Group Sequential and Adaptive Methods in the Design of Clinical Trials

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

https://cjennison1.github.io

Instituts für Medizinische Statistik

Universitätsmedizin Göttingen

November 2026

©2026 Jennison, Turnbull

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

There is 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) or Independent Data Monitoring Committee (IDMC) 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 the course

1. **Group sequential tests**

   Basic theory and computation

   Error spending tests

   Examples of group sequential designs:

      With a normal response,

      With a survival endpoint

   Futility boundaries: conditional and predictive power

   Group sequential tests with a delayed response

   Bayes group sequential tests

   Inference on termination of a group sequential trial

   Estimation and testing for a secondary endpoint

Outline of the course

2. Multiple testing procedures

Introduction to multiple testing

Closed testing procedures

Graphical representation of multiple testing procedures

Combining multiple testing and group sequential tests

Application to testing a secondary endpoint after a group sequential test

Outline of the course

3. Adaptive clinical trial designs with a single hypothesis

- Combination tests
- The conditional probability of rejection principle
- Sample size re-estimation:
  - For a nuisance parameter
  - For an estimated treatment effect
  - The promising zone approach
- Inference on termination of an adaptive trial

Outline of the course

4. Adaptive clinical trial designs with multiple hypotheses

- Combining multiple testing and combination tests
- Applications
  - Seamless Phase 2/3 designs
  - Multi-armed multi-stage (MAMS) designs
  - Enrichment designs
- Inference on termination of an adaptive trial

Reference

Group Sequential and Adaptive Methods for Clinical Trials (2025), by C. Jennison & B.W. Turnbull, CRC Press. (Hereafter, JT)

Part 1. Group sequential tests

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 survival endpoint

1.7. Futility boundaries: conditional and predictive power

1.8. Group sequential tests with a delayed response

1.9. Bayes group sequential tests

1.10. Inference on termination of a group sequential trial

1.11. Estimation and testing for a secondary 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

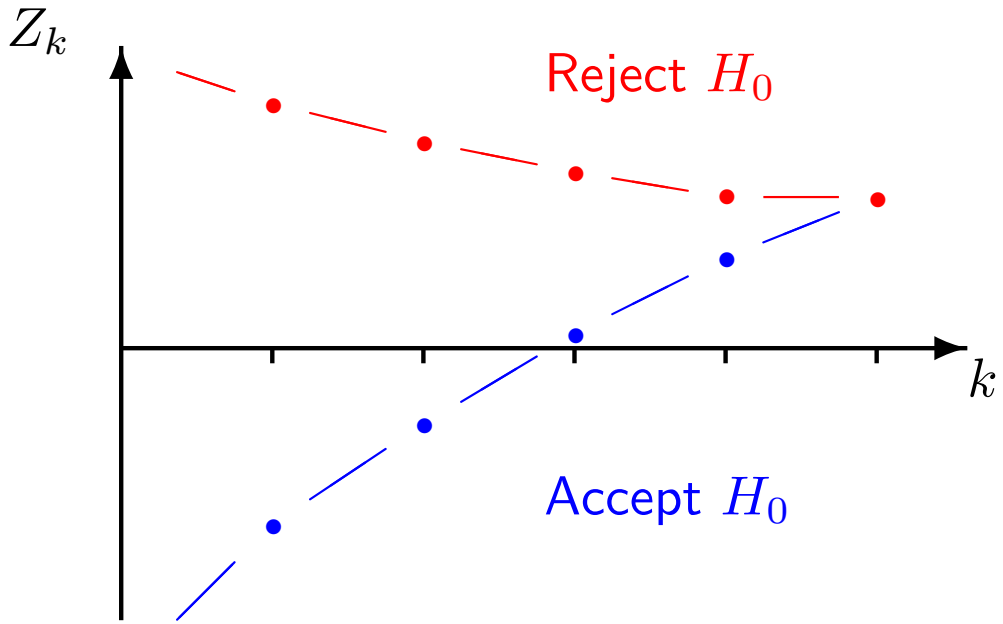
$$P_{\theta=0}\{\text{Reject } H_0\} = \alpha,$$

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

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, $b_1, \ldots, b_K$ leads to early stopping for a positive outcome, rejecting H_0 in favour of $\theta > 0$.

Crossing the lower boundary $a_1, \ldots, a_K$ 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: JT, Chapters 3 and 15.

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 $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$

Example: A single normal mean and 2 analyses

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 $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$

Example: A two-treatment comparison and K analyses

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 $\text{Cov}(\widehat{\theta}_{k_1}, \widehat{\theta}_{k_2}) = \text{Var}(\widehat{\theta}_{k_2})$ 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.

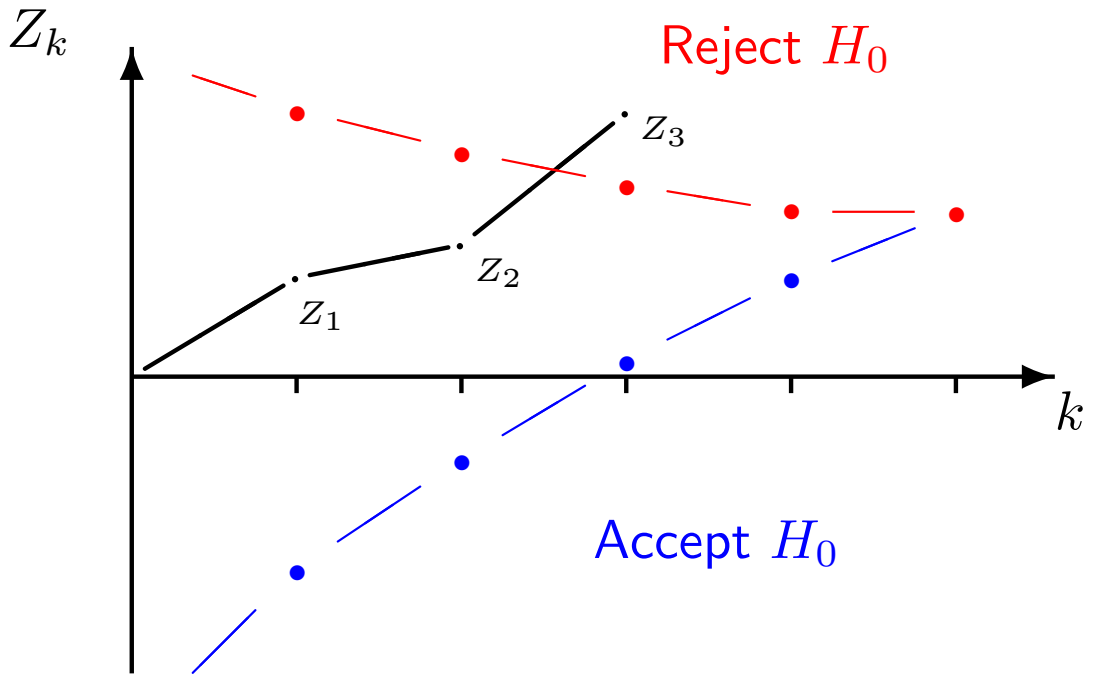
For survival data, observed information is roughly proportional to the number of failures.

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)

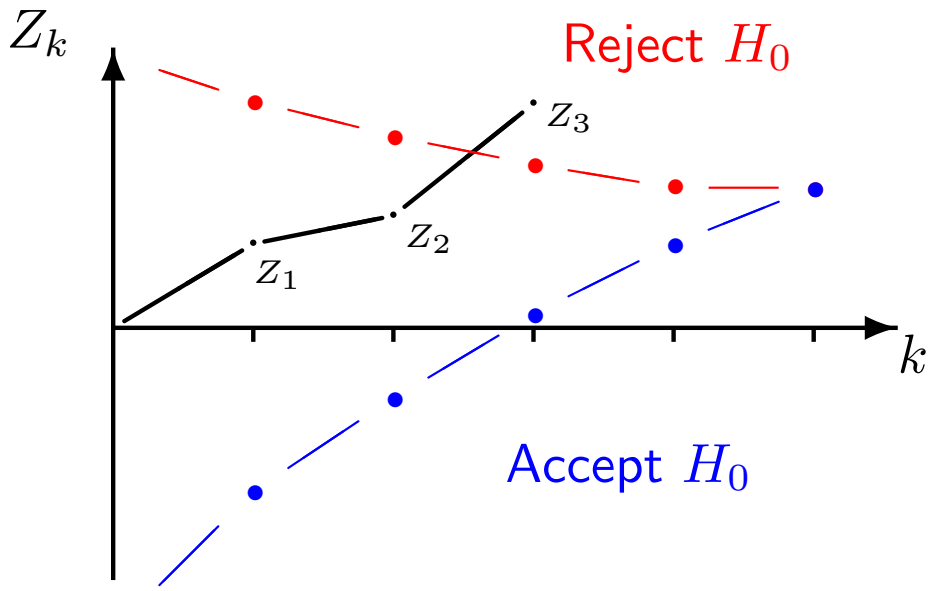
Reference: JT, Section 3.5 and Chapter 31.



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



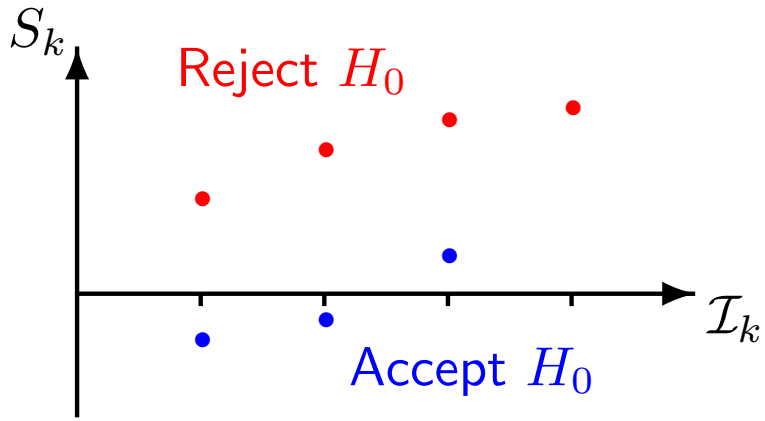
Probabilities such as $P_\theta\{a_1 < Z_1 < b_1,\ \ a_2 < Z_2 < b_2,\ \ Z_3 > b_3\}$ can be computed by repeated numerical integration.

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 Pampallona & Tsiatis test with parameter Δ, boundaries on the score statistic scale are

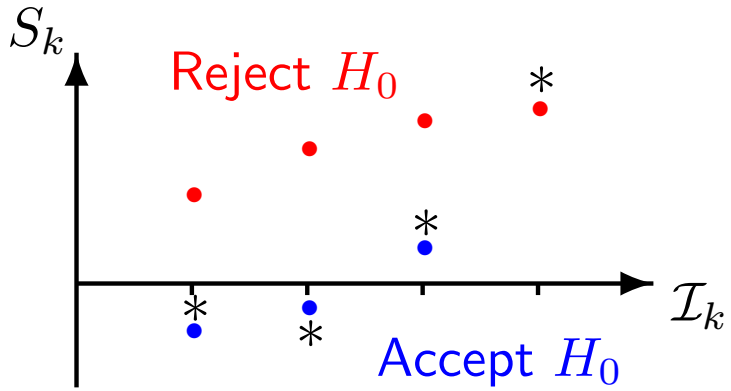
$$a_k = \mathcal{I}_k \, \delta - C_2 \, \mathcal{I}_k^{\Delta}, \quad b_k = C_1 \, \mathcal{I}_k^{\Delta}.$$

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

Reference: JT, Chapter 12.

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.

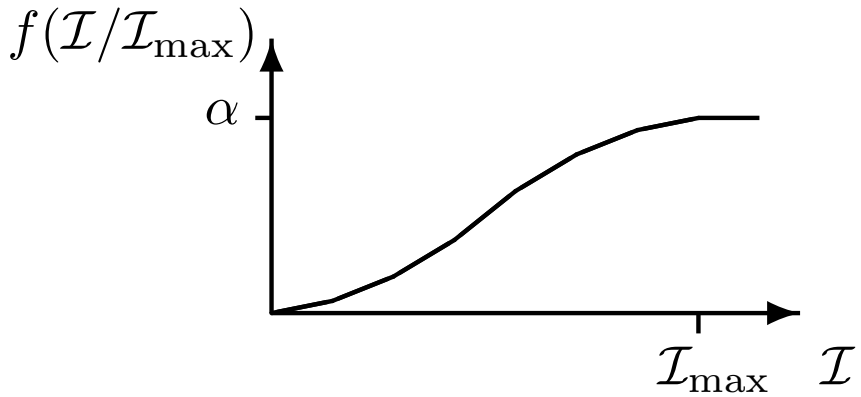
1.5 Error spending tests

Reference: JT, Chapter 8.

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.

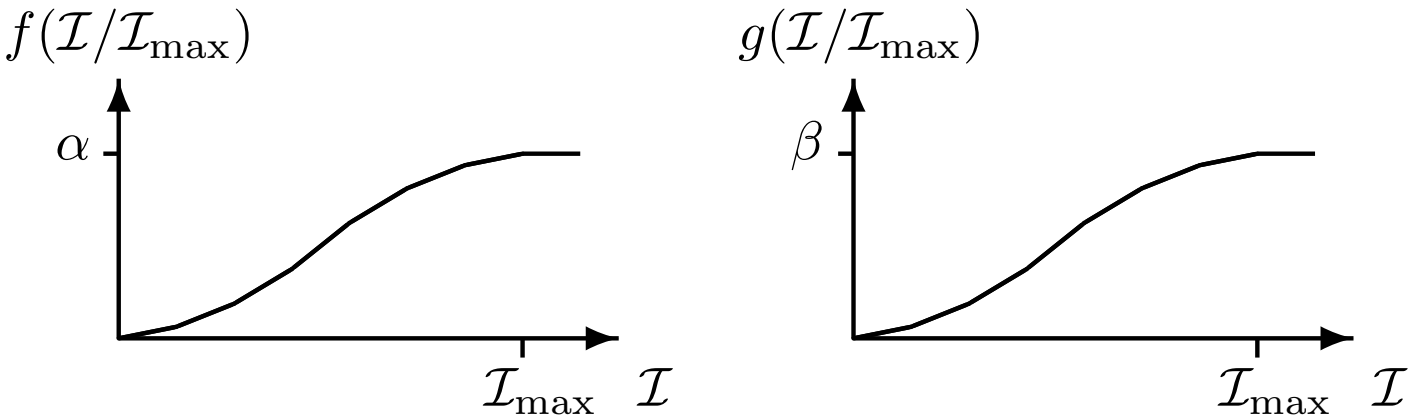
One-sided error spending tests

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:

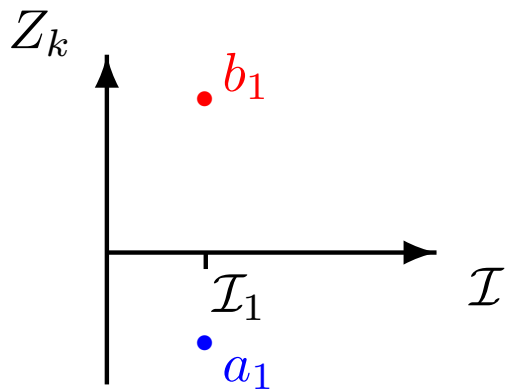
Observed information $\mathcal{I}_1$.

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 $\mathcal{I}_2$

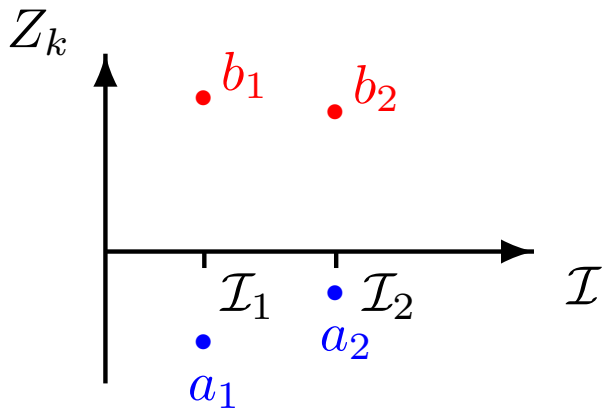
Reject H_0 if $Z_2 > b_2$, where

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

— *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

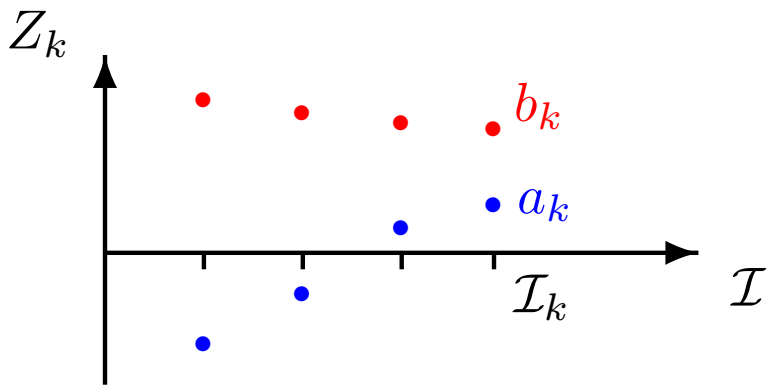
Analysis k: Observed information $\mathcal{I}_k$

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

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

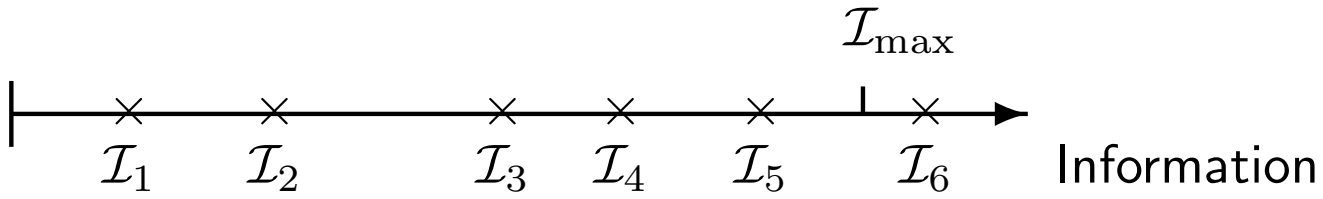


Remarks on error spending tests

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

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

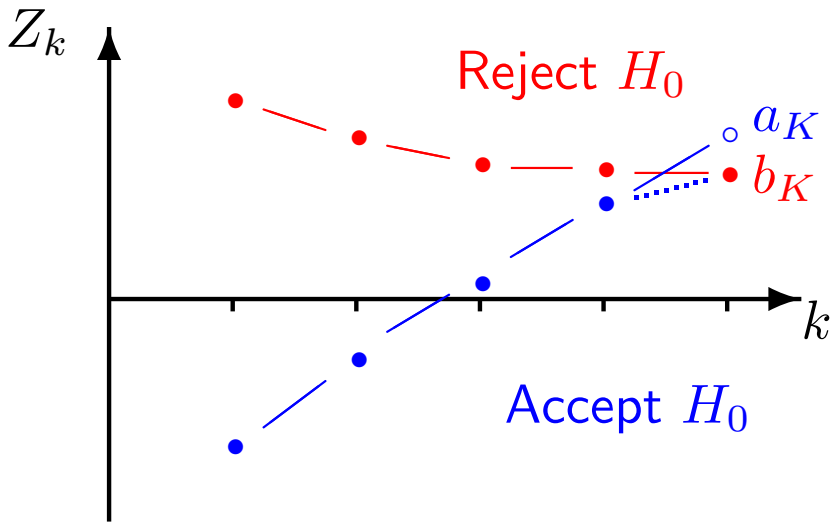
The value of ρ determines the inflation factor $R = \mathcal{I}_{\max}/\mathcal{I}_{fix}$.

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 $a_K > b_K$.



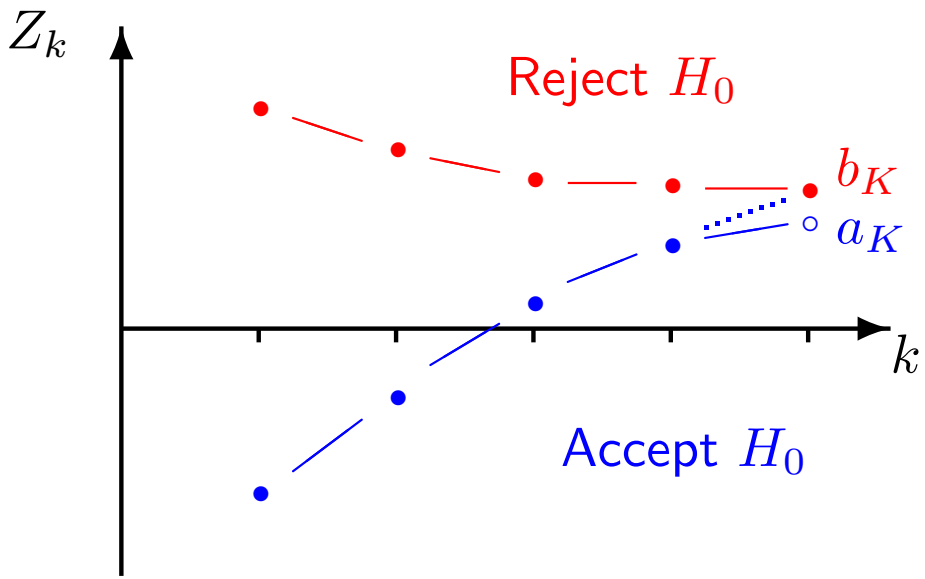
The value calculated for b_K guarantees type I error probability equal to α. So, reduce a_K to b_K — and gain extra power.

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

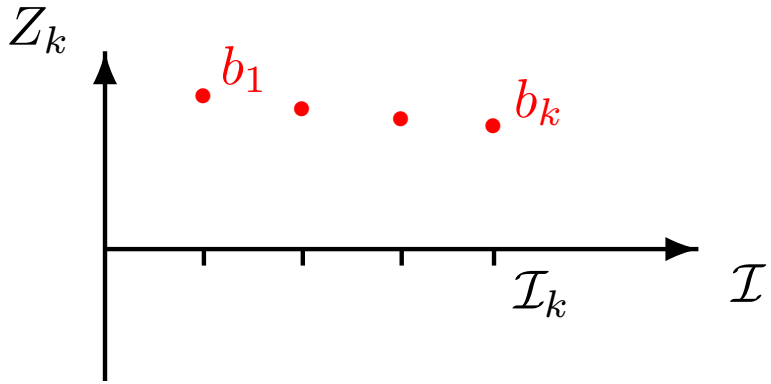
If the futility boundary is treated as non-binding, computation of the efficacy boundary only involves the type I error spending function $f(\mathcal{I}/\mathcal{I}_{\max})$.

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

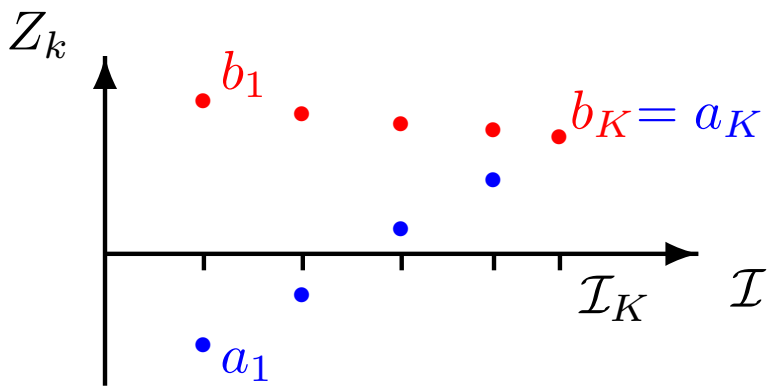
A futility boundary can be added through a type II error spending function $g(\mathcal{I}/\mathcal{I}_{\max})$.

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 $R = 1.133$ (see JT, Table 8.12).

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	*Cumulative sample size*	$\mathcal{I}_k$	*Boundary*	
k	$n_A + n_B$		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

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

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 $T_k < \tilde{a}_k$.

Software Demonstration: Normal data

In the Practical Session we shall suppose 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 survival data

Example: Oropharynx Clinical Trial Data

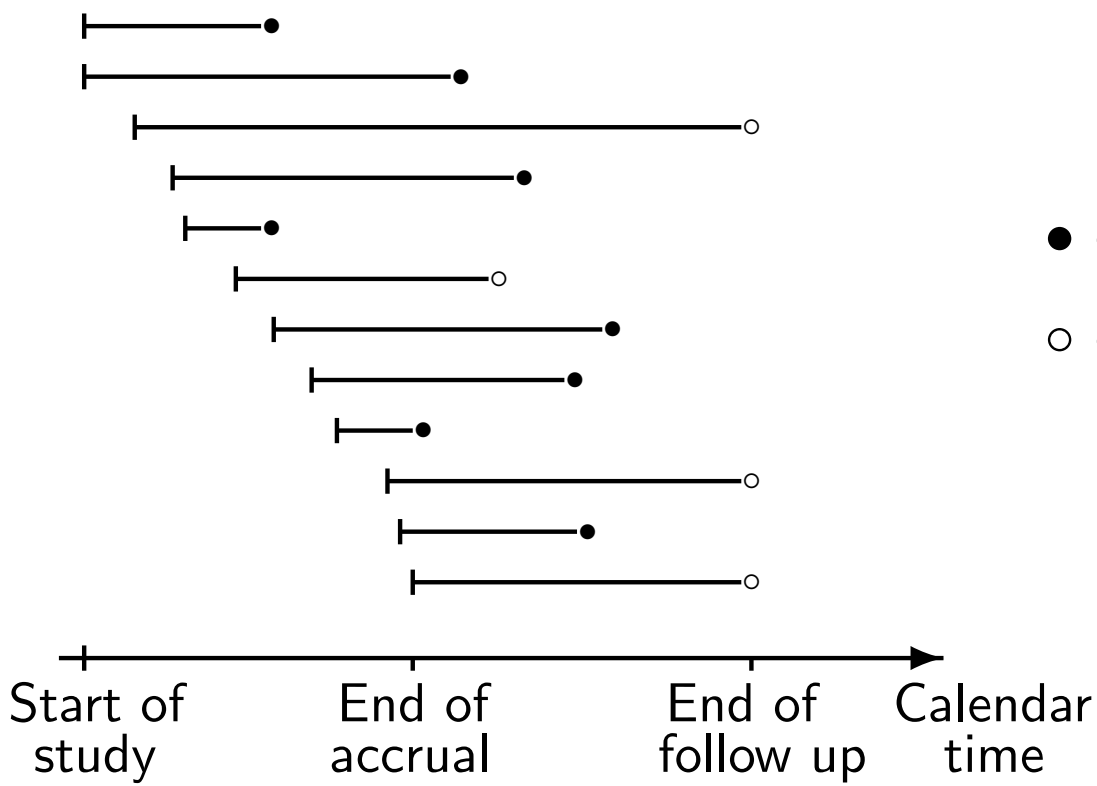
Reference: JT, Section 17.6

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.

Accrual and follow up in a survival study



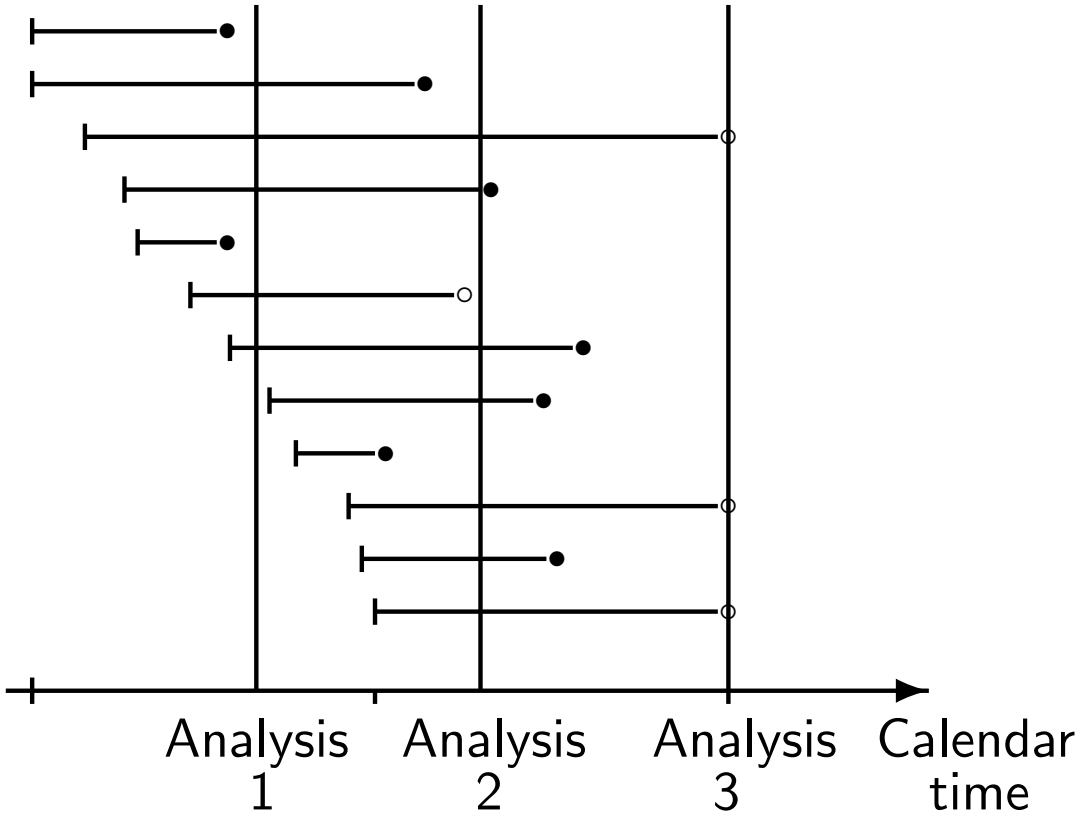
Key

- • death time observed
- ○ censored observation

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

Survival is measured from entry to the study.

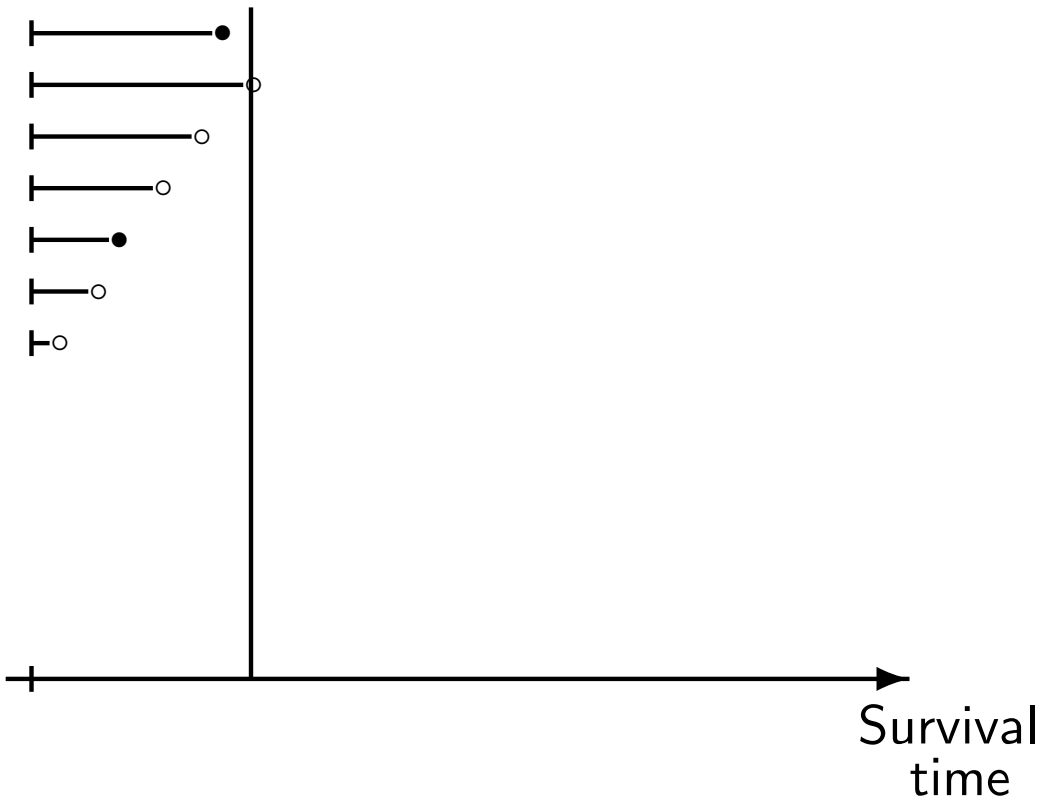
Interim analyses



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

Information on such patients continues to accrue at later analyses.

