The Design of Group Sequential Clinical Trials that Test Multiple Endpoints

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

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

Bruce Turnbull

Department of Statistical Science,

Cornell University

http://www.orie.cornell.edu/∼bruce

Takeda, London

February 2015

1. Group sequential tests

Suppose two treatments are to be compared in a Phase III trial.

The treatment effect θ represents the advantage of Treatment A over Treatment B, with a positive value meaning that Treatment A is more effective.

In a group sequential trial, data are examined on a number of occasions during the course of the study.

These analyses may be at pre-specified time points — or they may be conducted when certain numbers of observations become available.

Standardised test statistics Z_1, Z_2, . . . , are computed at interim analyses and these are used to define a stopping rule for the trial.

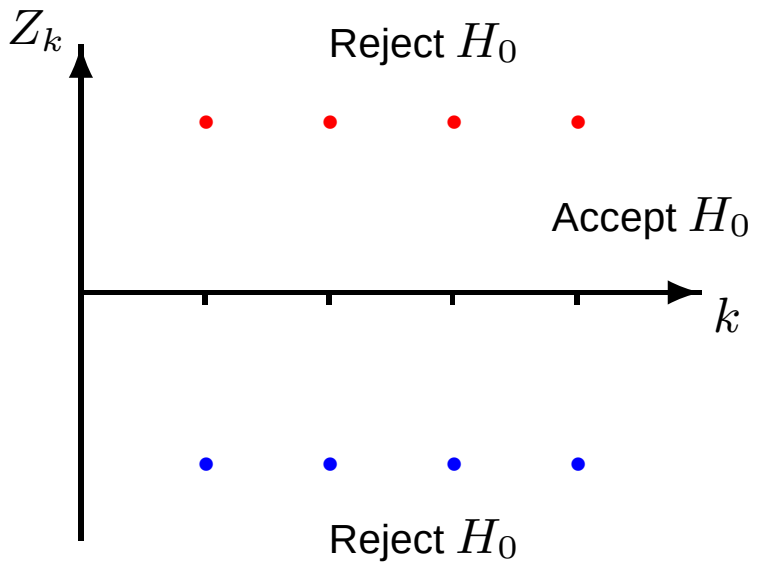
Pocock's repeated significance test (*Biometrika*, 1977)

To test H_0: $\theta = 0$ against $\theta \neq 0$, where θ represents the treatment difference.

Stop to reject H_0 at analysis k if

$$|Z_k| > c.$$

If H_0 is not rejected by analysis K, stop and accept H_0.



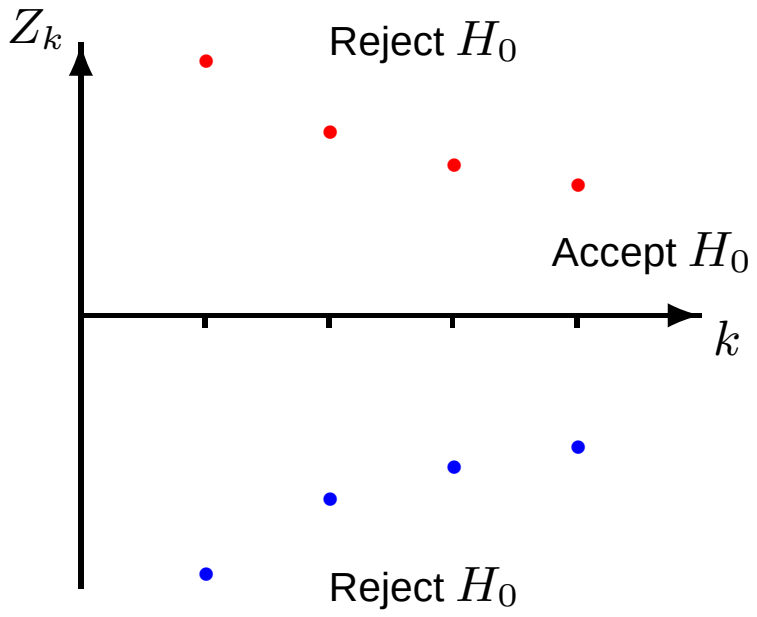
O'Brien & Fleming's test (*Biometrics*, 1979)

To test H_0: $\theta = 0$ against $\theta \neq 0$, where θ represents the treatment difference.

Stop to reject H_0 at analysis k if

$$|Z_k| > c' \sqrt{\frac{K}{k}}.$$

If H_0 is not rejected by analysis K, stop and accept H_0.



A one-sided hypothesis test

Suppose a new treatment is being compared to a placebo or positive control in a Phase III trial.

Now, the treatment effect θ represents the advantage of the new treatment over the control, with a positive value meaning that the new treatment is effective.

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

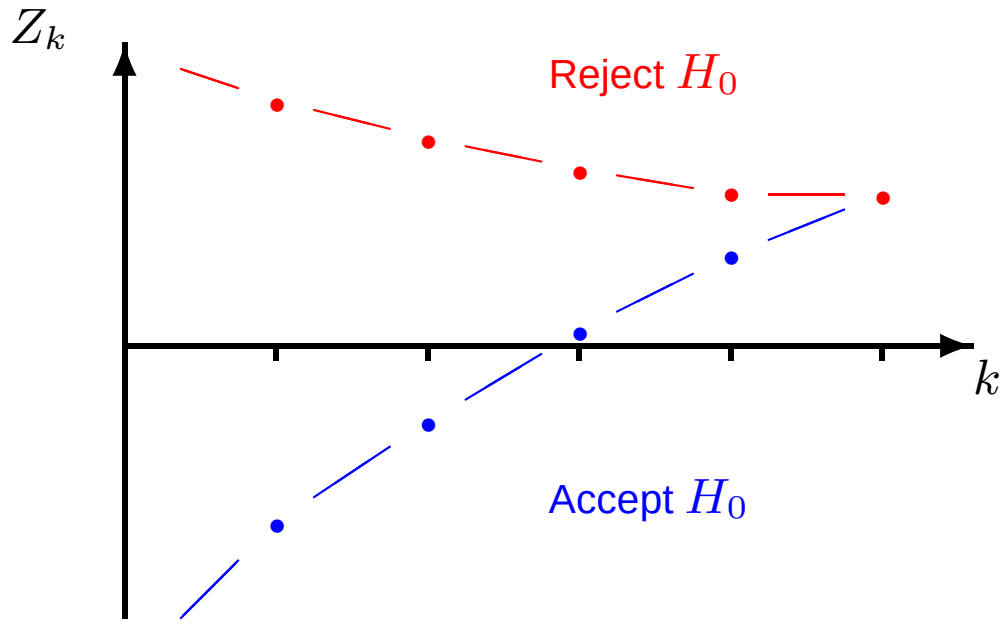
$$P_{\theta=0}\{\text{Reject } H_0\} = \alpha,$$

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

Standardised test statistics $Z_1, Z_2, \ldots,$ are computed at interim analyses and these are used to define a stopping rule for the trial.

Group sequential one-sided tests

A typical boundary for a one-sided test 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.

Joint distribution of parameter estimates

Reference: Sec. 3.5 and 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 demonstrated 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.

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 (score statistics) for comparing two survival curves

— and to Z-statistics formed from these.

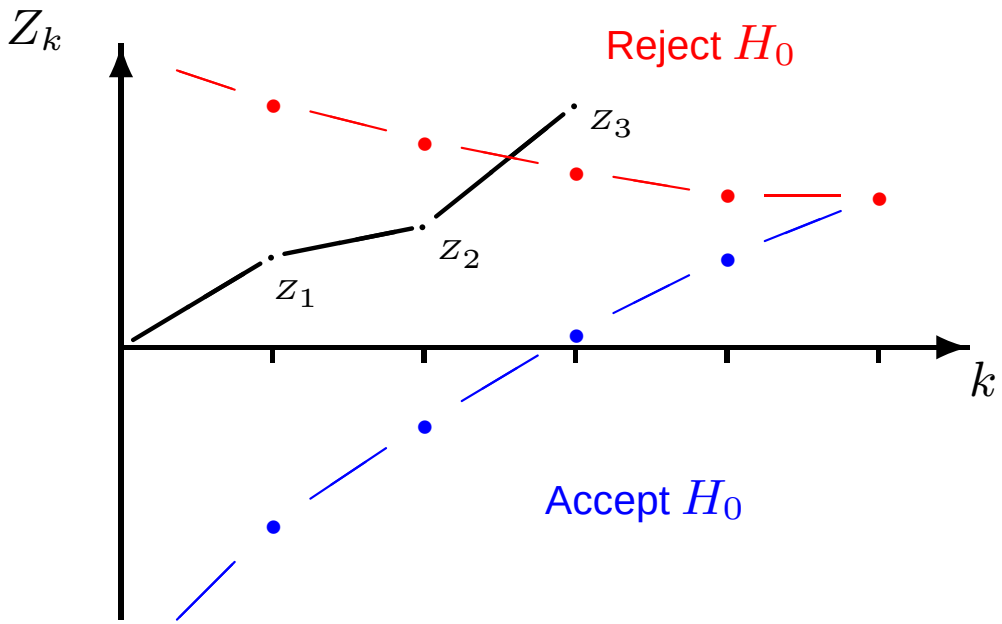
For survival data, observed information is roughly proportional to the number of failures.

