Group Sequential and Adaptive Clinical Trial Designs

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

http://people.bath.ac.uk/mascj

PSI One day Scientific meeting, South West:

Designing and Analysing Adaptive Trial Design Studies

Bath, 24 June 2019

Motivation: Phase 3 clinical trials

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

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

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

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

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

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

Motivation: Interim monitoring

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

Ethics

Administration (accrual, compliance, ...)

Economics

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

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

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

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

Motivation: Repeated hypothesis tests during a study

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

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

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

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

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

Motivation: Adaptive clinical trial designs

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

An adaptive trial could:

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

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

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

Outline of talk

1. **Group sequential tests**

   Sequential distribution theory

   Monitoring a survival study

   Computations for group sequential tests

   Benefits of group sequential testing

   Error spending tests

   Example 1: Normal response

   Example 2: Binary response

   Example 3: Survival endpoint

Outline of talk

2. Adaptive trial designs

A survival trial with treatment selection

Protecting the type I error rate

Multiple hypothesis testing: Closed Testing Procedures

Combination tests

Avoiding error rate inflation in an adaptive survival trial

Choosing an adaptive design and assessing its benefits

1.1 Group sequential tests: Introduction

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

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

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

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

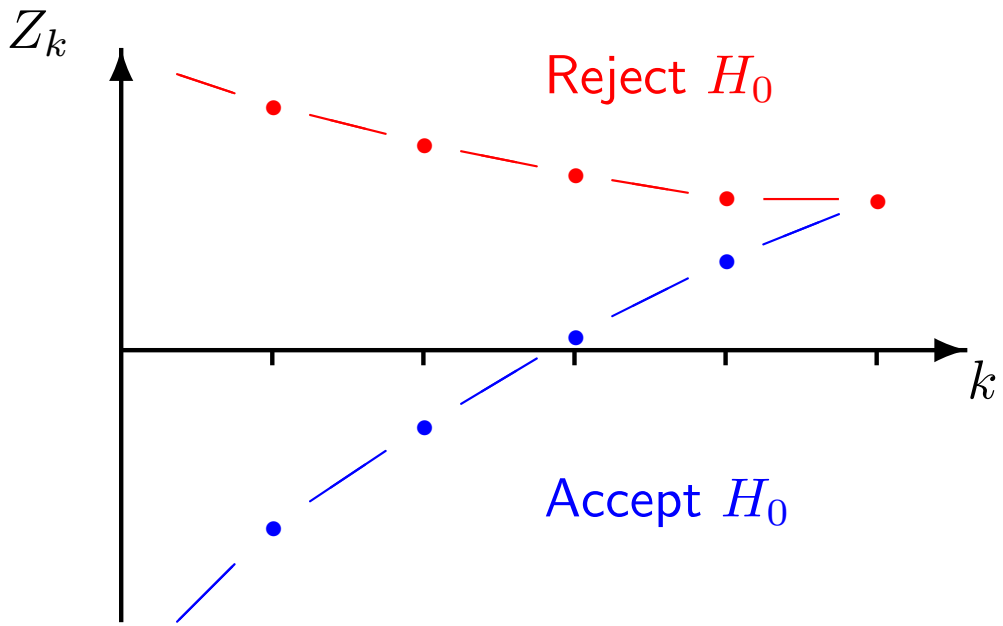
$$P_{\theta=0}\{\text{Reject } H_0\} = \alpha,$$

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

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

Group sequential tests

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



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

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

1.2 Joint distribution of parameter estimates

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

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

The information for θ at analysis k is

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

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

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

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

and

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

Sequential distribution theory

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

θ a single normal mean,

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

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

a general normal linear model,

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

Thus, theory supports general comparisons, including:

crossover studies,

analysis of longitudinal data,

comparisons adjusted for covariates.

Canonical joint distribution of 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 $\mathcal{I}_1, \ldots, \mathcal{I}_K$.

1.3 Survival data

The canonical joint distributions also arise for

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

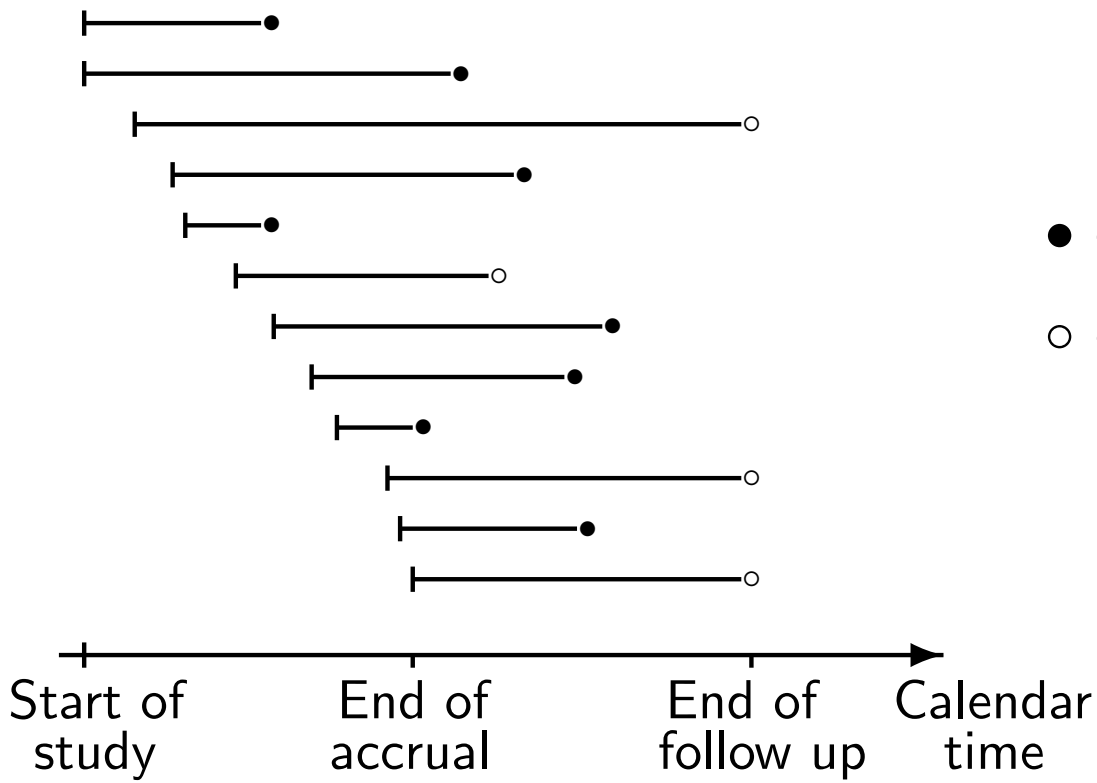
b) log-rank statistics for comparing two survival curves.

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).

Accrual and follow up in a survival study



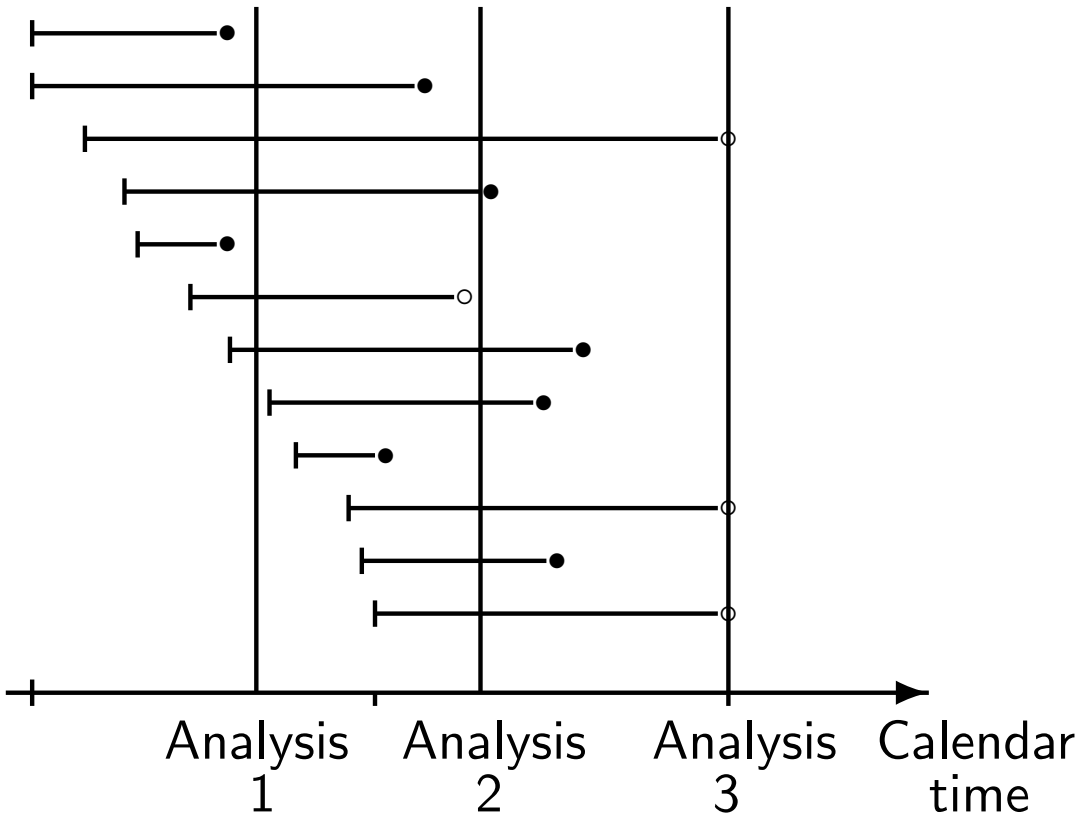
Key

- • death time observed
- ○ censored observation

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

Survival is measured from entry to the study.