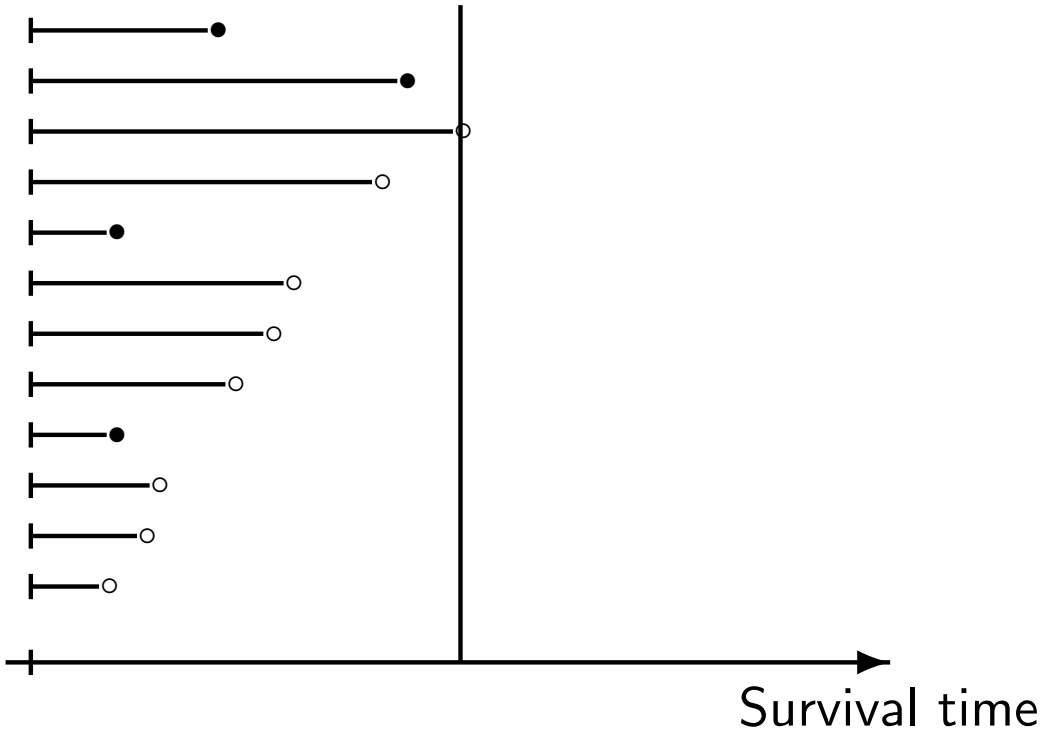
Interim analysis 1



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

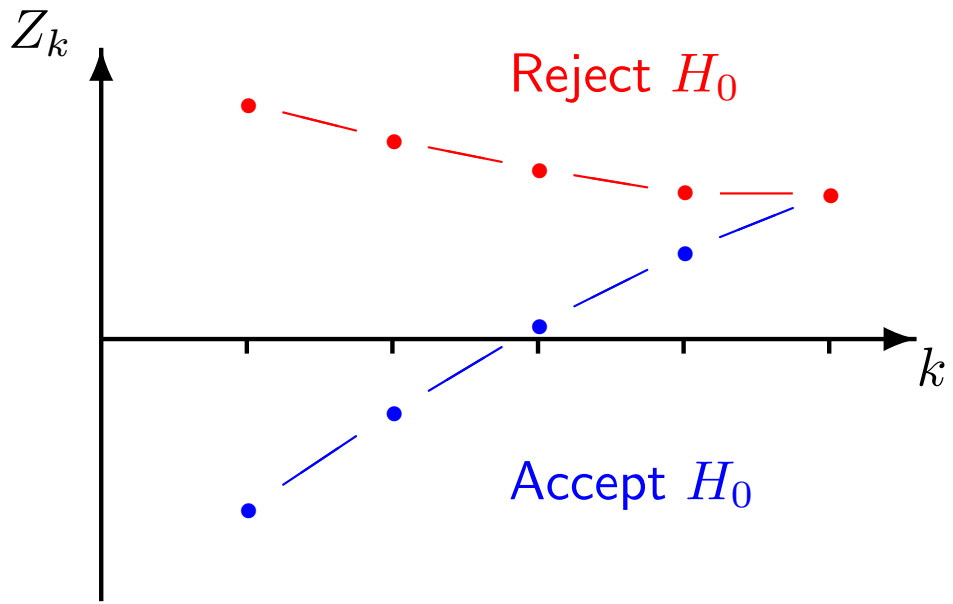
$$\mathcal{I}_{fix} = \frac{\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2}{0.5^2} = 31.40.$$

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 $\theta = \ln(\lambda) = 0.5$,

Non-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 $R = 1.133$.

So, we require $\mathcal{I}_{\max} = 1.133 \times 31.40 = 35.58$, which needs a total of $4 \times 35.58 \approx 142$ observed deaths.

A one-sided, error spending design

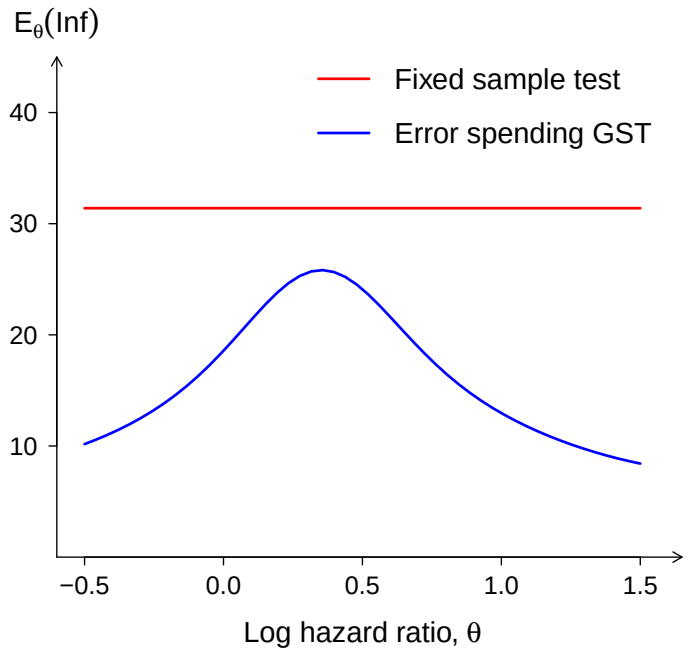
Suppose that, as assumed when planning the trial, information levels are equally spaced up to $\mathcal{I}_5 = \mathcal{I}_{\max} = 35.58$.

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	7.12	-1.08	3.09
2	14.23	-0.02	2.71
3	21.35	0.76	2.47
4	28.46	1.43	2.28
5	35.58	2.11	2.11

A one-sided, error spending design

If information levels $\mathcal{I}_k = (k/5)\,35.58,\ k = 1, \ldots, 5,$ are observed, and the futility boundary is applied consistently, 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	-1.44	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

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

An error spending test with a binding futility boundary

If a **binding** futility boundary had been used, the required maximum information level would have been a little lower at 34.48.

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	-1.41	3.23	-1.04
2	126	58	12.58	-0.20	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	2.06	2.06	-0.87

Again, the trial would have terminated at the second analysis 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	2.05	2.05	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.

Software Demonstration: Survival data

In the Practical Session we shall suppose 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.

RPACT Demonstration: Normal data

Suppose 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.

RPACT Demonstration: Survival data

Suppose 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.

1.7 Futility boundaries

Group sequential designs allow a clinical trial to be terminated

For efficacy:

When there is overwhelming evidence that the new treatment is effective,

For futility:

When it is clear the trial is unlikely to reach a positive conclusion.

In retrospective analyses of 72 ECOG cancer studies, Rosner & Tsiatis (*Statist. in Med.*, 1989) found that, had group sequential stopping rules been applied, early stopping (mostly to accept H_0) would have occurred in around 80% of cases.

Many clinical trials have a formal stopping rule for efficacy but informal guidelines for futility stopping. Why should these issues be treated differently — and is this a good idea?

Defining a stopping rule for efficacy

Consider a Phase 3 clinical trial comparing a new treatment against a standard.

Let θ denote the "effect size", a measure of the improvement in the new treatment over the standard.

We shall test the null hypothesis H_0: $\theta \leq 0$ against $\theta > 0$.

Rejecting H_0 allows us to conclude the new treatment is superior.

We allow type I error probability α for rejecting H_0 when it is true.

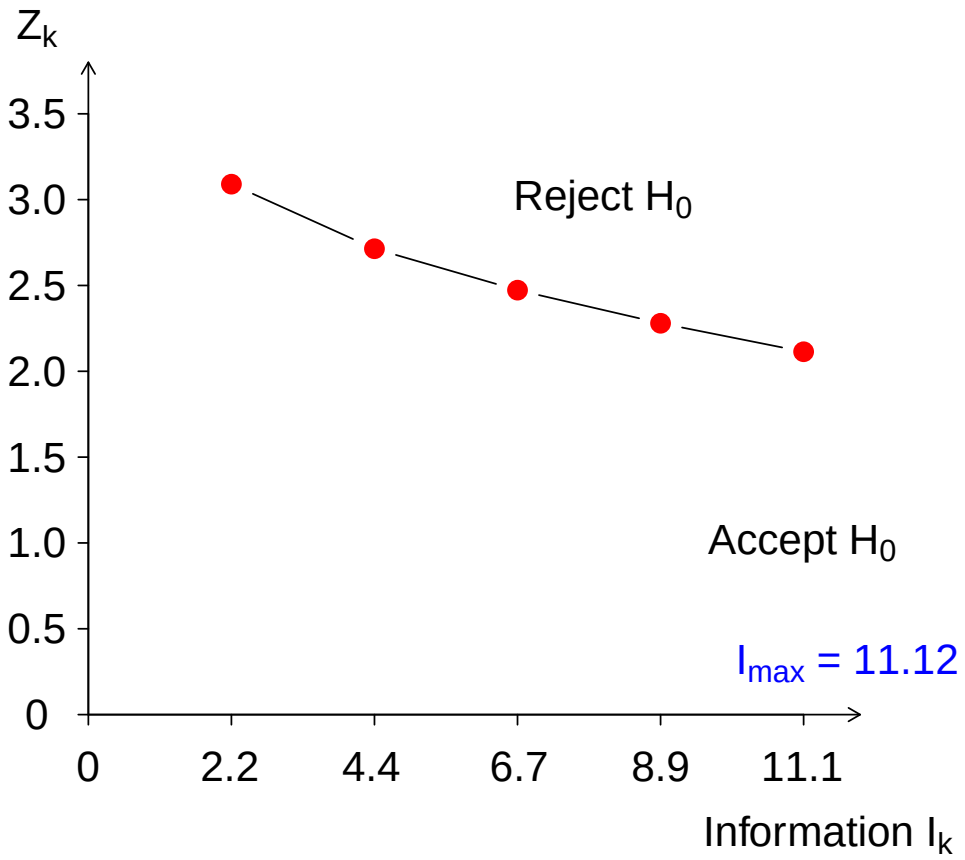
We specify power $1 - \beta$ as the probability that H_0 should be rejected when $\theta = \delta$.

Here δ is, typically, the minimal clinically significant treatment difference.

Example: A trial design with an efficacy boundary only

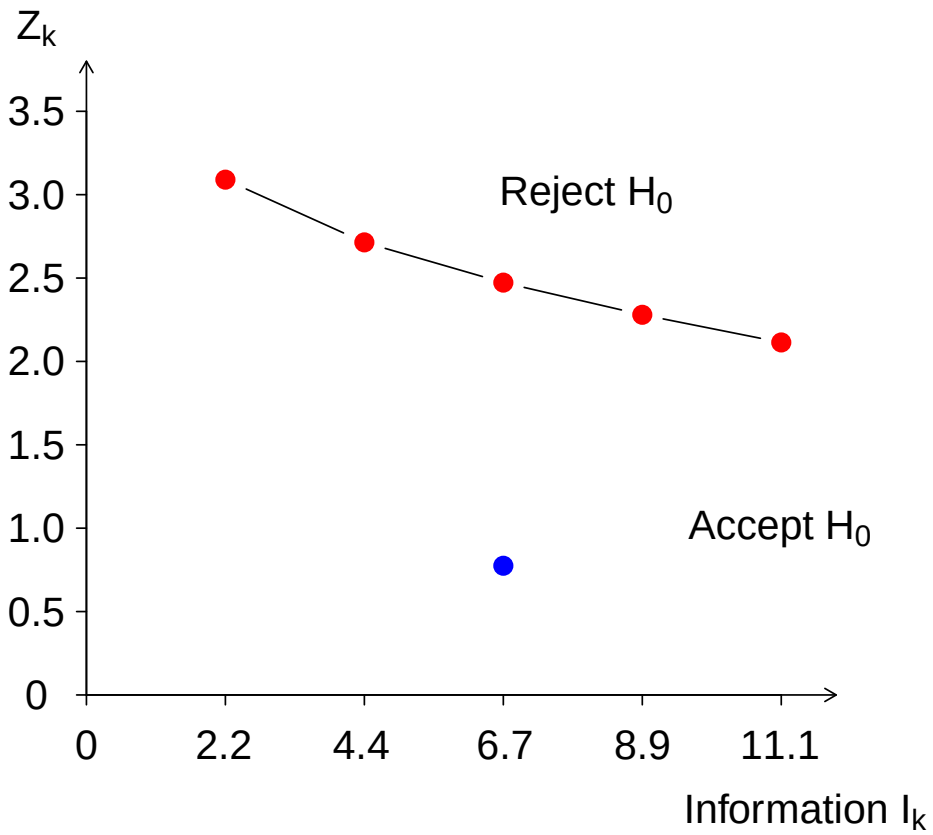
A test with 5 planned analyses, type I error probability $\alpha = 0.025$, power 0.9 if $\theta = \delta = 1$, and type I error spending function

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



Using conditional power to guide stopping for futility

Suppose the trial has reached analysis 3, $\hat{\theta}_3 = 0.3$ and $Z_3 = 0.78$.



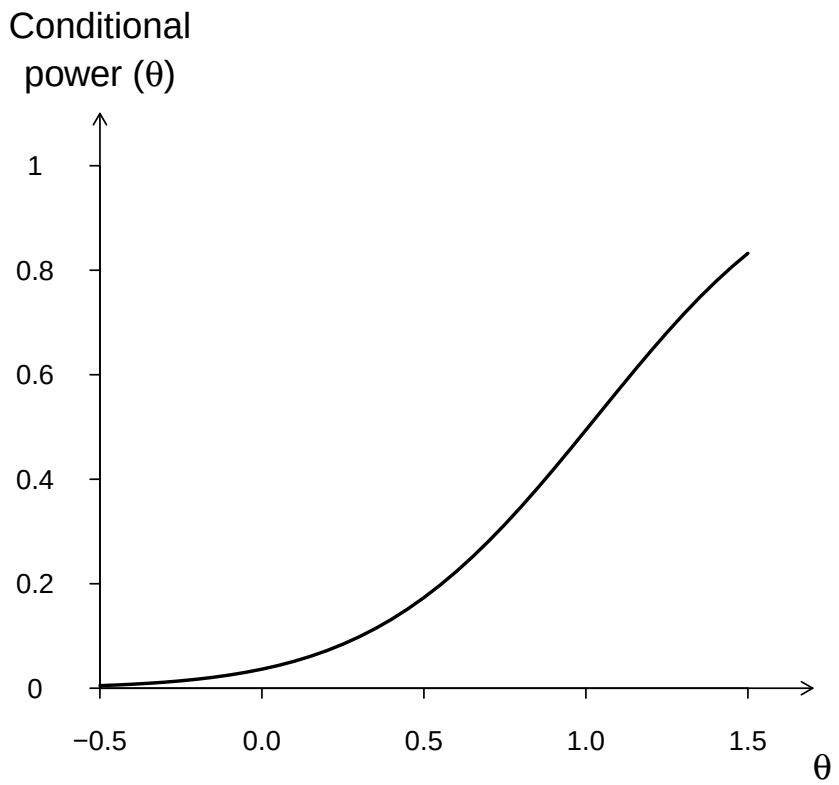
One may ask

"How likely is it that the trial's final outcome will be positive?"

Using conditional power to guide stopping for futility

We can compute the "conditional power function",

$$P_\theta\{H_0 \text{ will be rejected at analysis 4 or 5} \mid Z_3 = 0.78\}.$$

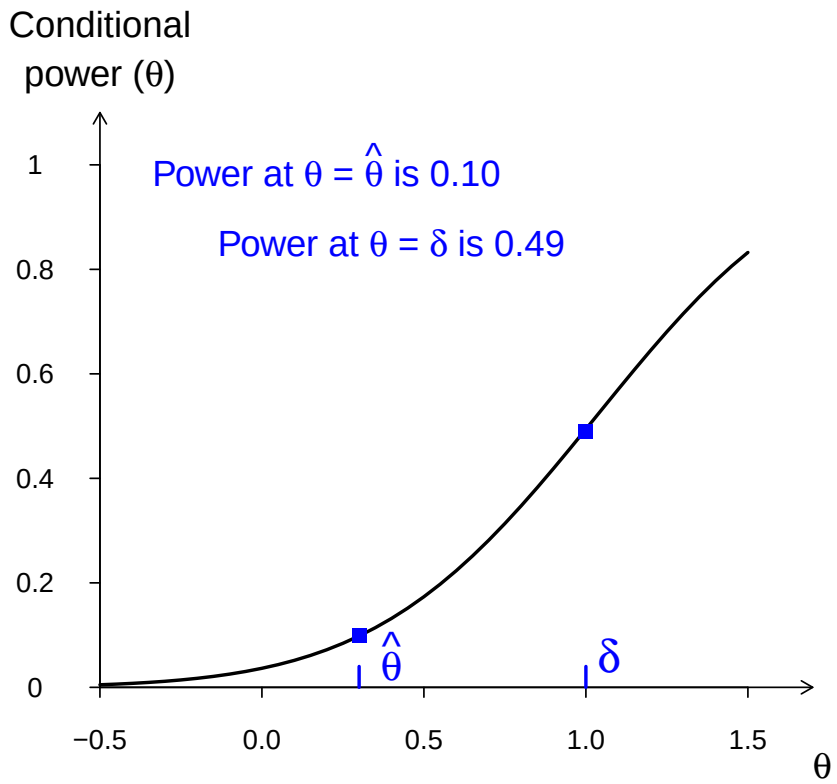


However, we do not know the true value of θ.

Using conditional power to guide stopping for futility

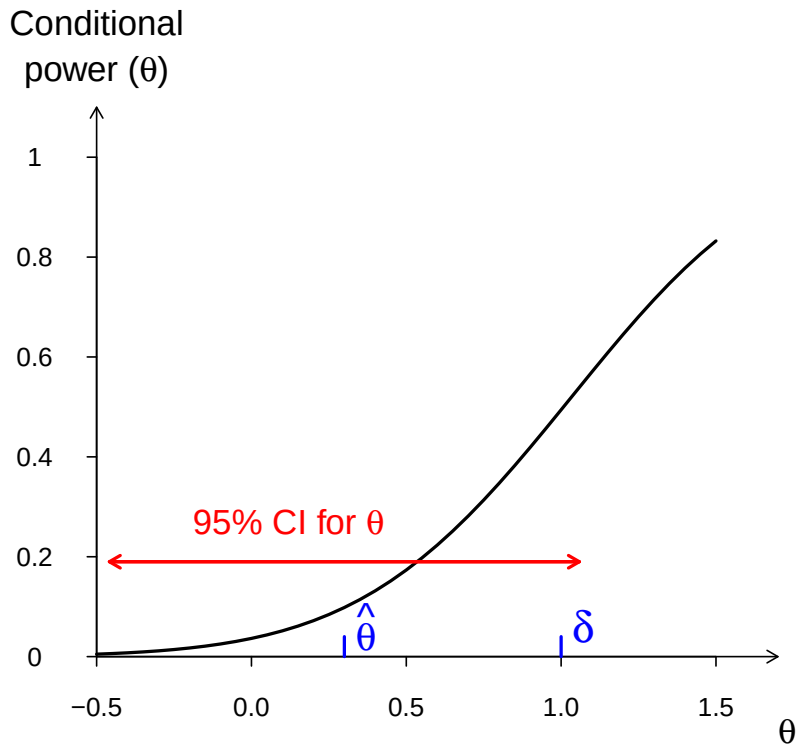
One may focus on the conditional power assuming θ is equal to the current estimate, $\hat{\theta}_3 = 0.3$.

Or, one may focus on conditional power assuming $\theta = \delta = 1$, the value used in the power calculation.



Using conditional power to guide stopping for futility

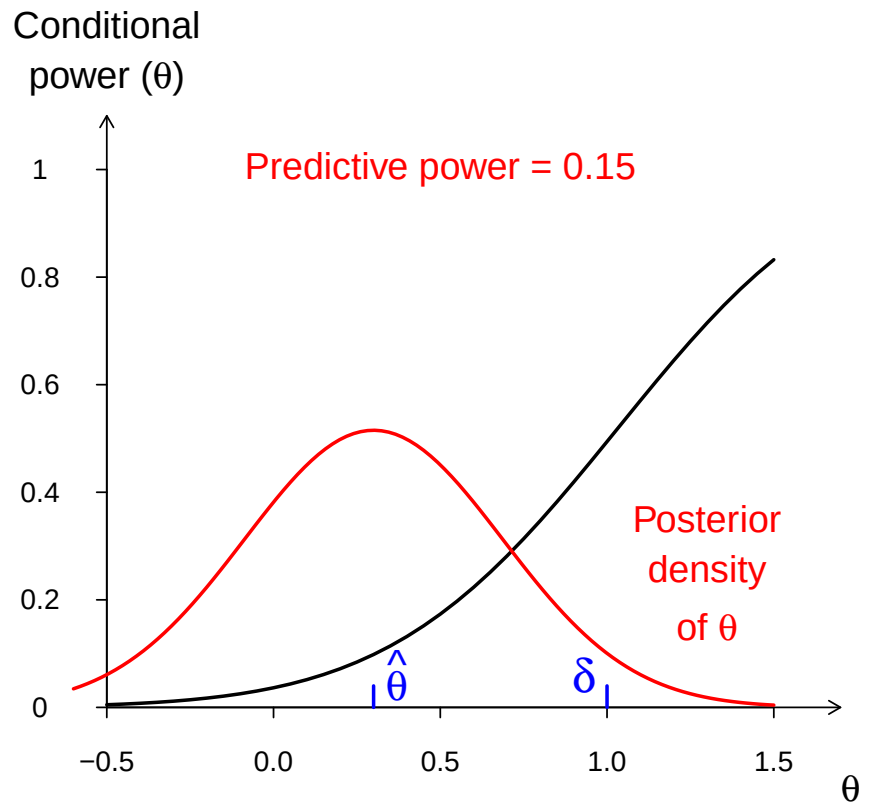
It is important to remember that the estimate of θ at an interim analysis has a high variance.



Adopting a Bayesian approach, one can integrate conditional power over a posterior distribution to obtain a "predictive power".

Using predictive power to guide stopping for futility

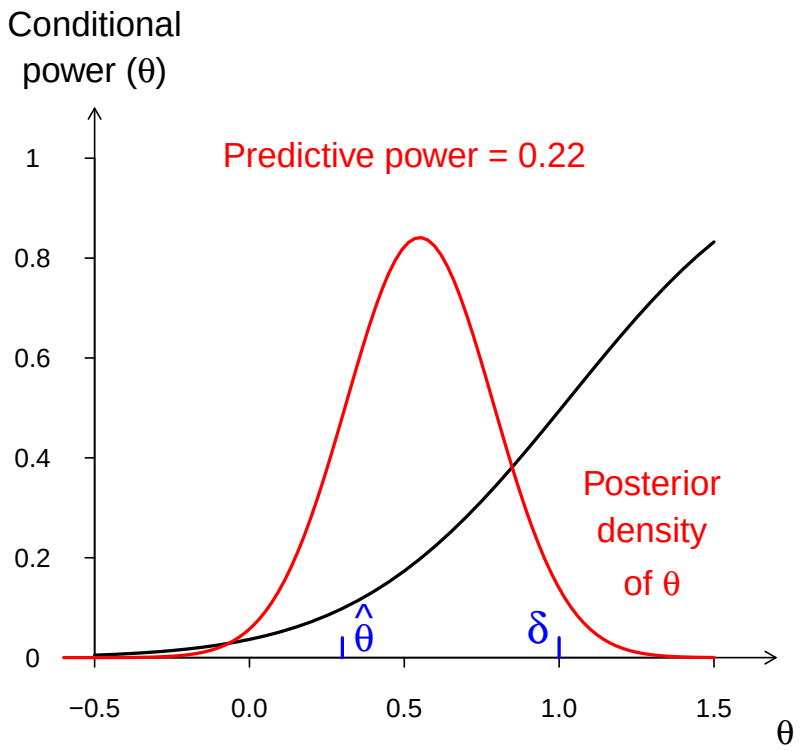
A common practice is to assume a flat (improper) prior for θ in calculating predictive power.



In this case the posterior distribution of θ is $N(0.3,\, 0.39^2)$.

Using predictive power to guide stopping for futility

Given the high variance of the interim estimate $\hat{\theta}_3$, the choice of prior can have a significant impact on the predictive power.



Under the prior $\theta \sim N(0.7,\, 0.3^2)$, the posterior distribution of θ is $\theta \mid \hat{\theta}_3 \sim N(0.55,\, 0.24^2)$, and predictive power rises to 0.22.

Using predictive power to guide stopping for futility

Once you have calculated your chosen conditional power or predictive power, the question remains:

How high should the conditional probability of success be to justify continuation of the trial?

Here, one needs to balance

- The benefits from saving resources in this study and moving on to conduct trials for other promising therapies,
- The risk of stopping the current trial prematurely when it would have gone on to produce a positive result.

Decision making is hard when conditional power is low but there is still a non-negligible chance the trial may still succeed.

Error spending tests with efficacy and futility boundaries

Recall, we want a group sequential test of H_0: $\theta \leq 0$ vs $\theta > 0$ with

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

$$P_{\theta=\delta}\{\text{Accept } H_0\} = \beta,$$

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

If we specify α, β, δ, K and $\mathcal{I}_{\max}$, we can find the stopping rule that minimises a criterion such as

$$\sum_i w_i E_{\theta_i}(\mathcal{I}) \quad \text{or} \quad \int w(\theta) E_\theta(\mathcal{I})\, d\theta.$$

See:

Barber & Jennison (*Biometrika*, 2002),

Öhrn (*PhD thesis, University of Bath*, 2011),

Jennison & Turnbull (*Kuwait J. Science*, 2013).

Error spending tests with efficacy and futility boundaries

Barber & Jennison (2002) and Öhrn (2011) observe that group sequential tests with error spending functions of the form

$$f(\mathcal{I}) = \min\{(\mathcal{I}/\mathcal{I}_{\max})^{\rho_1}, 1\}\,\alpha \quad \text{(type I error)}$$

and

$$g(\mathcal{I}) = \min\{(\mathcal{I}/\mathcal{I}_{\max})^{\rho_2}, 1\}\,\beta \quad \text{(type II error)}$$

have high efficiency for a variety of optimality criteria.

The values of ρ_1 and ρ_2 determine $\mathcal{I}_{\max}$ and, hence, the trial's maximum sample size.

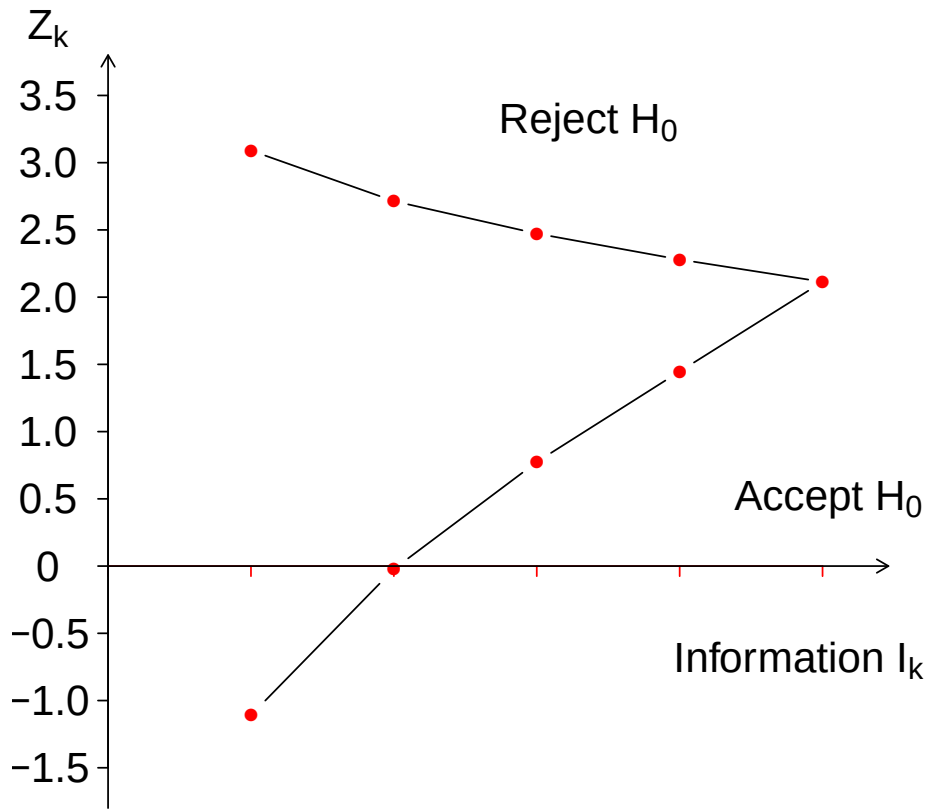
The resulting designs are efficient for this value of $\mathcal{I}_{\max}$.

The monitoring committee can treat the (non-binding) futility boundary as a guideline, allowing them to consider safety data or secondary endpoints in deciding whether to stop for futility.