Special types of group sequential test are needed to handle unpredictable and unevenly spaced information levels: see *error spending tests*.

Reference:

"Group-sequential analysis incorporating covariate information", Jennison and Turnbull (*J. American Statistical Association*, 1997).

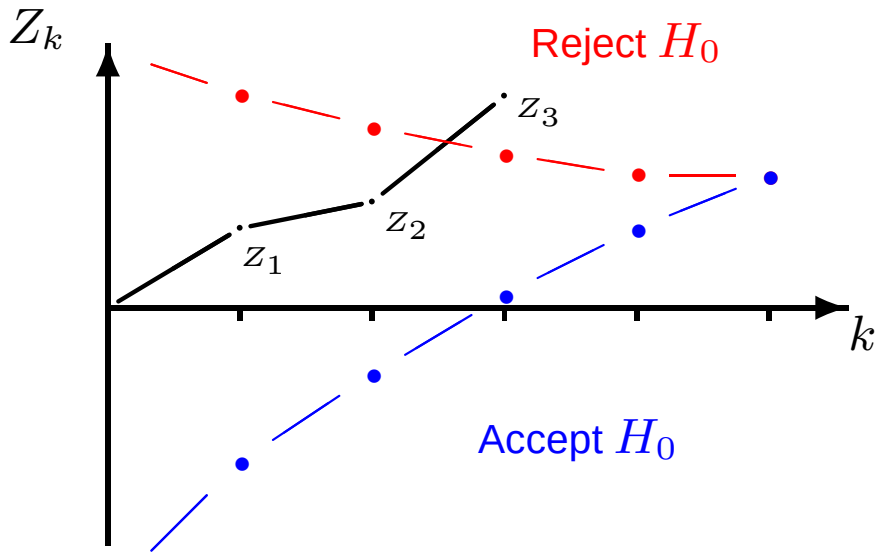
Computations for group sequential tests



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 (see JT, Ch. 19).

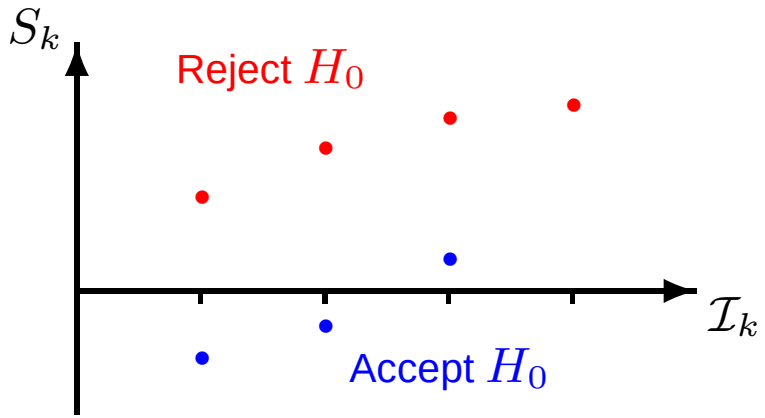
Combining such probabilities yields properties of a group sequential boundary.

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

Pampallona & Tsiatis (*J. Statistical Planning and Inference*, 1994).

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 $\theta = \delta$.



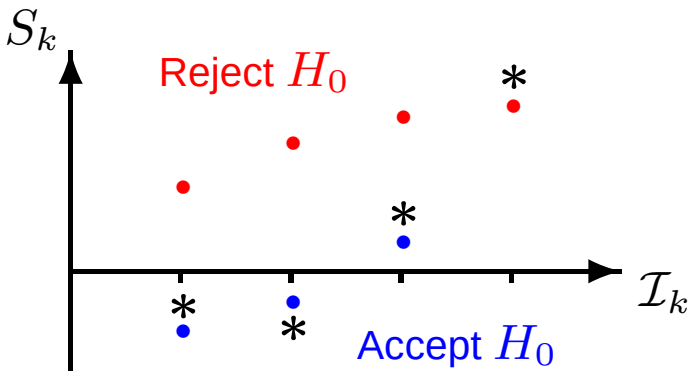
The computational methods just described can be used to define tests with parametric stopping boundaries meeting the design criteria.

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

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

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 study 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.

For planning purposes, power and expected sample size should be computed assuming the futility boundary will be obeyed.

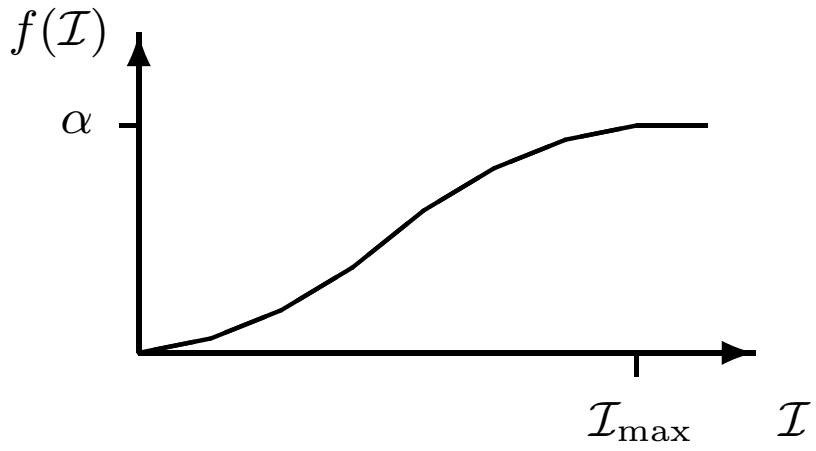
Constants can be computed in this way for, say, a Pampallona & Tsiatis test.

Error spending tests

Since the sequence $\mathcal{I}_1, \mathcal{I}_2, \ldots$ is often unpredictable, it is good to have a group sequential design that can adapt to the observed information levels.

Lan & DeMets (*Biometrika*, 1983) presented two-sided tests of H_0: $\theta = 0$ against $\theta \neq 0$ which "spend" type I error as a function of observed information.

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



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

The null hypothesis, H_0, is accepted if $\mathcal{I}_{\max}$ is reached without rejecting H_0.

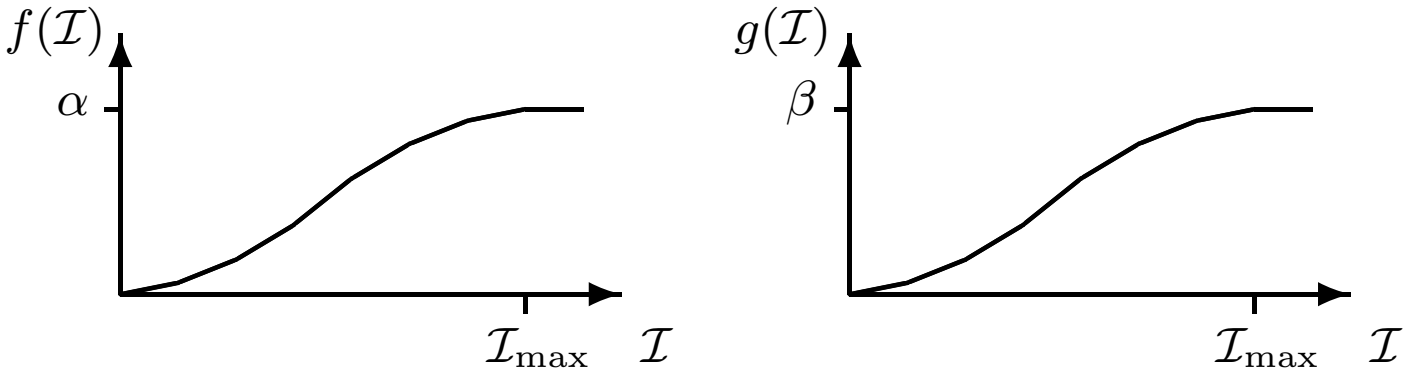
One-sided error spending tests

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:

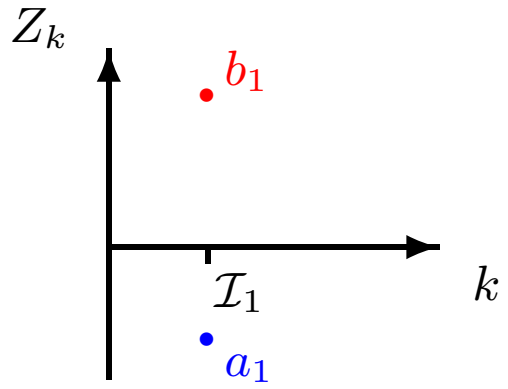
Observed information $\mathcal{I}_1$.