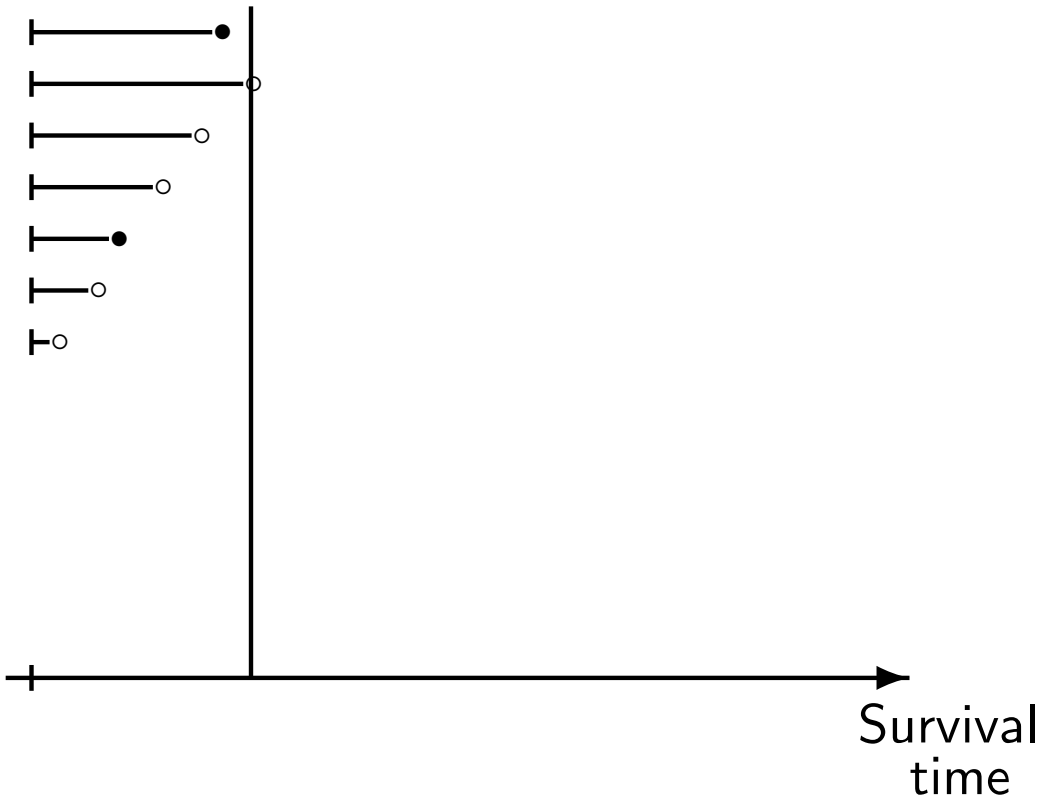
Interim analyses



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

Information on such patients continues to accrue at later analyses.

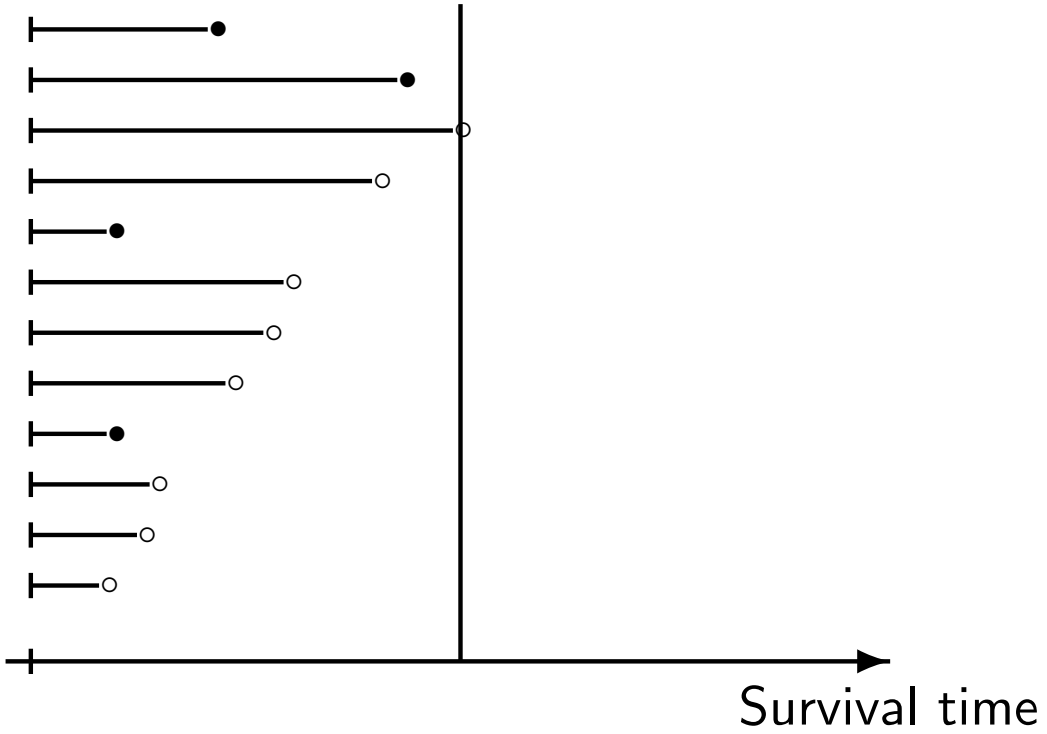
Interim analysis 1



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 $\{\mathcal{I}_1, \ldots, \mathcal{I}_K\}$ 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:

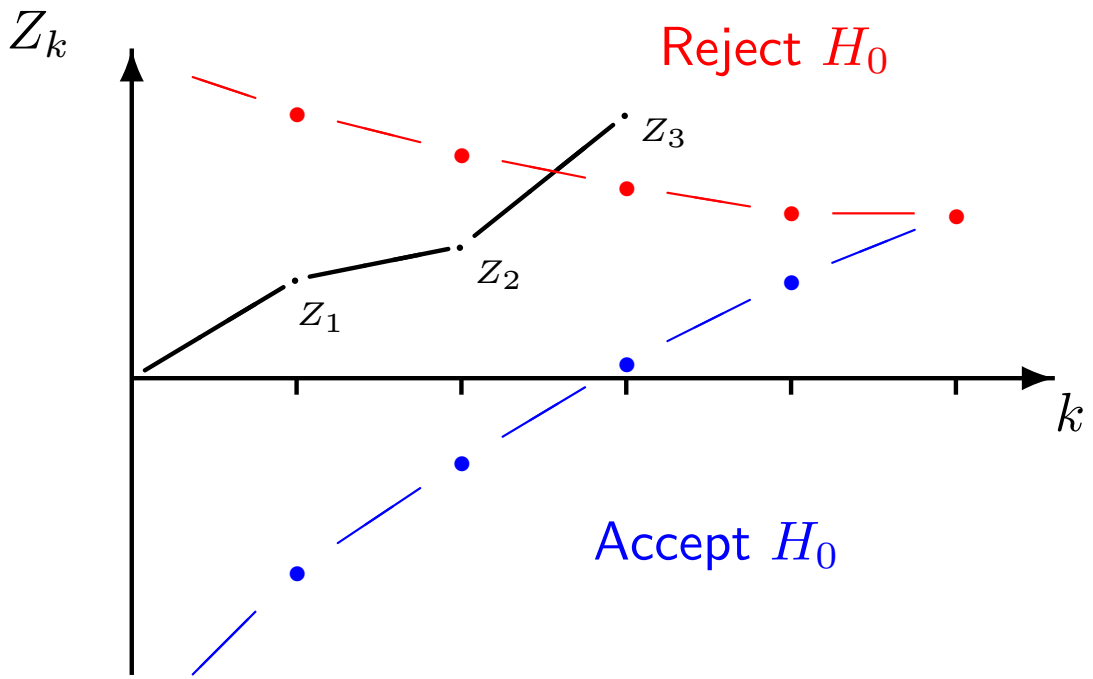
Since $Z_k \sim N(\theta\sqrt{\mathcal{I}_k},\, 1)$, we can estimate θ at analysis k by

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

It follows that

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

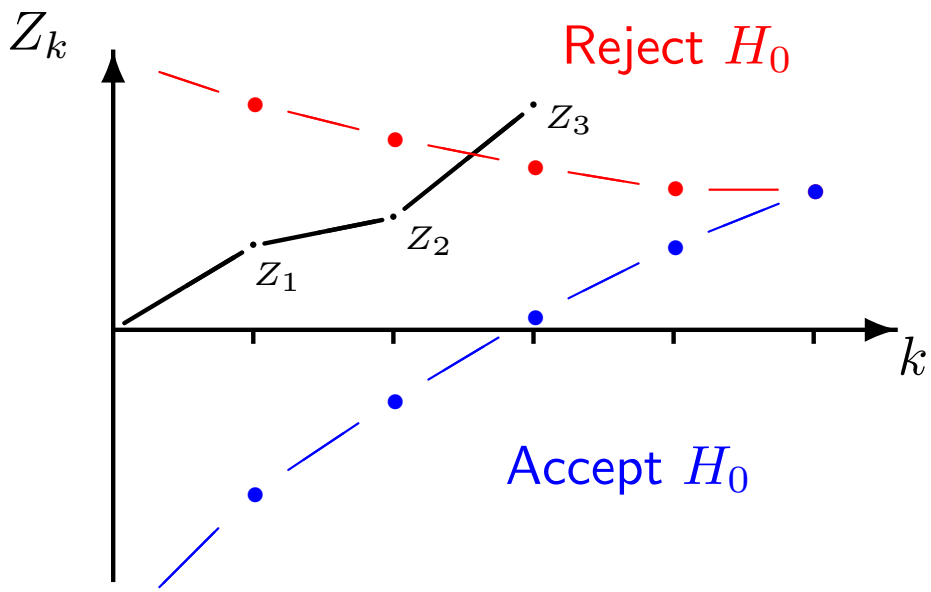
1.4 Computations for group sequential tests