An efficacy boundary and non-binding futility boundary

A test with 5 planned analyses, type I error probability $\alpha = 0.025$, power 0.9 if $\theta = \delta = 1$, and type I and II error spending functions

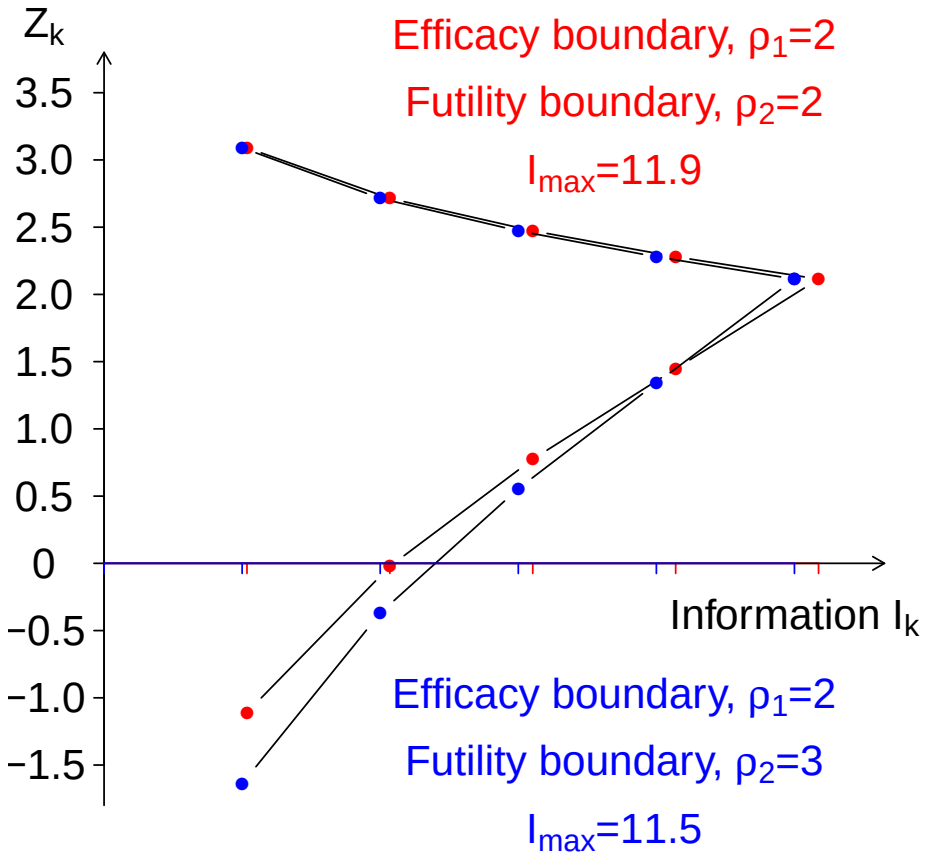
$$f(\mathcal{I}) = \min\{(\mathcal{I}/\mathcal{I}_{\max})^2, 1\}\,\alpha, \quad g(\mathcal{I}) = \min\{(\mathcal{I}/\mathcal{I}_{\max})^2, 1\}\,\beta.$$



An efficacy boundary and non-binding futility boundary

Contrast: A test with type I and type II error spending functions

$$f(\mathcal{I}) = \min\{(\mathcal{I}/\mathcal{I}_{\max})^2, 1\}\,\alpha, \quad g(\mathcal{I}) = \min\{(\mathcal{I}/\mathcal{I}_{\max})^3, 1\}\,\beta.$$



Resources saved by early stopping

Tests with $\alpha = 0.025$, $1 - \beta = 0.9$ and 5 equally spaced analyses.

Values of $\mathcal{I}_{\max}$ and $E_\theta(\mathcal{I})$, expressed as a percentage of $\mathcal{I}_{fix}$.

		$E_\theta(\mathcal{I})$			
Design	$\mathcal{I}_{\max}$	$\theta=0$	$\theta=0.5$	$\theta=1.0$	$\theta=1.5$
E: $\rho_1 = 2$ only	106	105.2	96.7	70.5	46.8
E: $\rho_1 = 2$, F: $\rho_2 = 2$	113	59.2	80.9	70.6	48.2
E: $\rho_1 = 2$, F: $\rho_2 = 3$	109	64.1	83.0	70.1	47.5

E: Efficacy boundary, F: Non-binding futility boundary

With $\rho_2 = 3$, the maximum sample size is smaller and expected sample size at low values of θ is larger.

The number of analyses and design parameters ρ_1 and ρ_2 can be chosen to give acceptable values of $\mathcal{I}_{\max}$ and $E_\theta(\mathcal{I})$.

Conclusions for futility stopping

1. It is common practice for trials to take an informal approach to stopping for futility and decision making is often guided by conditional power calculations.

Just how one should use "conditional power" or "predictive power" in deciding whether to stop a trial is not very clear.

2. We can create a group sequential design with a non-binding futility boundary: the ρ-family of error spending designs provides a simple way to define efficient procedures.

If the monitoring committee wishes, this futility boundary can be expressed in terms of conditional power or predictive power (see examples in JT, Table 11.2).

The monitoring committee should be guided by this futility boundary — but they can still use their discretion in deciding whether to recommend stopping the trial.

1.8 Group sequential tests with a delayed response

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

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: H&J 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.

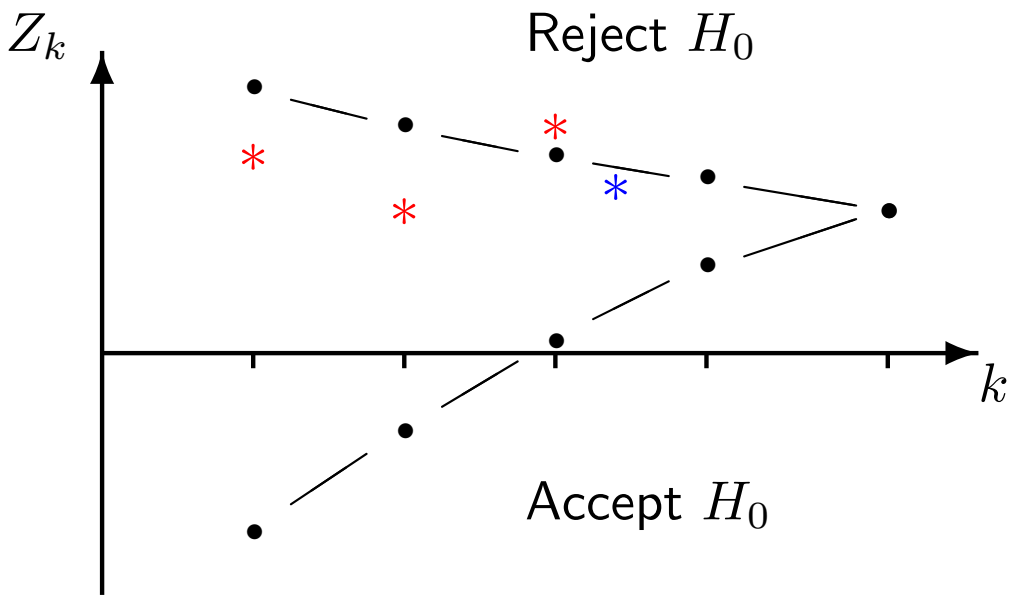
Example 2: Consider a clinical trial with a time-to-event endpoint.

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.

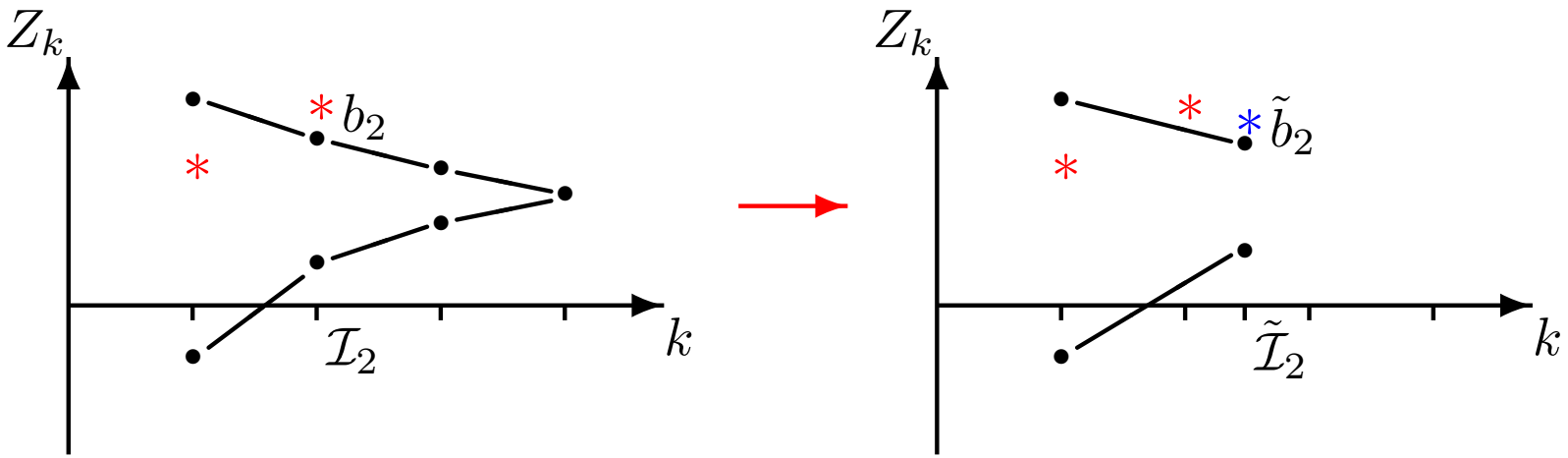
The trial stops but, with the pipeline data included, $Z = 2.1$.

Can the investigators claim significance at level α?

Incorporating delayed observations after a GST terminates

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

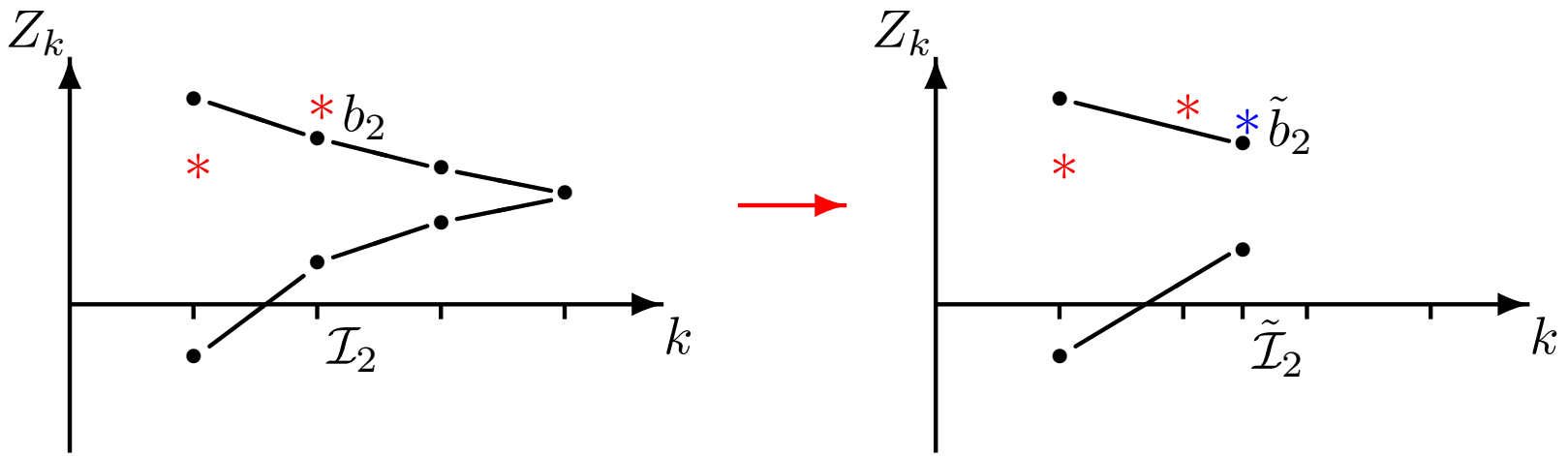
The analysis k at which termination occurs is deleted and one behaves as if analysis k had occurred with the information level $\tilde{\mathcal{I}}_k$ arising from the final set of responses.



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:

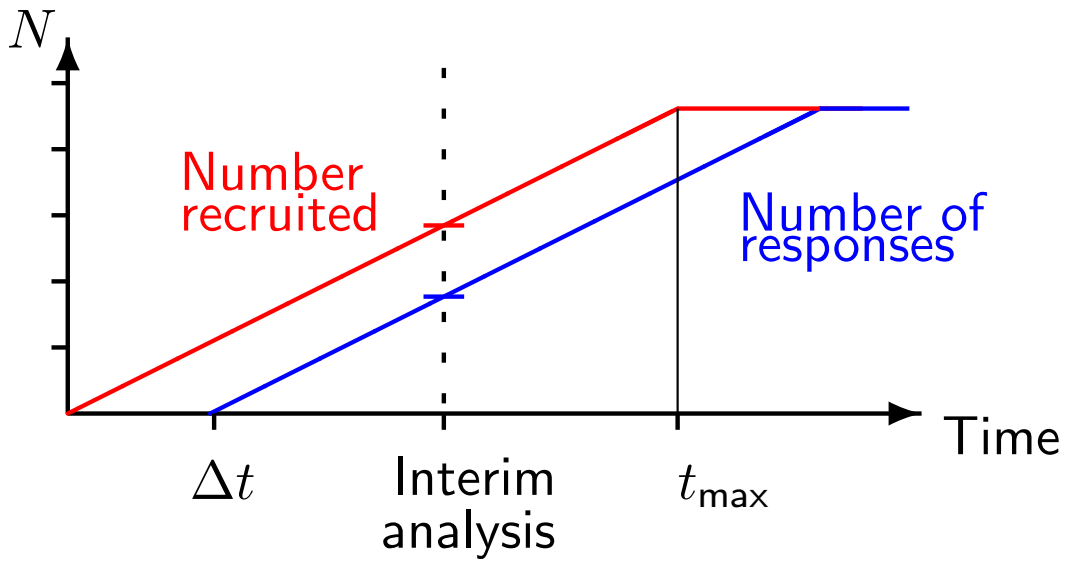
1. Stop recruitment of any more subjects,
2. After responses have been observed for all recruited subjects, make a decision to accept or reject H_0.

Defining a group sequential test with delayed responses

For now, we assume, as in Example 1:

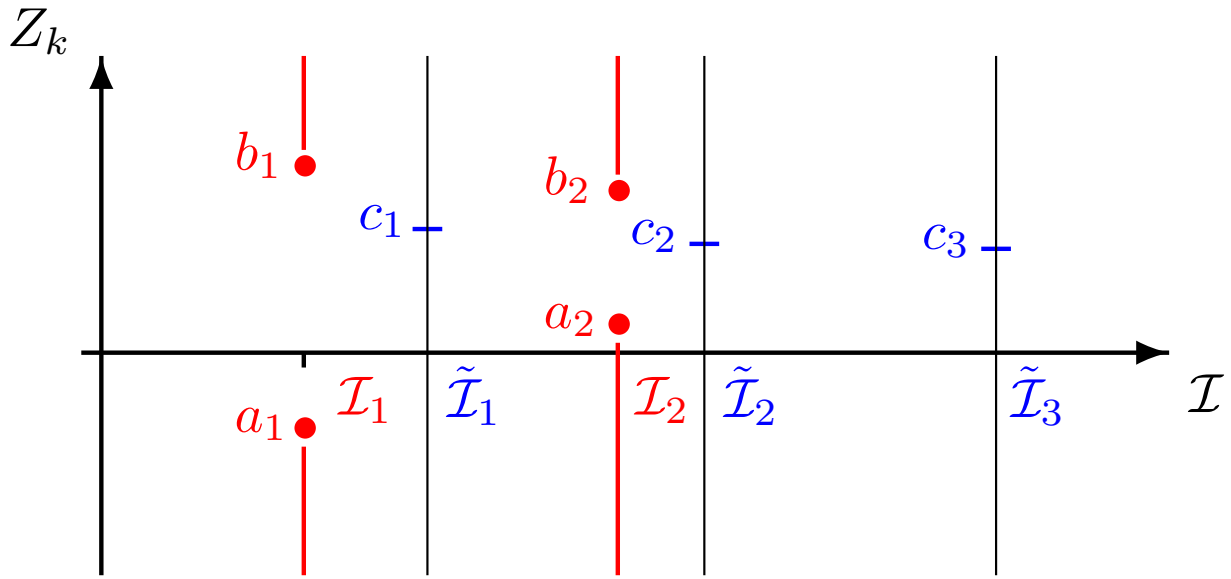
- The primary endpoint is measured a fixed time after treatment commences,
- The endpoint will be known (eventually) for all treated subjects,
- If recruitment is stopped, it cannot be re-started.

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

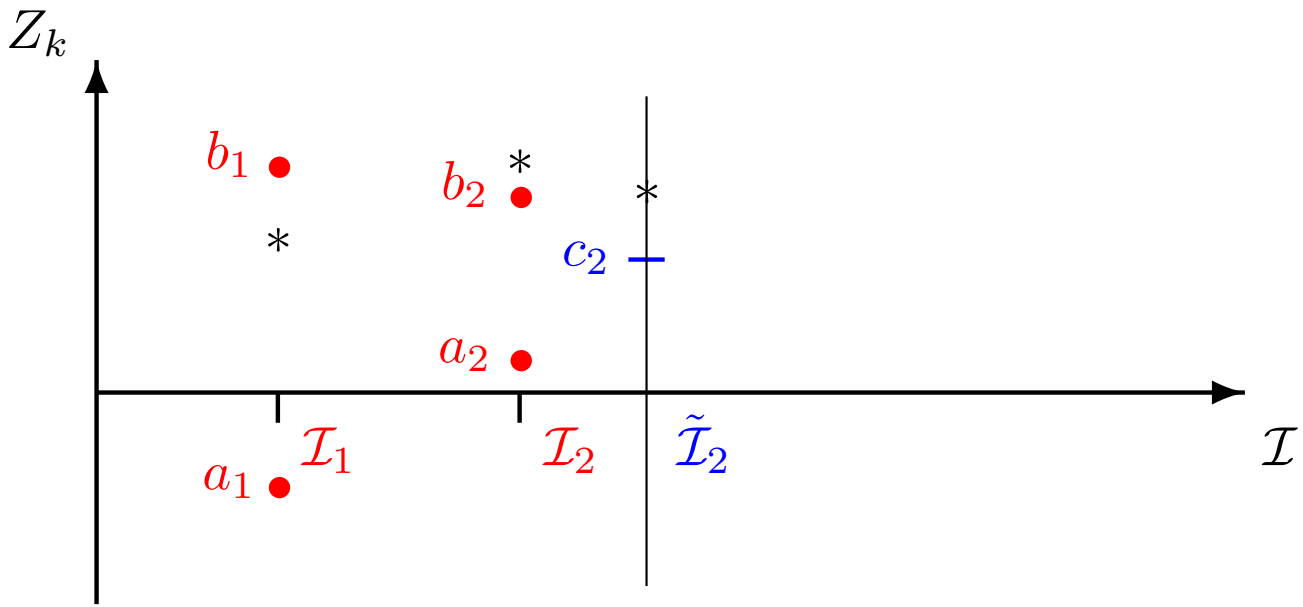
If we reach the **final** analysis K, we reject H_0 if $\tilde{Z}_K > c_K$.

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.

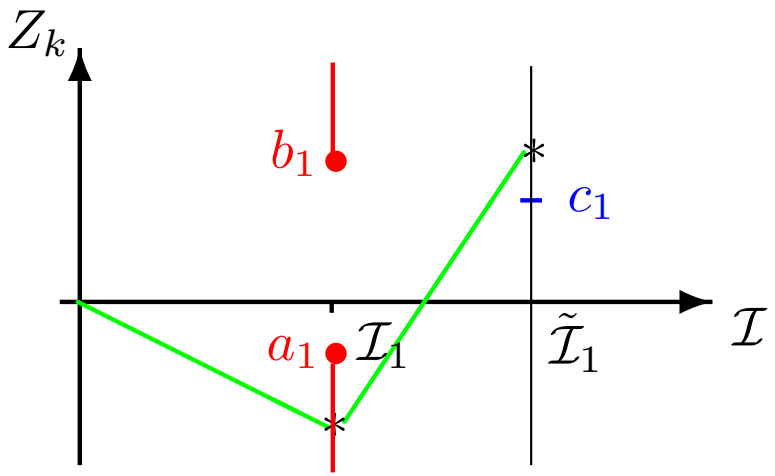
The value of information from pipeline subjects

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.

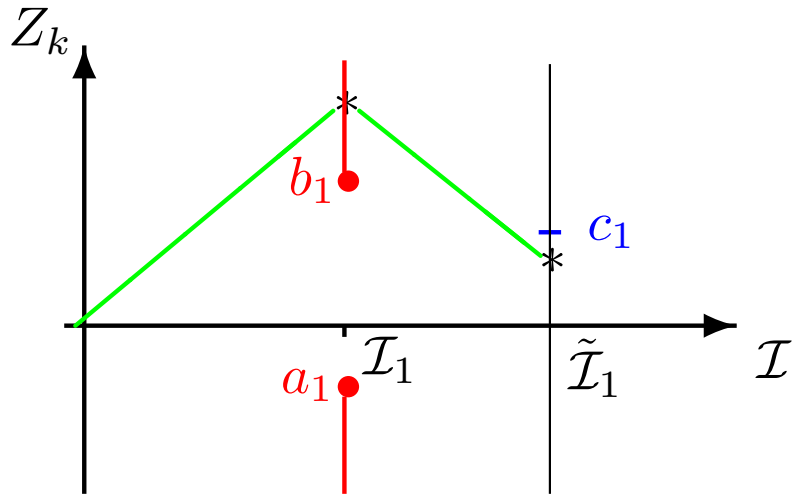
We could, for example, observe:



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,

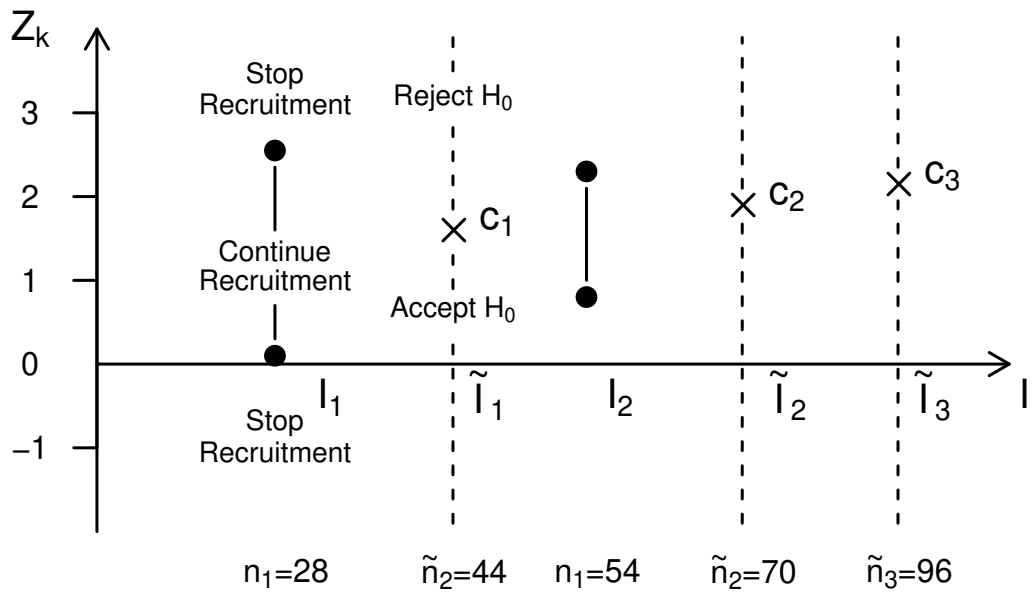
$\tilde{n}_3 = 96$ responses if there is no early stopping.

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. For credibility, these can be raised to 1.96 with little change to the design's power curve.

Error spending Delayed Response GSTs

H&J also construct error spending Delayed Response GSTs.

Here is a variation on these methods that allows a non-binding futility boundary; see Baayen et al. (*Statist. in Med.*, 2025).

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.

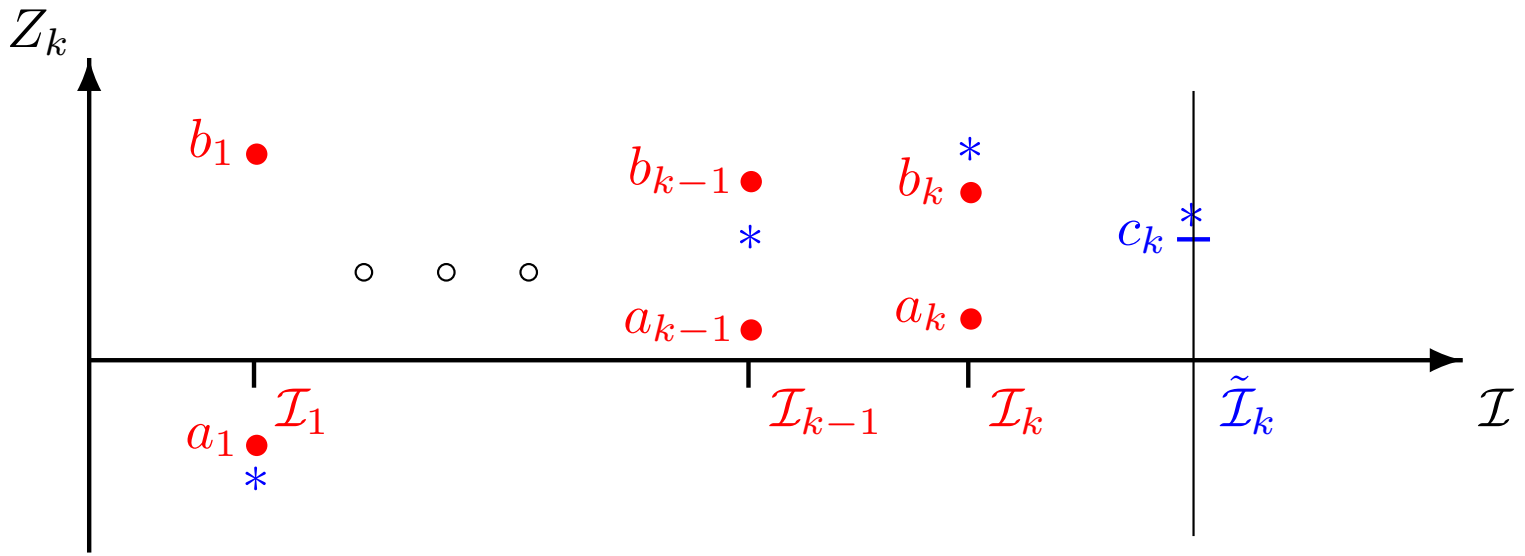
Recruitment stops when the target information $\mathcal{I}_{max}$ is reached (or will be reached with the responses from pipeline subjects).

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

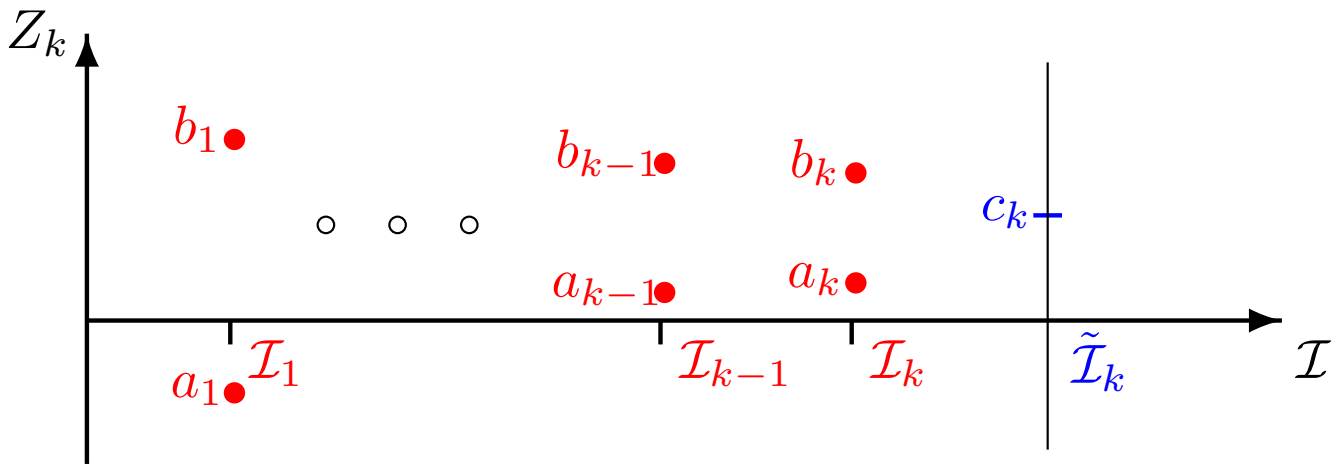


Non-binding futility boundary:

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

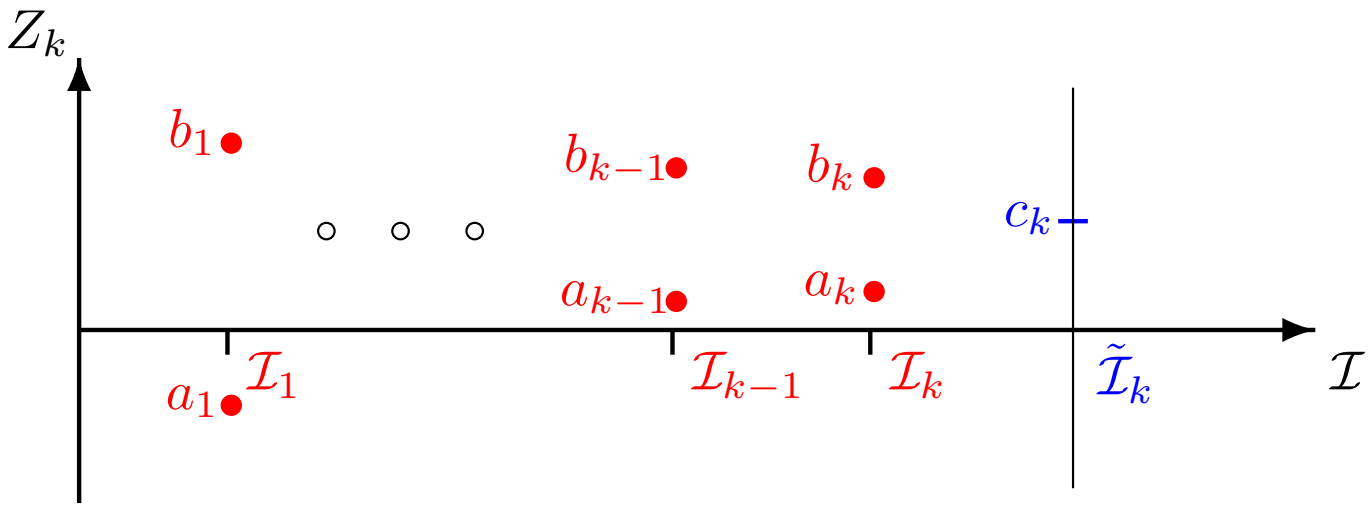
$$P_{\theta=0}\{Z_1 < b_1, \ldots, Z_{k-1} < b_{k-1}, Z_k > b_k, \tilde{Z}_k > c_k\}$$
$$= f(\mathcal{I}_k/\mathcal{I}_{\max}) - f(\mathcal{I}_{k-1}/\mathcal{I}_{\max}),$$

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



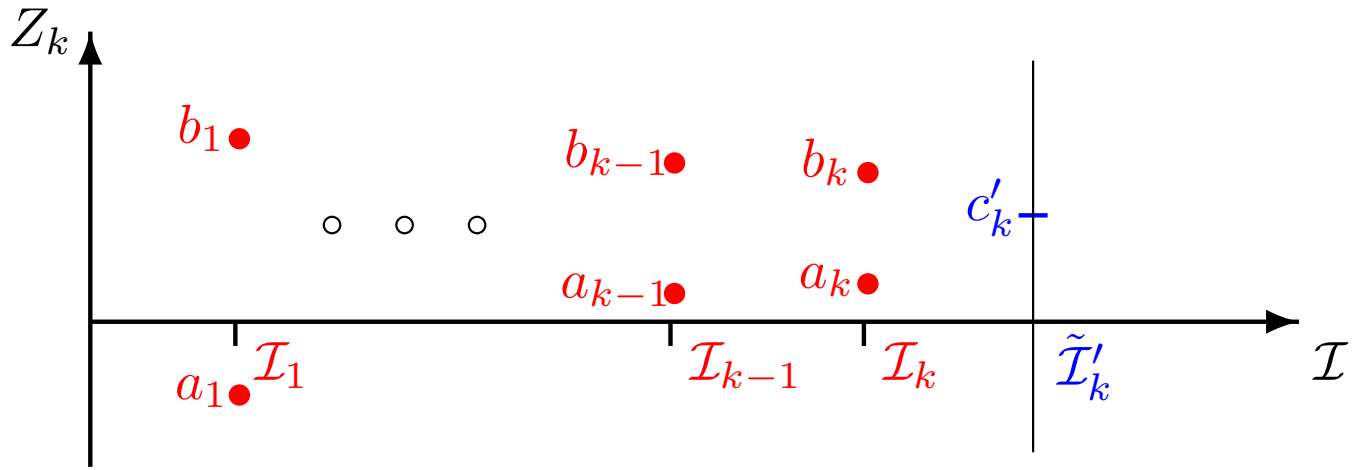
H&J noted their optimal Delayed Response GSTs with $\alpha = 0.025$ often had values of $c_1, \ldots, c_{K-1}$ less than $\Phi^{-1}(1-\alpha) = 1.96$.