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 $\mathcal{I}_2$

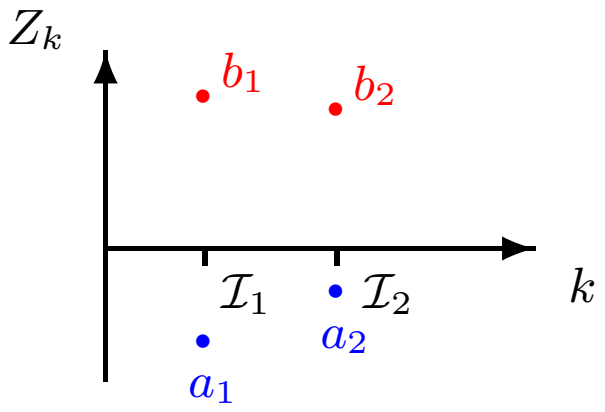
Reject H_0 if $Z_2 > b_2$, where

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

— 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

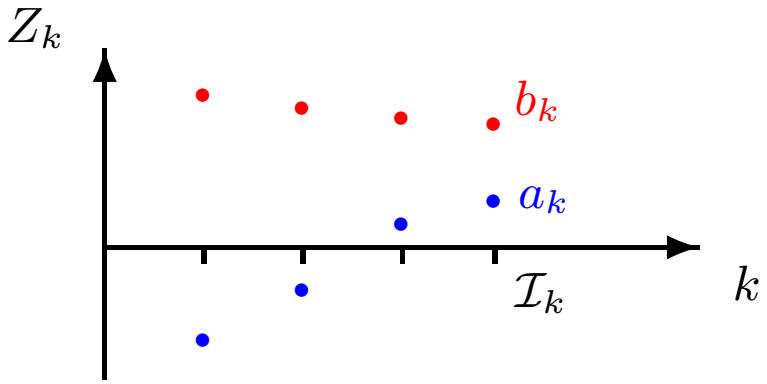
Analysis k: Observed information $\mathcal{I}_k$

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

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



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

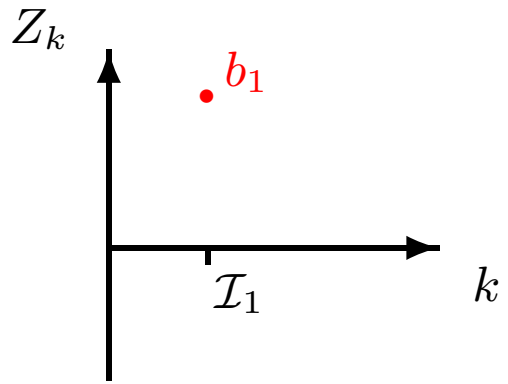
If the futility boundary is treated as non-binding, computation of the error-spending efficacy boundary only involves the type I error spending function $f(\mathcal{I})$.

Analysis 1:

Observed information $\mathcal{I}_1$.

Reject H_0 if $Z_1 > b_1$, where

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

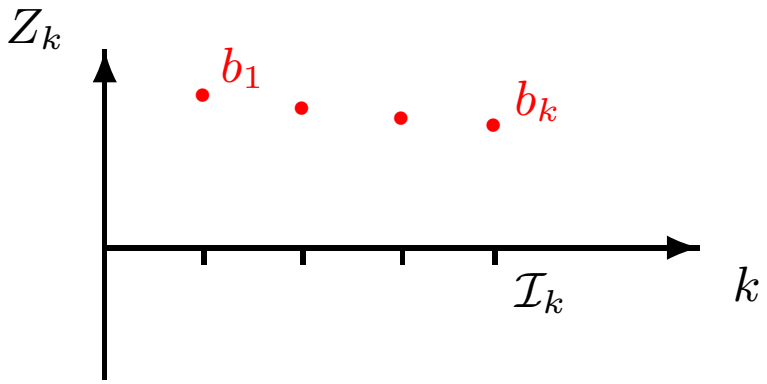


One-sided error-spending tests: Non-binding futility boundary

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 boundary

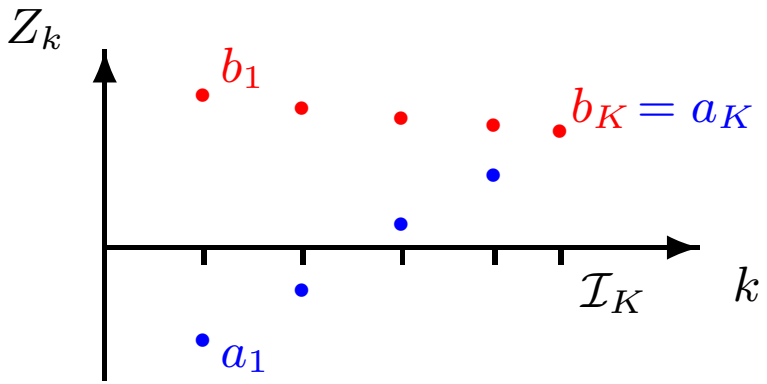
The futility boundary can be added through a type II error spending function $g(\mathcal{I})$.

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 $a_K = b_K$.



Example: An error spending test with normal response

Consider a two-treatment comparison with responses $X_{Ai} \sim N(\mu_A,\, \sigma^2)$ and $X_{Bi} \sim N(\mu_B,\, \sigma^2)$ on treatments A and B, respectively. Let $\theta = \mu_A - \mu_B$.

Suppose it is desired 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 $\rho = 2$.

This test spends type I error probability as

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

and type II error probability as

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

One-sided error spending test with non-binding futility boundary

Information

Suppose it is known that $\sigma^2 = 0.64$.

When the total numbers of observations are 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.

Assuming 5 equally spaced analyses, calculation shows the ρ-family error spending test with $\rho = 2$ 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 that at analysis 1 we observe $\widehat{\theta}_1 = 0.10$ 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.

Successive analyses proceed along the same lines . . . .

Applying a ρ-family error spending test

After further analyses using a ***non-binding*** futility boundary, for the data and testing boundary shown below the trial stops to reject H_0 at analysis 4.

Analysis k	*Cumulative sample size* $(n_A + n_B)$	$\mathcal{I}_k$	*Boundary* a_k, b_k	$\hat{\theta}_k$	*s.e.* $(\hat{\theta}_k)$	Z_k
1	40	15.6	-1.038, 3.061	0.10	0.253	0.395
2	76	29.7	-0.032, 2.721	0.06	0.184	0.327
3	114	44.5	0.769, 2.475	0.21	0.150	1.401
4	152	59.4	1.441, 2.282	0.31	0.130	2.389
5	190	74.2	2.113, 2.113	(0.33)	(0.116)	(2.843)

A ρ-family error spending test with binding futility boundary

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

The upper boundary is now lower — but only noticeably so at analyses 4 and 5:

Analysis k	*Cumulative sample size* $(n_A + n_B)$	$\mathcal{I}_k$	*Boundary* a_k, b_k	$\hat{\theta}_k$	s.e. $(\hat{\theta}_k)$	Z_k
1	40	15.6	-1.038, 3.061	0.10	0.253	0.395
2	76	29.7	-0.032, 2.721	0.06	0.184	0.327
3	114	44.5	0.769, 2.475	0.21	0.150	1.401
4	152	59.4	1.441, 2.277	0.31	0.130	2.389
5	190	74.2	2.041, 2.041	(0.33)	(0.116)	(2.843)

A ρ-family error spending test with binding futility boundary

With a non-binding futility boundary, power under $\theta = 0.4$ is 0.900.