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

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

Computations for group sequential tests



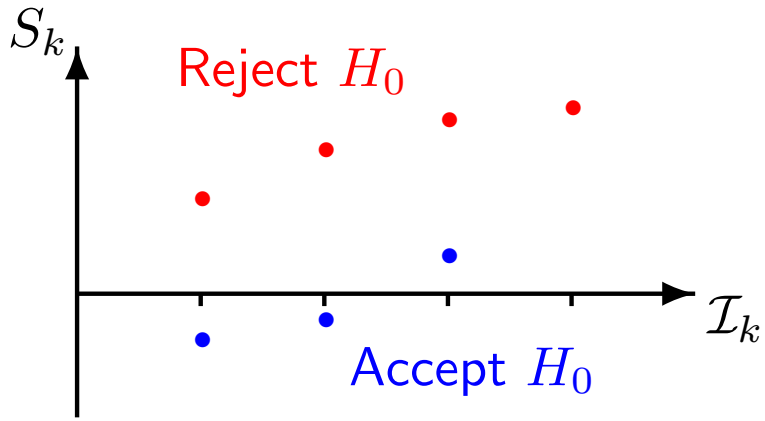
Probabilities such as $P_\theta\{a_1 < Z_1 < b_1,\ \ a_2 < Z_2 < b_2,\ \ Z_3 > b_3\}$ can be computed by repeated numerical integration (JT, Ch. 19).

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

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

One-sided tests: The Pampallona & Tsiatis family

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



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

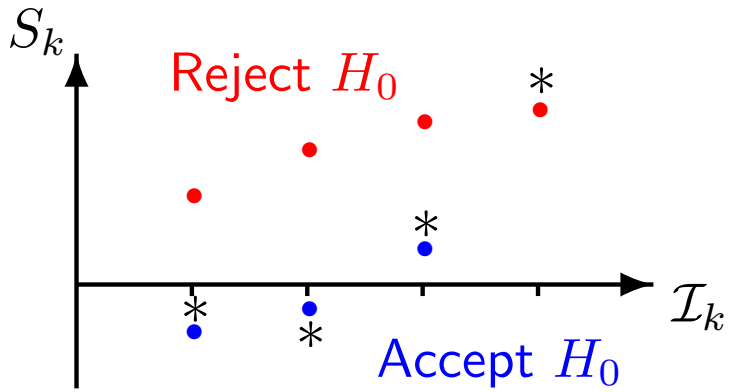
$$a_k = \mathcal{I}_k \, \delta - C_2 \, \mathcal{I}_k^{\Delta}, \quad b_k = C_1 \, \mathcal{I}_k^{\Delta}.$$

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

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

One-sided tests with a non-binding futility boundary

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



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

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

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

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

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

Optimal group sequential tests

We can seek a group sequential test 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).

Benefits of group sequential testing

One-sided tests 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

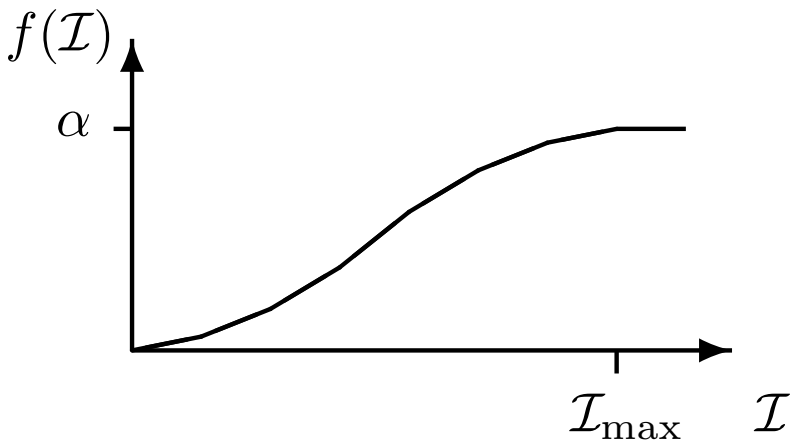
Note: $\mathbb{E}(\mathcal{I}) \searrow$ as $K \nearrow$ but with diminishing returns,
$\mathbb{E}(\mathcal{I}) \searrow$ as $R \nearrow$ up to a point.

1.6 Error spending tests (JT Ch. 7)

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

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

Maximum information design with error spending function $f(\mathcal{I})$



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

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

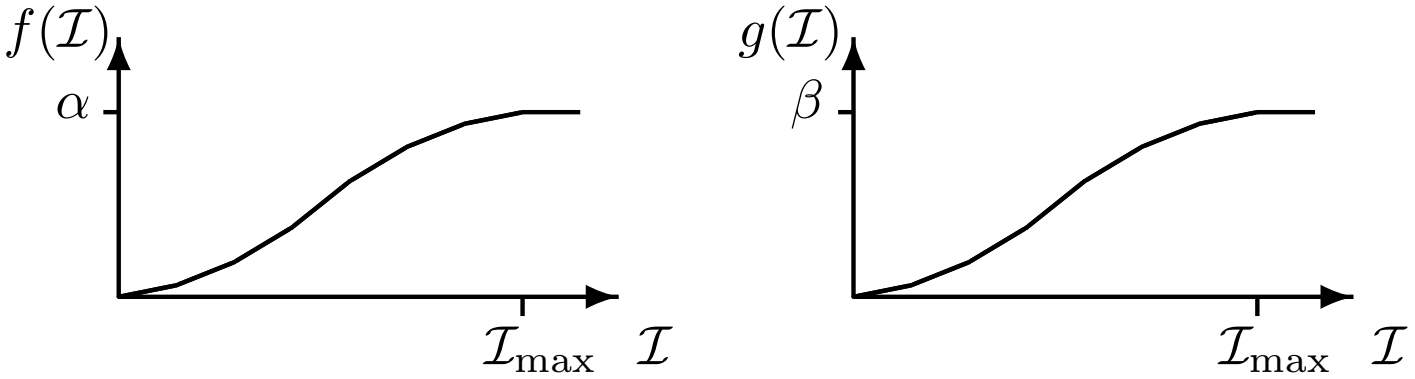
One-sided error spending tests

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 $f(\mathcal{I})$, and type II error probability β according to $g(\mathcal{I})$.

One-sided error-spending tests

Analysis 1:

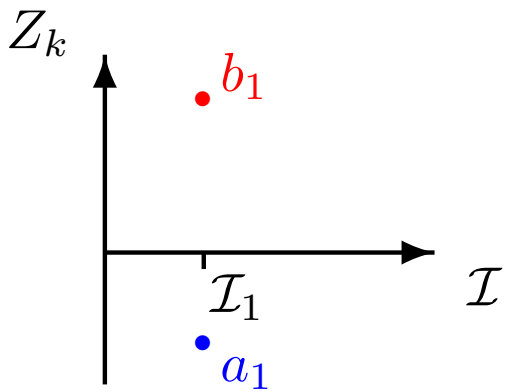
Observed information $\mathcal{I}_1$.

Reject H_0 if $Z_1 > b_1$, where

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

Accept H_0 if $Z_1 < a_1$, where

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



One-sided error-spending tests

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

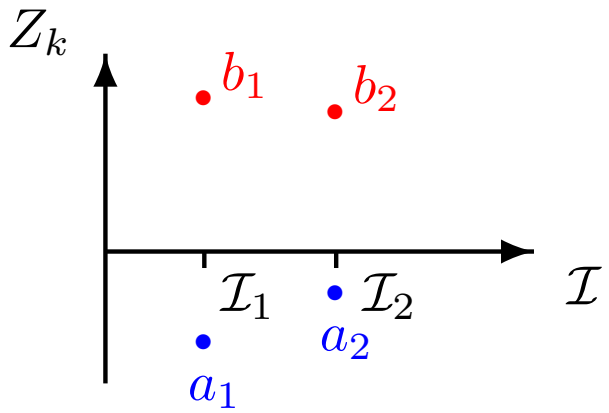
Reject H_0 if $Z_2 > b_2$, where

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

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

Accept H_0 if $Z_2 < a_2$, where

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



One-sided error-spending tests

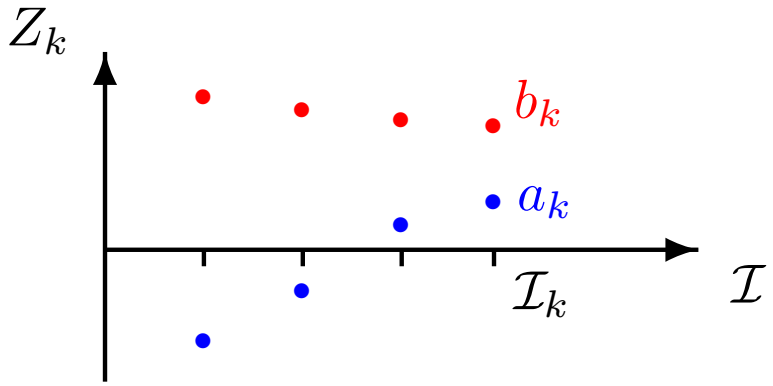
Analysis k: Observed information $\mathcal{I}_k$