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

H&J 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

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

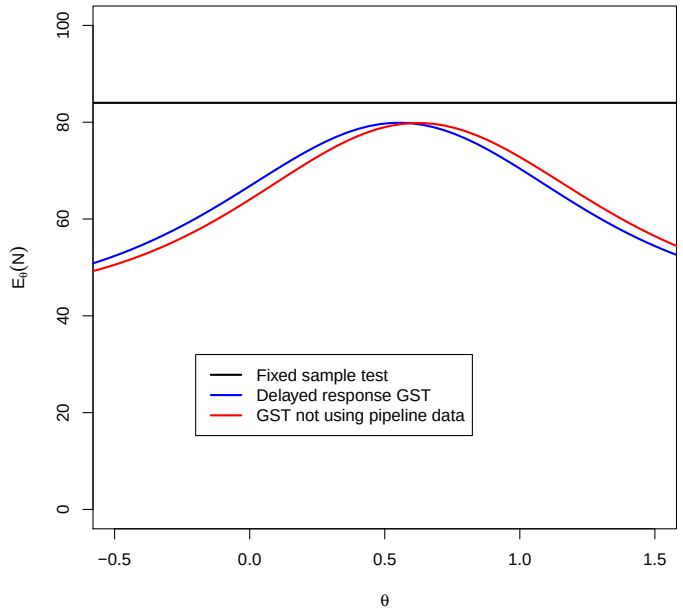
$$a_2 = 0.664, b_2 = 2.244, c_2 = 1.960;$$

$$c_3 = 2.069.$$

Example 1: A ρ-family error spending GST

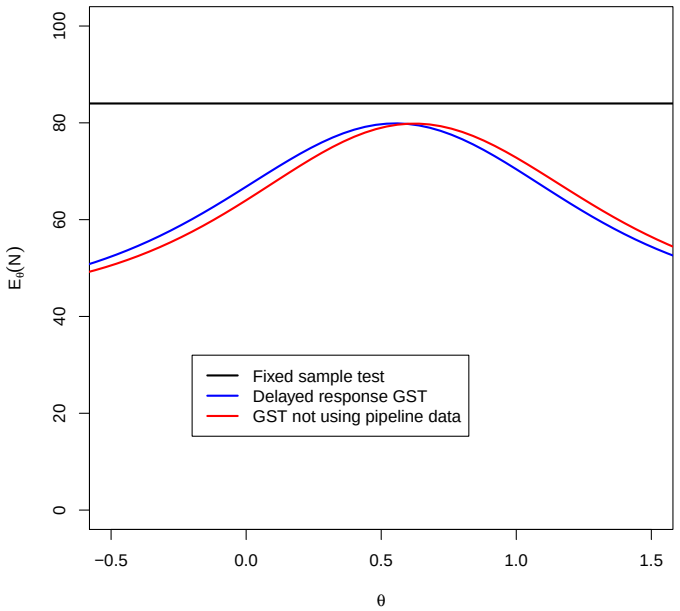
The figure shows $\mathbb{E}_\theta(N)$ for:

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 restore power of 0.9.

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 our examples show this is possible!

Further development of Delayed Response GSTs

A variety of optimality criteria

H&J 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 H&J's paper commented on the nature of "pipeline" data and H&J categorised possible cases in their response.

See Asikanius et al. (*Trials*, 2025) for a discussion of terminology.

Further development of Delayed Response GSTs

Inference on termination

H&J 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.

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

1.9 Bayes group sequential tests

Consider a Phase III clinical trial comparing a new treatment against a control.

We denote the treatment effect by θ.

Examples: θ could be the difference in mean response or, in a time-to-event study, θ could be the log hazard ratio.

We wish to decide whether $\theta > 0$, in which case the new treatment is superior.

The trial will have K analyses.

The information for θ at analysis k will be $\mathcal{I}_k$.

Marginally,

$$\widehat{\theta}_k \sim N(\theta, \mathcal{I}_k^{-1})$$

and the score statistics, $S_k = \widehat{\theta}_k \mathcal{I}_k$, have independent increments.

Decision theoretic formulation: Berry & Ho (*Bmcs*, 1988)

Prior distribution $\pi(\theta)$.

Possible decisions

$$d_1: \quad \text{Do not pursue drug approval,}$$

$$d_2: \quad \text{Pursue drug approval.}$$

Loss function for taking decision d_1 or d_2 when the true value of the treatment effect is θ

$$L(\theta, d_1) = 0 \qquad \text{for all } \theta,$$

$$L(\theta, d_2) = \begin{cases} -K\theta & \text{if } \theta > 0, \\ L & \text{if } \theta \leq 0. \end{cases}$$

Sampling cost: 1 per subject in the trial, N in total, say.

Aim: with x the observed data, minimise the expected loss

$$\int \pi(\theta) \int f(x \mid \theta) \{L(\theta, d(x)) + N(x)\} \, \mathrm{d}x \, \mathrm{d}\theta$$

Bayes Decision theoretic formulation

Aim: minimise the expected loss

$$\int \pi(\theta) \int f(x \mid \theta) \{L(\theta, d(x)) + N(x)\} \, \mathrm{d}x \, \mathrm{d}\theta$$

Berry & Ho found the optimal stopping rule and decision rule by dynamic programming.

They presented examples with priors

$$\theta \sim N(-1, 2), \quad \theta \sim N(0, 2), \quad \theta \sim N(1, 2),$$

$$K = 5,000 \quad \text{and} \quad L = 2,000,$$

and showed results for designs with 2 and 3 analyses.

Lorden (*Ann. Statistics*, 1976) had applied the numerical optimisation scheme described by Lai (*Ann. Statistics*, 1973) to solve similar problems in a frequentist setting.

A Bayes stopping rule: Spiegelhalter et al. (1994)

Spiegelhalter, Freedman & Parmar (*JRSS, A*, 1994) proposed the following sequential procedure. Given a prior $\pi(\theta)$,

At each analysis $k = 1, \ldots, K-1$

if $Pr\{\theta > 0 \mid \textit{data}\} > 1 - \epsilon$	stop, declare $\theta > 0$
if $Pr\{\theta < 0 \mid \textit{data}\} > 1 - \epsilon$	stop, declare $\theta \leq 0$
otherwise	continue to group $k+1$,

after group K

if $Pr\{\theta > 0 \mid \textit{data}\} > 1 - \epsilon$	stop, declare $\theta > 0$
otherwise	stop, do not declare $\theta > 0$.

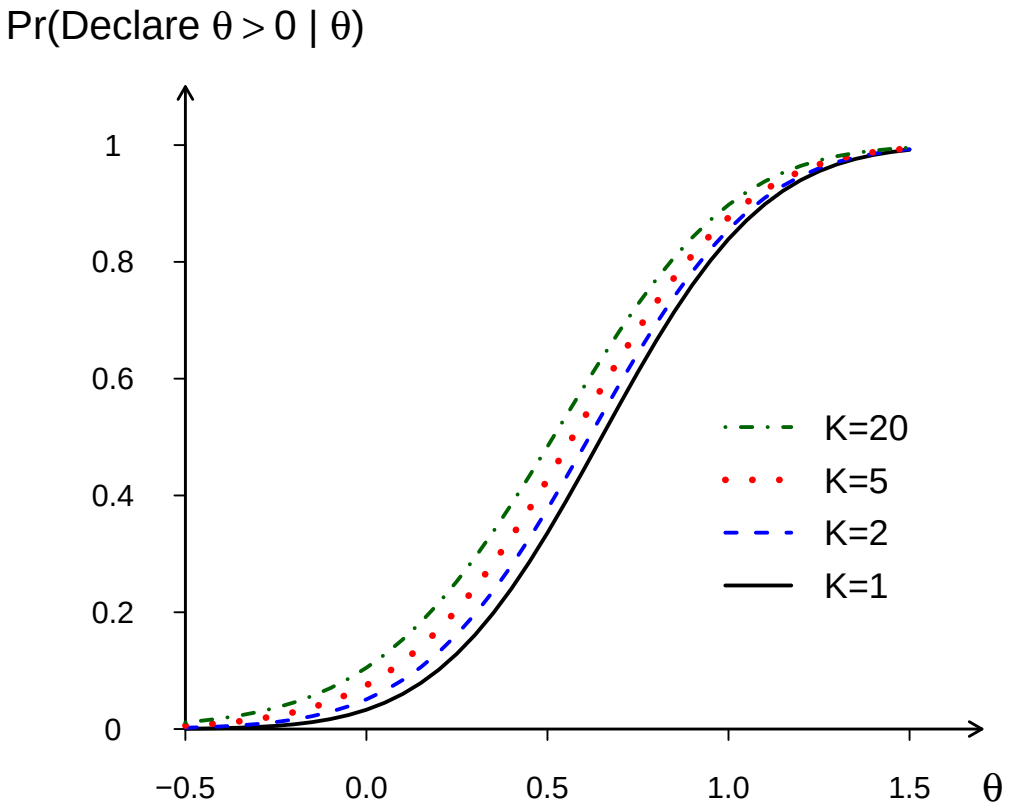
The trial stops early if the $1 - 2\epsilon$ credible interval for θ does not contain zero.

The credible interval is not affected by the fact that earlier analyses have been conducted.

A stopping rule based on the posterior distribution of θ

However, the frequentist properties of the Bayes procedure **are** affected by the number of analyses conducted.

Probability of declaring $\theta > 0$ for procedures with prior $\theta \sim N(0.5, 0.5)$, $\epsilon = 0.025$, and 1, 2, 5 and 20 analyses:



A calibrated Bayesian procedure

Spiegelhalter et al. (1994) proposed use of a "handicap prior", chosen so that the procedure has a particular type I error rate.

If the number of analyses is K and maximum information is $\mathcal{I}_K$, the handicap prior is

$$\theta \sim N(0, (h\mathcal{I}_K)^{-1}).$$

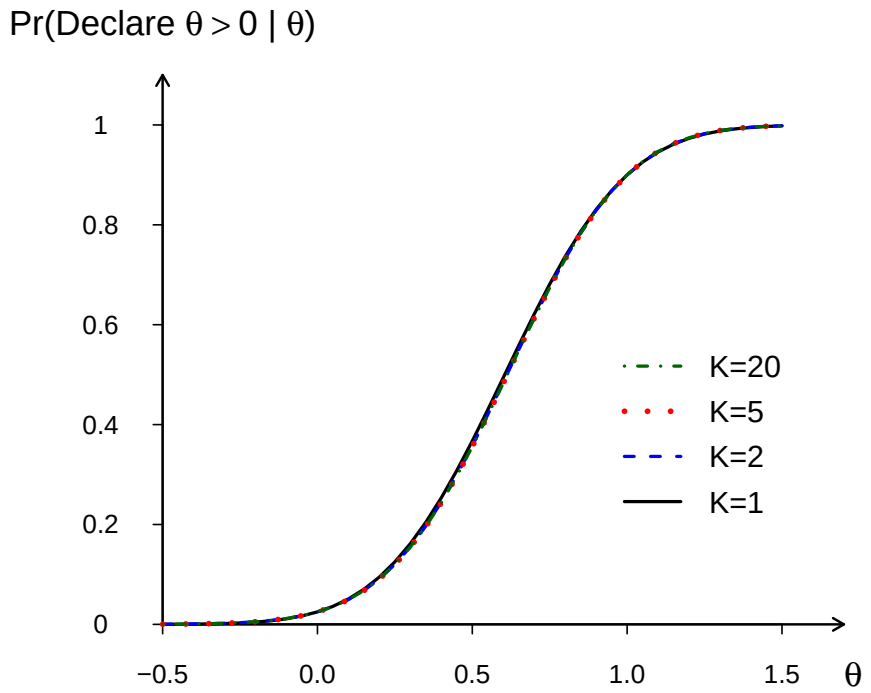
The "handicap" h depends on the number of analyses. Here are values for $\alpha = 0.025$.

Number of analyses K	*Handicap* h
1	0.000
2	0.163
5	0.271
20	0.382

A calibrated Bayesian procedure

We can choose $\mathcal{I}_K$ so that the procedure with a handicap prior has a specific power if the treatment effect is $\theta = \delta$.

Power functions of designs with $Pr\{\text{Reject } H_0 \,|\, \theta = 1\} = 0.9$



Note: When power curves are matched at $\theta = 0$ and $\theta = 1$, they are just about indistinguishable everywhere.

A frequentist solution: Barber & Jennison (*Bmka*, 2002)

We wish to test H_0: $\theta \leq 0$ against $\theta > 0$.

We specify type I and type II error rates

$$Pr_{\theta=0}\{\text{Reject } H_0\} \leq \alpha, \quad Pr_{\theta=\delta}\{\text{Reject } H_0\} \geq 1-\beta.$$

So, a fixed sample size test requires information

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

Aim: in a group sequential test with K analyses and with $\mathcal{I}_{max} = R\mathcal{I}_{fix}$, minimise

$$\int f(\theta)\, \mathbb{E}(\mathcal{I}_T)\, \mathrm{d}\theta,$$

where $f(\theta)$ is a $N(\delta/2, (\delta/2)^2)$ density and $\mathcal{I}_T$ is the observed information on termination.

Barber & Jennison found optimal stopping and decision rules by dynamic programming.

A frequentist solution

Minimum possible values of $\int f(\theta) E_\theta(\mathcal{I}_T)\, \mathrm{d}\theta$, where $f(\theta)$ is the density of a $N(\delta, \delta^2/4)$ distribution, for group sequential tests with K equally sized groups, $\mathcal{I}_{max} = R\,\mathcal{I}_{fix}$, type I error probability $\alpha = 0.025$, power 0.9 at $\theta = \delta$.

Minimum values of $\int f(\theta) E_\theta(\mathcal{I}_T)\, \mathrm{d}\theta$, as a percentage of $\mathcal{I}_{fix}$

	R					
K	1.01	1.05	1.1	1.2	1.3	*Minimum over R*
2	79.3	74.7	**73.8**	74.8	77.1	73.8 at R=1.11
3	74.8	69.0	67.0	**66.1**	66.6	66.1 at R=1.20
5	71.1	65.1	62.7	60.9	**60.5**	60.5 at R=1.32
10	68.2	62.1	59.5	57.5	**56.7**	56.4 at R=1.46
20	66.8	60.6	58.0	55.8	**54.8**	54.2 at R=1.59

Recommend: $K = 3$, 4 or 5 $R = 1.05$ or 1.1.

A frequentist solution

For practical application:

Error spending tests with type I and type II error spending functions of the form

$$f(\mathcal{I}) = \alpha\,(\mathcal{I}/\mathcal{I}_{\max})^{\rho}, \quad g(\mathcal{I}) = \beta\,(\mathcal{I}/\mathcal{I}_{\max})^{\rho}$$

are almost optimal.

Error spending designs adapt to observed information levels, controlling the type I error rate and maintaining efficiency.

Rho-family designs with $\rho = 2$ have a sample size "inflation factor" around $R = 1.1$.

Designs with $\rho = 3$ have an "inflation factor" around $R = 1.05$.

Derivation of optimal frequentist tests

The problem is to minimise

$$\int f(\theta)\, E_\theta(\mathcal{I}_T)\, \mathrm{d}\theta,$$

subject to

$$Pr_{\theta=0}\{\text{Reject } H_0\} \leq \alpha, \quad Pr_{\theta=\delta}\{\text{Reject } H_0\} \geq 1-\beta. \qquad (1)$$

Following the Lagrangian approach, we use dynamic programming to minimise

$$\int f(\theta)\, E_\theta(\mathcal{I}_T)\, \mathrm{d}\theta + \lambda_1\, Pr_{\theta=0}\{\text{Reject } H_0\} + \lambda_2\, Pr_{\theta=\delta}\{\text{Accept } H_0\}.$$

Then we search for values λ_1 and λ_2 so that (1) is satisfied — and we have the solution to the problem with error rate constraints.

Derivation of optimal frequentist tests

In minimising

$$\int f(\theta)\, E_\theta(\mathcal{I}_T)\, \mathrm{d}\theta + \lambda_1\, Pr_{\theta=0}\{\text{Reject } H_0\} + \lambda_2\, Pr_{\theta=\delta}\{\text{Accept } H_0\},$$

we have solved a Bayes decision problem with prior

$$\begin{array}{ll} \theta = 0 & \text{with probability } 1/3 \\ \theta = \delta & \text{with probability } 1/3 \\ \theta \sim N(\delta/2, \delta^2/4) & \text{with probability } 1/3 \end{array}$$

and a cost of sampling $\mathcal{I}_T$ when $\theta \sim N(\delta/2, \delta^2/4)$.

This is a Bayes optimal procedure — but for a rather odd looking problem.

A "hybrid" Bayesian proposal: Ventz & Trippa (2015)

Ventz & Trippa (*Biometrics*, 2015) proposed optimising and calibrating at the same time. Their problem formulation has:

Possible decisions

$$d_1: \quad \text{Do not pursue drug approval,}$$
$$d_2: \quad \text{Pursue drug approval.}$$

A "gain function" or "utility" comprising

$$G(\theta, d_1) = 0 \qquad \text{for all } \theta,$$
$$G(\theta, d_2) = \begin{cases} K(\theta) & \text{if } \theta > 0, \\ -L(\theta) & \text{if } \theta \leq 0, \end{cases}$$

plus a term

$$B(\theta, d_i, T)$$

denoting the additional benefit from reaching decision d_i at analysis T minus the cost of treating patients in the trial.

A "hybrid" Bayesian proposal

Assuming a prior distribution $\pi(\theta)$, Ventz & Trippa seek to maximise the expected gain

$$-\int_{-\infty}^{0} L(\theta)\, Pr(D = d_2 \,|\, \theta)\, \pi(\theta)\, d\theta + \int_{0}^{\infty} K(\theta)\, Pr(D = d_2 \,|\, \theta)\, \pi(\theta)\, d\theta + \int_{-\infty}^{\infty} E\{B(\theta, d_i, T)\}\, \pi(\theta)\, d\theta.$$

subject to error rate constraints

$$Pr(D = d_2 \,|\, \theta = 0) = \alpha \quad \text{and} \quad Pr(D = d_1 \,|\, \theta = \delta) = \beta$$

for a specified value of δ.

Note that type I and type II errors feature twice, in different guises.

A "hybrid" Bayesian proposal

We have just seen an example of how, to a high degree of accuracy, power curves belong to a two parameter family.

Thus, the constraints

$$Pr(D = d_2 \,|\, \theta = 0) \;=\; \alpha \quad \text{and} \quad Pr(D = d_1 \,|\, \theta = \delta) \;=\; \beta$$

are essentially equivalent to

$$\int_{-\infty}^{0} L(\theta)\, Pr(D = d_2 \,|\, \theta)\, \pi(\theta)\, d\theta \;=\; p_1$$

and

$$\int_{0}^{\infty} K(\theta)\, Pr(D = d_2 \,|\, \theta)\, \pi(\theta)\, d\theta \;=\; p_2$$

for certain values of p_1 and p_2.

A "hybrid" Bayesian proposal

Thus, when Ventz & Trippa maximise

$$-\int_{-\infty}^{0} L(\theta)\, Pr(D = d_2 \,|\, \theta)\, \pi(\theta)\, d\theta \;+\; \int_{0}^{\infty} K(\theta)\, Pr(D = d_2 \,|\, \theta)\, \pi(\theta)\, d\theta$$

$$+\; \int_{-\infty}^{\infty} E\{B(\theta, d_i, T)\}\, \pi(\theta)\, d\theta$$

subject to

$$Pr(D = d_2 \,|\, \theta = 0) \;=\; \alpha \quad \text{and} \quad Pr(D = d_1 \,|\, \theta = \delta) \;=\; \beta.$$

they are effectively maximising

$$\int_{-\infty}^{\infty} E\{B(\theta, d_i, T)\}\, \pi(\theta)\, d\theta.$$

subject to

$$\int_{-\infty}^{0} L(\theta)\, Pr(D = d_2 \,|\, \theta)\, \pi(\theta)\, d\theta = p_1 \;\text{ and }\; \int_{0}^{\infty} K(\theta)\, Pr(D = d_2 \,|\, \theta)\, \pi(\theta)\, d\theta = p_2.$$

A "hybrid" Bayesian proposal

Given the equivalence of the two types of constraint on the power function, they are also maximising

$$\int_{-\infty}^{\infty} E\{B(\theta, d_i, T)\}\, \pi(\theta)\, d\theta.$$

subject to

$$Pr(D = d_2 \,|\, \theta = 0) \;=\; \alpha \quad \text{and} \quad Pr(D = d_1 \,|\, \theta = \delta) \;=\; \beta.$$

This is exactly the type of problem that Jennison and others have tackled from a frequentist perspective.

A remark on the peculiar prior for frequentist designs

I noted that the designs produced by Barber & Jennison were Bayes procedures for rather strange looking priors.

However, we can replace the constraint

$$Pr(D = d_2 \,|\, \theta = 0) \;=\; \alpha \quad \text{and} \quad Pr(D = d_1 \,|\, \theta = \delta) \;=\; \beta$$

by an (almost) equivalent constraint of the form

$$\int_{-\infty}^{0} Pr(D = d_2 \,|\, \theta)\, f(\theta)\, d\theta = p_1 \;\text{ and }\; \int_{0}^{\infty} Pr(D = d_2 \,|\, \theta)\, f(\theta)\, d\theta = p_2,$$

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

Then, we will obtain essentially the same optimal design by solving a more reasonable looking Bayesian problem.

CONVERGENCE!

Conclusions

Theoretical underpinnings

The convergence between frequentist and calibrated Bayes procedures demonstrates the "complete class theorems" of Brown, Cohen & Strawderman (*Ann. Statistics*, 1980) which state

$$\{\text{The set of admissible frequentist procedures}\}$$

$$= \{\text{The set of Bayes optimal procedures}\}.$$

Practical consequences

Both schools can learn from each other — we are (or should be) solving the same problems.

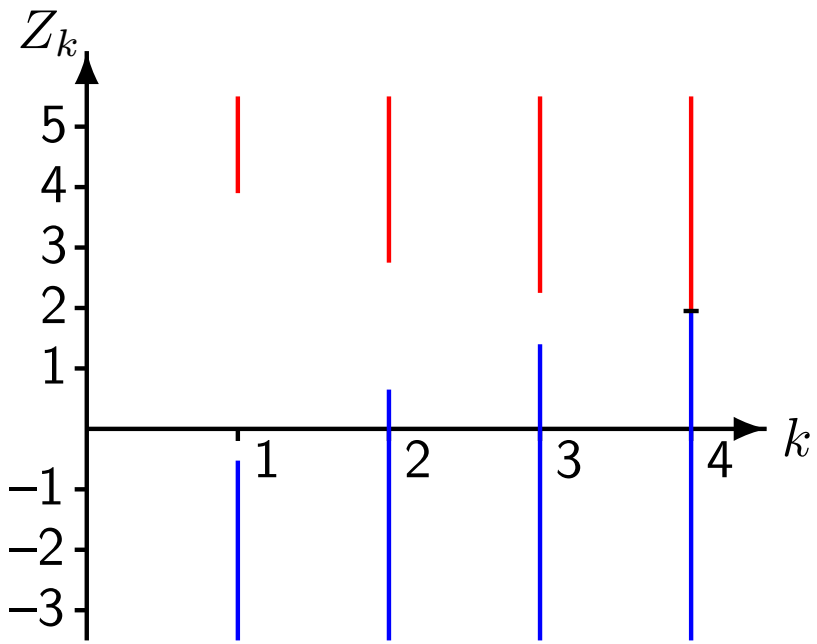
1.10 Inference on termination of a group sequential test

Reference: JT, Chapter 9

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 $Z_3 = 2.7$.

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

$$P\{N(0,1) > 2.70\} = 0.0035.$$

But, using this definition, we would also get a P-value ≤ 0.0035 by

stopping at stage 1 with $Z_1 > 3.90$,

stopping at stage 2 with $Z_2 > 2.76$,

stopping at stage 3 with $Z_3 > 2.70$,

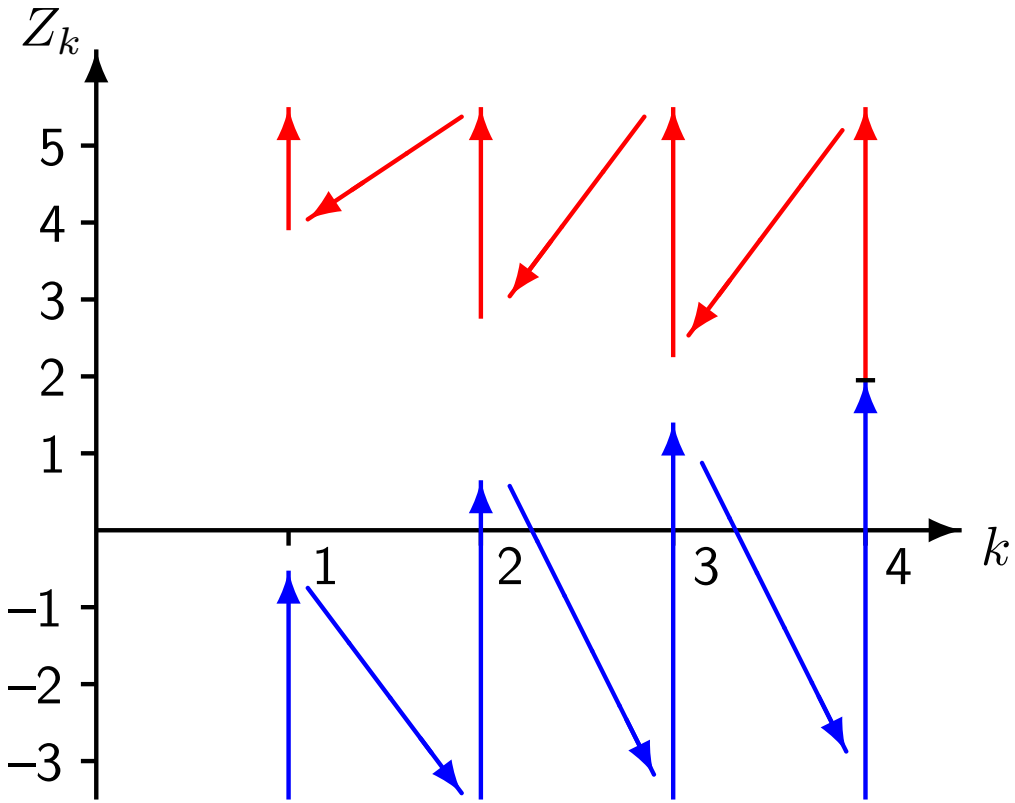
stopping at stage 4 with $Z_4 > 2.70$,

and the total probability under $\theta = 0$ of "$P \leq 0.0035$" is 0.0061.

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

Analysis on termination of a group sequential test (GST)

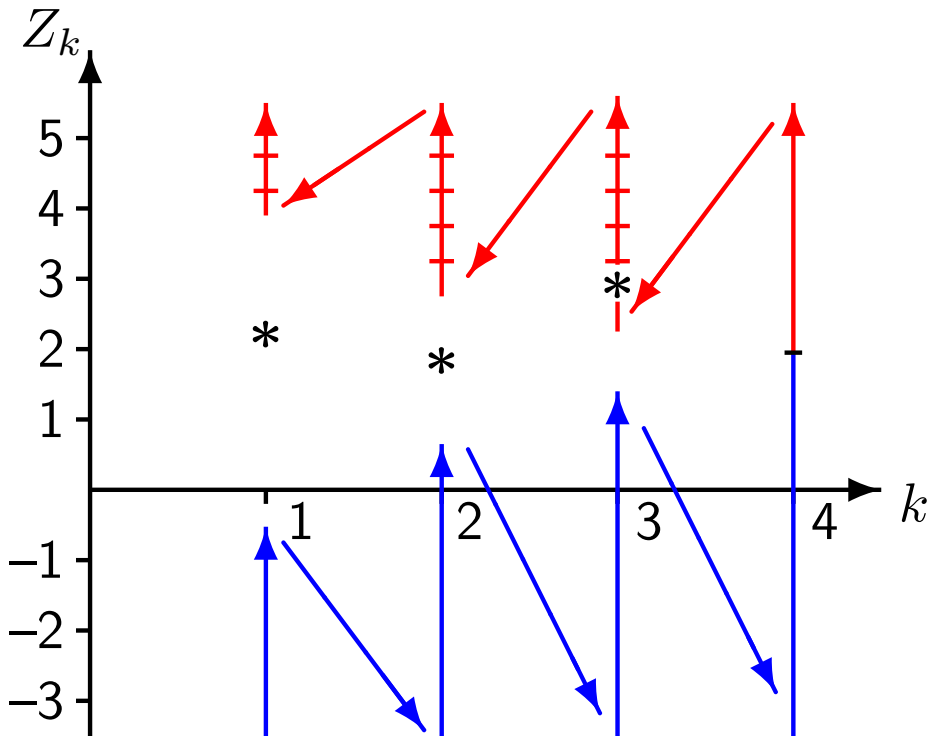
For proper frequentist inference, we first order the sample space.



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 $Z_3 = 2.70$, the 1-sided P-value for H_0: $\theta \leq 0$ is

$$P_{\theta=0}\{\text{Stop with } Z_1 \geq 3.90 \text{ or } Z_2 \geq 2.76 \text{ or } Z_3 \geq 2.70\} = 0.0053.$$

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: $\theta = 0$.

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: $\theta = \tilde{\theta}$.

A confidence interval on termination

Example:

Suppose the trial stops at analysis 3 with $Z_3 = 2.7$.

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

$$(0.24,\ 1.81)$$

In contrast:

The "naive" fixed sample CI would be $(0.29,\ 1.82)$.

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.

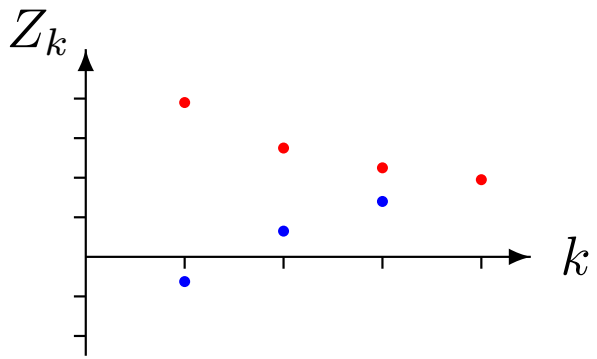
This happens automatically if outcomes for which we reject H_0 are at the top end of the sample space ordering — and any sensible ordering does this.

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.

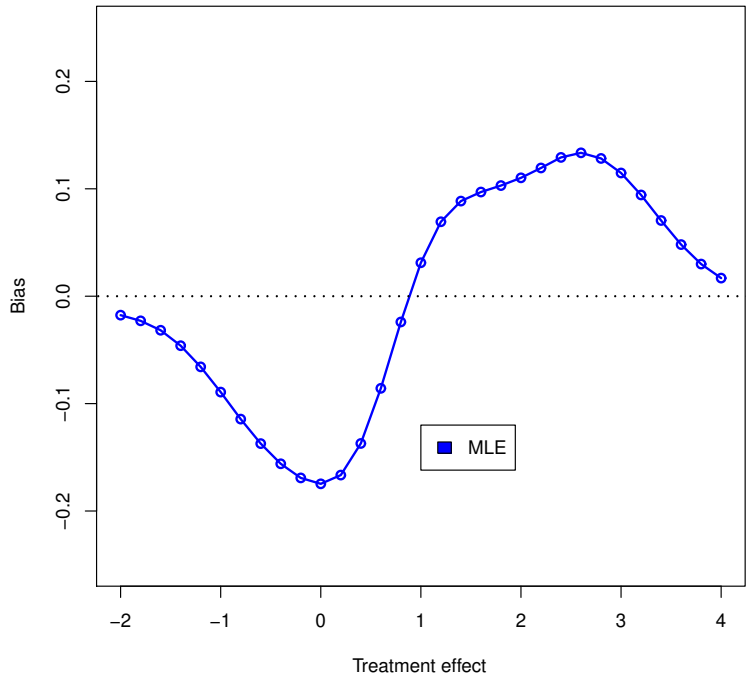
Thus, the MLE is biased with $E_\theta(\widehat{\theta}_M) > \theta$ for high values of θ.