With a binding futility boundary, the lower efficacy boundary gives higher power: when $\theta = 0.4$, the power is 0.906.

Alternatively, if a binding futility boundary is used, the trial can be designed with $\mathcal{I}_{max} = 72.26$ to give power 0.900 when $\theta = 0.4$.

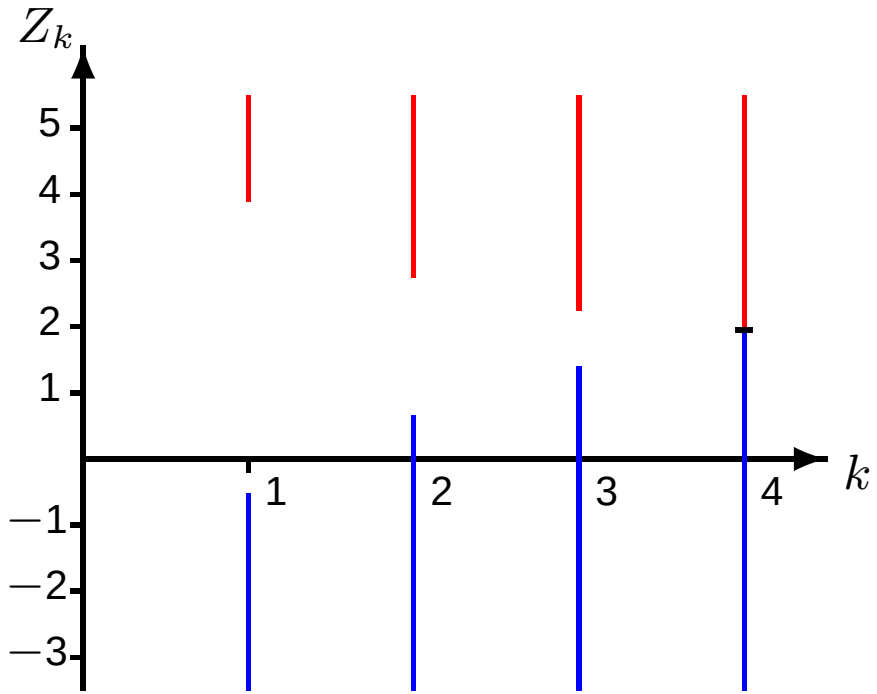
The ρ-family error spending function with $\rho = 2$ spends error slowly at the first few analyses. The boundaries are wide, making it difficult to cross one boundary and then the other, so the differences between binding and non-binding cases are small.

These differences can be greater when error is spent more rapidly, e.g., for a ρ-family error spending design with $\rho = 1$.

2. Analysis on termination of a group sequential trial (JT, Ch. 8)

A group sequential test's sample space is all possible pairs (k, Z_k) on termination.

The figure shows this sample space for a Pampallona & Tsiatis test with $\Delta = 0$, $K = 4$ analyses, type I error rate $\alpha = 0.025$ and power $1 - \beta = 0.8$ at $\theta = 1$.



Frequentist inference is based on probabilities over the *sample space of the study*.

The need for special methods

Suppose our 4 stage study with a Pampallona & Tsiatis boundary ends at stage 3 with $Z_3 = 2.6$. It is tempting to quote a 1-sided P-value of

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

But then, we would also get a P-value ≤ 0.0047 by

stopping at stage 1 with $Z_1 > 3.90$,

stopping at stage 2 with $Z_2 > 2.76$,

stopping at stage 3 with $Z_3 > 2.60$,

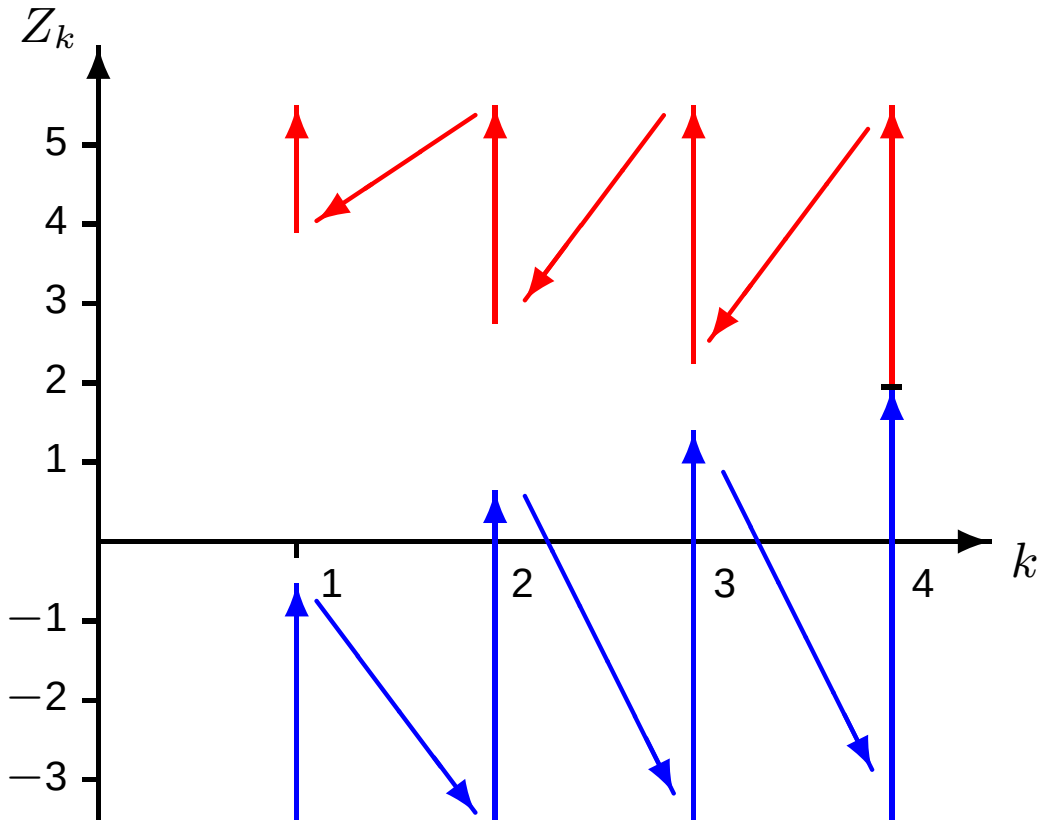
stopping at stage 4 with $Z_4 > 2.60$,

and the total probability under $\theta = 0$ of a "P-value" ≤ 0.0047 would be 0.0076.

So, this "P-value" is *not* distributed as $U(0,1)$.

Analysis on termination

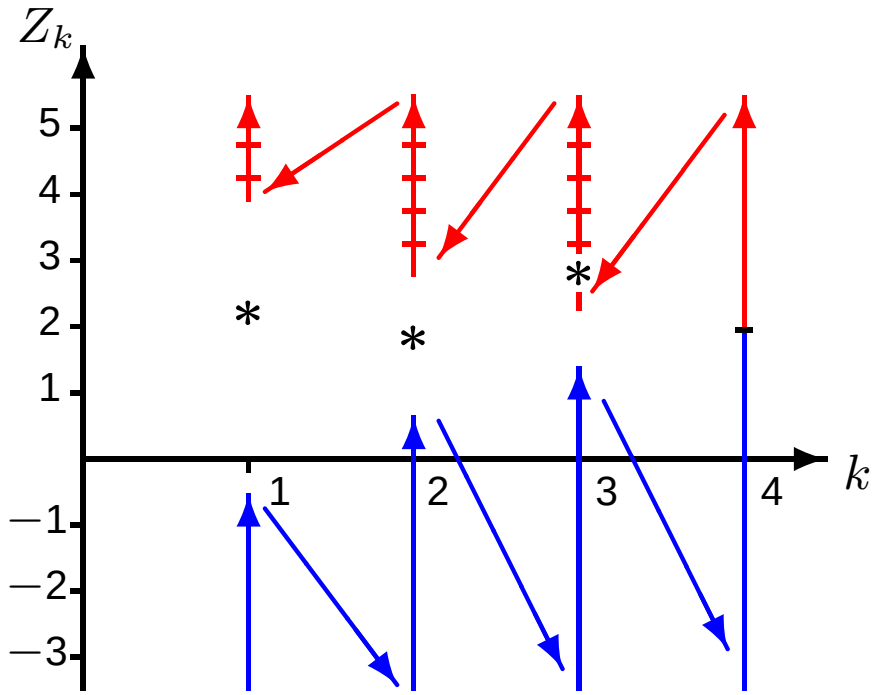
For proper frequentist inference, we first order the sample space.