Find a_k and b_k to satisfy

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

and

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

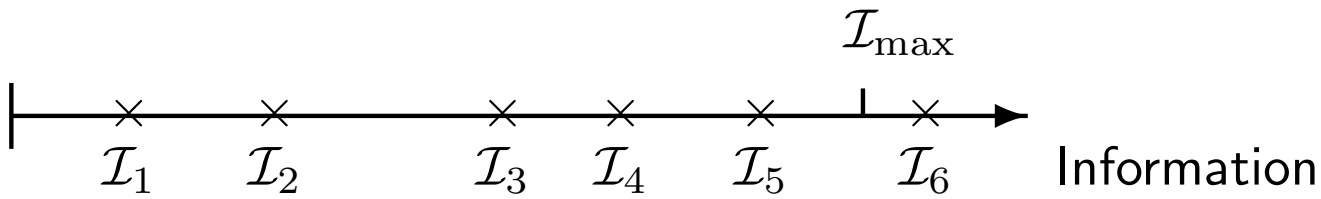


Remarks on error spending tests

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 $f(\mathcal{I}_K)$ 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

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

and type II error probability as

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

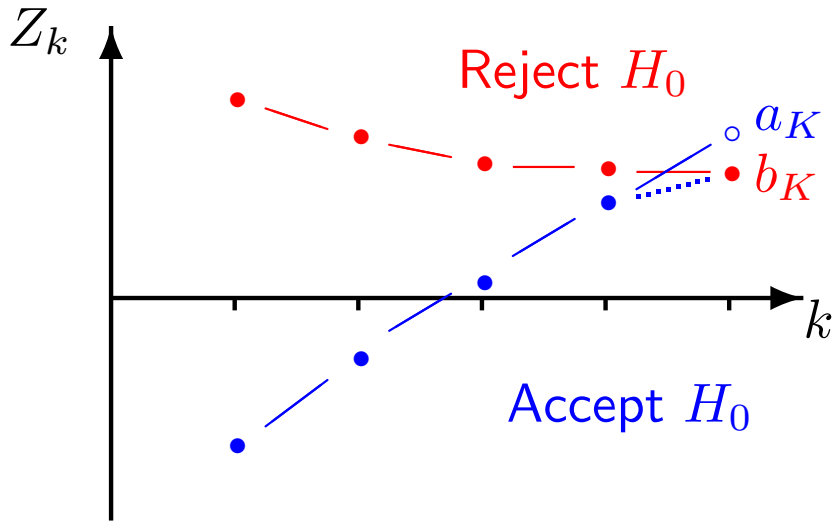
The value of ρ determines the inflation factor $R = \mathcal{I}_{\max}/\mathcal{I}_{fix}$.

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 $a_K > b_K$.



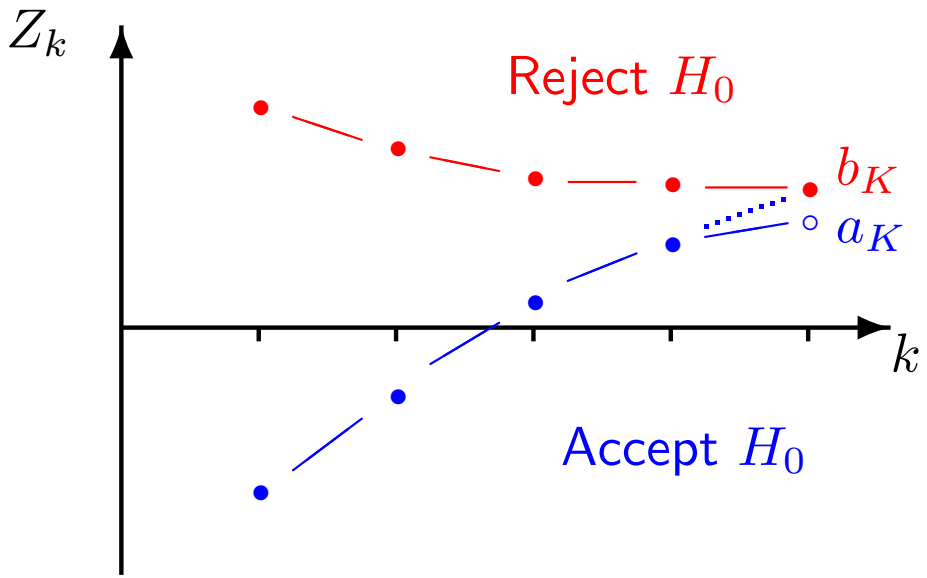
The value calculated for b_K guarantees type I error probability equal to α. So, reduce a_K to b_K — and gain extra power.

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

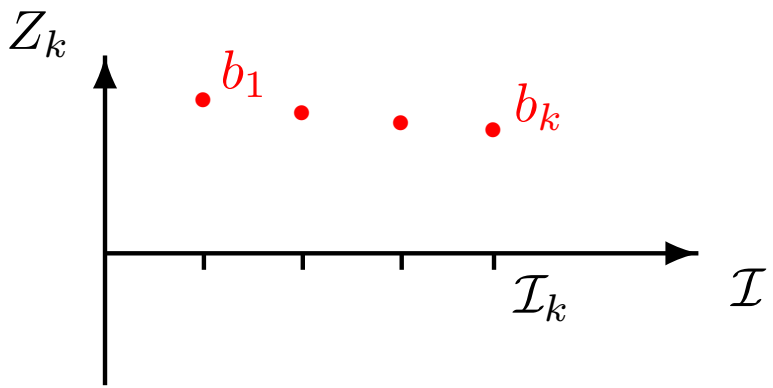
If the futility boundary is treated as non-binding, computation of the efficacy boundary only involves the type I error spending function $f(\mathcal{I})$.

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

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



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

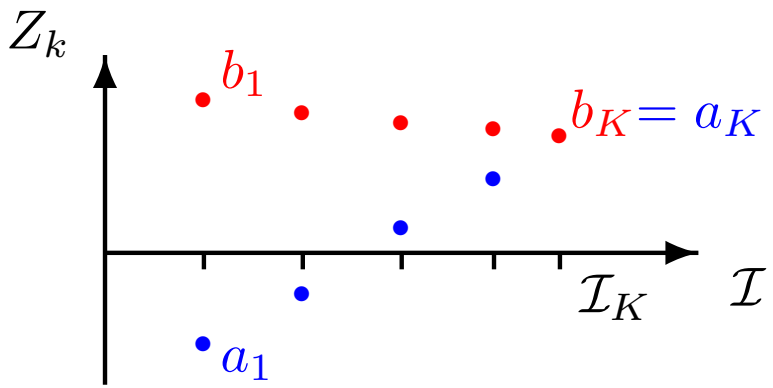
A futility boundary can be added through a type II error spending function $g(\mathcal{I})$.

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

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

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

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



1.7 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

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

and type II error probability as

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

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 — see JT Ch. 14.3.2 for the case of unknown σ^2.)

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.

The ρ-family error spending test with $\rho = 2$, 5 equally spaced analyses, and a **non-binding** futility boundary needs $\mathcal{I}_{\max} = 74.39$ $(n_A = n_B = 95)$ to satisfy type I error and power requirements.

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

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

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

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.

The attained type I error rate is 0.023 and the design gives power 0.898 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.

1.8 An error spending test with binary data

Treatment for heart failure

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

The primary endpoint

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

The current treatment

has a re-admission rate of 25%.

Testing for superiority

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

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

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

Binary example: The testing problem

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

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

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

Let

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

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

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

For large n_t and n_c we have, approximately,

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

Binary example: A fixed sample test

A fixed sample test requires information

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

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

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

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

$$n = 1461$$

subjects per treatment arm.

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

Binary example: A group sequential design

Suppose investigators choose:

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

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

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

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

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

Binary example: Applying the error spending test

At analysis k:

Using current estimates $\hat{p}_c$ and $\hat{p}_t$, calculate observed information

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

and Z-statistic

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

Compute boundary values a_k and b_k using error spending functions

$$f(\mathcal{I}) = 0.025\,\min\{1,\,(\mathcal{I}/\mathcal{I}_{\max})^3\},\;\; g(\mathcal{I}) = 0.1\,\min\{1,\,(\mathcal{I}/\mathcal{I}_{\max})^3\}.$$

Apply the stopping rule

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

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

Binary example: Information monitoring

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

These rates can be re-estimated from observed data.

Information is related to sample size per treatment by

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

At an interim analysis, estimate γ as

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

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

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

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

Binary example: Illustrative data

Analysis 1

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

Analysis 2

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

Binary example: Illustrative data

Analysis 3

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

Analysis 4

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

— Stop, reject H_0 —

Binary example: Illustrative data

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

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

The upper boundary is crossed at analysis 4 out of 5.

The null hypothesis H_0: $\theta \leq 0$ is rejected at analysis 4.

1.9 An error spending test with survival data

Example: Oropharynx Clinical Trial Data

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

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

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

See also JT, Ch. 13.

Canonical joint distribution of logrank-statistics

Recall that in the **Proportional Hazards Model:**

We assume that 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.