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 $b(\widehat{\theta}_M)$.

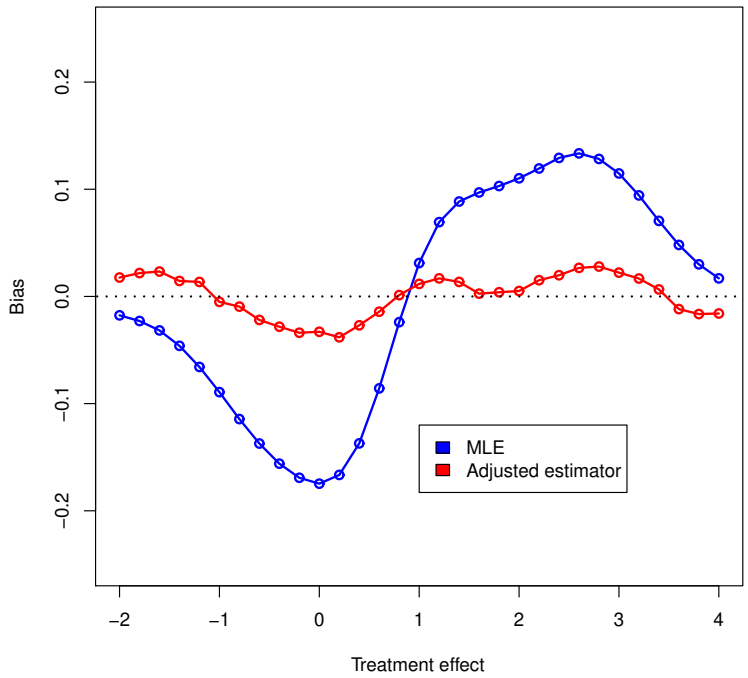
The adjusted estimator is then

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

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.

Unbiased estimates after a group sequential tests

The International Council for Harmonisation has issued

ICH E20: Adaptive Designs for Clinical Trials (draft)

Key principle number 4 concerns "Reliability of estimation".

It is noted that, while controlling the chances of false positive efficacy conclusions is expected,

> *In addition, reliable estimation of treatment effects for the primary efficacy endpoint and other key efficacy and safety outcomes is important.*
>
> *In the trade-off between bias and variance, the expectation is generally for limited to no bias in the primary estimate of the treatment effect.*

It may be surprising to suppose it is possible to give an estimate with "no bias" after a group sequential or adaptive trial.

Unbiased estimates after a group sequential tests

Emerson & Fleming (*Biometrika*, 1990) noted that $\widehat{\theta}_1$, the MLE based on the data at analysis 1, is unbiased for θ.

Applying "Rao-Blackwellization", we can calculate the conditional expectation of $\widehat{\theta}_1$ given the final set of data to obtain an unbiased estimate of θ.

This is the Uniform Minimum Variance Unbiased Estimate (UMVUE) among estimators that do not require knowledge of future information levels.

The numerical methods used to compute properties of a group sequential test can be adapted to compute this UMVUE.

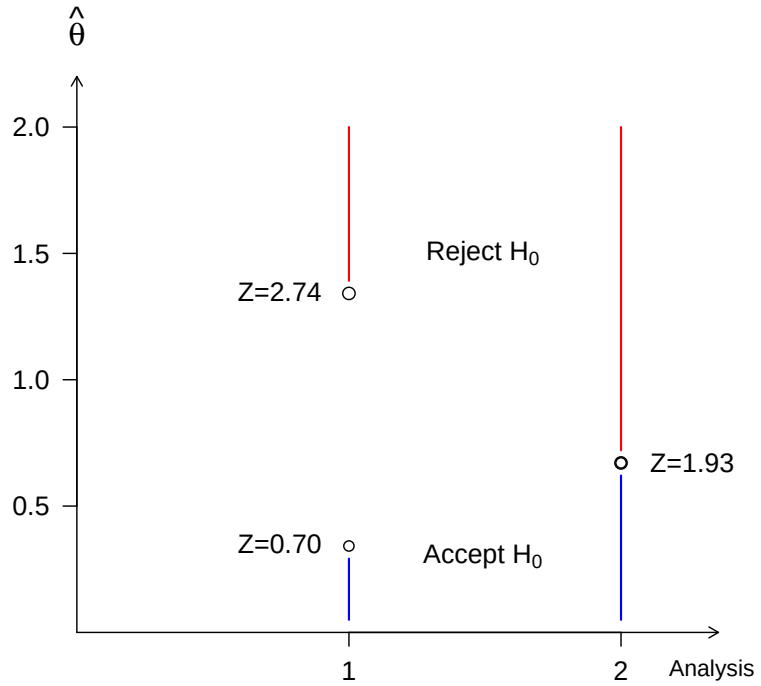
However, unbiased estimators may have a large variance and a relatively high mean square error.

The UMVUE may also be rather strange!

Unbiased estimates after a group sequential tests

Clinical trial designs with just two analyses are common.

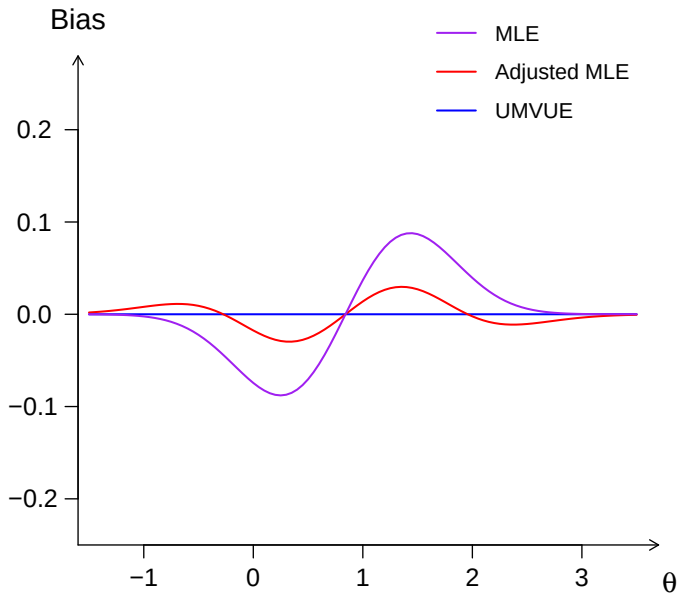
Consider the following one-sided, group sequential test of H_0: $\theta \leq 0$ against $\theta > 0$.



This test has type I error probability $\alpha = 0.025$ and power 0.8 is achieved if $\theta = 1$.

Unbiased estimates after a group sequential tests

The bias of several estimates:



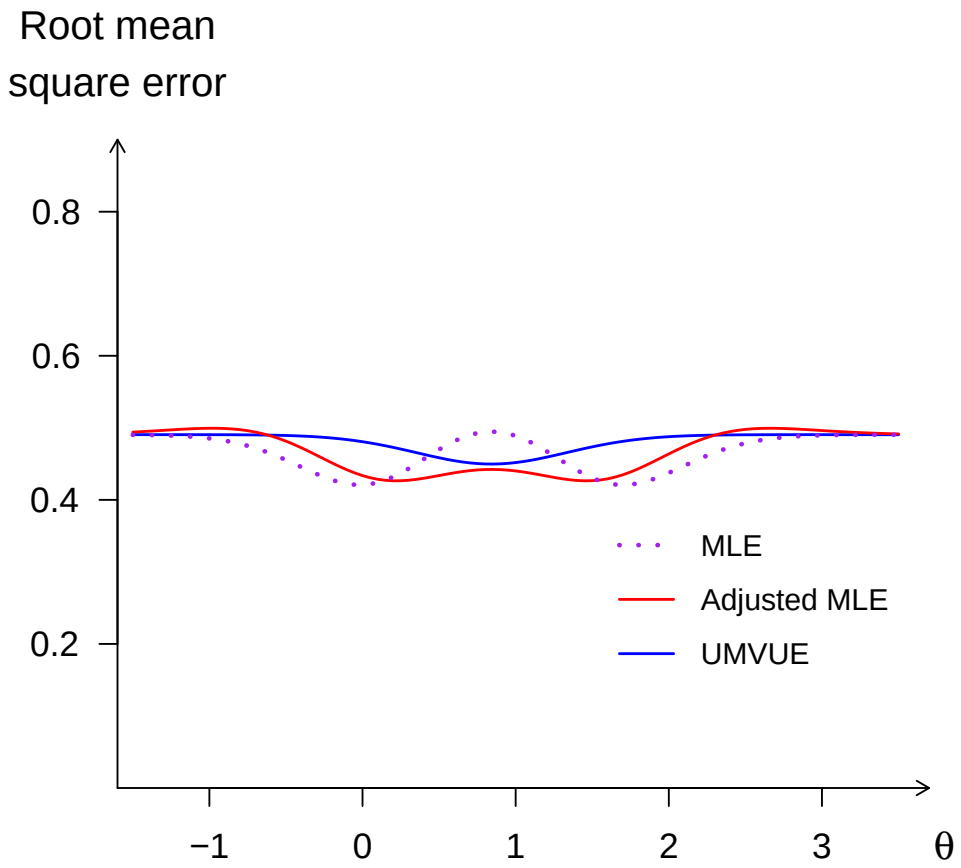
Denote the maximum likelihood estimate (MLE) on termination by $\widehat{\theta}_M$ and its bias function by $b(\widehat{\theta})$.

The Adjusted MLE is $\widehat{\theta}_{adj} = \widehat{\theta}_M - b(\widehat{\theta}_M)$.

The UMVUE is the conditional expectation of $\widehat{\theta}_1$ given the final set of data.

Unbiased estimates after a group sequential tests

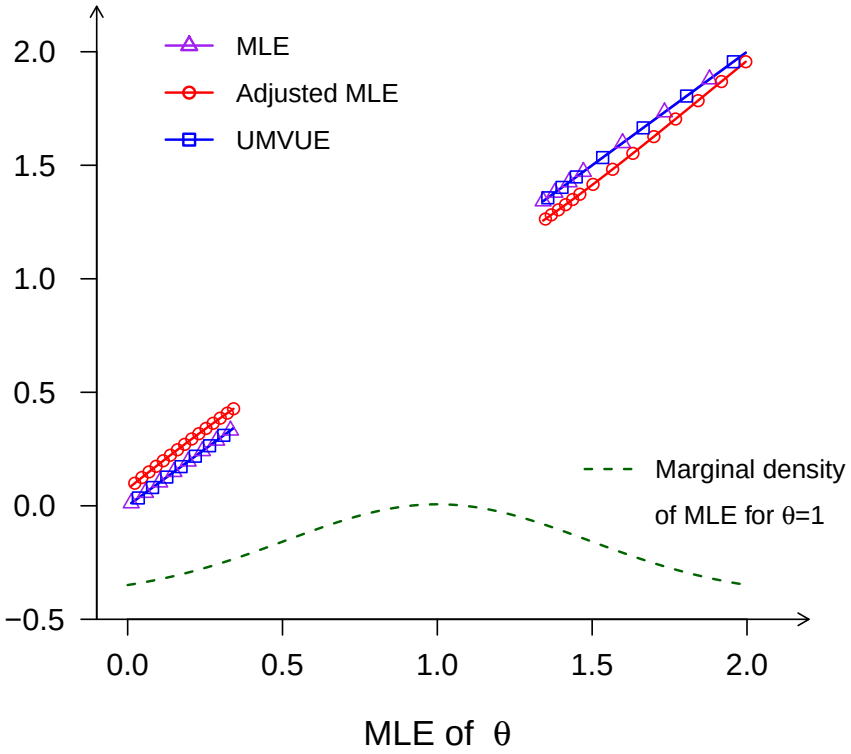
The root mean square error of several estimates:



The UMVUE has a higher variance than the Adjusted MLE, and this results in a higher mean square error.

Unbiased estimates after a group sequential tests

Estimates on termination at analysis 1

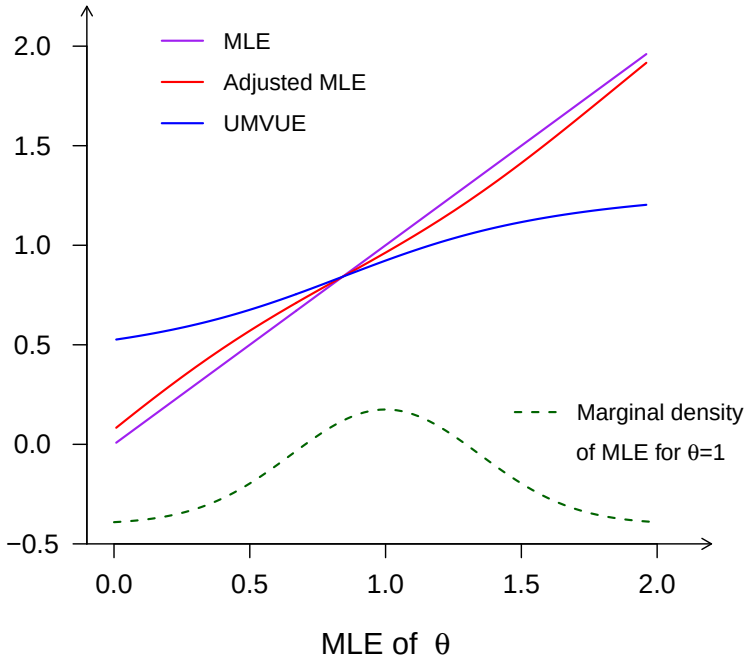


If the trial stops at analysis 1 the UMVUE is $\widehat{\theta}_1$, the MLE of θ.

So there is no "adjustment for bias" in the UMVUE!

Unbiased estimates after a group sequential tests

Estimates on termination at analysis 2



When stopping at analysis 2, the UMVUE can be substantially lower than the MLE.

If the MLE is $\widehat{\theta}_2 = 1.5$, the UMVUE is only 1.12 — but the bias in the MLE is at most 0.09 for any value of θ.

Unbiased estimates after a group sequential tests

Looking ahead: in more complex adaptive designs, bias may arise

(i) from selecting a treatment arm or patient sub-population based on promising early results,

(ii) from early stopping on a "random high".

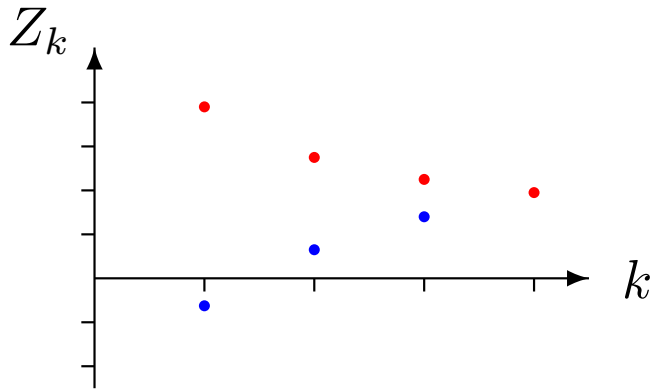
Some of the methods proposed for estimation after such trials also use Rao-Blackwellization to find UMVUE or Uniform Minimum Variance Conditionally Unbiased Estimates.

Given the behaviour of the UMVUE estimate in our simple example of a two-stage group single trial with a single parameter to estimate, we should look more closely at how these estimates may behave.

Adjusted estimates with a small bias may well be preferable.

1.11 Analysis of 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

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

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

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

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

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

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.

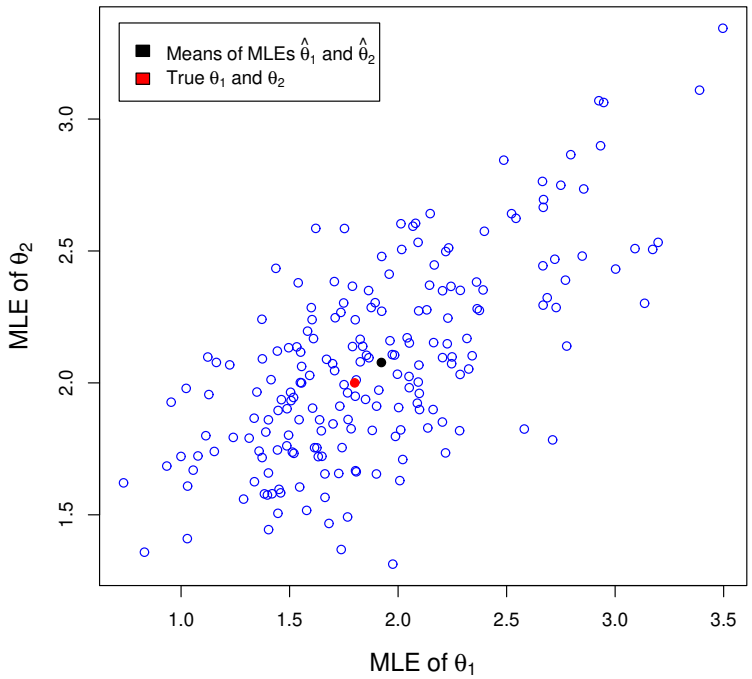
As for the primary endpoint, we can adjust the MLE, $\hat{\theta}_2$, by subtracting an estimate of its bias, $(\rho\,\sigma_2/\sigma_1)\, b_1(\hat{\boldsymbol{\theta}})$.

Estimation for a secondary endpoint: Example

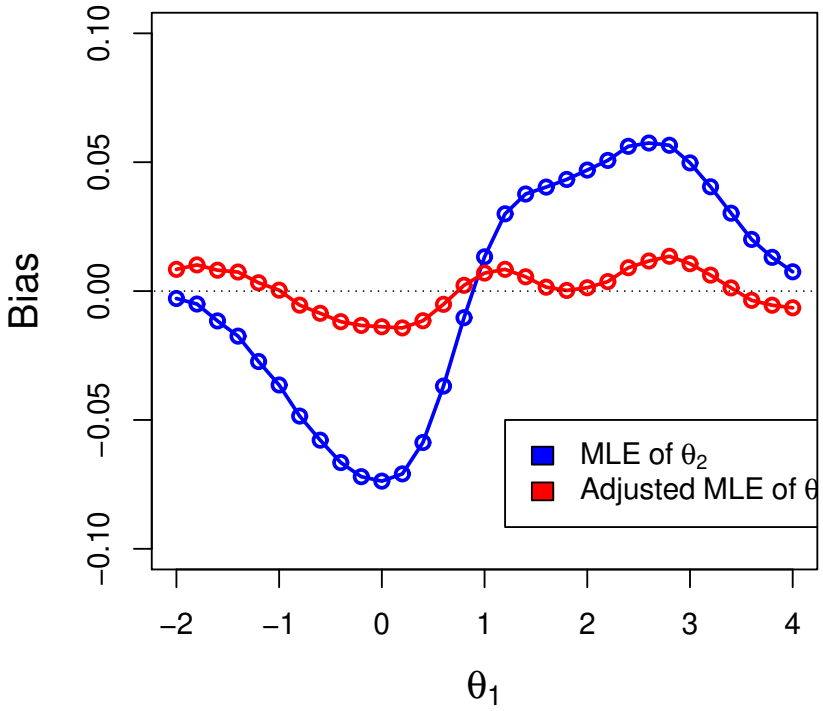
Suppose a trial tests its primary endpoint, using a Pampallona & Tsiatis GST with $\Delta = 0$, $\alpha = 0.025$ and power 0.8 at $\theta_1 = 1$.

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

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

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 $\theta_1 = 1$.

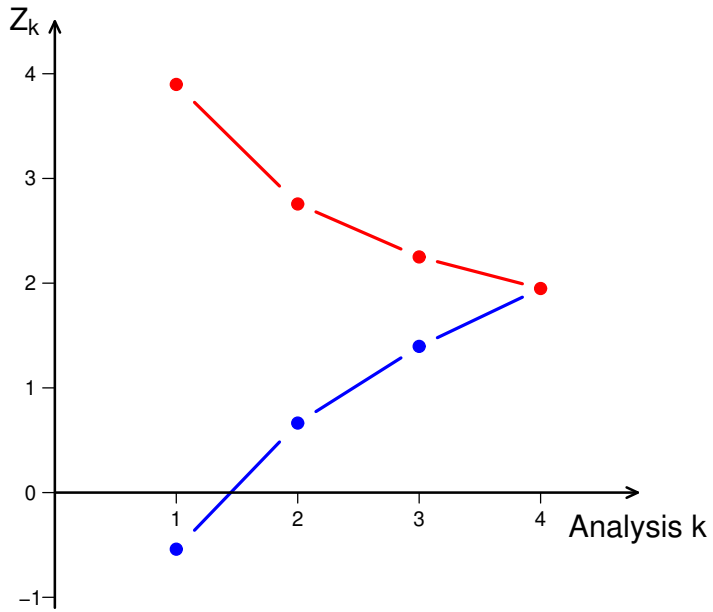
This test has upper boundary:

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

and lower boundary

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

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

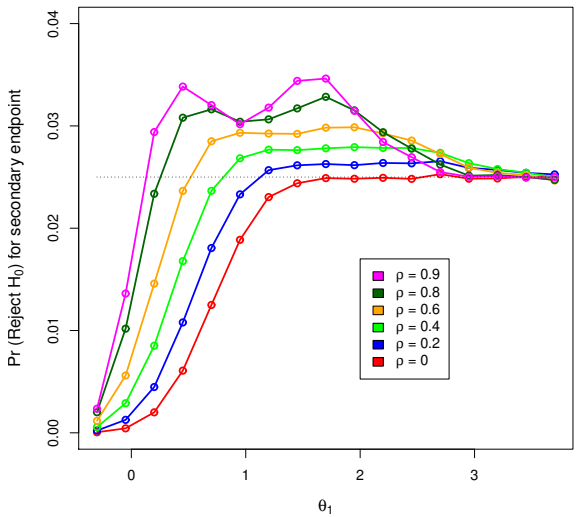


If the upper boundary is crossed, the secondary endpoint is tested in a level α, fixed sample size test, using current data.

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.
- Software is available to carry out the necessary calculations.
- 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.**

Recapitulation: Group sequential tests

See JT (2025) for more on:

- Two-sided Tests (Chapters 2, 4, 6)
- Tests for Equivalence and Non-inferiority (Ch. 7)
- Repeated Confidence Intervals (Ch. 10)
- Binary Data (Ch. 16)
- Multivariate Endpoints (Ch. 19)
- Response Adaptive Randomisation in GSTs (Ch. 30)

Part 2. Multiple testing procedures

Reference: JT, Chapter 20

2.1. Introduction: The familywise error rate

2.2. Bonferroni tests

2.3. Recycling type I error probability

2.4. Example: Primary and secondary endpoints

2.5. Closed testing procedures

2.6. Graphical representation of multiple testing procedures

2.7. Combining multiple testing and group sequential design

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

2.1 Multiple testing: The familywise error rate

A clinical trial may involve

Multiple primary endpoints:

A positive outcome required for at least one endpoint

Two primary endpoints:

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

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

2.2 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 two 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

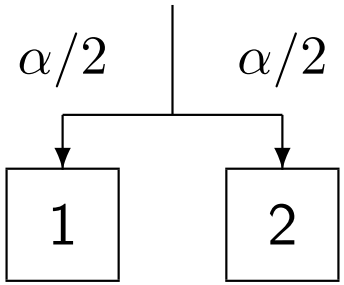
$$Z_1 > \Phi^{-1}(1 - \alpha/2) = 2.24,$$

An effect on Expanded MACE is declared if

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

Example: Two 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.

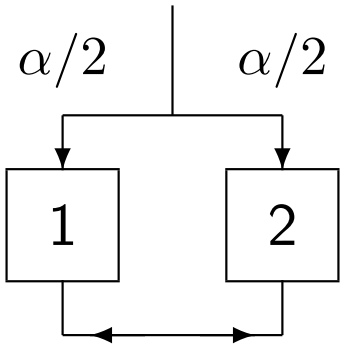
Also, if one hypothesis is false, the familywise type I error probability is at most $\alpha/2$.

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

2.3 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.

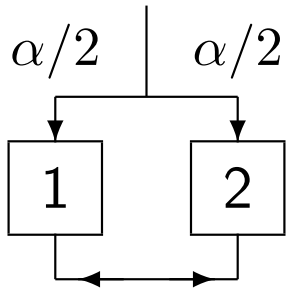
This method can be represented as:



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.

2.4 Example: Primary and secondary endpoints

A hierarchical testing or "gatekeeping" procedure

Consider a trial where

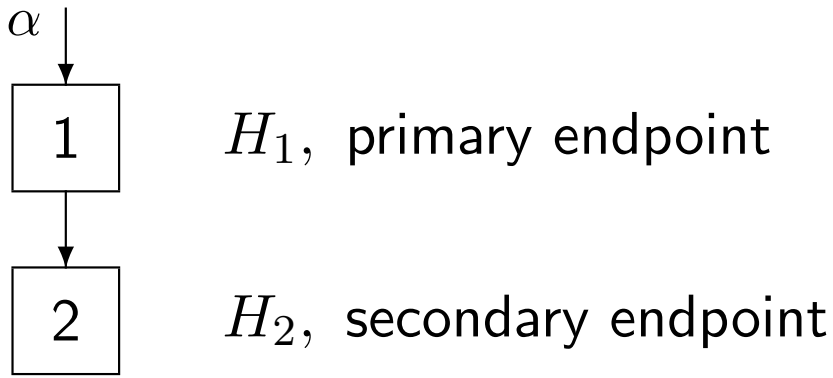
The null hypothesis H_1 concerns the primary endpoint,

The null hypothesis H_2 relates to a secondary endpoint,

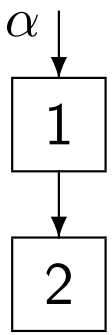
and H_2 will only be tested if H_1 has already been rejected.

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

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

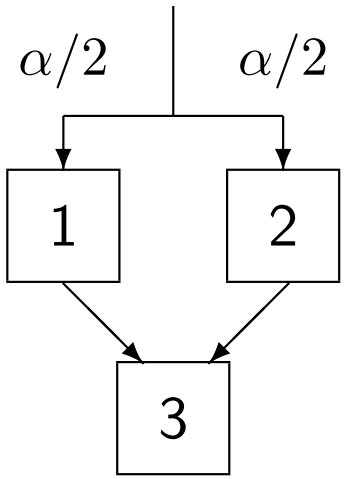
If H_1 is false and H_2 is true,

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

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

Example: Testing two primary and one secondary endpoint

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



H_1, H_2: two primary endpoints

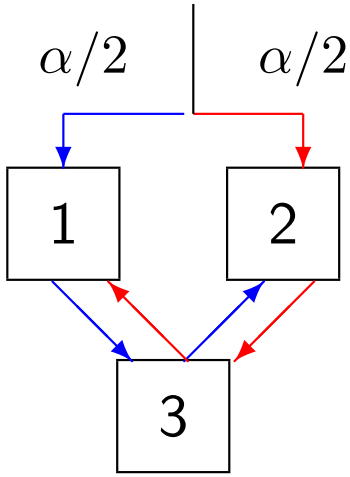
H_3: secondary endpoint

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 two primary and one secondary endpoint

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



H_1, H_2: two 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 two primary and one secondary endpoint

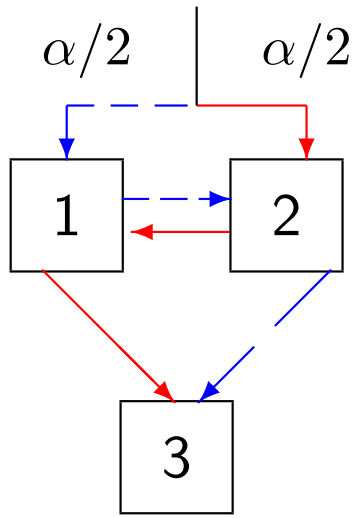
We may prefer to gain maximum power for tests of two primary endpoints before testing a secondary endpoint.

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: two 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.

2.5 Closed testing procedures

Reference: JT, Section 20.3

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

Reference: JT, Sections 20.4 and 21.4

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 $\{Z_i,\, i \in I\}$.

2.6 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 certain 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.*

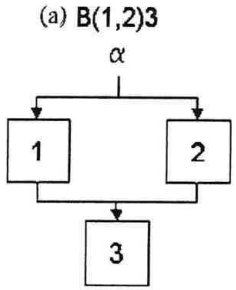
These procedures are closed testing procedures in which the tests of intersection hypotheses are weighted Bonferroni tests.

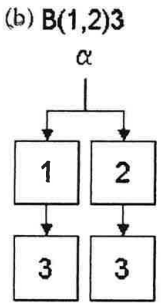
It is implicit in their method of construction that these procedures provide strong control of the FWER.

A figure from Burman et al. (2009)

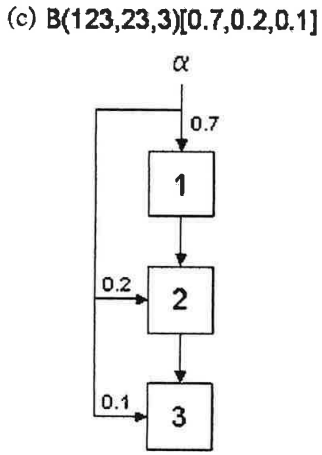
The following diagrams illustrate the graphical representations of multiple testing procedures used by Burman et al.

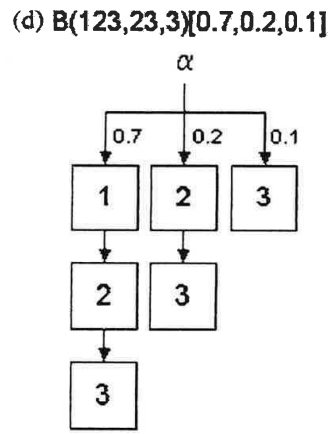
(a) and (b) A parallel gatekeeping procedure





(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.