We shall define P-values and confidence intervals with respect to this ordering.

A P-value on termination

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



So, on stopping at analysis 3 with $Z_3 = 2.60$, the 1-sided P-value for H_0: $\theta \leq 0$ is

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

A P-value on termination

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

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

If the group sequential test has one-sided type I error probability α,
the P-value is $\leq \alpha$ precisely when the test stops with rejection of H_0,
i.e., in the part of the sample space coloured red.

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

A confidence interval on termination

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

A $100(1-2\alpha)\%$, equal-tailed confidence interval for θ contains precisely those values θ for which the observed outcome (k^*, Z^*) is in the middle $(1-2\alpha)$ of the probability distribution of outcomes under θ.

This can be seen to be the interval $(\theta_1,\, \theta_2)$ where

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

and

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

This follows from the relation between a $100(1-2\alpha)\%$ lower confidence limit for θ and the family of level 2α, two-sided tests of hypotheses H: $\theta = \tilde{\theta}$.

A confidence interval on termination

Example:

If the trial stops at analysis 3 with $Z_3 = 2.6$, the 95% confidence interval for θ is

$$(0.22,\ 1.77)$$

using our specified ordering.

In contrast:

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

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

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

Consistency of hypothesis testing and CI on termination

Suppose a group sequential study 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.

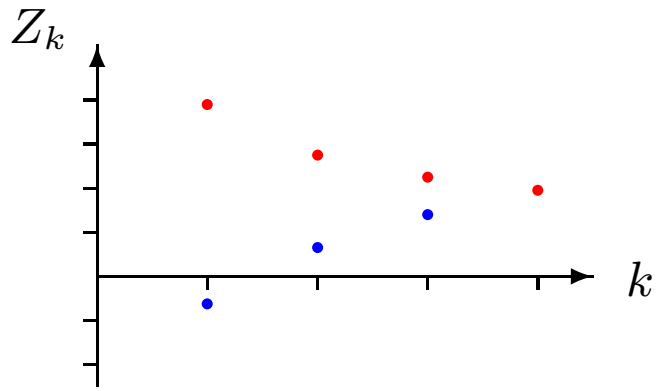
This happens automatically if outcomes for which we reject H_0 are at the top end of the sample space ordering — and any sensible ordering does this.

Why the naive approach does not work

Note that 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.

Because of the multiple testing effect, the probability of such an outcome is greater than the desired level α.

Estimating θ after a group sequential test



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

$$\widehat{\theta}_M = \sum_{i=1}^{n_k} (X_{Ai} - Y_{Bi})/n_k.$$

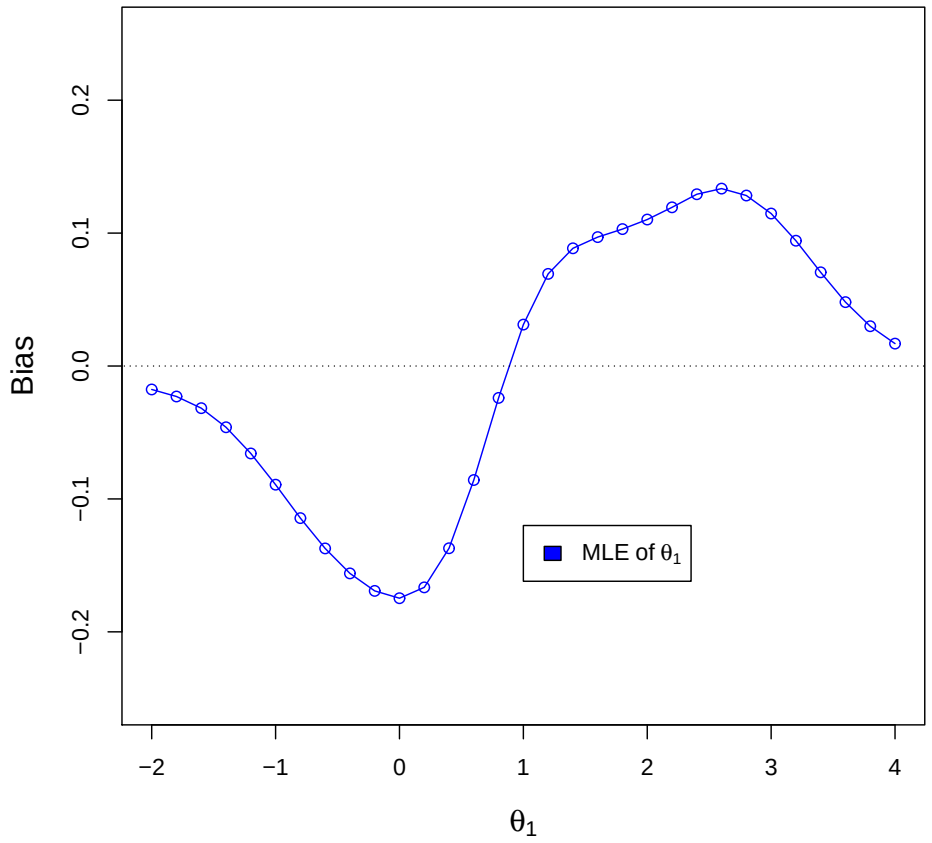
For large, positive values of θ, high values of $\widehat{\theta}$ lead to early stopping, while lower values lead to collection of more data and the chance for $\widehat{\theta}$ to increase.

This results in an upward bias of the MLE, so $E_\theta(\widehat{\theta}_M) > \theta$ for larger values of θ.

Similarly, $E_\theta(\widehat{\theta}_M) < \theta$ for lower values of θ.

Bias of the MLE of θ after a 4 group Pampallona & Tsiatis test

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



In our example, sample size is chosen to give power 0.8 for detecting a treatment effect of 1, and 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.

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

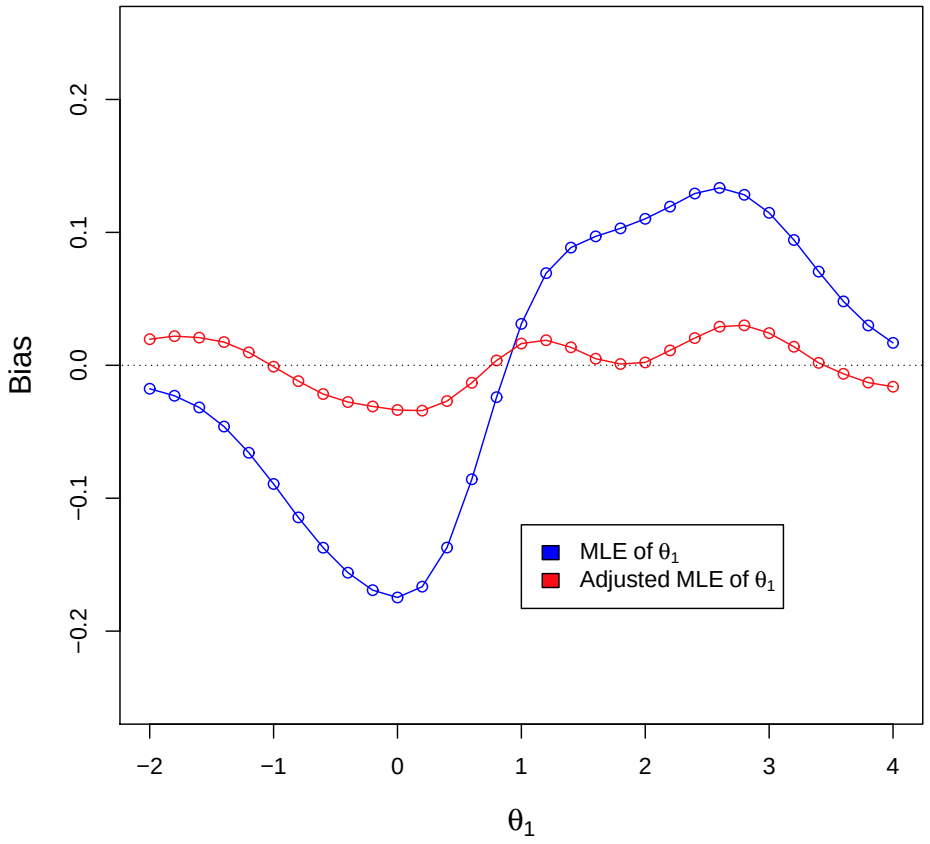
The adjusted estimator is then

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

Bias of the MLE of θ after a 4 group Pampallona & Tsiatis test

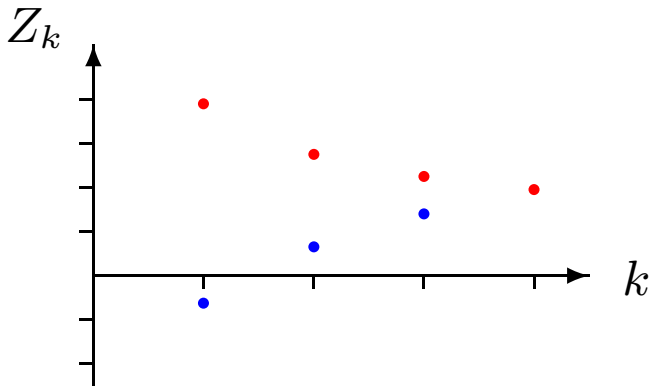
Simulation results show that Whitehead's adjusted estimator has much smaller bias than the MLE on which it is based.