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

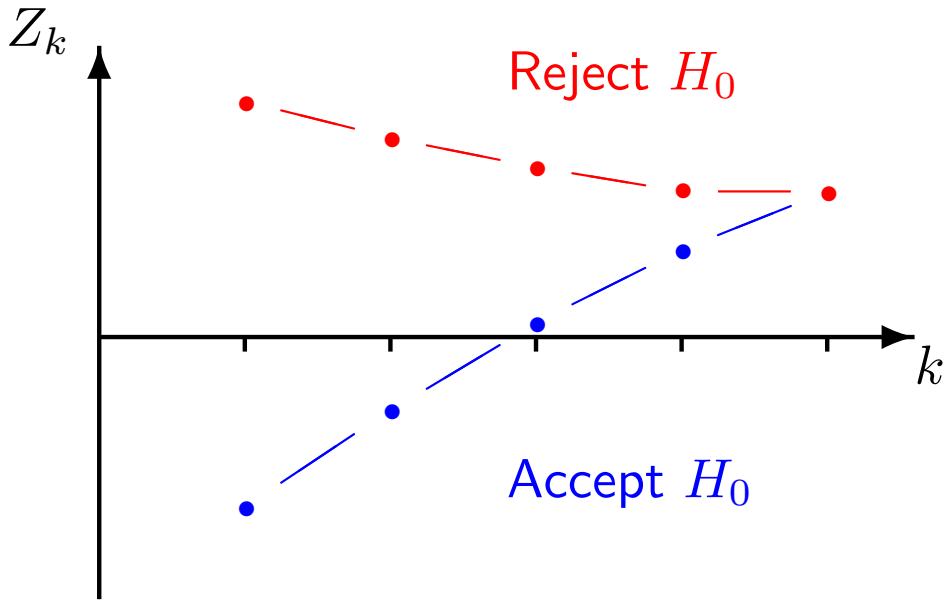
$$\mathcal{I}_f = \frac{\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2}{0.5^2} = 31.40.$$

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

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

Design of the Oropharynx trial

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



However, increments in information between analyses are unpredictable.

So, an error spending design is a natural choice.

A one-sided, error spending design

Specification:

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

Type I error probability $\alpha = 0.025$,

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

Binding futility boundary.

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

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

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

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

A one-sided, error spending design

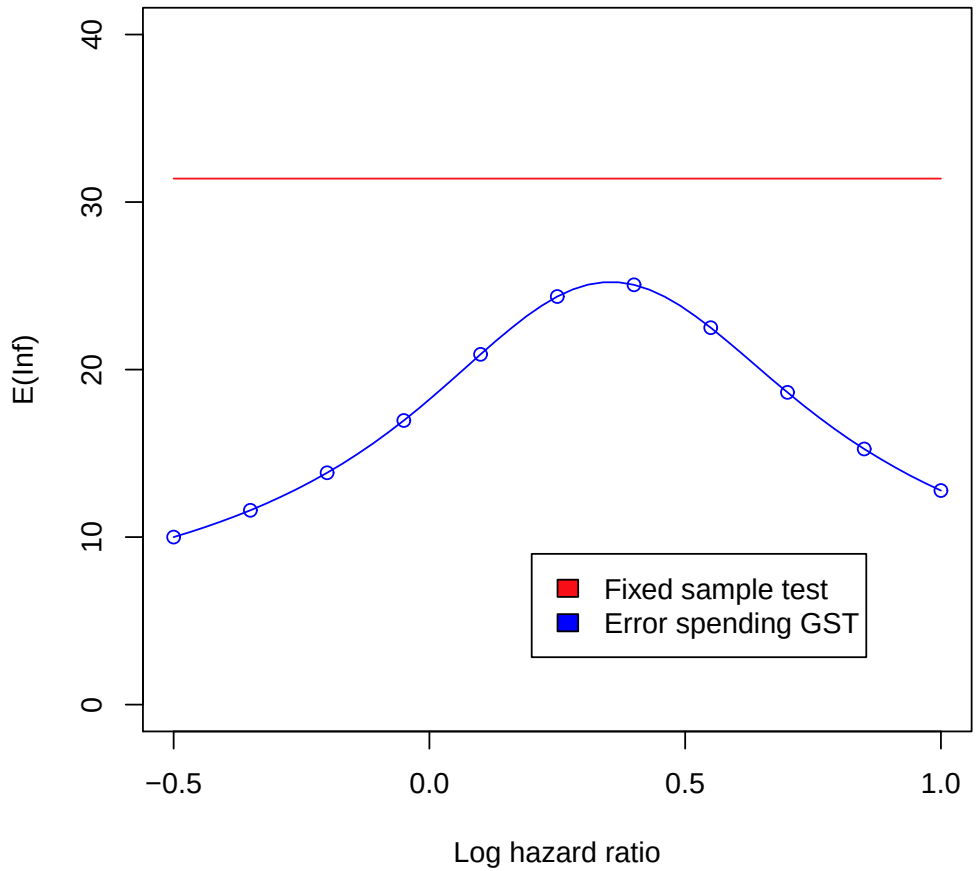
Suppose that, as assumed when planning the trial, information levels are equally spaced up to $\mathcal{I}_5 = \mathcal{I}_{\max} = 34.48$.

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

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

A one-sided, error spending design

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



Summary data and critical values for the Oropharynx trial

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

This gives the following boundary values $(a_1, b_1), \ldots, (a_5, b_5)$ for the standardised logrank statistics $Z_1, \ldots, Z_5$.

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

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

An error spending test with a non-binding futility boundary

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

Applying this design to the observed information levels gives:

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

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

Covariate adjustment in the Oropharynx trial

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

A proportional hazards regression model includes

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

Modelling the hazard rate for patient i as

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

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

Covariate adjustment in the Oropharynx trial

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

We are interested in the treatment effect β_1.

At stage k we have

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

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

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

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

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

Covariate-adjusted analysis of the Oropharynx trial

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

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

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

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

Recapitulation: Group sequential tests

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

A case study:

A survival trial with treatment selection

Protecting the type I error rate

- Multiple hypothesis testing: Closed Testing Procedures
- Combination Tests
- Avoiding error rate inflation in an adaptive survival trial

Choosing an adaptive design and assessing its benefits

2.1 A survival trial with treatment selection

Consider a Phase 3 trial of cancer treatments comparing

Experimental Treatment 1: Intensive dosing

Experimental Treatment 2: Slower dosing

Control treatment

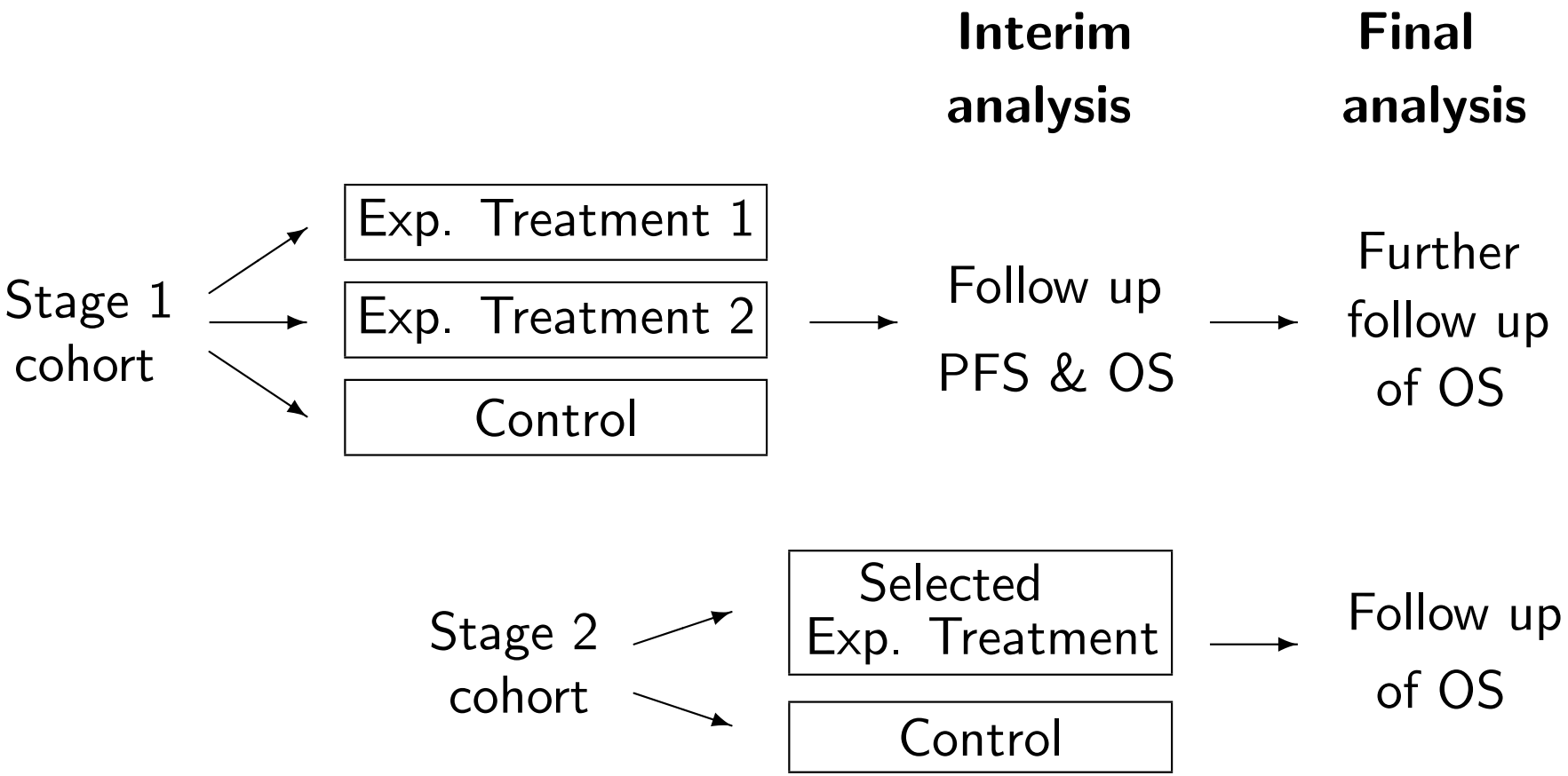
The primary endpoint is Overall Survival (OS).