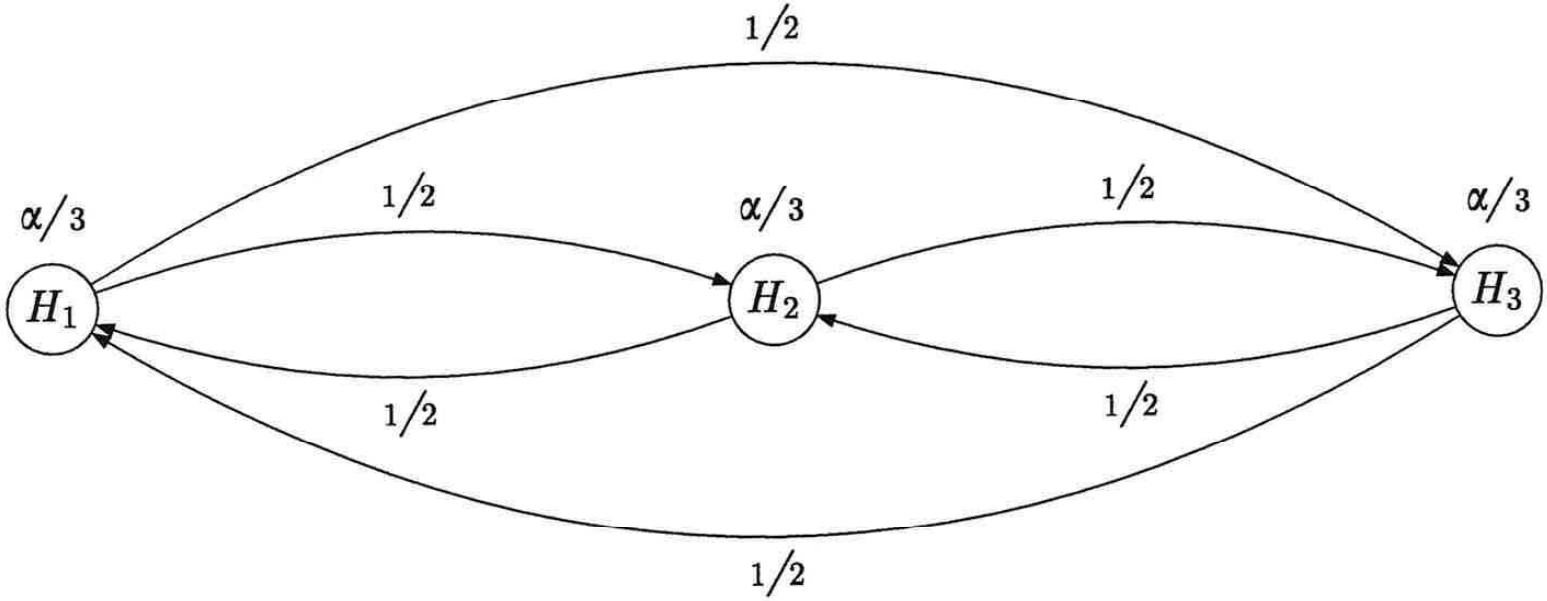


Figure 3. Graphical illustration of the Bonferroni–Holm procedure with $m = 3$ hypotheses and initial allocation $\boldsymbol{\alpha} = (\alpha/3, \alpha/3, \alpha/3)$.

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

2.7 Multiple testing within a group sequential design

Maurer & Bretz (*Statistics 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

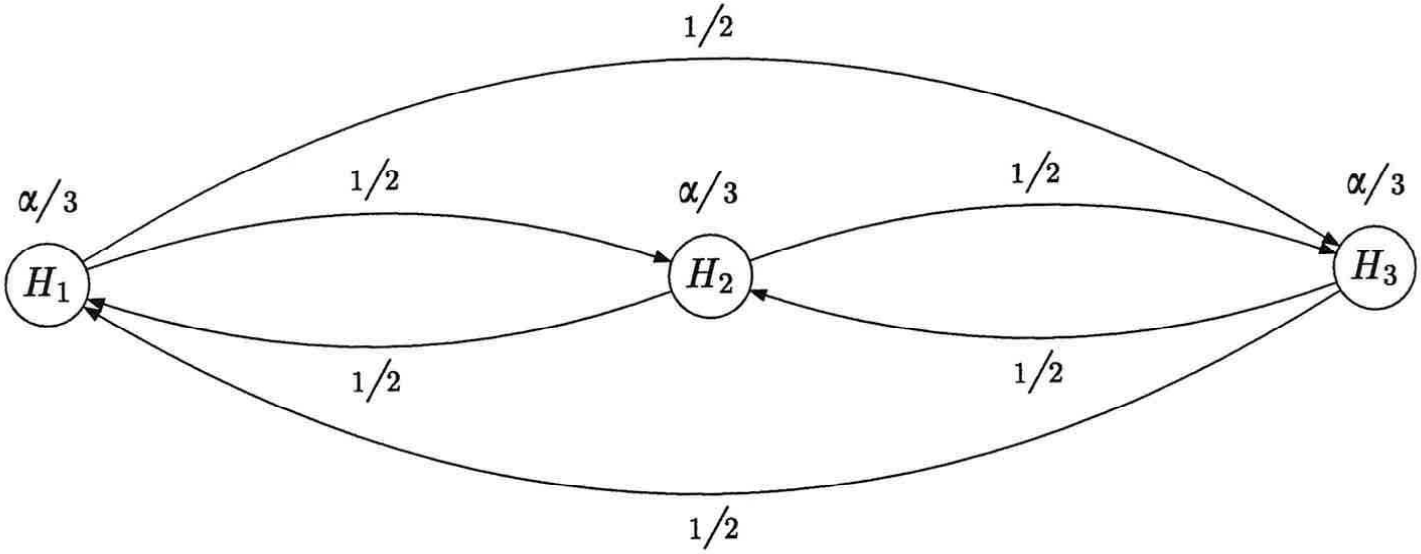


Figure 3. Graphical illustration of the Bonferroni–Holm procedure with $m = 3$ hypotheses and initial allocation $\boldsymbol{\alpha} = (\alpha/3, \alpha/3, \alpha/3)$.

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.

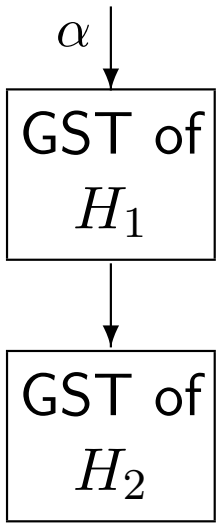
2.8 Testing a secondary endpoint after a sequential test

Reference: JT, Section 20.10

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, futility 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

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

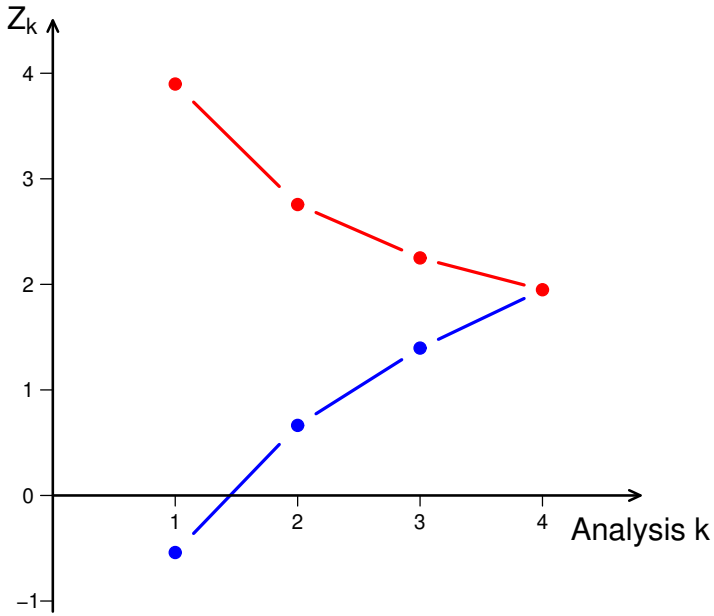
so the FWER is protected conservatively.

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

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

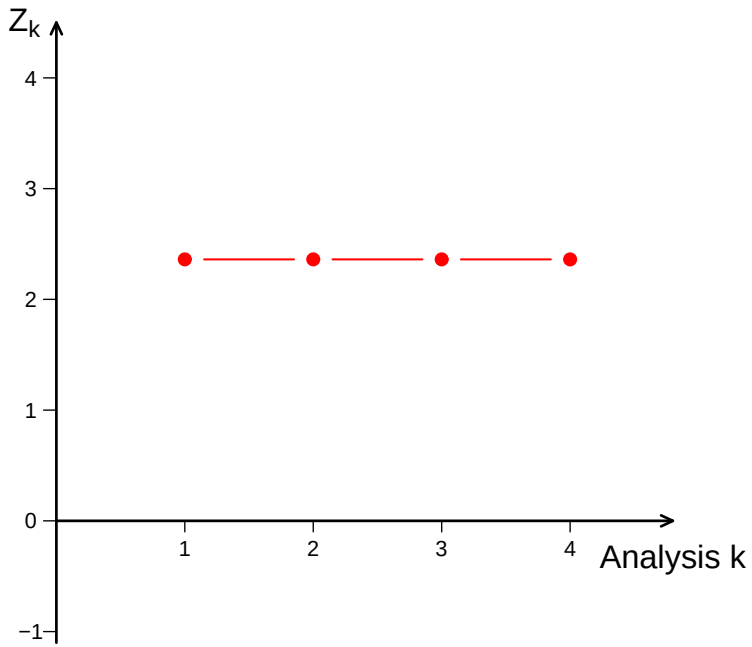
We consider two options for this test of H_2.

Example: Testing primary and secondary endpoints

Two options for the group sequential test of H_2.

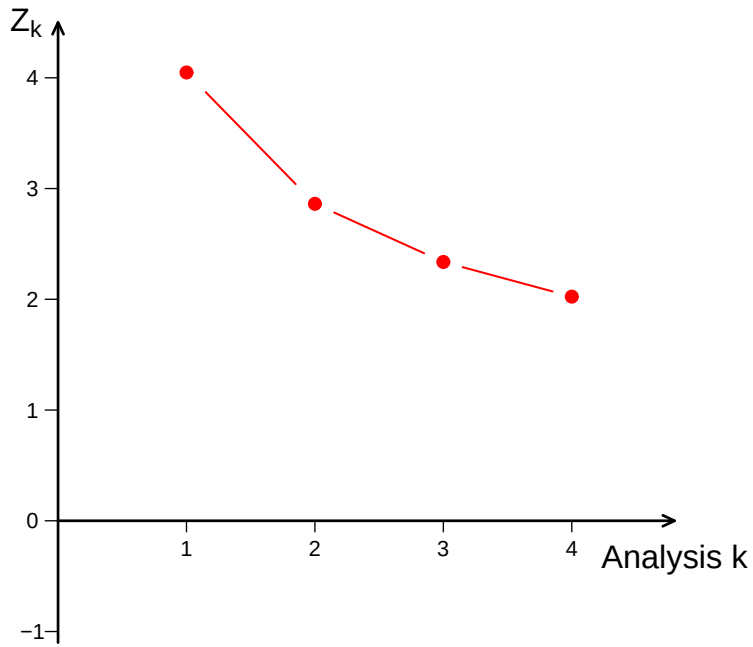
A: Pocock boundary for H_2

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



B: OBF boundary for H_2

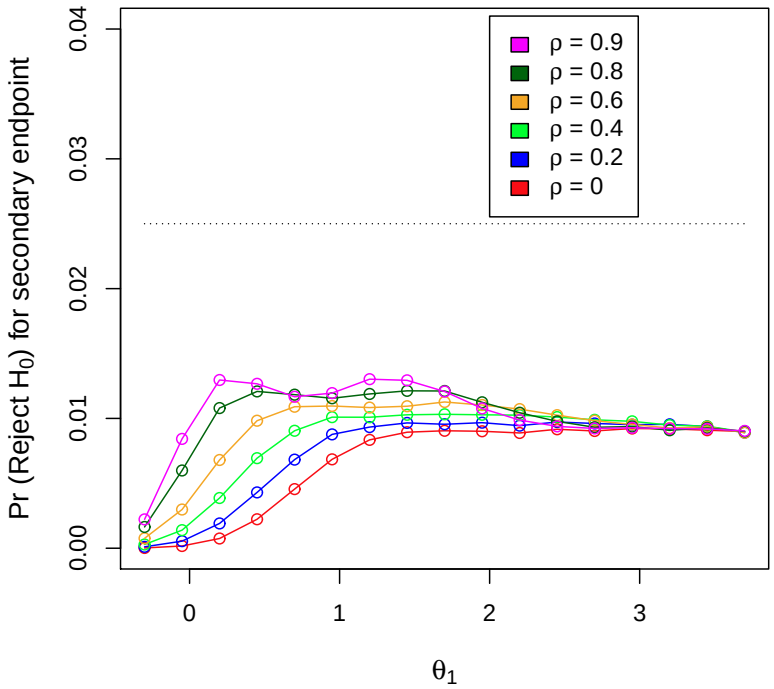
$$c_k = 2.024\sqrt{4/k}, \quad k = 1, \ldots, 4.$$



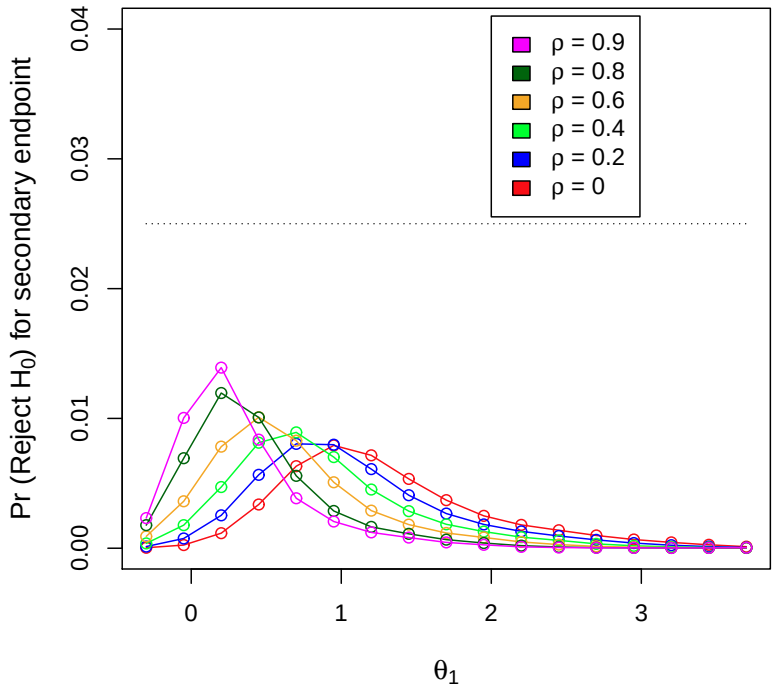
Note: The O'Brien & Fleming boundary requires a very high value of Z_{2,k^*} to reject H_2 if the GST of H_1 stops at the first analysis.

Type I error probability for testing H_2

A: Pocock boundary for H_2



B: OBF boundary for H_2

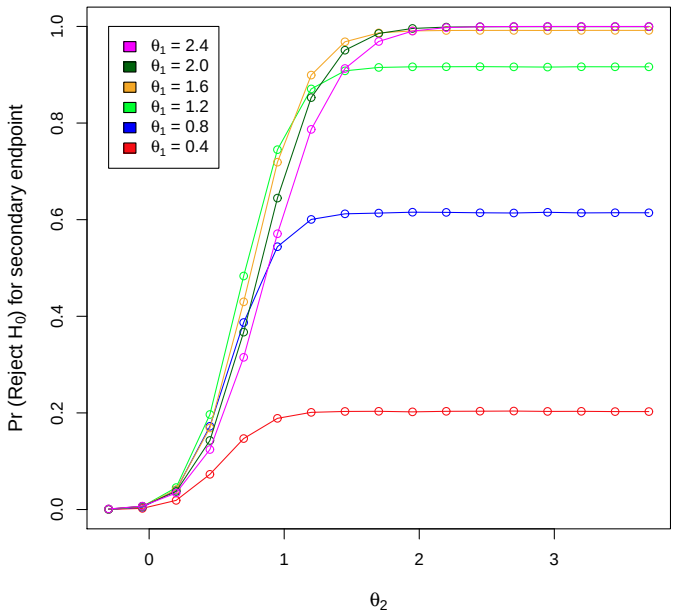


Type I error probabilities are calculated under $\theta_2 = 0$, but they also depend on θ_1 and the correlation, ρ, between the primary and secondary endpoints.

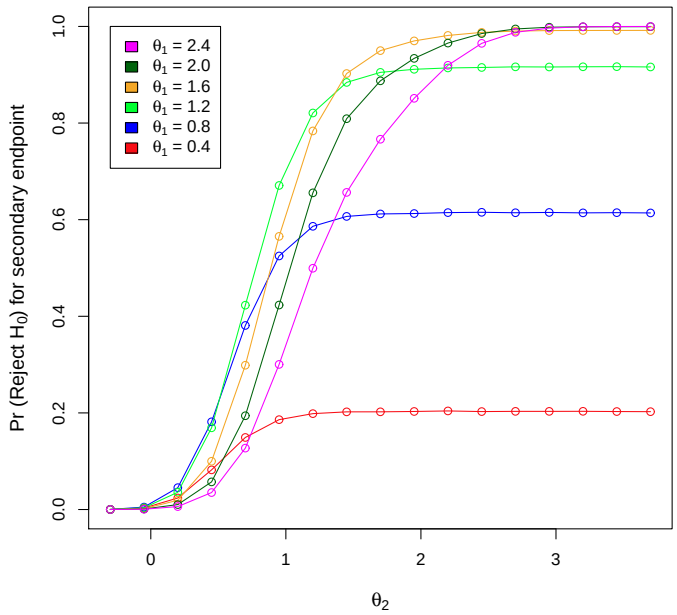
The OBF test of H_2 is particularly conservative when θ_1 is large.

Power for testing H_2, $\rho = 0.25$

A: Pocock boundary for H_2



B: OBF boundary for H_2



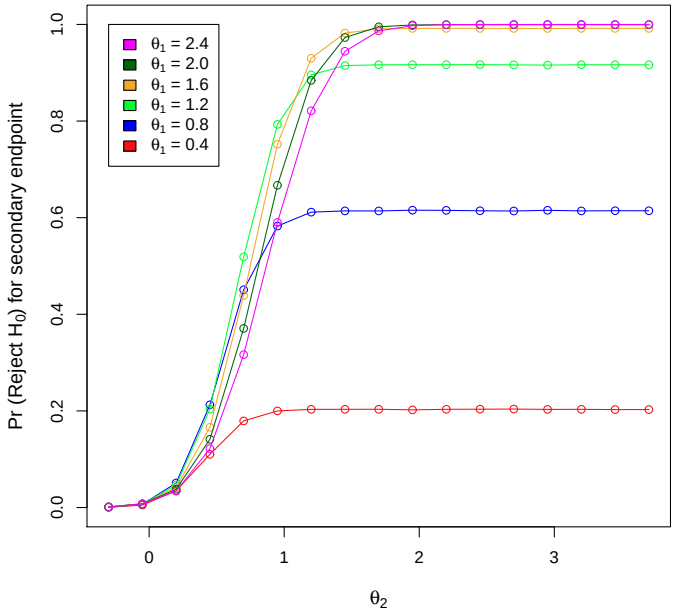
Results are shown for the case that the variance of the secondary response is 0.5 times that for the primary response.

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

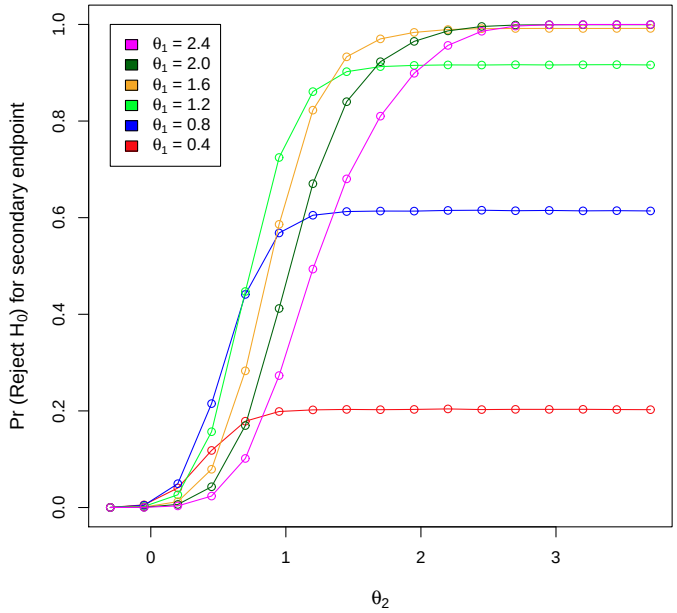
The Pocock boundary for H_2 deals better with the trial's uncertain termination time — which depends significantly on the value of θ_1.

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.

- Closed testing procedures provide a systematic way to construct tests that control the familywise error rate.

- Graphical representations (Statist. in Med. 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.

Recapitulation: Testing multiple hypotheses

JT (2025) has more on multiple hypotheses in:

- GSTs for Multivariate Endpoints (Chapter 19)
- Multi-arm Multi-stage Trials (Chapter 21)
- Adaptive Trials Testing Multiple Hypotheses (Chapter 25)
- Seamless Phase II/III Trials (Chapter 26)
- Adaptive Multi-arm Multi-stage Phase III Trials (Chapter 27)
- Enrichment Designs (Chapter 28)
- A Case Study with Survival Endpoints (Chapter 29)

We shall cover some of these topics in Part 4 of the course.

Part 3. Adaptive trial designs with a single hypothesis

3.1. Motivation for adaptive designs

3.2. Combination tests

In a two-stage trial; in multi-stage trial

3.3. The conditional probability of rejection principle

3.4. Sample size re-assessment based on estimates of a nuisance parameter

3.5. Sample size re-assessment in response to external information

3.6. Sample size re-assessment based on estimates of the treatment effect

The promising zone approach

3.7. Inference on termination of an adaptive trial

3.1 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 set 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.

ICH E20: Adaptive Designs for Clinical Trials

The document presents key principles under five headings:

- Adequacy Within the Development Program
- Adequacy of Trial Planning
- Limiting the Chances of Erroneous Conclusions
- Reliability of Estimation
- Maintenance of Trial Integrity

There is an increased emphasis on estimation as well as testing.

This is motivated by the need to inform benefit-risk comparisons.

Avoiding bias in estimates will also help healthcare providers when they decide whether to pay for a new treatment.

3.2 Combination tests.

Reference JT: Chapters 22 and 23

We have seen that regulatory guidelines stress the importance of controlling the type I error rate.

Suppose we run a clinical trial adaptively in two stages:

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

Analyse results from Stage 1,

Consider external information, if appropriate.

Set the design for 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.

Suppose we wish to 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.

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

Now, 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.

Ideally, w_1 and w_2 should be proportional to the square roots of the Stage 1 and Stage 2 sample sizes.

However, **it is crucial that w_1 and w_2 are pre-specified** and not changed in response to observed data.

A combination test with more than two groups

Reference: Lehmacher & Wassmer, *Biometrics*, 1999).

A combination test can be applied in a group sequential trial.

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

Specify weights $w_1, \ldots, w_K$ and define cumulative Z-statistics

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

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

$$\mathcal{I}_k \;=\; \sqrt{\frac{w_1^2 + \ldots + w_k^2}{w_1^2 + \ldots + w_K^2}} \quad \text{for } 1 \leq k \leq K.$$

Since the weights $w_1, \ldots, w_K$ are pre-specified, this joint distribution is fixed, regardless of any adaptations that may occur.

A combination test with more than two groups

Let $(a_1, b_1), \ldots, (a_K, b_K)$ be the boundary of a group sequential, one-sided test of H_0: $\theta \leq 0$ against $\theta > 0$, with a non-binding futility boundary and type I error rate α at $\theta = 0$.

Then,

$$P\{Z_k > b_k \text{ for some } 1 \leq k \leq K\} = \alpha.$$

Hence, we can compare Z_k with the critical values a_k and b_k, stopping to

Reject H_0 if $Z_k > b_k$,

Accept H_0 if $Z_k < a_k$.

Under $\theta = 0$, we have

$$P\{\text{Reject } H_0\} \leq \alpha,$$

with equality if the futility boundary is applied rigorously.

3.3 The conditional probability of rejection principle

Reference: JT, Sections 23.5 and 23.6

Consider a fixed sample study designed to test H_0: $\theta \leq 0$ against $\theta > 0$ at significance level $\alpha = 0.025$.

The planned sample size per treatment arm is $N = 200$ and H_0 will be rejected if

$$Z_{200} > \Phi^{-1}(1-\alpha) = 1.96.$$

At an interim point after $n_1 = 120$ subjects per treatment arm have been observed, an unplanned analysis is performed and the observed value of the Z-statistic is $Z_1 = z_1$,

Based on this information, the investigators wish to increase the remaining sample size from 80 to 130 subjects per treatment arm in order to increase the power of their study.

How can the type I error rate be protected after this unplanned adaptation?

The conditional probability of rejection principle

We can calculate the conditional type I error probability before the change in sample size,

$$p^*(z_1) = P_{\theta=0}\{Z_{200} > 1.96 \,|\, Z_1 = z_1\} = \Phi\left\{\frac{z_1\sqrt{120} - 1.96\sqrt{200}}{\sqrt{80}}\right\}.$$

Suppose the additional subjects produce a Z-value $Z^{(2)}$ and H_0 is rejected if

$$P^{(2)} \;=\; 1 - \Phi(Z^{(2)}) \;<\; p^*(z_1).$$

Then, the overall probability of a type I error remains at $\alpha = 0.025$.

Proschan & Hunsberger (*Biometrics*, 1995) use the "conditional error function", $p^*(z_1)$, to define two-stage tests with sample size adaptation.

Müller & Schäfer (*Biometrics*, 2001; *Statist. in Med.*, 2004) propose use of the "conditional probability of rejection principle" to analyse data after an unplanned adaptation.

The conditional probability of rejection principle

In our example, the rule to reject H_0 if

$$P^{(2)} = 1 - \Phi(Z^{(2)}) < p^*(z_1) = \Phi\{(z_1\sqrt{120} - 1.96\sqrt{200})/\sqrt{80}\}$$

produces exactly the same result as an inverse normal combination test with weights $w_1 = \sqrt{120/200}$ and $w_2 = \sqrt{80/200}$.

The "conditional probability of rejection principle" is attractive as it offers the flexibility to make unplanned adaptations — although recent regulatory guidances discourage these.

But, calculating the conditional probability of rejecting H_0 requires knowledge of the future information level or levels.

Suppose, in our example, the sample size of 120 per treatment arm was observed at the half way point of the trial. Should one use $N = 200$ or $N = 240$ in calculating $p^*(z_1)$?

Predicting information in a survival study is particularly difficult.

3.4 Sample size re-assessment: for a nuisance parameter

Reference, JT, Section 23.3

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

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

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 (2) 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 — or maybe replaces — 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.

Blinded variance estimation:

Friede & Miller (*Applied Statistics*, 2012) show that, for normally distributed data, sample size modification based on a certain blinded estimate of σ^2 leads to almost zero type I error 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 group sequential test

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.

Whatever the size of the kth group of subjects, 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.

3.5 Responding to external information

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.

Investigators decide to increase the second stage sample size to increase the 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.

3.6 Sample size re-assessment: in response to $\widehat{\theta}$

Reference JT: Chapter 24

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.

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 a maximum possible sample size but the actual sample size can be smaller than this — just like a GST.

Such designs can be viewed as group sequential tests in which group sizes are chosen based on the data observed so far.

The decision rule to reject or accept H_0 must take account of the data-dependent group sizes in order to protect the type I error rate.

Norbert Schmitz (1993) had proposed similar designs, which he called "Sequentially planned sequential decision procedures".

Sample size re-assessment in response to $\widehat{\theta}$

JT (*Statistics in Medicine*, 2003 and 2006) discussed proposals for adaptive designs that incorporate sample size re-assessment.

They claimed one could usually do better by

- Thinking carefully about power when designing the trial,
- Planning a maximum sample size to achieve this power,
- Using a GST to stop early when possible.

Proposals for adaptive designs with sample size re-assessment continued to appear.

JT (*Biometrika*, 2006) computed properties of optimal adaptive GSTs ("Schmitz" designs) and optimal GSTs.

They showed that familiar group sequential designs provide almost all the available efficiency gains — whereas some proposals for adaptive designs with sample size re-assessment are inefficient.

Mehta & Pocock's "Promising zone" design

Mehta & Pocock (*Statistics in Medicine*, 2011) published the paper

> *"Adaptive increase in sample size when interim results are promising: a practical guide with examples".*

Mehta & Pocock (MP) refer to the work of JT, saying

> *"These results are of great theoretical interest but of limited practical value for sponsors of industry trials."*

Jennison & Turnbull (*Statistics in Medicine*, 2015) discussed the MP designs and demonstrated various inefficiencies.

They proposed new, decision theoretic rules for sample size re-assessment and a different final hypothesis test.

Hsiao, Liu & Mehta (*Biometrical Journal*, 2019) wrote the paper "*Optimal promising zone designs*". They refer to JT's proposals as the "gold standard" and propose variations on these methods.

Mehta & Pocock's "Promising zone" design

The following slides summarise the content of JT (*SiM*, 2015)

> *"Adaptive sample size modification in clinical trials: start small then ask for more?"*

JT compare **adaptive trial designs** that

- Start with a fixed sample size design,
- Examine interim data,
- Add observations to increase power when appropriate

with **group sequential designs** that

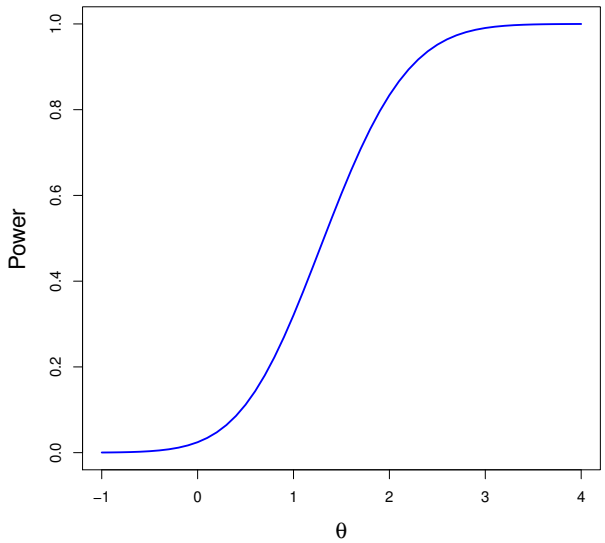
- Specify the desired power function,
- Set the maximum sample size appropriately,
- Stop at an interim analysis if data support an early conclusion.

Mehta & Pocock's "Promising zone" design

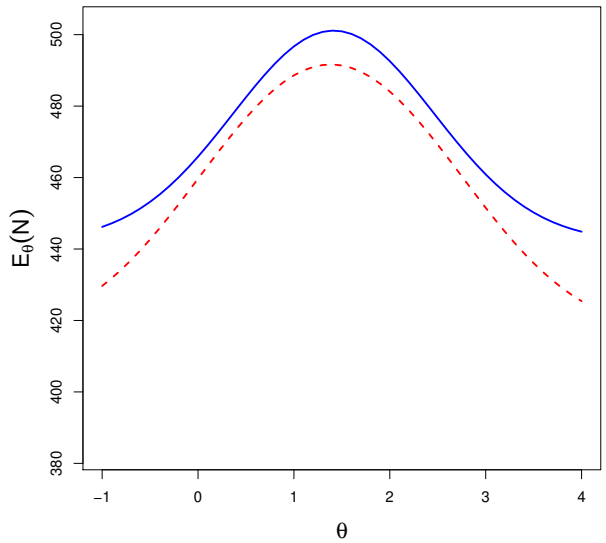
JT note that the "unconditional" properties of a trial design are important.

Suppose the treatment effect is denoted by θ and a trial is conducted to test H_0: $\theta \leq 0$ vs $\theta > 0$.

Common power curve



$E_\theta(N)$ curves



If two designs have the same power curve, then the design with the lower expected sample size function is to be preferred.

MP's Example 1: Phase 3 trial of schizophrenia treatments

A new drug is to be tested against an active comparator.

The efficacy endpoint is improvement in the Negative Symptoms Assessment score from baseline to week 26.

The initial plan is for a total of 442 patients (221 per treatment) which gives power 0.61 if $\theta = 1.6$ — but power 0.8 is desired.

An interim analysis is planned after observing $n_1 = 208$ responses (104 on each treatment).

With staggered accrual and the 26-week time to response, another 208 pipeline subjects will have been randomised but followed up for less than 26 weeks by the interim analysis time.

At the interim analysis, the total sample size will be revised to a value n_2, where n_2 is in the range 442 to 884.

The choice of n_2 will be based on the conditional power under the current estimate of θ.

MP's Example 1: Protecting the type I error probability

Chen, DeMets & Lan (*Statistics in Medicine*, 2004) consider a trial testing H_0: $\theta \leq 0$ vs $\theta > 0$ with type I error probability α.

Suppose an interim analysis is performed after a certain fraction of the total sample size has been observed.

Define the conditional power under treatment effect θ as

$$CP(\theta) = Pr_\theta\{H_0 \text{ will be rejected} \,|\, \text{Interim estimate} = \widehat{\theta}_1\}.$$

Chen, DeMets & Lan (CDL) show that if

$CP(\widehat{\theta}_1) > 0.5$,

The sample size is increased,

A standard fixed sample analysis is carried out,

then the type I error probability will be no greater than α.