For our example:



Estimating the treatment effect on a secondary endpoint after a group sequential test

Stopping boundary for the primary endpoint



Denote the treatment effect on the primary endpoint by θ_1.

Suppose the trial terminates with rejection of 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 subject, the primary and secondary responses are correlated.

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

Estimation for a secondary endpoint after a group sequential test

Suppose responses for an individual subject are bivariate normal with correlation ρ.

For a patient on Treatment A,

Primary endpoint $\quad X_1 \sim N(\mu_{A1},\, \sigma_1^2),$

Secondary endpoint $\quad X_2 \sim N(\mu_{A2},\, \sigma_2^2).$

Similarly, for a patient on Treatment B,

Primary endpoint $\quad X_1 \sim N(\mu_{B1},\, \sigma_1^2),$

Secondary endpoint $\quad X_2 \sim N(\mu_{B2},\, \sigma_2^2).$

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 group sequential test

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

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

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

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

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

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

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

Since this bias is a multiple of $b_1(\theta)$, it depends on θ_1 — and not on θ_2.

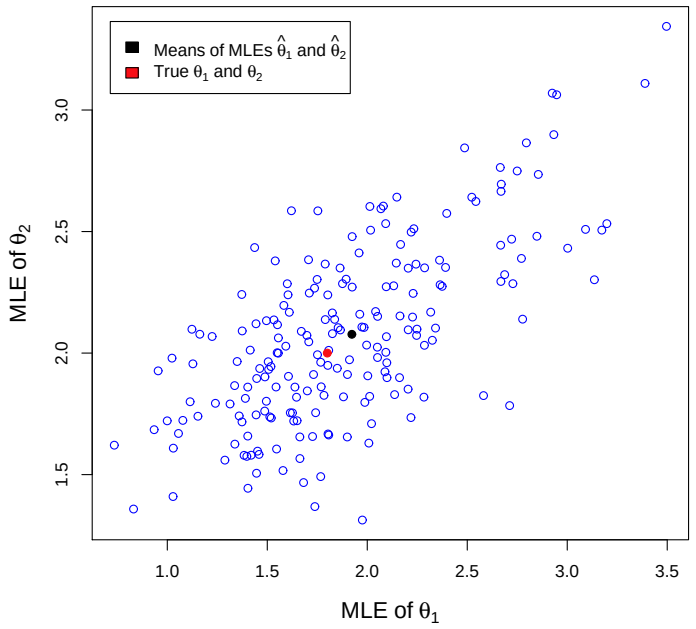
We can follow the same approach as for the primary endpoint and adjust the MLE, $\hat{\theta}_2$, by subtracting an estimate of its bias, $(\rho\,\sigma_2/\sigma_1)\,b_1(\hat{\theta})$.

Estimation for a secondary endpoint: Example

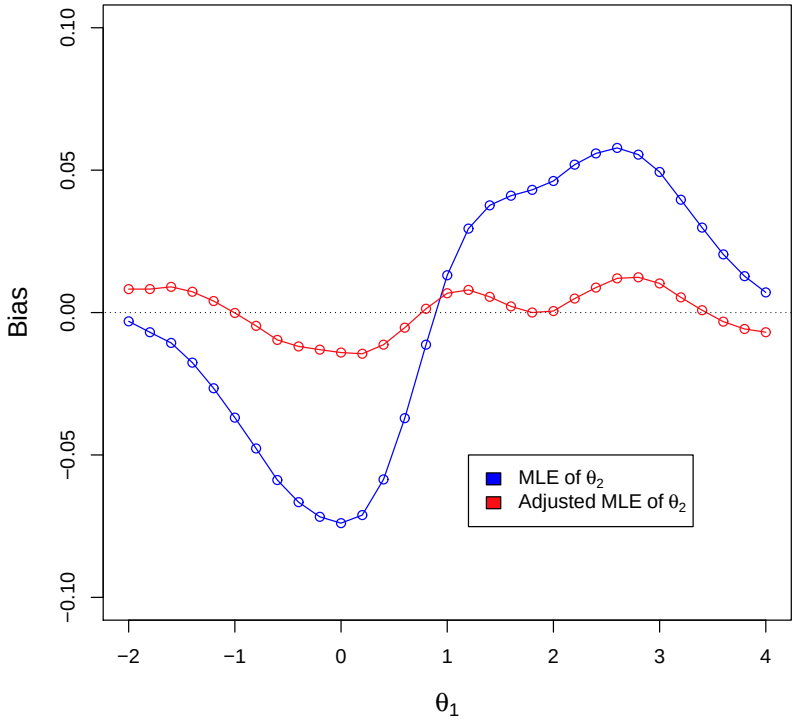
As previously, suppose a trial is designed for the primary endpoint using a Pampallona & Tsiatis test with $\Delta = 0, 4$ analyses, type I error rate $\alpha = 0.025$ and power 0.8 at $\theta_1 = 1$.

Assume responses are bivariate normal with correlation $\rho = 0.6$ 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 Pampallona & Tsiatis test.



Estimation for a secondary endpoint: Example



The plot shows the bias in the MLE $\hat{\theta}_2$ is largely eliminated in the adjusted estimator

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

3. Testing multiple endpoints in a single group sequential trial

Consider a trial comparing treatments A and B where the treatment effect for the primary endpoint is θ_1.

The trial has a group sequential design with a Pampallona & Tsiatis test with $\Delta = 0$, 4 analyses, $\alpha = 0.025$ and power 0.8 at $\theta_1 = 1$.

If the trial has a positive outcome, rejecting H_1: $\theta_1 \leq 0$ in favour of $\theta_1 > 0$, a secondary endpoint with treatment effect θ_2 is analysed.

The investigators believe it is appropriate to carry out a fixed sample size, level α test of H_2: $\theta_2 \leq 0$ against $\theta_2 > 0$.

Suppose that for an individual patient, the primary and secondary responses are bivariate normal with correlation ρ.

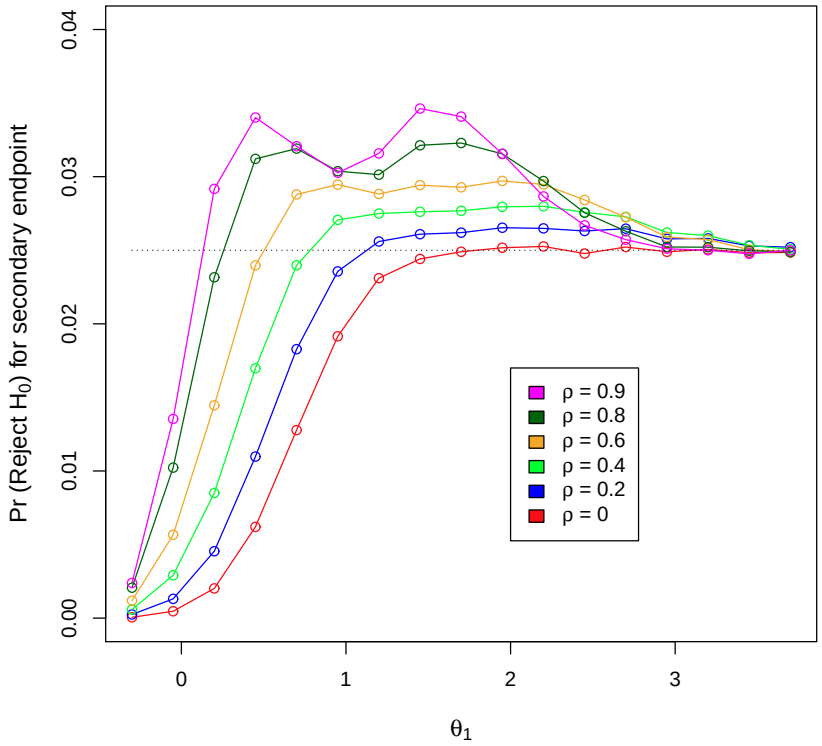
Is this approach to testing the two endpoints valid?

