection hypothesis H_{12}

We have null hypotheses H_1: $\theta_1 \leq 0$ and H_2: $\theta_2 \leq 0$.

In the closed testing procedure, we must also test

$$H_{12} = H_1 \cap H_2 : \ \theta_1 \leq 0 \ \text{ and } \ \theta_2 \leq 0.$$

We could test H_{12} by pooling the Exp Trt 1 and Exp Trt 2 patients and carrying out a logrank test vs the Control group.

Alternatively we could use a **Simes** test or a **Dunnett** test.

Simes' test:

Given observed values $p_1^{(1)}$ and $p_2^{(1)}$ of $P_1^{(1)}$ and $P_2^{(1)}$ in Stage 1, Simes' test of H_{12} yields the P-value

$$P_{12}^{(1)} = \min\left(2\min(p_1^{(1)}, p_2^{(1)}), \max(p_1^{(1)}, p_2^{(1)}) \right).$$

Simes' test protects type I error conservatively when $P_1^{(1)}$ and $P_2^{(1)}$ are independent or positively associated.

Dunnett's test of an intersection hypothesis

Dunnett's test for comparisons with a common control

Suppose $Z_1^{(1)}$ and $Z_2^{(1)}$ are the Stage 1 Z-values for logrank tests of Exp Trt 1 vs control and Exp Trt 2 vs Control.

If $z_1^{(1)}$ and $z_2^{(1)}$ are the observed values of $Z_1^{(1)}$ and $Z_2^{(1)}$, the Dunnett test of H_{12} yields the P-value

$$P_{12}^{(1)} = P(\max(Z_1^{(1)}, Z_2^{(1)}) \geq \max(z_1^{(1)}, z_2^{(1)}))$$

where $(Z_1^{(1)}, Z_2^{(1)})$ is bivariate normal with $Z_1^{(1)} \sim N(0,1)$, $Z_2^{(1)} \sim N(0,1)$ and $\text{Cov}(Z_1^{(1)}, Z_2^{(1)}) = 0.5$.

We shall see from comparisons of different methods that the Dunnett test of the intersection hypothesis leads to the most efficient versions of both adaptive and non-adaptive designs.

Comparing adaptive and non-adaptive trial designs

With selected values of ψ_1, θ_1, ψ_2, θ_2 and ρ, we simulate logrank statistics from their large sample distributions.

For the adaptive design, we define

$$P(1) = P(\text{Select Treatment 1 and Reject } H_1 \text{ overall})$$

$$P(2) = P(\text{Select Treatment 2 and Reject } H_2 \text{ overall})$$

For the non-adaptive design, we set

$$P(1) = P(\widehat{\theta}_1 > \widehat{\theta}_2 \text{ and } H_1 \text{ is rejected overall})$$

$$P(2) = P(\widehat{\theta}_2 > \widehat{\theta}_1 \text{ and } H_2 \text{ is rejected overall})$$

Hence, we define the overall expected "Gain" or utility measure

$$\mathbb{E}(\text{Gain}) = \theta_1 \times P(1) + \theta_2 \times P(2).$$

Comparing tests of the intersection hypothesis

Intersection tests produce $Z_{12}^{(1)}$ in an adaptive trial design with

$$\psi_1 = \theta_1, \quad \psi_2 = \theta_2, \quad \lambda = 1, \quad \rho = 0.6, \quad \alpha = 0.025.$$

		$P(1)$			$\mathbb{E}(\text{Gain})$		
θ_1	θ_2	Pooled	Simes	Dunnett	Pooled	Simes	Dunnett
0.3	0.0	0.77	0.85	0.86	0.232	0.254	0.259
0.3	0.1	0.78	0.81	0.82	0.238	0.245	0.247
0.3	0.2	0.68	0.68	0.69	0.238	0.237	0.238
0.3	0.25	0.58	0.58	0.58	0.250	0.249	0.249
0.3	0.295	0.48	0.47	0.47	0.275	0.274	0.274

All simulation results are based on 1,000,000 replicates.

The Dunnett test has the highest power. Unlike the pooled test, it is well aligned (consonant) with individual tests of H_1 and H_2.

Comparing adaptive and non-adaptive trial designs

We compare designs using a Dunnett test for H_{12} with

$$\psi_1 = \theta_1, \quad \psi_2 = \theta_2, \quad \lambda = 1, \quad \rho = 0.6, \quad \alpha = 0.025.$$

		Non-adaptive			Adaptive		
θ_1	θ_2	$P(1)$	$P(2)$	$\mathbb{E}$(Gain)	$P(1)$	$P(2)$	$\mathbb{E}$(Gain)
0.3	0.0	0.78	0.00	0.235	0.86	0.00	0.259
0.3	0.1	0.78	0.01	0.234	0.82	0.02	0.247
0.3	0.2	0.70	0.11	0.234	0.69	0.16	0.238
0.3	0.25	0.60	0.26	0.244	0.58	0.30	0.249
0.3	0.295	0.47	0.43	0.267	0.47	0.44	0.274

Here, $\lambda = 1$ implies there are 300 PFS events at the interim analysis.