MP's Example 1: The "Promising zone" design

MP control the type I error rate using an extension of the CDL result, due to Gao, Ware & Mehta (*Biopharm. Statistics*, 2008).

The total sample size, n_2, is a function of $CP(\widehat{\theta}_1)$, the conditional power under the current treatment effect estimate .

In MP's Example 1, the sample size rule is

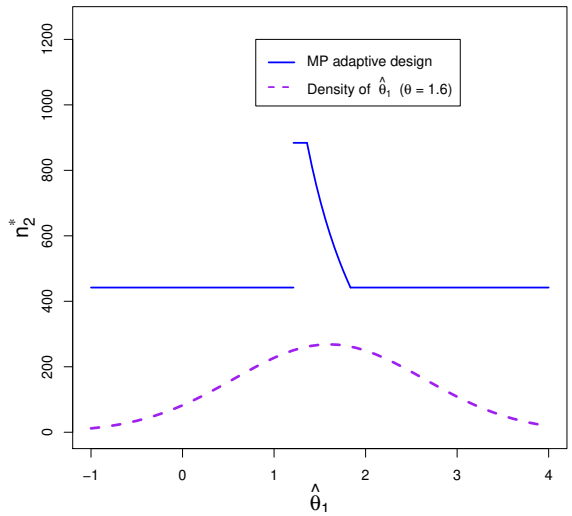
Favorable	$CP(\widehat{\theta}_1) \geq 0.8$	Continue to $n_2 = 442$,
Promising	$0.365 \leq CP(\widehat{\theta}_1) \leq 0.8$	Increase n_2,
Unfavourable	$CP(\widehat{\theta}_1) < 0.365$	Continue to $n_2 = 442$.

The "Promising zone" is simply the region where $CP(\widehat{\theta}_1) \leq 0.8$ and the Gao et al. extension of the CDL result can be employed.

In this zone, n_2 is increased to give $CP(\widehat{\theta}_1) = 0.8$, subject to a cap of $n_2 = 884$.

MP's Example 1: The "Promising zone" design

The sample size re-assessment rule for MP's "Promising zone" design is shown below.



Note that the region in which changes to the sample size occur is small when compared to the distribution of $\widehat{\theta}_1$.

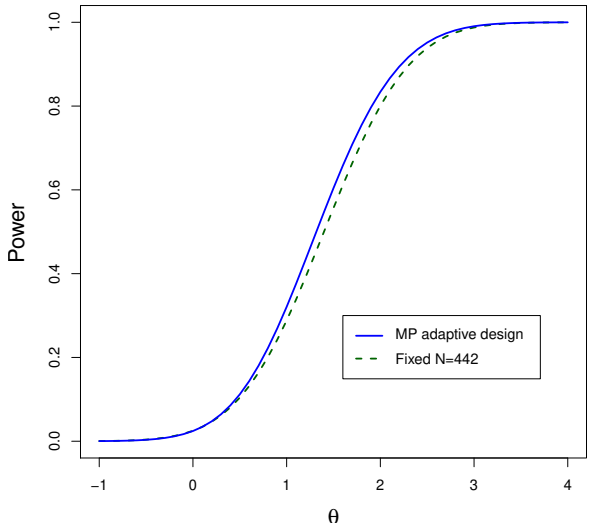
The estimate $\widehat{\theta}_1$ has a double role in $CP(\widehat{\theta}_1)$: it is both the current data and the value of θ at which conditional power is calculated.

This helps explain the high sensitivity of n_2 to $\widehat{\theta}_1$.

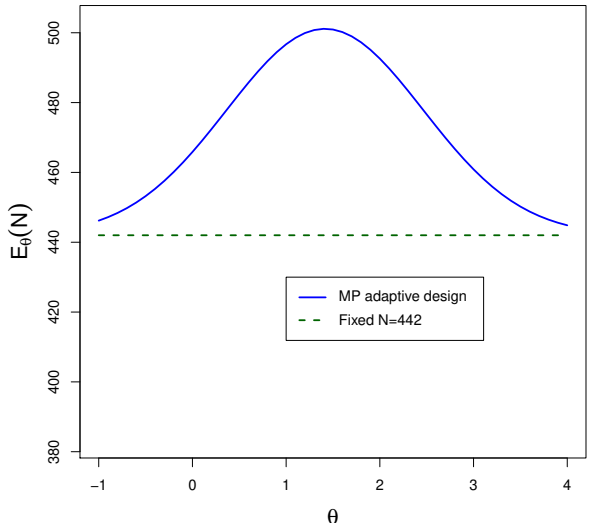
MP's Example 1: The "Promising zone" design

We can compare the overall, "unconditional" properties of the MP design with those of the fixed sample size design with $N = 442$.

Power curves



$E_\theta(N)$ curves



Although it is stated that power 0.8 at $\theta = 1.6$ would be desirable, power at this effect size has only risen from 0.61 to 0.66.

The cost of this increase in power is a considerably higher expected sample size function.

MP's Example 1: What might we do instead?

We could opt for:

(i) A fixed sample design with $N = 490$.

(ii) A group sequential design with $n_1 = 208$ and stopping rule:

At analysis 1

If $Z_1 \geq 2.54$	Stop, reject H_0
If $Z_1 \leq 0.12$	Stop, accept H_0
If $0.12 < Z_1 < 2.54$	Continue to $n_2 = 514$

At analysis 2

If $Z_2 \geq 2.00$ Reject H_0

If $Z_2 < 2.00$ Accept H_0.

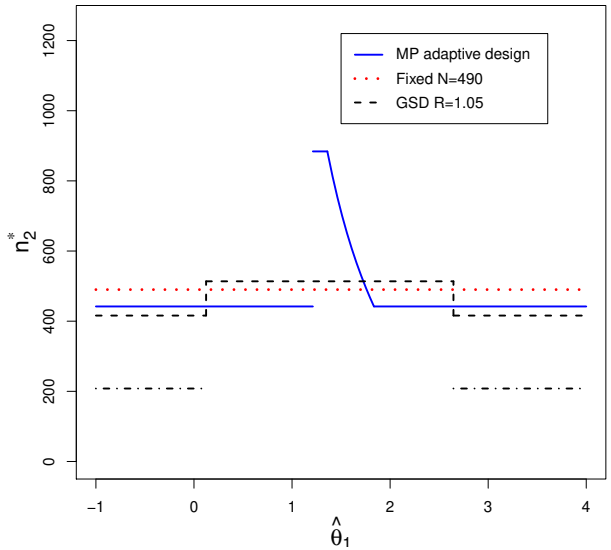
Error spending GST,
Rho-family, $\rho = 2$,
Power 0.8 at $\theta = 1.9$.

If the trial stops at analysis 1, "pipeline subjects" are not used in hypothesis testing, but they will contribute to $E_\theta(N)$.

MP's Example 1: What might we do instead?

Sample size rules for

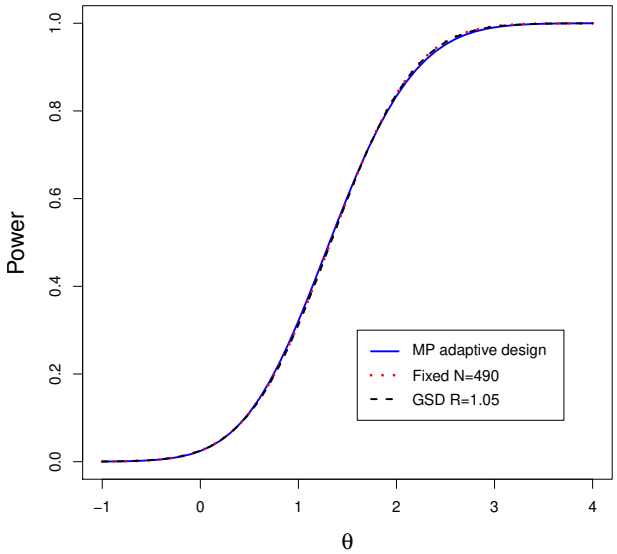
- Mehta & Pocock design (MP),
- Fixed sample size trial (N = 490),
- Group sequential design (GSD).



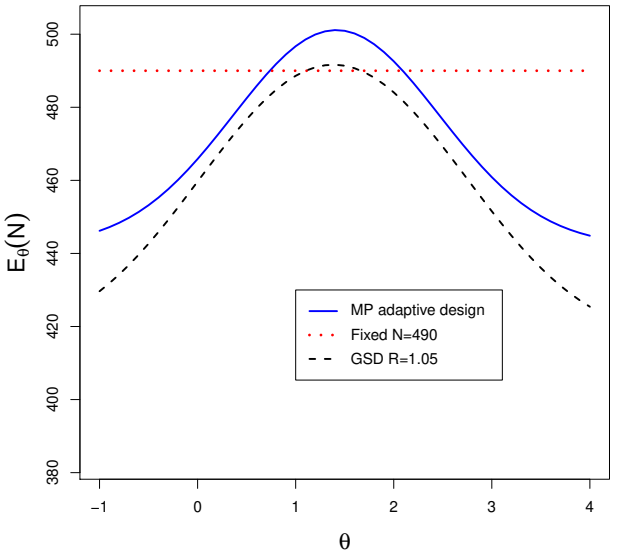
Two black lines show $n_1 = 208$, the number of observations, and $n_1^* = 416$, the number of patients enrolled at analysis 1.

MP's Example 1: Properties of three trial designs

Power curves



$E_\theta(N)$ curves



By construction, the fixed sample design and GSD have power curves matching that of the MP design.

The fixed sample size design is more efficient than the MP design at important values of θ between 1 and 2.

Despite the "pipeline" patients, the GSD is more efficient than the MP design at all values of θ.

MP's Example 1: Refining the "Promising zone" design

Jennison & Turnbull (2015) discuss how to modify Mehta & Pocock's design to improve its efficiency.

They recommend dropping the CDL methodology.

Instead, they propose using an inverse normal combination test to combine data from before and after sample size re-assessment.

JT derive a sample size rule in which n_2 is chosen to maximise a combined objective

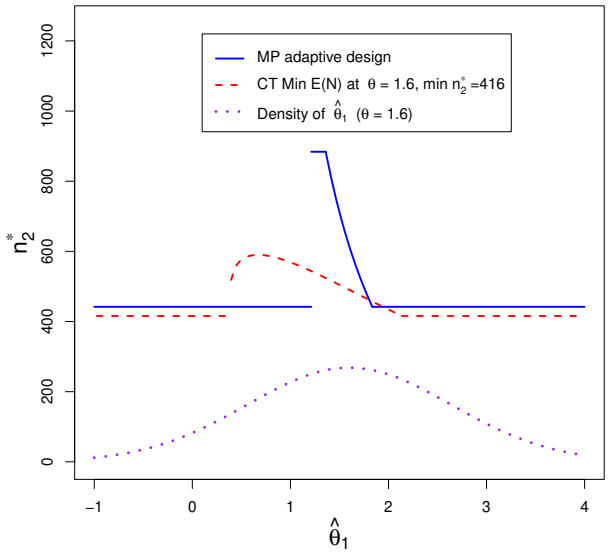
$$P_{\theta=1.6}\{\text{Reject } H_0 \,|\, \widehat{\theta}_1, n_2\} \,-\, \gamma(n_2 - 442).$$

Here, the tuning parameter $\gamma > 0$ represents a "rate of exchange" between conditional power and sample size.

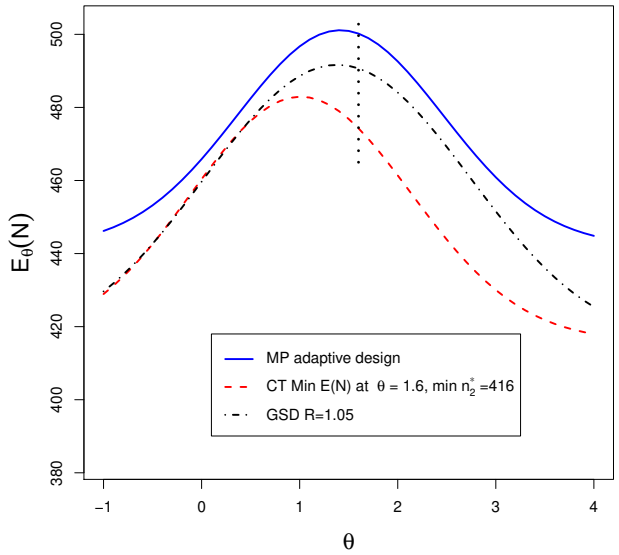
JT prove that this sample size rule achieves the minimum possible $E_{\theta=1.6}(N)$ among rules that achieve the same power when $\theta = 1.6$.

MP's Example 1: Refining the "Promising zone" design

Sample size rules



$E_\theta(N)$ curves



JT's refined design with $\gamma = 0.000622$ has the same power curve as the MP design.

It increases n_2 by a smaller amount of a wider range of $\widehat{\theta}_1$.

In the refined design, there is no stopping at the interim analysis.

Responses of "pipeline" patients are included in the final analysis.

The refined design reduces $E_\theta(N)$ over a wide range of θ values.

Sample size re-assessment in response to $\widehat{\theta}$: Conclusions

Group sequential designs

GSDs offer an excellent way to adapt sample size to observed data.

Efficient error spending designs are available — and we have good methods for inference on termination.

Designs with sample size re-assessment

Designs with SSR offer a little extra efficiency, but many proposed SSR designs do not achieve this.

If you want an SSR design, consider the approach of JT (2015).

Can you start small and ask for more?

Both GSDs and SSR designs require a commitment to obtain their implied maximum sample size, should this be needed.

Dealing with pipeline data:

See "Group sequential tests for delayed responses" by Hampson & Jennison (*JRSS, B*, 2013).

3.7 Inference on termination of an adaptive trial

Reference: JT, Section 23.8

Consider a 2-stage adaptive design with an inverse normal combination test, or a multi-stage Lehmacher-Wassmer design with sample size re-assessment.

Methods similar to those for GSTs in Section 1.7 can be used to derive a

Bias-adjusted estimate of θ,

P-value for H_0: $\theta \leq 0$,

Confidence interval for θ.

The methods grow more elaborate as we increase the number of stages or the degree of flexibility in the trial design.

Inference after a combination test: Point estimation

Estimate of θ after a 2-stage design with SSR

Consider a two treatment comparison with normal observations.

With θ denoting the difference in mean responses, it is desired to test H_0: $\theta \leq 0$ against $\theta > 0$ with type I error rate $\alpha = 0.025$.

Assuming a known variance, $\sigma^2 = 9$, observing 140 subjects per treatment arm will give power 0.8 if $\theta = 1$.

In a 2-stage design, the first stage with $n_1 = 70$ subjects per treatment produces the estimate $\hat{\theta}^{(1)}$.

The second stage sample size per arm is set as

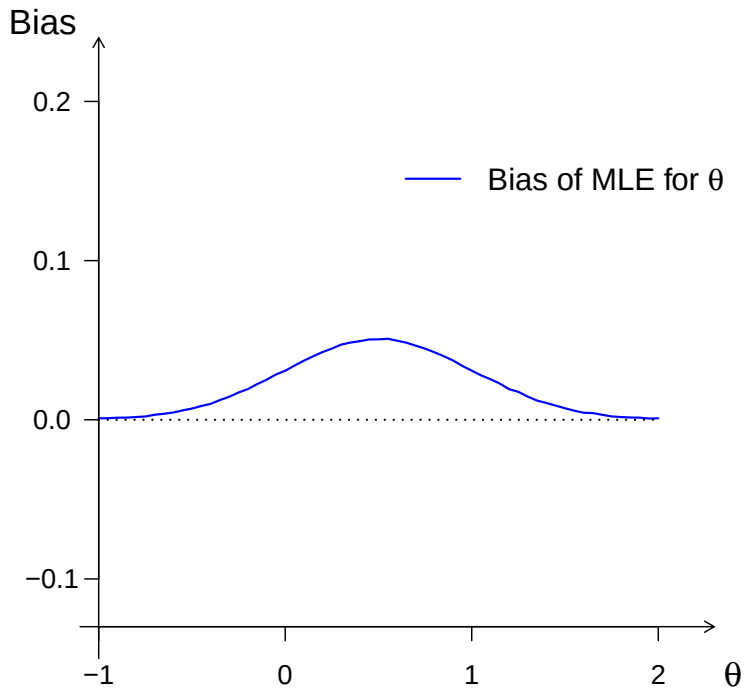
$$n_2 = 70 \text{ if } \hat{\theta}^{(1)} \geq 0.5, \quad n_2 = 210 \text{ if } \hat{\theta}^{(1)} < 0.5.$$

On termination, the MLE of θ is

$$\widehat{\theta}_{ML} = \frac{n_1\hat{\theta}^{(1)} + n_2\hat{\theta}^{(2)}}{n_1 + n_2}.$$

Inference after a combination test: Point estimation

The MLE is positively biased with a maximum bias of 0.05.



However, for any value of u, the estimate

$$\widehat{\theta} = u\,\hat{\theta}^{(1)} + (1-u)\,\hat{\theta}^{(2)}$$

is unbiased.

Inference after a combination test: Point estimation

Suppose an inverse normal combination test is to be applied with weights

$$w_1 = \sqrt{n_1^*/(n_1^* + n_2^*)} \quad \text{and} \quad w_2 = \sqrt{n_2^*/(n_1^* + n_2^*)},$$

where n_1^* and n_2^* denote "anticipated" stage 1 and 2 sample sizes.

If we set

$$\widehat{\theta} = w_1^2\,\hat{\theta}^{(1)} + w_2^2\,\hat{\theta}^{(2)},$$

this estimator will be the MLE of θ whenever the anticipated sample sizes, n_1^* and n_2^*, arise.

In our example, we could set

$$w_1 = \sqrt{70/(70+140)} \quad \text{and} \quad w_2 = \sqrt{140/(70+140)}$$

to give a reasonably efficient test and an estimate fairly close to the MLE for both the cases $n_2 = 70$ and $n_2 = 210$.

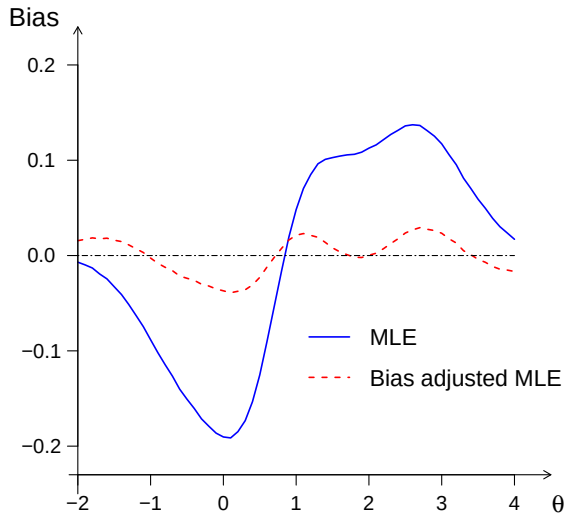
Inference after a combination test: Point estimation

Estimate of θ after a GST with SSR

Just as for a non-adaptive GST, we can adjust the maximum likelihood estimator of θ by estimating the bias function and subtracting the estimated bias from the MLE.

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

But now, if $1.3 \leq Z_2 \leq 2.0$, the size of the final two groups is increased by a factor of 1.5.

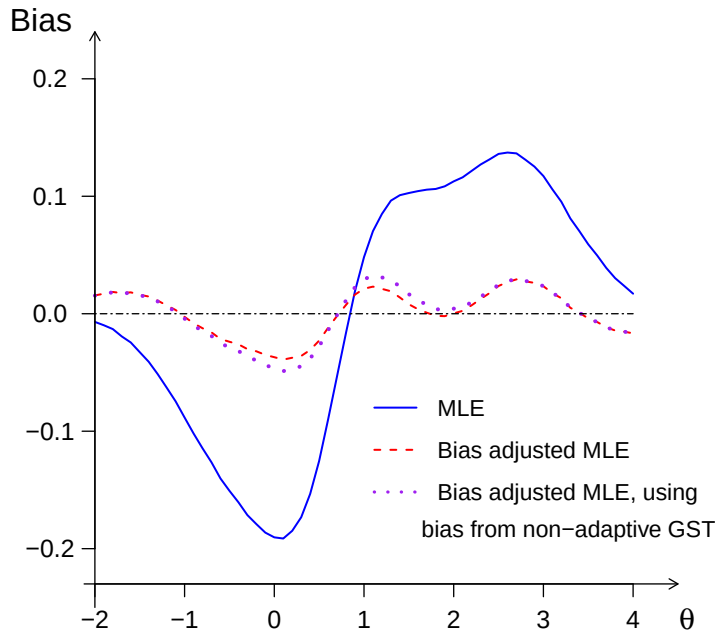


The bias adjustment of the MLE is just as effective as it was for a non-adaptive GST.

Inference after a combination test: Point estimation

In a flexible design where the sample size rule is not completely fixed, it is not possible to calculate the bias function.

However, it may suffice to use an approximation to this function.



The dotted line in this graph shows the bias of an adjusted MLE using the bias function from the initial GST, assuming no adaptation occurs.

Inference after a combination test: P-value

P-Value after a 2-stage design with SSR and early stopping

Suppose an inverse normal combination test is to be applied with weights w_1 and w_2 and boundary values a_1 and b_1 at analysis 1.

Denote the stage-wise Z-values by $Z^{(1)}$ and $Z^{(2)}$ and the final combined Z-value for testing H_0: $\theta \leq 0$ against $\theta > 0$ by

$$Z_2 = w_1 Z^{(1)} + w_2 Z^{(2)}.$$

Define the outcome as (T, Z) when the trial stops at analysis T, where $Z = Z^{(1)}$ if $T = 1$ or $Z = (Z^{(1)}, Z^{(2)})$ if $T = 2$.

Order the sample space $\{(T, Z)\}$ by:

- The boundary crossed,
- The stage at which the trial stops,
- The Z-statistic, $Z^{(1)}$ within stage 1 or Z_2 within stage 2.

Inference after a combination test: P-value

Specifically, we say $(t', z') \succ (t, z)$ if any one of the following four conditions holds:

(i) $t' = t = 1$ and $z'^{(1)} \geq z^{(1)}$,

(ii) $t' = t = 2$ and $w_1 z'^{(1)} + w_2 z'^{(2)} \geq w_1 z^{(1)} + w_2 z^{(2)}$,

(iii) $t' = 1,\ t = 2$ and $z'^{(1)} \geq b_1$,

(iv) $t' = 2,\ t = 1$ and $z^{(1)} \leq a_1$.

We calculate the P-value as

$$P\{(T, Z) \succ (t^*, z^*)\},$$

where (t^*, z^*) is the observed value of (T, Z).

Inference after a combination test: Confidence intervals

CI after a 2-stage design with SSR and no early stopping

In constructing the CI for θ, we test hypotheses $H\colon\ \theta = \theta_0$ for all values θ_0. If $H\colon\ \theta = \theta_0$ is not rejected, θ_0 is included in the CI.

The test of $H\colon\ \theta = \theta_0$ combines Z-values

$$Z_0^{(1)} = \frac{\widehat{\theta}^{(1)} - \theta_0}{\sqrt{\{\mathrm{Var}(\widehat{\theta}^{(1)})\}}} \quad \text{and} \quad Z_0^{(2)} = \frac{\widehat{\theta}^{(2)} - \theta_0}{\sqrt{\{\mathrm{Var}(\widehat{\theta}^{(2)})\}}},$$

which have independent $N(0,1)$ distributions if $\theta = \theta_0$.

If no early stopping is permitted at the interim analysis, a $100(1-2\alpha)\%$ CI comprises the set of values θ_0 for which

$$|w_1\, Z_0^{(1)} + w_2\, Z_0^{(2)}| \ \leq\ \Phi^{-1}(1-\alpha).$$

Inference after a combination test: Confidence intervals

CI after a 2-stage design with SSR and early stopping

Reference: JT, last part of §23.8.2.2.

In constructing a $100(1-2\alpha)\%$ CI for θ, we need a level 2α test of H: $\theta = \theta_0$ vs $\theta \neq \theta_0$ for each θ_0.

Given θ_0, if the trial stops at analysis T, we define the outcome as (T, Z_0), where $Z_0 = Z_0^{(1)}$ if $T = 1$ or $Z_0 = (Z_0^{(1)}, Z_0^{(2)})$ if $T = 2$.

Order the sample space $\{(T, Z_0)\}$ by saying $(t', z_0') \succ (t, z_0)$ if any one of the following four conditions holds:

(i) $t' = t = 1$ and $z_0'^{(1)} \geq z_0^{(1)}$,

(ii) $t' = t = 2$ and $w_1 z_0'^{(1)} + w_2 z_0'^{(2)} \geq w_1 z_0^{(1)} + w_2 z_0^{(2)}$,

(iii) $t' = 1,\ t = 2$ and $z_0'^{(1)} \geq b_1$,

(iv) $t' = 2,\ t = 1$ and $z_0^{(1)} \leq a_1$.

Inference after a combination test: Confidence intervals

For a given a value θ_0 and the above ordering of outcomes (T, Z_0), find $(t_u(\theta_0), z_u(\theta_0))$ and $(t_\ell(\theta_0), z_\ell(\theta_0))$ such that

$$P_{\theta=\theta_0}\{(T, Z_0) \succeq (t_u(\theta_0), z_u(\theta_0))\} = \alpha$$

and

$$P_{\theta=\theta_0}\{(T, Z_0) \preceq (t_\ell(\theta_0), z_\ell(\theta_0))\} = \alpha.$$

Then, the acceptance region for a test of $H\colon\ \theta = \theta_0$ is

$$A(\theta_0) = \{(t, z_0)\colon (t_\ell(\theta_0), z_\ell(\theta_0)) \prec (t, z_0) \prec (t_u(\theta_0), z_u(\theta_0))\}.$$

Hence, a two-sided $100(1-2\alpha)\%$ level confidence interval for θ is

$$\{\theta_0\colon (T, Z_0) \in A(\theta_0)\}.$$

Inference after a combination test: Confidence intervals

CI after a multi-stage group sequential design with SSR

Reference: JT, last part of §23.8.2.4.

Consider a trial designed to test H_0: $\theta \leq 0$ vs $\theta > 0$ using a Lehmacher-Wassmer multi-stage combination test.

Suppose this design has:

- More than two analyses where stopping can occur,
- One "adaptive" analysis where future group sizes may be modified in a flexible manner,
- Two or more analyses where stopping can occur after the adaptive analysis.

In constructing a $100(1-2\alpha)\%$ CI for θ, we need to define a level 2α test of H: $\theta = \theta_0$ vs $\theta \neq \theta_0$ for each θ_0.

This is not a straightforward task!

Inference after a combination test: Confidence intervals

The method we applied for a 2-stage design does not have a simple extension to multi-stage designs.

JT present a solution based on tests of each hypothesis H: $\theta = \theta_0$, related to the method of Gao, Liu & Mehta (*Statist. Med.*, 2013).

JT define a "reference design" with a specified sequence of information levels, and level α tests of H: $\theta = \theta_0$ vs $\theta > \theta_0$ and H: $\theta = \theta_0$ vs $\theta < \theta_0$ appropriate to this design.

If the adaptive analysis is reached, one calculates the conditional probabilities under $\theta = \theta_0$ that these tests will reject H: $\theta = \theta_0$ when the trial follows the reference design.

These probabilities are then used as the type I error rates in tests of H: $\theta = \theta_0$ vs $\theta > \theta_0$ and H: $\theta = \theta_0$ vs $\theta < \theta_0$ after adaptation.

The $100(1-2\alpha)\%$ CI for θ comprises the set of values θ_0 for which H: $\theta = \theta_0$ is not rejected.

Recapitulation: Adaptive designs with a single hypothesis

- **Combination tests** or the **conditional probability of rejection principle** can be used to create tests that protect the type I error rate in an adaptive trial.
- Sample size can be re-assessed based on estimates of parameters that affect the required sample size.
- Sample size can be re-assessed based on estimates of the treatment effect θ — but it is advisable to

  Consider whether a simple group sequential is preferable,

  Think carefully about the sample size rule.
- Inference after an adaptive trial is possible, producing a

  Point estimate, P-value, or Confidence interval.

Part 4 Adaptive trial designs with multiple hypotheses

4.1. Motivation

4.2. Protecting the familywise error rate

Closed testing procedures

Testing an intersection hypothesis

4.3. Using combination tests in a closed testing procedure

4.4. Seamless Phase 2/3 designs

4.5. Enrichment designs: testing for treatment effects in sub-populations

4.6. Inference on termination of an adaptive trial with multiple hypotheses

4.1 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*

Multi-arm multi-stage trial: *Several treatments compared to a control with elimination of poorly performing treatments.*

Enrichment trial: *Possible restriction to patients in a pre-specified sub-group*

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).

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.

Closed testing procedures

Reference: JT, Section 20.3

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.

Testing an intersection hypothesis

Testing $H_I = \cap_{i \in I} H_i$, the intersection of m simple hypotheses.

For each $i \in I$, let P_i be the 1-sided P-value for testing H_i and denote the ordered values of the P_i by $P_{[1]} \leq P_{[2]} \leq \ldots \leq P_{[m]}$.

The **Bonferroni** P-value for H_I is

$$P_I \; = \; m \, P_{[1]}.$$

The **Simes** P-value for H_I is

$$P_I \; = \; \min_{k=1,\ldots,m} (m \, P_{[k]}/k).$$

Comparing m treatments with a control, with Z-statistics Z_i.

The **Dunnett** P-value for H_I is

$$P\{\max_{i \in I} Z_i > z^*\}$$

where z^* is the observed value of $\max_{i \in I} Z_i$.

4.3 Using combination tests in a closed testing procedure

Reference: JT, Chapter 25

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 or more 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 stage k must be defined before commencing stage k, but it may depend on responses from stages 1 to $k-1$.

4.4 Seamless Phase 2/3 designs

Reference: JT, Chapter 26

In 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.

In 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, (Institutional Review Boards, FDA, . . .)

Run Phase 2b, analyse data, reach conclusions.

Write Phase 3 protocol, seek ethical and regulatory approval, (Institutional Review Boards, FDA, . . .)

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 Data Monitoring Committee 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 p dose levels plus the control in Phase 2.

Let θ_i, $i = 1, \ldots, p$, denote the effect size of dose level i *vs* the control treatment and define $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$.

We consider the null hypotheses H_i: $\theta_i \leq 0$, for $i = 1, \ldots, p$.

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 p 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 $\hat{\theta}_i^{(1)}$, $i = 1, \ldots, p$,

Phase 3

Treatment effect estimate $\hat{\theta}_{i^*}^{(2)}$.