At an interim analysis, information on OS, Progression Free Survival (PFS), PK measurements and safety will be used to choose between the two experimental treatments.

Note that PFS is useful here as it is more rapidly observed.

After the interim analysis, patients will only be recruited to the selected treatment and the control.

Overall plan of the trial



At the final analysis, we test the null hypothesis that OS on the selected treatment is no better than OS on the control treatment.

Protecting the type I error rate

We shall assume a proportional hazards model for OS with

$$\lambda_1 = \text{Hazard ratio, Control } \textit{vs} \text{ Exp Treatment 1}$$

$$\lambda_2 = \text{Hazard ratio, Control } \textit{vs} \text{ Exp Treatment 2}$$

$$\theta_1 = \log(\lambda_1), \quad \theta_2 = \log(\lambda_2).$$

We test null hypotheses

$H_{0,1}$: $\theta_1 \leq 0$ vs $\theta_1 > 0$ *(Exp Treatment 1 superior to control)*,

$H_{0,2}$: $\theta_2 \leq 0$ vs $\theta_2 > 0$ *(Exp Treatment 2 superior to control)*.

In order to control the "familywise error rate", we require

$$P_{(\theta_1,\theta_2)}\{\text{Reject any true null hypothesis}\} \leq \alpha$$

for all (θ_1, θ_2).

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 $\theta = (\theta_1, \ldots, \theta_h)$ is

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

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

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

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

Marcus et al. (*Biometrika*, 1976) introduced a **Closed Testing Procedure** which provides strong control of FWER by combining level α tests of each H_i and of intersections of these hypotheses.

Suppose we have null hypotheses H_i, $i = 1, \ldots, h$.

For each subset I of $\{1, \ldots, h\}$, define the intersection hypothesis

$$H_I = \cap_{i \in I} H_i.$$

Construct a level α test of each intersection hypothesis H_I, i.e., a test which rejects H_I with probability at most α whenever all hypotheses specified in H_I are true.

Closed Testing Procedure

The simple hypothesis H_j: $\theta_j \leq 0$ is rejected overall if, and only if, H_I is rejected for every set I containing index j.

Closed Testing Procedures

Proof of strong control of familywise error rate

In the Closed Testing Procedure, overall rejection of the simple hypothesis H_j can only occur if H_I is rejected for every set I containing index j.

Let $\tilde{I}$ be the set of indices of all true hypotheses H_i.

Since $H_{\tilde{I}}$ is true, $P\{\text{Reject } H_{\tilde{I}}\} = \alpha$.

For a familywise error to be committed, $H_{\tilde{I}}$ must be rejected.

Hence, the probability of a familywise error is no greater than α.

Testing an intersection hypothesis

Suppose the intersection hypothesis $H_I = \cap_{i \in I} H_i$ is the intersection of m simple hypotheses.

For each $i \in I$, let P_i be the 1-sided P-value for testing H_i.

Denote the ordered values of the P_i by $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$.

There are several ways to test an intersection hypothesis.

Bonferroni adjustment

The overall P-value for testing H_I is $P_I = m\, P_{(1)}$.

Simes' method (Biometrika, 1986):

The overall P-value for H_I is

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

Bonferroni and Simes' methods

The Bonferroni adjustment is simple, but conservative.

In the definition of Simes' P-value,

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

the term for $k = 1$ is $mP_{(1)}$, i.e., the Bonferroni adjusted P-value.

Other low P-values can reduce the overall result, e.g., if $P_{(2)}$ is only a little higher than $P_{(1)}$ so $P_{(2)}/2 < P_{(1)}$, then this will reduce P_I.

The Simes method is valid — and still slightly conservative — when the P_i are independent or positively dependent.

Such positive dependence arises in a comparison of m treatments with a common control.

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

A Closed Testing Procedure for our 3-arm survival trial

Define level α tests of

$$H_{0,1}: \ \theta_1 \leq 0,$$

$$H_{0,2}: \ \theta_2 \leq 0$$

and a level α test of the intersection hypothesis

$$H_{0,12} \ = \ H_{0,1} \cap H_{0,2}: \ \ \theta_1 \leq 0 \ \text{ and } \ \theta_2 \leq 0.$$

Then:

Reject $H_{0,1}$ **overall** *if the above tests reject* $H_{0,1}$ *and* $H_{0,12}$,

Reject $H_{0,2}$ **overall** *if the above tests reject* $H_{0,2}$ *and* $H_{0,12}$.

The requirement to reject $H_{0,12}$ compensates for testing multiple hypotheses and the "selection bias" in choosing the treatment to focus on in Stage 2.

2.3 Combining data across stages

Consider testing a generic null hypothesis H_0: $\theta \leq 0$ against $\theta > 0$.

Suppose Stage 1 data produce Z_1 where

$$Z_1 \sim N(0,1) \quad \text{if } \theta = 0.$$

On adaptation, Stage 2 data yield Z_2 *conditionally* distributed as

$$Z_2 \sim N(0,1) \quad \text{if } \theta = 0,$$

while Z_2 is stochastically smaller than $N(0,1)$ if $\theta < 0$.

Weighted inverse normal Combination Test

With pre-specified weights w_1 and w_2 satisfying $w_1^2 + w_2^2 = 1$,

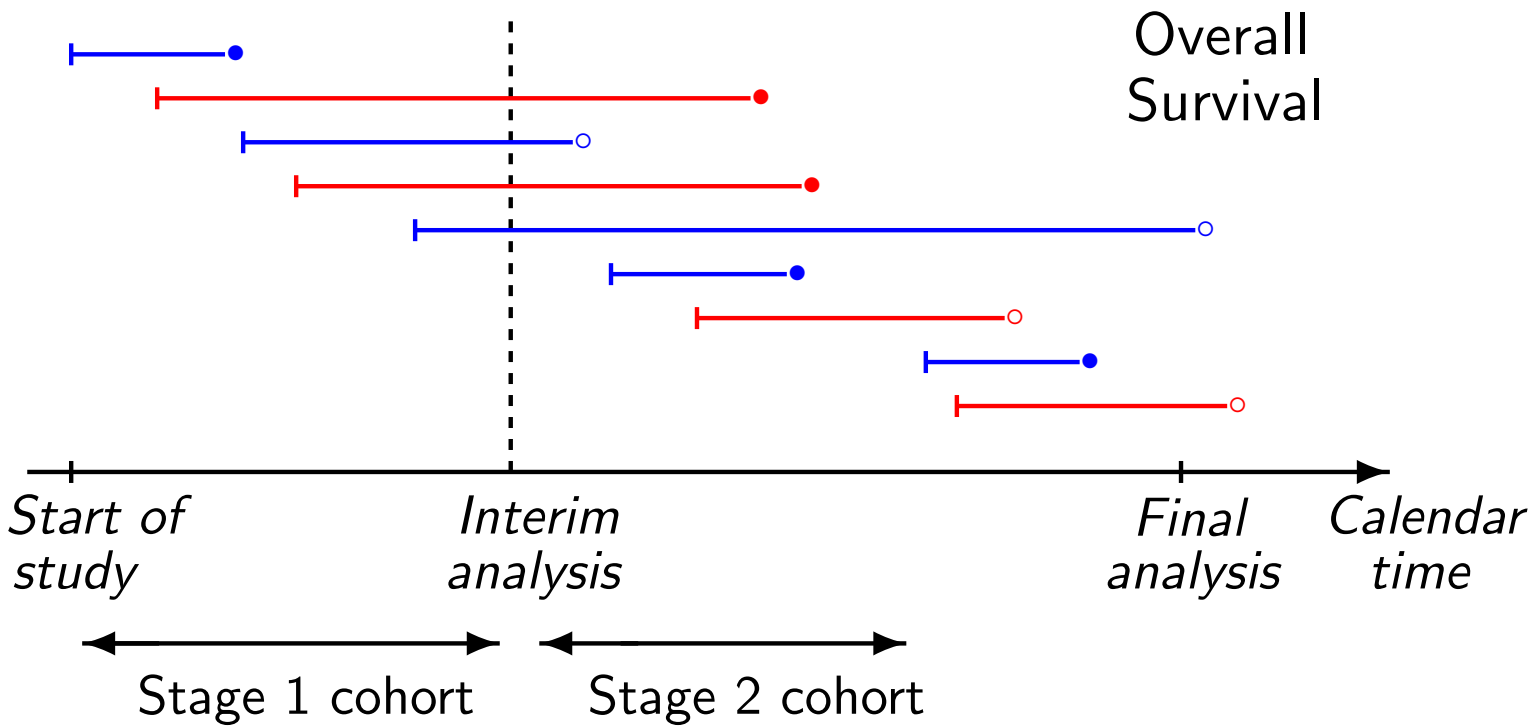
$$Z \;=\; w_1 Z_1 + w_2 Z_2 \;\sim\; N(0,1) \quad \text{if } \theta = 0,$$

and Z is stochastically smaller than $N(0,1)$ if $\theta < 0$.