The adaptive design has higher $P(1)$ when θ_1 is well above θ_2.

With θ_1 and θ_2 closer, the adaptive design still has higher $\mathbb{E}$(Gain).

Comparing adaptive and non-adaptive trial designs

The adaptive design can only succeed if there is adequate information to select the correct treatment at the interim analysis:

Treatment effects on PFS should be be reliable indicators of treatment effects on OS,

There must be good information on PFS at the interim analysis.

We have investigated varying the parameters γ and λ where

$\psi_1 = \gamma \times \theta_1,\ \psi_2 = \gamma \times \theta_2,$ with $\theta_1 = 0.3$ and $\theta_2 = 0.1$

Final number of OS events for Stage 1 cohort $= 300$ (over 3 arms)

Number of OS events for Stage 2 cohort $= 300$ (over 2 or 3 arms)

Number of PFS events at interim analysis $= \lambda \times 300.$

NB It is quite plausible that γ should be greater than 1, i.e., a larger treatment effect on PFS than on OS.

Comparing adaptive and non-adaptive trial designs

We compare designs with $\theta_1 = 0.3,\ \theta_2 = 0.1,\ \rho = 0.6,\ \alpha = 0.025,$

PFS log hazard ratios: $\psi_1 = \gamma\,\theta_1,\ \psi_2 = \gamma\,\theta_2,$

Number of PFS events at interim analysis $= \lambda \times 300.$

		Non-adaptive			Adaptive		
γ	λ	$P(1)$	$P(2)$	$\mathbb{E}$(Gain)	$P(1)$	$P(2)$	$\mathbb{E}$(Gain)
1.5	1.2				0.88	0.00	0.264
1.2	1.1				0.85	0.01	0.256
1.0	**1.0**	**0.78**	**0.01**	**0.234**	**0.82**	**0.02**	**0.247**
0.9	0.9	for all γ and λ			0.78	0.03	0.238
0.8	0.8	(PFS is not used)			0.74	0.04	0.225
0.7	0.7				0.68	0.05	0.208

Adaptation works well when there is enough PFS information for treatment selection at the interim analysis.

Related work

1. Friede et al. (*Statistics in Medicine*, 2011) consider a seamless phase II/III trial design with treatment selection based on both short-term and long-term responses.

They give an example of a trial comparing treatments for multiple sclerosis. When the treatment selection decision is made, only a short-term response is available for some subjects but these patients will go on to provide a long-term response later.

As for a survival endpoint, follow-up of patients on the selected treatment is likely to produce results that are biased towards a positive treatment effect, since the treatment selection decision was based on promising short-term response data.

Friede et al. follow a similar approach to Jenkins, Stone & Jennison (2011) and apply a combination test to the long-term response data from the *cohorts* of patients admitted before and after the interim decision point.

Related work

2. Irle & Schäfer (*JASA*, 2012) propose similar adaptive designs for survival data.

Changes to the design and critical values for test statistics preserve the conditional probability of rejecting a null hypothesis.

As the "Conditional Probability of Rejection" principle is related to combination tests, the method has much in common with that of Jenkins, Stone & Jennison (2011).

Irle & Schäfer's method imposes the same requirement of a fixed length of follow-up for "Cohort 1" patients.

Determining the conditional probability of a future event can be problematic, since the final information level (in a log-rank statistic, say) is not known at the time this probability is calculated.

We recommend our combination test approach as simpler to explain and easier to implement.

Conclusions about the benefits of the adaptive design

- The adaptive design offers the chance to select the better treatment and focus on this in the second stage of the trial.
- Overall, adaptation is beneficial as long as there is sufficient information to make a reliable treatment selection decision.
- Other evidence may be used in reaching this decision:

  Safety data

  Pharmacokinetic data

  Overall survival

- In addition to reaching a final decision, both non-adaptive and adaptive trials compare the two forms of treatment: the conclusions from this comparison may be more broadly useful.

Recapitulation: Adaptive clinical trial designs

- It is desirable to adapt a clinical trial design as information becomes available on parameters that were initially unknown.
- Methods are available to create adaptive designs that will protect the overall type I error rate.
- Combination tests allow results from different stages of the trial to be merged.
- Closed Testing Procedures allow tests of multiple hypotheses, or of a single hypothesis selected in a data-dependent manner.
- It should not be assumed that introducing adaptation will automatically make a trial design more efficient.
- Critical appraisal of trial designs is crucial and, where feasible, it is advisable to define an objective function and optimise for this criterion within a chosen class of designs.