Hung, Wang and O'Neill (*J. Biopharm. Statis.*, 2007) explain why it is not valid.

Testing a secondary endpoint after a group sequential test

The plot shows the overall probability of rejecting H_2: $\theta_2 \leq 0$ (which requires rejection of H_1 first), when $\theta_2 = 0$. Values of θ_1 range from below 0 to above 3.

As ρ increases, the type I error rate for testing H_2 exceeds the nominal 0.025.



The type I error rate for the test of H_2 is inflated for the same reason that the MLE of θ_2 is biased upon conclusion of a group sequential test of θ_1.

Testing multiple hypotheses

A clinical trial may involve

Co-primary endpoints

Positive outcomes required for at least one endpoint

Positive outcomes required on all endpoints

Secondary endpoints, tertiary endpoints, . . .

The trial may have

Multiple treatments,

Pre-defined sub-populations of patients.

The trial design may be

Of fixed sample size,

Group sequential.

The familywise error rate

Suppose we have h null hypotheses, H_i: $\theta_i \leq 0$ for $i = 1, \ldots, h$.

A testing procedure yields a decision to accept or reject each of the h hypotheses.

The procedure's ***familywise error rate*** under a set of values $(\theta_1, \ldots, \theta_h)$ is

$$Pr\{\text{Reject } H_i \text{ for some } i \text{ with } \theta_i \leq 0\} \;=\; Pr\{\text{Reject any true } H_i\}.$$

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

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

Using such a procedure, we can choose to focus on a parameter θ_{i*} and claim significance for a test of the null hypothesis H_{i*}, without having to worry that this choice of hypothesis was based on observed data.

Bonferroni adjustment

Carlo Bonferroni (1892–1960) is associated with a simple, but very useful, result.

Suppose we have h null hypotheses and test each one at significance level α/h.

Then, if all h null hypotheses are true,

$$Pr\{\text{Reject at least one of } H_1 \ \ldots \ H_h\}$$

$$\leq \ Pr\{\text{Reject } H_1\} \ + \ \ldots \ + \ Pr\{\text{Reject } H_h\} \ = \ h \ \frac{\alpha}{h} \ = \ \alpha.$$

If only some of the h null hypotheses are true,

$$Pr\{\text{Reject at least one true } H_i\} \ < \ \alpha.$$

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

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

A Bonferroni test for co-primary endpoints

Example: Co-primary endpoints

A trial compares a new treatment against control with respect to two endpoints

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

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

One-sided type I error probability $\alpha = 0.025$ is divided between the endpoints.

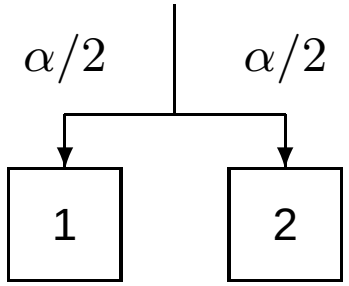
With Z-statistics Z_1 and Z_2 for endpoints 1 and 2,

An effect on Core MACE is declared if $Z_1 > \Phi^{-1}(1-\alpha/2) = 2.24$,

An effect on Expanded MACE is declared if $Z_2 > \Phi^{-1}(1-\alpha/2) = 2.24$.

Example: Co-primary endpoints

This Bonferroni procedure can be represented graphically as:



Familywise type I error is protected conservatively as there is a positive correlation between the two tests, due to the common aspects of the two endpoints.

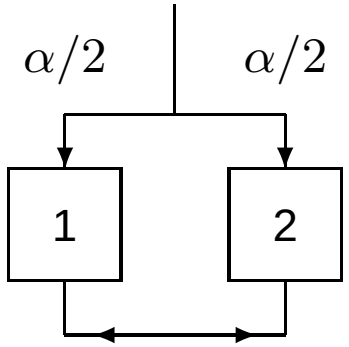
Suppose we have rejected H_1, might it be permissible to test H_2 at significance level α rather than $\alpha/2$?

If H_1 is false, we only need to worry about type I errors concerning H_2.

If H_1 is true, we have already made a type I error, so it will not increase the familywise error rate if we make another!

Bonferroni procedure with recycling (the Holm procedure)

We can represent a new version of the Bonferroni procedure, which "recycles" error probability after rejecting H_1 or H_2, as:

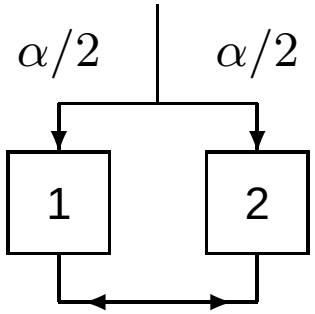


Proof that FWER is protected

If H_1 and H_2 are both true,

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

Bonferroni procedure with recycling (the Holm procedure)



Proof that FWER is protected (continued)

If H_1 is true and H_2 is false,

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

Similarly, if H_2 is true and H_1 is false,

$$\text{FWER} \;=\; Pr\{\text{Reject } H_2\} \;\leq\; Pr\{Z_2 > \Phi^{-1}(1-\alpha)\} \;=\; \alpha.$$

If H_1 and H_2 are both false,

A type I error cannot be made so FWER $= 0$.

A hierarchical testing or "gatekeeping" procedure

Example: Primary and secondary endpoints

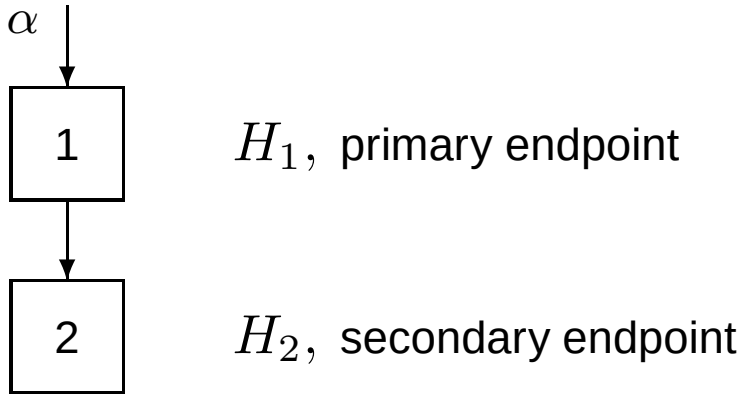
Consider a trial where

The null hypothesis H_1 concerns the primary endpoint,

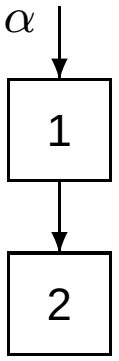
The null hypothesis H_2 relates to a secondary endpoint.

Suppose H_2 will only be tested if H_1 has already been rejected — O'Neill (*Controlled Clinical Trials*, 1997) states this is the only time one should test H_2.

We test H_1 first at significance level α. If H_1 is rejected, we continue to test H_2 at significance level α.



Example: Primary and secondary endpoints



H_1, primary endpoint

H_2, secondary endpoint

Proof that FWER is protected

If H_1 is true, a family-wise error occurs if H_1 is rejected (regardless of H_2),

$$\text{FWER} \;=\; Pr\{\text{Reject } H_1\} \;=\; Pr\{Z_1 > \Phi^{-1}(1-\alpha)\} \;=\; \alpha.$$

If H_1 is false and H_2 is true,

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

If H_1 and H_2 are both false, a type I error cannot be made and FWER $= 0$.

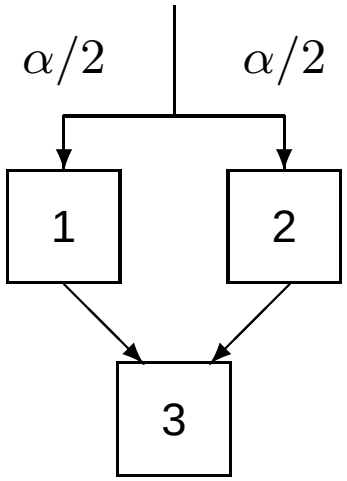
Co-primary and secondary endpoints

Suppose we wish to test a secondary endpoint if a positive result is obtained on either primary endpoint.