So, for a level α test, we reject H_0 if $Z > \Phi^{-1}(1-\alpha)$.

Applying a Combination Test to survival data

For now, consider Experimental Treatment 1 vs Control.



Key:
- Subjects randomised to Exp Treatment 1
- Subjects randomised to Control
- ● Death observed
- ○ Censored observation

Properties of log-rank tests

Comparing Experimental Treatment 1 vs Control, define

S_1 = Unstandardised log-rank statistic at interim analysis,

$\mathcal{I}_1$ = Information for θ_1 at interim analysis $\approx$ (Number of deaths)/4

S_2 = Unstandardised log-rank statistic at final analysis,

$\mathcal{I}_2$ = Information for θ_1 at final analysis $\approx$ (Number of deaths)/4

Here, "Number of deaths" refers to the total number of deaths on Experimental Treatment 1 and Control arms only.

Then, approximately,

$$S_1 \sim N(\mathcal{I}_1 \theta_1, \mathcal{I}_1),$$

$$S_2 - S_1 \sim N(\{\mathcal{I}_2 - \mathcal{I}_1\} \theta_1, \{\mathcal{I}_2 - \mathcal{I}_1\})$$

and S_1 and $(S_2 - S_1)$ are **independent** (independent increments).

Reference: Tsiatis (*Biometrika*, 1981).

A Combination Test for survival data

We create Z statistics

Based on data at the interim analysis:

$$Z_1 = \frac{S_1}{\sqrt{\mathcal{I}_1}},$$

Based on data accrued **between** the interim and final analyses:

$$Z_2 = \frac{S_2 - S_1}{\sqrt{\mathcal{I}_2 - \mathcal{I}_1}}.$$

If $\theta_1 = 0$, then $Z_1 \sim N(0,1)$ and $Z_2 \sim N(0,1)$ are independent.

If $\theta_1 < 0$, Z_1 and Z_2 are stochastically smaller than this.

So, we can use $Z = w_1 Z_1 + w_2 Z_2$ in an inverse normal Combination Test of $H_{0,1}$: $\theta_1 \leq 0$.

A Combination Test for survival data: Caution!

The above distribution theory for logrank statistics of a single comparison requires

$$Z_2 = \frac{S_2 - S_1}{\sqrt{\mathcal{I}_2 - \mathcal{I}_1}} \sim N(0,1) \quad \text{under } \theta_1 = 0,$$

regardless of decisions taken at the interim analysis.

Bauer & Posch (*Statistics in Medicine*, 2004) note this implies that the conduct of the second part of the trial should not depend on the prognosis of Stage 1 patients at the interim analysis.

Suppose prognoses are better for patients on Exp Treatment 1 than for those on Control, and the Stage 2 cohort size is reduced while follow up of Stage 1 patients is extended: then, the distribution of Z_2 could be biased upwards.

Our example has another potential source of bias, depending on how the Stage 2 statistic for testing $H_{0,12}$ is defined.

2.4 Analysing an adaptive survival trial

In applying a Closed Testing Procedure, we require level α tests of

$H_{0,1}$: $\theta_1 \leq 0$,

$H_{0,2}$: $\theta_2 \leq 0$,

$H_{0,12}$: $\theta_1 \leq 0$ and $\theta_2 \leq 0$.

Combination Tests for these hypotheses are formed from:

	Stage 1 data	*Stage 2 data*
$H_{0,1}$	$Z_{1,1}$	$Z_{2,1}$
$H_{0,2}$	$Z_{1,2}$	$Z_{2,2}$
$H_{0,12}$	$Z_{1,12}$	$Z_{2,12}$

The question is how we should define $Z_{1,1}$, $Z_{2,1}$, etc?

Analysing an adaptive survival trial

A natural choice is to:

Base $Z_{1,1}$, $Z_{1,2}$ and $Z_{1,12}$ on data at the interim analysis,

Base $Z_{2,1}$, $Z_{2,2}$ and $Z_{2,12}$ on the additional information accruing between interim and final analyses.

We could take $Z_{1,1}$ and $Z_{1,2}$ to be standardised log-rank statistics, and $Z_{2,1}$ and $Z_{2,2}$ standardised increments between analyses.

For the intersection hypothesis: $Z_{1,12}$ is formed from $Z_{1,1}$ and $Z_{1,2}$, **while $\boldsymbol{Z_{2,12} = Z_{2,j}}$, where $\boldsymbol{j}$ is the selected treatment.**

However, treatment j is selected because it has better PFS outcomes at the interim analyses, so it is likely that future OS for these patients will also be better.

This approach would lead to a bias in the null distribution of $Z_{2,12}$.

The method of Jenkins, Stone & Jennison (2011)

If we base a Combination Test on the two parts of the data accrued before and after the interim analysis, bias can result:

	Z_1	Z_2
Stage 1 cohort	Overall survival (during Stage 1)	**Overall survival (during Stage 2)**
Stage 2 cohort		Overall survival (during Stage 2)

Instead, we divide the data into the parts from the two cohorts:

Stage 1 cohort	Overall survival (during Stage 1)	Overall survival (during Stage 2)	Z_1
Stage 2 cohort		Overall survival (during Stage 2)	Z_2

Partitioning data for a Combination Test

To avoid bias: All patients in the Stage 1 cohort are followed for overall survival up to a fixed time, shortly before the final analysis.

"Stage 1" statistics are based on Stage 1 cohort's **final** OS data

$Z_{1,1}$ from log-rank test of Exp Tr 1 vs Control

$Z_{1,2}$ from log-rank test of Exp Tr 2 vs Control

$Z_{1,12}$ from pooled log-rank test, or a Simes or Dunnett test.

"Stage 2" statistics are based on OS data for the Stage 2 cohort

If Exp Treatment 1 is selected:

$Z_{2,1}$ from log-rank test of Exp Tr 1 vs Control, $Z_{2,12} = Z_{2,1}$

If Exp Treatment 2 is selected:

$Z_{2,2}$ from log-rank test of Exp Tr 2 vs Control, $Z_{2,12} = Z_{2,2}$.

Partitioning data for a Combination Test

Discussion

Jenkins, Stone & Jennison (2011) introduced the proposed method in a design where a choice is made between testing for an effect in the full population or a sub-population.

They stipulated that the amount of follow up for the Stage 1 cohort should be fixed at the outset to avoid any risk of inflating the type I error rate.

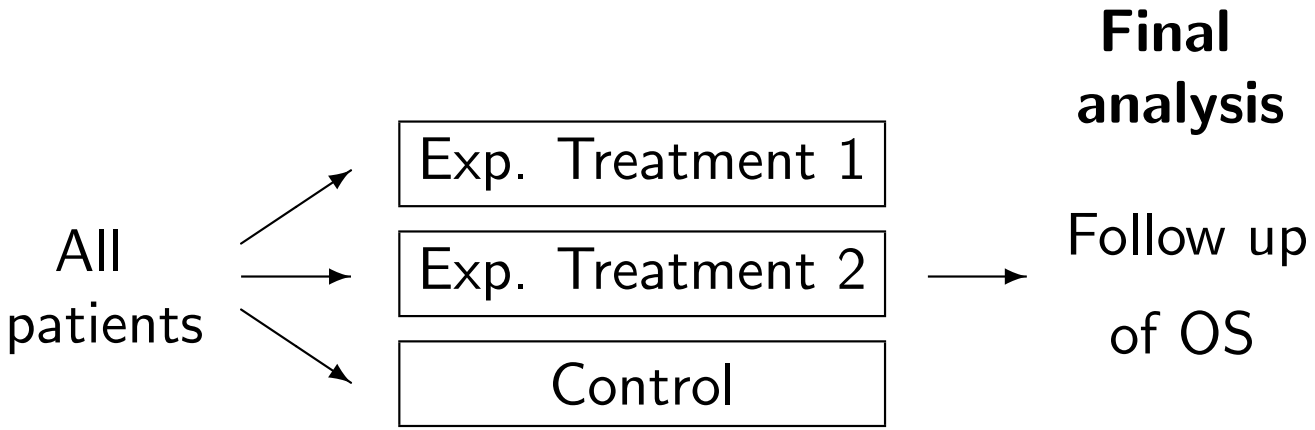
Some adaptive designs allow an early decision based on summaries of "Stage 1" data at an interim analysis.

In our three-treatment design, the statistics $Z_{1,1}$, $Z_{1,2}$ and $Z_{1,12}$ are not known at the time of the interim analysis, so we cannot define a formal stopping rule.

However, with only a little OS data available at the interim analysis, this is not a serious limitation.

2.5 Choosing an adaptive design and assessing its benefits

We compare the adaptive design with a non-adaptive trial in which randomisation is to both experimental treatments and control *throughout* the trial:



A Closed Testing Procedure is used to control familywise error rate.

When the total numbers of patients and lengths of follow-up are the same in adaptive and non-adaptive designs,

Does the adaptive design provide higher power?

Are there other advantages?

Assessing the adaptive design: Model assumptions

Overall Survival

	Log hazard ratio
Exp Treatment 1 vs control	θ_1
Exp Treatment 2 vs control	θ_2

Logrank statistics are correlated due to the common control arm.

Progression Free Survival

	Log hazard ratio
Exp Treatment 1 vs control	ψ_1
Exp Treatment 2 vs control	ψ_2

Denote correlation between logrank statistics for OS and PFS by ρ.