Here, dose i^* is selected based on a high observed treatment effect in Phase 2, $\hat{\theta}_{i^*}^{(1)}$, 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 $\hat{\theta}_{i^*}^{(1)}$ as *the highest estimated effect from p 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 4.2 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 $\hat{\theta}_{i^*}^{(2)}$ — 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

$$-\log(P_I^{(1)}\, P_{i^*}^{(2)}),$$

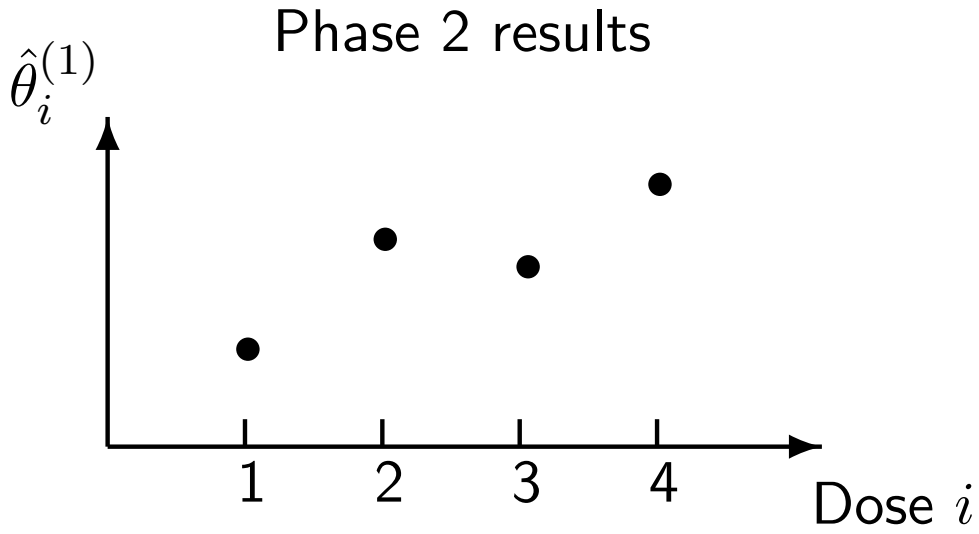
while the inverse normal combination test rejects H_I for high values of

$$w_1\Phi^{-1}(1-P_I^{(1)}) + w_2\Phi^{-1}(1-P_{i^*}^{(2)}).$$

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

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 in the set I 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 in this set I 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

With $P_1^{(1)} = 0.2,\ P_2^{(1)} = 0.04,\ P_3^{(1)} = 0.05,\ P_4^{(1)} = 0.03.$

Four-element sets I containing $i^* = 4$

There is just one four-element set, $I = \{1, 2, 3, 4\}$.

The ordered P-values are $P_{[1]}^{(1)} = 0.03,\ P_{[2]}^{(1)} = 0.04,\ P_{[3]}^{(1)} = 0.05$ and $P_{[4]}^{(1)} = 0.2$, so

$$P_I^{(1)} \ = \ \min_{j=1,\ldots,4} (4\, P_{[j]}^{(1)}/j) \ = \ 4 \times 0.05/3 \ = \ 0.067.$$

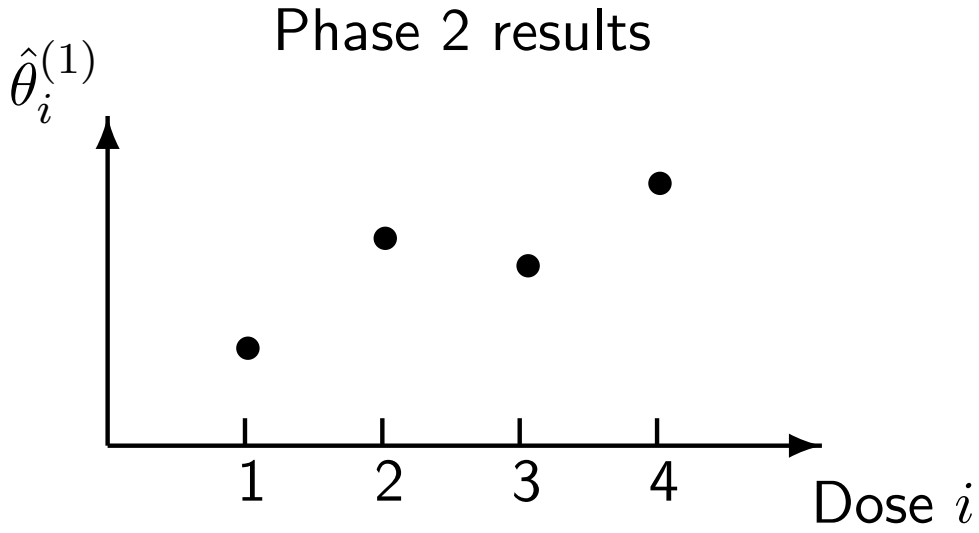
Conclusion:

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

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

$$Z_4^{(1)} \;=\; \Phi^{-1}(1 - 0.03) \;=\; 1.88.$$

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 $\hat{\theta}_i^{(1)}$.

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 $i = 1, \ldots, p$ and control,

Identify $\hat{\theta}_{i^*}^{(1)}$, the maximum of the effect estimates $\hat{\theta}_i^{(1)}$,

If $\hat{\theta}_{i^*}^{(1)} < C_1$, stop and accept H_0: $\theta_1 \leq 0, \ldots, \theta_p \leq 0$,

If $\hat{\theta}_{i^*}^{(1)} \geq C_1$, select dose i^* and proceed to Phase 3.

Phase 3

Take m_2 observations on dose i^* and the control,

Combine data in $T_{i^*} = (m_1\, \hat{\theta}_{i^*}^{(1)} + m_2\, \hat{\theta}_{i^*}^{(2)})/(m_1 + m_2)$,

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 final statistic

$$T_{i^*} = (m_1\hat{\theta}_{i^*}^{(1)} + m_2\hat{\theta}_{i^*}^{(2)})/(m_1 + m_2)$$

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 are chosen to satisfy type I error and power requirements.

TSE set their type I error probability α when $\boldsymbol{\theta} = (0, \ldots, 0)$.

JT (*J. Biopharmaceutical Statistics*, 2007) express the TSE method as a Closed Testing Procedure and deduce that the familywise error rate is protected strongly at level α.

They explain how to adapt the method when a treatment i^* is selected with

$$\hat{\theta}_{i^*}^{(1)} \leq \max_{i=1,\ldots,p} \hat{\theta}_i^{(1)},$$

in which case the type I error rate is controlled conservatively.

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 p-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.

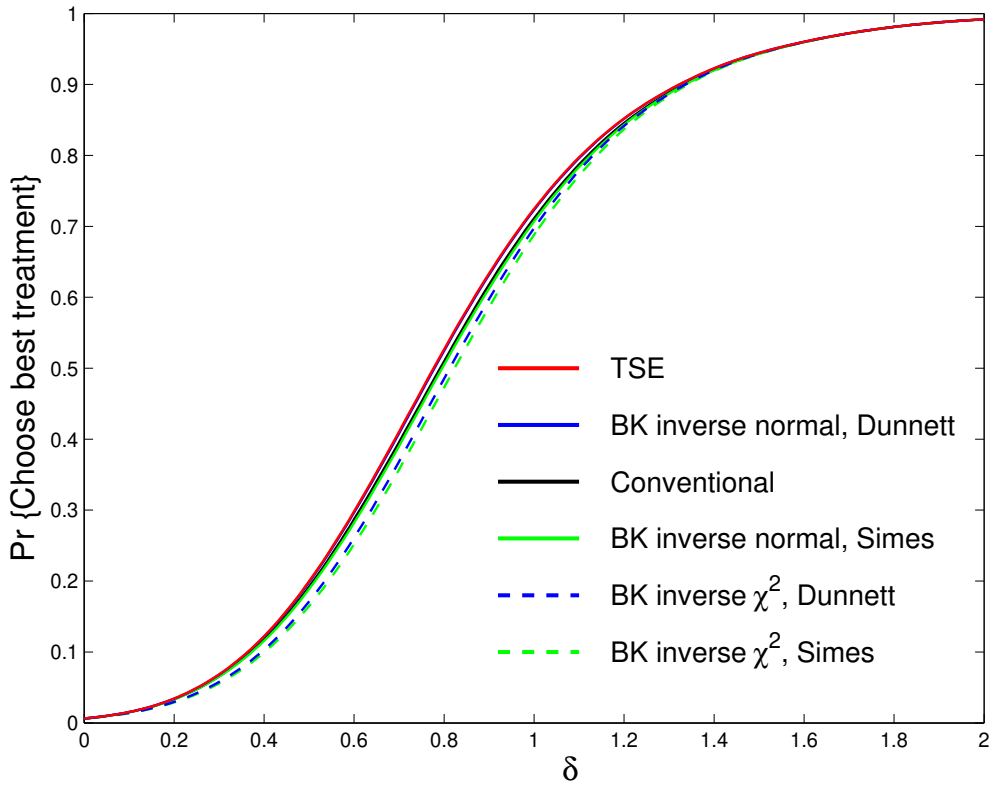
They considered situations when the vector of treatment effects is of the form $\boldsymbol{\theta} = (0, 0, 0, \delta)$.

Hampson & Jennison, *Statistics in Medicine, 2015* (H&J) extended 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).

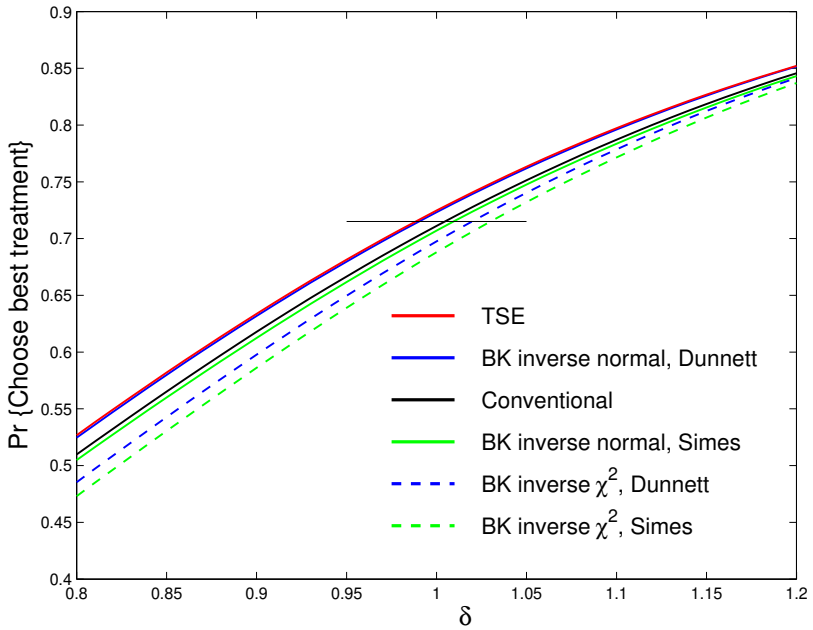
Power functions of six selection and testing procedures

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%.

Assessing the benefits of Seamless Phase 2/3 designs

H&J created a decision theoretic formulation of the testing problem at the end of a Phase 2/3 trial and searched 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.

They explain why Simes' rule can perform poorly when comparing multiple treatments with a control, due to the positive correlation between treatment effect estimates.

H&J compared the power of seamless designs with that of a "conventional" design with a higher Phase 3 sample size.

They conclude that Phase 2 data on dose i^* and the control are, typically, worth about half their face value in the final analysis.

Multi-arm Multi-stage trials

In a Phase 2/3 trial, one might consider carrying forward more than one treatment to the Phase 3 stage.

In addition, clearly inferior treatments could be dropped during the Phase 2 stage.

The trial would then fall in the general class of multi-arm multi-stage (MAMS) designs.

More broadly, platform trials or "umbrella trials" that study multiple treatments for a specific disease are examples of MAMS designs.

In a MAMS trial, one can follow the same approach using a closed testing procedure and combination tests to construct a testing procedure that protects the familywise error rate.

See JT, Chapter 27 for more details and illustrative examples.

4.5 Enrichment designs

Reference, JT, Chapter 28

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.

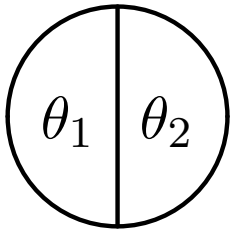
If the new treatment appears effective in the full population, continue as before.

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 $\theta_3 = \lambda_1\theta_1 + \lambda_2\theta_2$.

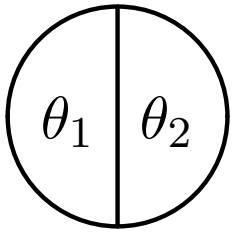
The null hypothesis for the sub-population is H_1: $\theta_1 \leq 0$.

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 $P_3^{(1)}$.

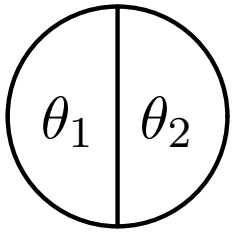
We use Simes' method to obtain a P-value from Stage 1 data for $H_{13} = H_1 \cap H_3$

$$P_{13}^{(1)} = \min\{2\min(P_1^{(1)}, P_3^{(1)}), \max(P_1^{(1)}, P_3^{(1)})\}.$$

(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 $P_1^{(2)}$.

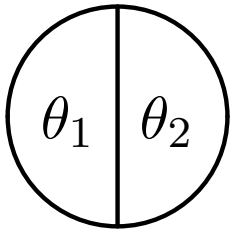
P-value for H_3 from the full Stage 2 data is $P_3^{(2)}$.

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 (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, continued

Stage 1 results

$$\text{Older men:} \quad P_1^{(1)} = 0.02$$

$$\text{Full population:} \quad P_3^{(1)} = 0.20$$

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, 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, 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

$$Z_1 = \sqrt{0.5}\,\Phi^{-1}(1-0.02) + \sqrt{0.5}\,\Phi^{-1}(1-0.03) = 2.78,$$
$$Z_{13} = \sqrt{0.5}\,\Phi^{-1}(1-0.04) + \sqrt{0.5}\,\Phi^{-1}(1-0.03) = 2.57,$$

$$P_1 = 1-\Phi(2.78) = 0.003,$$
$$P_{13} = 1-\Phi(2.57) = 0.005.$$

$\boldsymbol{H_1}$ **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.

Assessing the possible benefits of an enrichment design

Reference Burnett & Jennison, "Adaptive enrichment trials: What are the benefits?" *Statistics in Medicine*, 2020.

Burnett & Jennison (hereafter, BJ) considered a two-arm trial with a specified sub-population in which the new treatment was expected to be particularly beneficial.

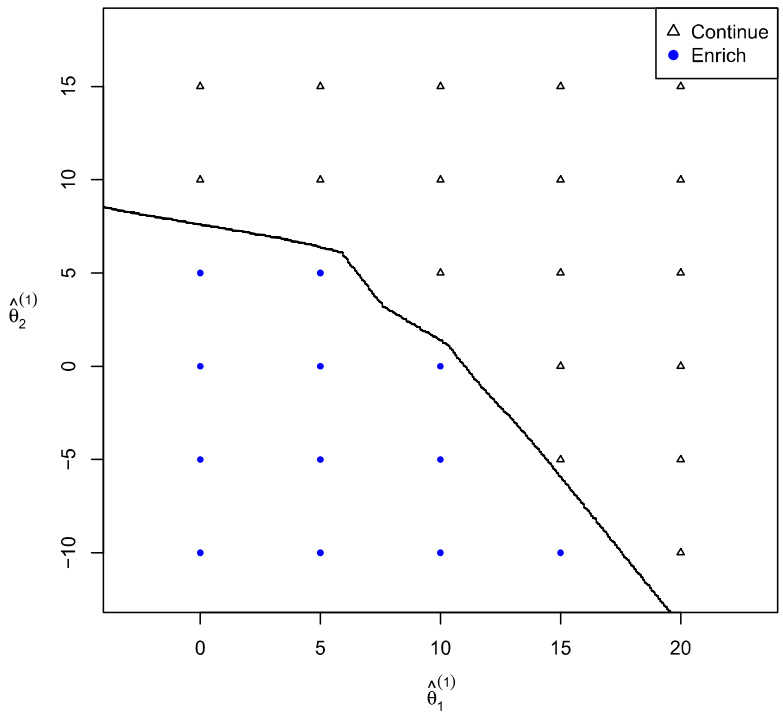
BJ studied adaptive enrichment designs using a closed testing procedure with Simes' test for the intersection hypothesis and inverse normal combination tests to combine data across stages.

A gain function was defined to measure the benefit from the trial outcome, depending on treatment effects in the full population and sub-population and which null hypotheses were rejected.

A prior distribution for the treatment effects was specified.

The benefits of an enrichment design

BJ derived the enrichment rule that maximised the expected, gain averaging over the prior distribution for the treatment effects.



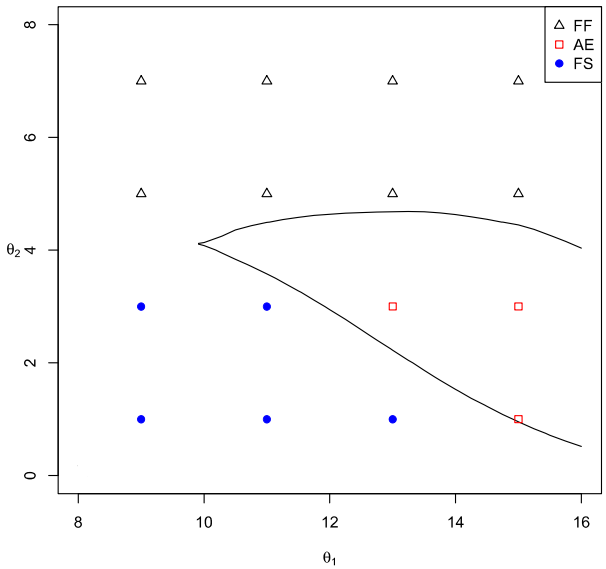
Optimal decision: Enrich if the interim estimate of (θ_1, θ_2) is below the line, otherwise continue to recruit from the full population.

The benefits of an enrichment design

BJ compared the optimised adaptive enrichment design against designs with the same total sample size:

- A simple trial design for the full population,
- A trial with recruitment only from the sub-population.

They found treatment effects (θ_1, θ_2) for which the enrichment design was superior to both non-adaptive designs — but by quite a small amount.



Optimal design

FF: Non-adaptive, full population

AE: Adaptive enrichment

FS: Sub-population only

The benefits of an enrichment design

A more positive conclusion:

BJ found cases where the adaptive enrichment design was clearly superior to the simple full population trial.

Although a trial confined to the sub-population from the outset could have even higher expected gain, investigators may be reluctant to focus only on the sub-population before seeing any data at all.

Also, regulators may require a new treatment to be tested in the full patient population — see the ICH E20 draft guidelines.

BJ's results show the importance of defining study objectives and comparing adaptive designs against simpler alternatives to make sure there are likely to be significant benefits from running a more complex study.

4.6 Inference after a trial with multiple hypotheses

Suppose $\theta_1, \ldots, \theta_p$ are the effects of different treatments for a given indication, or the treatment effects for a single treatment in different sub-populations.

A clinical trial is conducted to test H_i: $\theta_i \leq 0$, $i = 1, \ldots, p$. On termination, we wish to report estimates and confidence intervals for certain θ_i.

As an example, in a seamless Phase 2/3 trial, we may wish to estimate θ_{i^*} when treatment i^* is selected for the Phase 3 stage.

In discussing the bias of an estimate $\hat{\theta}_{i^*}$, we have to recognise that the index i^* is data dependent — in different realisations of the trial we may estimate a different treatment effect.

In constructing a confidence interval for θ_{i^*}, we need to consider the familywise coverage rate of intervals that we might report for different parameters θ_i.

Estimation after a seamless Phase 2/3 trial

We shall develop methods in the case of a seamless Phase 2/3 trial.

Here, there is just one parameter of interest after the trial ends: the treatment effect θ_{i^*} when treatment i^* is selected for the Phase 3 stage.

Bias of $\hat{\theta}_{i^*}$

We define the bias of $\hat{\theta}_{i^*}$ to be

$$\text{Bias}(\hat{\theta}_{i^*}) = \mathbb{E}(\hat{\theta}_{i^*} - \theta_{i^*}),$$

noting that both i^* and $\hat{\theta}_{i^*}$ are random variables.

Suppose we estimate θ_{i^*} as

$$\hat{\theta}_{i^*} = w_1 \hat{\theta}_{i^*}^{(1)} + w_2 \hat{\theta}_{i^*}^{(2)},$$

where $w_1 + w_2 = 1$ and w_1 and w_2 are proportional to the sample sizes per treatment in Phase 2 and Phase 3.

Estimation after a seamless Phase 2/3 trial

We have

$$\begin{aligned}
\text{Bias}(\hat{\theta}_{i^*}) &= \mathbb{E}\{w_1\,\hat{\theta}_{i^*}^{(1)} + w_2\,\hat{\theta}_{i^*}^{(2)} - \theta_{i^*}\} \\
&= \mathbb{E}\{w_1\,(\hat{\theta}_{i^*}^{(1)} - \theta_{i^*}) + w_2\,(\hat{\theta}_{i^*}^{(2)} - \theta_{i^*})\} \\
&= \mathbb{E}\{w_1\,(\hat{\theta}_{i^*}^{(1)} - \theta_{i^*})\}.
\end{aligned}$$

Let μ_i denote the mean response on treatment i, $i = 1, \ldots, p$, and μ_c the mean response on the control arm.

Thus, $\theta_i = \mu_i - \mu_c$ and

$$\begin{aligned}
\mathbb{E}\{\hat{\theta}_{i^*}^{(1)} - \theta_{i^*}\} &= \mathbb{E}\{(\hat{\mu}_{i^*}^{(1)} - \mu_{i^*}) - (\hat{\mu}_c^{(1)} - \mu_c)\} \\
&= \mathbb{E}\{\hat{\mu}_{i^*}^{(1)} - \mu_{i^*}\}.
\end{aligned}$$

Estimation after a seamless Phase 2/3 trial

Suppose that treatment i^* with the largest estimated effect, $\hat{\theta}_i^{(1)}$, is always selected after Phase 2.

Then, assessing the bias of $\hat{\theta}_{i^*}$ reduces to the problem of assessing

$$\mathbb{E}\{\hat{\mu}_{i^*}^{(1)} - \mu_{i^*}\},$$

where $\mu_i^{(1)}$, $i = 1, \ldots, p$, are independent and $i^* = \text{argmax}_i\ \hat{\mu}_i$.

In particular, the correlation between the Phase 2 estimates $\hat{\theta}_i^{(1)}$, $i = 1, \ldots, p$, has no effect on bias and it suffices to consider the problem of comparing treatments that yield independent effect estimates.

This problem has a long history and was studied from a theoretical perspective in the 1980s.

There has been fresh interest and activity, given the relevance to current applications.

Estimation after a seamless Phase 2/3 trial

Reference: JT, Section 26.6.1

Cohen & Sackrowitz (*Stat. and Prob. Letters*, 1989) consider a two-stage trial in which one normal distribution is selected after stage 1 and the mean of this distribution is to be estimated.

They present a Uniformly Minimum Variance Conditionally Unbiased Estimate (UMCVUE), where the conditioning is on the rank order of the stage 1 means.

Bretz, Koenig et al. (*Statist. Med.*, 2009) generalise this method to enable application to a Phase 2/3 trial with different sample sizes in Phases 2 and 3.

The UMCVUE can be represented as the MLE estimate

$$\hat{\theta}_{i*} = w_1 \hat{\theta}_{i*}^{(1)} + w_2 \hat{\theta}_{i*}^{(2)}$$

minus an adjustment which is a function of $\hat{\theta}_{i*}$ and the second largest Phase 2 estimate $\hat{\theta}_i^{(1)}$.

Estimation after a seamless Phase 2/3 trial

Further estimates have been proposed by

Luo, Wu & Xiong (*Biometrical Journal*, 2010)

Kimani, Todd & Stallard (*Statist. Med.*, 2013)

Carreras & Brannath (*Statist. Med.*, 2013)

Bowden, Brannath & Glimm (*Statist. Med.*, 2014)

Pickard & Chang (*Statist. Biopharm. Research*, 2014)

and, for a time to event endpoint, by

Brückner, Titman & Jaki (*Statist. Med.*, 2017)

Khan, Kimani, Glimm & Stallard (*Statist. Med.*, 2023).

Carreras & Brannath report simulations comparing several estimators which show that, while the UMVCUE has zero bias, it has a high MSE compared to other estimators.

Confidence interval for θ_{i^*} after a seamless Phase 2/3 trial

Reference: JT, Sections 26.6.2 and 26.6.3

A confidence interval for $\boldsymbol{\theta_{i^*}}$

Suppose we wish to construct an upper confidence interval (θ_{L,i^*}, ∞) for θ_{i^*}, for which

$$P_{\boldsymbol{\theta}}\{\theta_{i^*} \in (\theta_{L,i^*}, \infty)\} \geq 1 - \alpha \quad \text{for all } \boldsymbol{\theta} = (\theta_1, \ldots, \theta_p).$$

Since i^* may take any value in $\{1, \ldots, p\}$, we need to consider all the intervals $(\theta_{L,i}, \infty)$ that could be reported.

Furthermore, an upper confidence interval for θ_i is equivalent to a family of tests $H_i(\tilde{\theta}_i)$: $\theta_i \leq \tilde{\theta}_i$ vs $\theta_i > \tilde{\theta}_i$.

Thus, we need a family of one-sided tests of the hypotheses

$$H_i(\tilde{\theta}_i)\text{: } \theta_i \leq \tilde{\theta}_i, \quad i = 1, \ldots, p, \quad \tilde{\theta}_i \in R,$$

with familywise error rate α.

Confidence interval for θ_{i^*} after a seamless Phase 2/3 trial

Magirr, Jaki, Posch & Klinglmueller (*Biometrika*, 2013) describe a general method for finding a confidence region for the parameter vector after a two-stage adaptive trial.

In this approach the confidence intervals are guaranteed to be compatible with the outcomes of the closed testing procedure.

Section 26.6.3 of JT describes the application of the Magirr et al. approach to a seamless Phase 2/3 design.

We shall present this method when Dunnett tests are used for intersection hypotheses in Phase 2 and an inverse normal combination test is applied to the Phase 2 and Phase 3 results.

We define a family of one-sided tests of the hypotheses

$$H_i(\tilde{\theta}_i)\colon\ \theta_i \leq \tilde{\theta}_i, \quad i = 1, \ldots, p, \ \ \tilde{\theta}_i \in R,$$

with familywise error rate α.

Confidence interval for θ_{i^*} after a seamless Phase 2/3 trial

For a given i, the hypotheses $H_i(\tilde{\theta}_i)$ are nested with

$$\cap_{j=1}^{k} H_i(\tilde{\theta}_j) = H_i(\min_{j=1,\ldots,k} \tilde{\theta}_j).$$

Hence, it suffices to consider intersection hypotheses of the form

$$H_J(\tilde{\boldsymbol{\theta}}_J) = \cap_{j \in J} H_j(\tilde{\theta}_j)$$

where J is a subset of $\{1, \ldots, p\}$ and $\tilde{\boldsymbol{\theta}}_J$ contains a set of values $\tilde{\theta}_j$ for $j \in J$.

We suppose treatment effect estimates based on the Phase 2 data are distributed as

$$\hat{\theta}_j^{(1)} \sim N(\theta_j, 1/\mathcal{I}^{(1)})$$

and, if treatment i^* is selected at the end of Phase 2, the Phase 3 data provide the estimate

$$\hat{\theta}_{i^*}^{(2)} \sim N(\theta_{i^*}, 1/\mathcal{I}^{(2)}).$$

Confidence interval for θ_{i^*} after a seamless Phase 2/3 trial

The hypothesis $H_{i^*}(\tilde{\theta}_{i^*})$: $\theta_{i^*} \leq \tilde{\theta}_{i^*}$ is rejected in the closed testing procedure if each intersection hypothesis

$$H_J(\tilde{\boldsymbol{\theta}}_J), \quad \text{where } i^* \in J \text{ and } \tilde{\boldsymbol{\theta}}_J \text{ includes } \theta_{i^*} = \tilde{\theta}_{i^*}$$

is rejected by its level α test.

The "worst case" for this arises in the intersection hypothesis involving all p treatments when $\tilde{\theta}_i \to \infty$ for each $i \neq i^*$.

Then, if

$$Z_{i^*}^{(1)} = \{\hat{\theta}_{i^*}^{(1)} - \tilde{\theta}_{i^*}\}\sqrt{\mathcal{I}^{(1)}}$$

has the observed value $z_{i^*}^{(1)}$, the Phase 2 Dunnett P-value is

$$P_{i^*}^{(1)}(\tilde{\theta}_{i^*}) = P\{\max_{i \in \{1,\ldots,p\}} Z_i > z_{i*}^{(1)}\}$$

where 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 $i \neq i'$.

Confidence interval for θ_{i^*} after a seamless Phase 2/3 trial

If the observed value of $Z_{i^*}^{(2)}(\tilde{\theta}_{i^*})$ is

$$z_{i^*}^{(2)}(\tilde{\theta}_{i^*}) = (\hat{\theta}_{i^*}^{(2)} - \tilde{\theta}_{i^*})\sqrt{\mathcal{I}^{(2)}},$$

the Phase III P-value is the same for all these intersection hypotheses $H_J(\tilde{\boldsymbol{\theta}}_J)$ and equal to

$$P_{i^*}^{(2)}(\tilde{\theta}_{i^*}) = 1 - \Phi\{z_{i^*}^{(2)}(\tilde{\theta}_{i^*})\}.$$

Applying the combination test, we reject $H_{i^*}(\tilde{\theta}_{i^*})$: $\theta_{i^*} \leq \tilde{\theta}_{i^*}$ if

$$w_1\,\Phi^{-1}(1 - P_{i^*}^{(1)}(\tilde{\theta}_{i^*})) + w_2\,\Phi^{-1}(1 - P_{i^*}^{(2)}(\tilde{\theta}_{i^*})) \geq \Phi^{-1}(1-\alpha).$$

The upper confidence interval for θ_{i^*} comprises the set of values $\tilde{\theta}_{i^*}$ for which $H_{i^*}(\tilde{\theta}_{i^*})$: $\theta_{i^*} \leq \tilde{\theta}_{i^*}$ is not rejected.

JT (Section 26.6.3) illustrate this method with a worked example.

Inference after MAMS and enrichment designs

References: JT,

Section 27.6, multi-arm multi-stage (MAMS)

Section 28.6, enrichment

A key difference in these applications is that more than treatment may prove superior to the control, or a positive result may be achieved in more than one sub-population.

Thus, inferences may be made about several treatment effects θ_i but the number of these is random.

A body of methodology exists but further work is needed to address the most general problems.

Inference after MAMS and enrichment designs

Bias: MAMS designs

Bowden & Glimm (*Biometrical J.*, 2014) derive conditionally unbiased estimates of the treatment effects, conditioning on the sets of treatments remaining after each analysis.

Stallard & Kimani (*Biometrika*, 2018) derive minimum variance conditionally unbiased estimates for treatments remaining at the end of the trial.

Bias: Enrichment designs

Kimani, Todd & Stallard (*Biometrical J.*, 2015) consider designs with a full population and one specific sub-population. They recommend a hybrid approach using an unbiased estimator when enrichment occurs or naive estimators otherwise.

Kunzmann, Benner & Kieser (*Statist. Med.*, 2017) investigate six estimators and note that in many scenarios, improved bias control comes at the price of a larger variance and mean square error.

Inference after MAMS and enrichment designs

Confidence intervals

Suppose p treatment effects, $\theta_1, \ldots, \theta_p$ are investigated.

As for seamless Phase 2/3 design, we can proceed by defining a family of one-sided tests of the hypotheses

$$H_i(\tilde{\theta}_i)\text{: } \theta_i \leq \tilde{\theta}_i, \quad i = 1, \ldots, p, \quad \tilde{\theta}_i \in R,$$

with familywise error rate α.

Then, the upper confidence interval for θ_i comprises the set of values $\tilde{\theta}_i$ for which $H_i(\tilde{\theta}_i)\text{: } \theta_i \leq \tilde{\theta}_i$ is not rejected.

If a closed testing procedure is defined to maximise power to reject a null hypothesis $H_i(0)\text{: } \theta_i \leq 0$ and tests of $H_i(\tilde{\theta}_i)\text{: } \theta_i \leq \tilde{\theta}_i$ are added later, "non-informative" confidence intervals can result:

$H_i(0)\text{: } \theta_i \leq 0$ is rejected but the upper CI for θ_i is $(0, \infty)$.

If confidence intervals are important, they should be built into the testing procedure from the beginning.

Recapitulation: Adaptive designs with multiple hypotheses

- **Combination tests** can be combined with a **closed testing procedure** to create tests that protect the familywise type I error rate in an adaptive trial.
- This facilitates designs for

  Inferentially seamless Phase 2-3 trials

  Multi-arm multi-stage trials

  Enrichment trials
- Inference after an adaptive trial is possible, at least in some cases, to give a

  Point estimate, P-value or Confidence interval

  — but further work is needed here.