In fact, hazard rates cannot be proportional for both endpoints.

However, it is the implications for the joint distribution of logrank statistics that matter, and it is convenient to describe these as if from two proportional hazards models.

Assessing the adaptive design: Model assumptions

Log hazard ratios for OS: $\theta_1,\ \theta_2$.

Log hazard ratios for PFS: $\psi_1,\ \psi_2$.

We suppose logrank statistics are distributed as if

$\psi_1 = \gamma \times \theta_1$ and $\psi_2 = \gamma \times \theta_2$

Final number of OS events for Stage 1 cohort $= 300$ (over 3 treatment arms)

Number of OS events for Stage 2 cohort $= 300$ (over 2 or 3 treatment arms)

Number of PFS events at interim analysis $= \lambda \times 300$.

When the log hazard ratio is θ, the standardised logrank statistic based on d observed events is, approximately, $N(\theta\sqrt{d/4}, 1)$.

Testing the intersection hypothesis $H_{0,12}$

We have null hypotheses $H_{0,1}$: $\theta_1 \leq 0$ and $H_{0,2}$: $\theta_2 \leq 0$.

In the Closed Testing Procedure, we must also test

$$H_{0,12} = H_{0,1} \cap H_{0,2} : \ \theta_1 \leq 0 \ \text{ and } \ \theta_2 \leq 0.$$

We could test $H_{0,12}$ by pooling the Exp Trt 1 and Exp Trt 2 patients and carrying out a logrank test vs the Control group.

Alternatively we could use a **Simes** test or a **Dunnett** test.

Simes' test:

Given observed values p_1 and p_2 of P_1 and P_2, Simes' test of $H_{0,12}$ yields the P-value

$$\min\left(2\min(p_1, p_2),\ \max(p_1, p_2) \right).$$

Simes' test protects type I error conservatively when P_1 and P_2 are independent or positively associated.

Dunnett's test of an intersection hypothesis

Dunnett's test for comparisons with a common control

Suppose Z_1 and Z_2 are the Z-values for logrank tests of Exp Trt 1 vs control and Exp Trt 2 vs Control.

If z_1 and z_2 are the observed values of Z_1 and Z_2, the Dunnett test of $H_{0,12}$ yields the P-value

$$P(\max(Z_1, Z_2) \geq \max(z_1, z_2))$$

where (Z_1, Z_2) is bivariate normal with $Z_1 \sim N(0,1)$, $Z_2 \sim N(0,1)$ and Corr$(Z_1, Z_2) = 0.5$.

We shall see from comparisons of different methods that the Dunnett test of the intersection hypothesis leads to the most efficient versions of both adaptive and non-adaptive designs.

Comparing adaptive and non-adaptive trial designs

With selected values of ψ_1, θ_1, ψ_2, θ_2 and ρ, we simulate logrank statistics from their large sample distributions.

For the adaptive design, we define

$$P(1) = P(\text{Select Treatment 1 and Reject } H_{0,1} \text{ overall})$$

$$P(2) = P(\text{Select Treatment 2 and Reject } H_{0,2} \text{ overall})$$

For the non-adaptive design, we set

$$P(1) = P(\widehat{\theta}_1 > \widehat{\theta}_2 \text{ and } H_{0,1} \text{ is rejected overall})$$

$$P(2) = P(\widehat{\theta}_2 > \widehat{\theta}_1 \text{ and } H_{0,2} \text{ is rejected overall})$$

Hence, we define the overall expected "Gain" or utility measure

$$E(\text{Gain}) = \theta_1 \times P(1) + \theta_2 \times P(2).$$

Comparing tests of the intersection hypothesis

Intersection tests produce $Z_{1,12}$ in an adaptive trial design with

$$\psi_1 = \theta_1, \quad \psi_2 = \theta_2, \quad \lambda = 1, \quad \rho = 0.6, \quad \alpha = 0.025.$$

		$P(1)$			$E(\text{Gain})$		
θ_1	θ_2	Pooled	Simes	Dunnett	Pooled	Simes	Dunnett
0.3	0.0	0.77	0.85	0.86	0.232	0.254	0.259
0.3	0.1	0.78	0.81	0.82	0.238	0.245	0.247
0.3	0.2	0.68	0.68	0.69	0.238	0.237	0.238
0.3	0.25	0.58	0.58	0.58	0.250	0.249	0.249
0.3	0.295	0.48	0.47	0.47	0.275	0.274	0.274

All simulation results are based on 1,000,000 replicates.

The Dunnett test has the highest power. Unlike the pooled test, it is well aligned (consonant) with individual tests of $H_{0,1}$ and $H_{0,2}$.

Comparing adaptive and non-adaptive trial designs

We compare designs using a Dunnett test for $H_{0,12}$ with

$$\psi_1 = \theta_1, \quad \psi_2 = \theta_2, \quad \lambda = 1, \quad \rho = 0.6, \quad \alpha = 0.025.$$

		Non-adaptive			Adaptive		
θ_1	θ_2	$P(1)$	$P(2)$	$E(\text{Gain})$	$P(1)$	$P(2)$	$E(\text{Gain})$
0.3	0.0	0.78	0.00	0.235	0.86	0.00	0.259
0.3	0.1	0.78	0.01	0.234	0.82	0.02	0.247
0.3	0.2	0.70	0.11	0.234	0.69	0.16	0.238
0.3	0.25	0.60	0.26	0.244	0.58	0.30	0.249
0.3	0.295	0.47	0.43	0.267	0.47	0.44	0.274

Here, $\lambda = 1$ implies there are 300 PFS events at the interim analysis.

The adaptive design has higher $P(1)$ when θ_1 is well above θ_2.

With θ_1 and θ_2 closer, the adaptive design still has higher $E(\text{Gain})$.

Comparing adaptive and non-adaptive trial designs

The adaptive design can only succeed if there is adequate information to select the correct treatment at the interim analysis:

Treatment effects on PFS should be be reliable indicators of treatment effects on OS,

There must be good information on PFS at the interim analysis.

We have investigated varying the parameters γ and λ where

$\psi_1 = \gamma \times \theta_1,\ \psi_2 = \gamma \times \theta_2,$ with $\theta_1 = 0.3$ and $\theta_2 = 0.1$

Final number of OS events for Stage 1 cohort $= 300$ (over 3 arms)

Number of OS events for Stage 2 cohort $= 300$ (over 2 or 3 arms)

Number of PFS events at interim analysis $= \lambda \times 300.$

NB It is quite plausible that γ should be greater than 1, i.e., a larger treatment effect on PFS than on OS.

Comparing adaptive and non-adaptive trial designs

We compare designs with $\theta_1 = 0.3,\ \theta_2 = 0.1,\ \rho = 0.6,\ \alpha = 0.025,$

PFS log hazard ratios: $\psi_1 = \gamma\,\theta_1,\ \psi_2 = \gamma\,\theta_2,$

Number of PFS events at interim analysis $= \lambda \times 300.$

		Non-adaptive			Adaptive		
γ	λ	$P(1)$	$P(2)$	$E(\text{Gain})$	$P(1)$	$P(2)$	$E(\text{Gain})$
1.5	1.2				0.88	0.00	0.264
1.2	1.1				0.85	0.01	0.256
1.0	**1.0**	**0.78**	**0.01**	**0.234**	**0.82**	**0.02**	**0.247**
0.9	0.9	for all γ and λ			0.78	0.03	0.238
0.8	0.8	(PFS is not used)			0.74	0.04	0.225
0.7	0.7				0.68	0.05	0.208

Adaptation works well when there is enough PFS information for treatment selection at the interim analysis.

2.6 Related work

1. Friede et al. (*Statistics in Medicine*, 2011) consider a seamless phase II/III trial design with treatment selection based on both short-term and long-term responses. They take a similar approach to Jenkins, Stone & Jennison (*Pharm. Statistics*, 2011) and apply a Combination Test to the long-term response data from the *cohorts* of patients admitted before and after the interim decision point.

2. Irle & Schäfer (*JASA*, 2012) propose similar adaptive designs for survival data. They determine critical values for test statistics through the "Conditional Probability of Rejection" principle. Since this is related to Combination Tests, the method has much in common with that of Jenkins, Stone & Jennison.

However, determining the conditional probability of rejection is problematic since the final information level (in a log-rank statistic, say) is not known at the time this probability is calculated.

Conclusions about the benefits of the adaptive design

1. The adaptive design offers the chance to select the better treatment and focus on this in the second stage of the trial.
2. Overall, adaptation is beneficial as long as there is sufficient information to make a reliable treatment selection decision.
3. Other evidence may be used in reaching this decision:

   Safety data

   Pharmacokinetic data

   Overall survival

4. In addition to reaching a final decision, both non-adaptive and adaptive trials compare the two forms of treatment: the conclusions from this comparison may be more broadly useful.

Recapitulation: Adaptive clinical trial designs

1. It is desirable to adapt a clinical trial design as information becomes available on parameters that were initially unknown.
2. Methods are available to create adaptive designs that will protect the overall type I error rate.
3. Combination Tests allow results from different stages of the trial to be merged.
4. Closed Testing Procedures allow tests of multiple hypotheses, or of a single hypothesis selected in a data-dependent manner.
5. It should not be assumed that introducing adaptation will automatically make a trial design more efficient.
6. Critical appraisal of trial designs is crucial and, where feasible, it is advisable to define an objective function and optimise for this criterion within a chosen class of designs.