To do this, we recycle family wise error probability from the primary endpoints.

The secondary endpoint is tested at significance level $\alpha/2$ if just one of H_1 and H_2 is rejected, and at level α if both H_1 and H_2 are rejected.

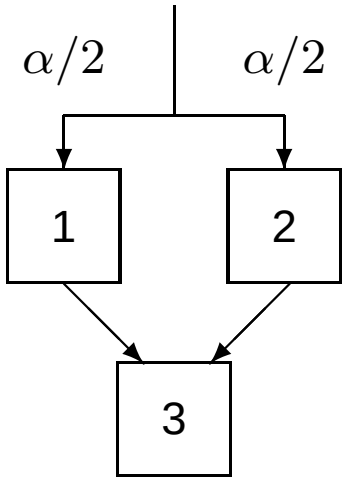
We can represent this testing procedure as:



$H_1,\ H_2$ co-primary endpoints

$H_3,$ secondary endpoint

Co-primary and secondary endpoints



$H_1,\ H_2$ co-primary endpoints

$H_3,$ secondary endpoint

There are eight different combinations of true and false values for H_1, H_2 and H_3.

Taking these eight cases in turn, it is quite easy to prove that FWER is at most α, whichever set of null hypotheses is true.

Questions?

1. Can we add more "recycling" to reduce conservatism and increase power?

2. Can we opt to recycle error between H_1 and H_2 before testing H_3 at all?

Closed testing procedures

In constructing and validating more elaborate forms of multiple testing, we can make use of "closed testing procedures" (Marcus et al, *Biometrika*, 1976).

The closed testing procedure

Suppose we have h null hypotheses, H_i: $\theta_i \leq 0$ for $i = 1, \ldots, h$.

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

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

which states that all hypotheses H_i are true, for $i \in I$.

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.

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 that a closed testing procedure provides strong control of the FWER at level α

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

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

Since $H_{\tilde{I}}$ is true,

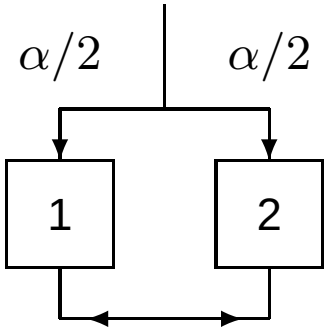
$$Pr\{\text{Reject } H_{\tilde{I}}\} = \alpha.$$

Thus,

$$Pr\{\text{Reject } H_j \text{ for at least one } j \in \tilde{I}\} \leq Pr\{\text{Reject } H_{\tilde{I}}\} = \alpha,$$

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

The Bonferroni test with recycling as a closed testing procedure



Let P_1 and P_2 denote P-values for simple tests of H_1 and H_2.

Write $H_{1,2}$ for the intersection hypothesis, $H_1 \cap H_2$.

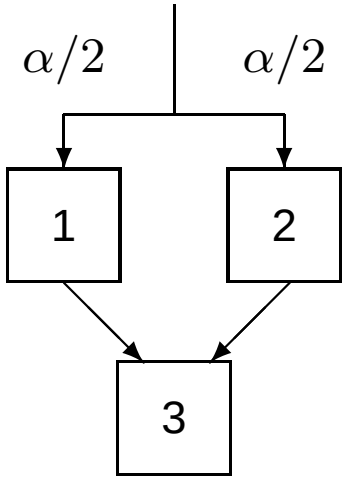
Using the closed testing procedure with the following set of tests is equivalent to the Bonferroni test with recycling.

Hypothesis	*Reject if*
H_1	$P_1 \leq \alpha$
H_2	$P_2 \leq \alpha$
$H_{1,2}$	$\min(P_1,\, P_2) \leq \alpha/2$

E.g., to reject H_1 overall we need individual tests to reject both H_1 and $H_{1,2}$, i.e.,

$$P_1 \leq \alpha \quad \text{and} \quad \min(P_1,\, P_2) \leq \alpha/2.$$

Testing co-primary and secondary endpoints as a closed testing procedure



$H_1,\ H_2$ co-primary endpoints

$H_3,$ secondary endpoint

Let P_1, P_2 and P_3 denote P-values for simple tests of H_1, H_2 and H_3.

The procedure depicted above is equivalent to a closed testing procedure with suitably defined tests of H_1, H_2 and H_3, and the related intersection hypotheses.

Co-primary and secondary endpoints: Closed testing procedure

Tests of intersection hypotheses are:

Hypothesis	*Reject if*
H_1	$P_1 \leq \alpha/2$
H_2	$P_2 \leq \alpha/2$
H_3	$P_3 \leq \alpha$
$H_{1,2}$	$\min(P_1,\, P_2) \leq \alpha/2$
$H_{1,3}$	$\min(P_1,\, P_3) \leq \alpha/2$
$H_{2,3}$	$\min(P_2,\, P_3) \leq \alpha/2$
$H_{1,2,3}$	$\min(P_1,\, P_2) \leq \alpha/2$

E.g., to reject H_3 overall needs rejection of H_3, $H_{1,3}$, $H_{2,3}$ and $H_{1,2,3}$, i.e.,

$$P_3 \leq \alpha,\ min(P_1, P_3) \leq \alpha/2,\ min(P_2, P_3) \leq \alpha/2,\ min(P_1, P_2) \leq \alpha/2,$$

which can be seen to agree with the procedure described earlier.

Answer to Question 1: A closed testing procedure with additional recycling

The tests of intersection hypotheses include:

Hypothesis	*Reject if*
H_1	$P_1 \leq \alpha/2$
H_2	$P_2 \leq \alpha/2$

This indicates conservatism. We can replace these tests by

Hypothesis	*Reject if*
H_1	$P_1 \leq \alpha$
H_2	$P_2 \leq \alpha$

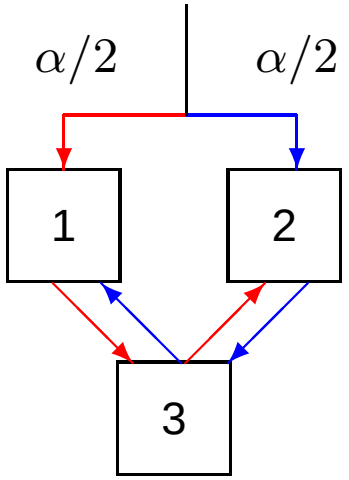
and their type I error rates will still be at most α.

This modification corresponds to recycling error probability from the test of H_3 back to whichever of H_1 and H_2 has not been rejected at level $\alpha/2$.

The extra opportunities to reject H_1 and H_2 give greater power.

Co-primary and secondary endpoints: A closed testing procedure with additional recycling

We can represent the testing procedure with additional recycling graphically.



$H_1,\ H_2$ co-primary endpoints

$H_3,$ secondary endpoint

The additional lines in the graph indicate that:

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

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

Answer to Question 2: Recycling between primary endpoints first

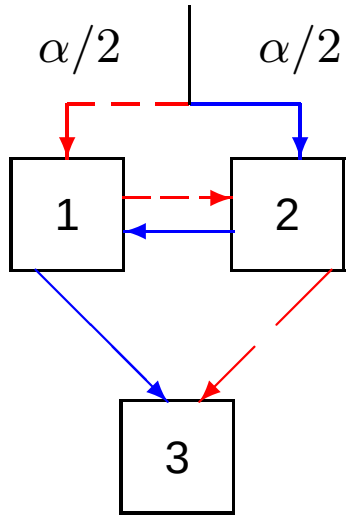
We may prefer to gain maximum power for tests of the co-primary endpoints before testing the secondary endpoint at all.

This is achieved by recycling error probability from H_1 to H_2, and vice versa, before allocating any error probability to a test of H_3.

A graphical representation is:

H_1, H_2 co-primary endpoints

H_3, secondary endpoint



One 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 H_3.

More complex procedures

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

It is still necessary to check that the familywise type I error rate is protected.

At the same time, we should try to avoid excessive conservatism.

We also want to be able to construct testing procedures that fit with:

A logical sequence for considering hypotheses, e.g., primary endpoint before secondary endpoint,

The relative impact of decisions to reject different hypotheses,

The perceived chance of being able to reject each hypothesis.

General methodology

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

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

The papers are:

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

and

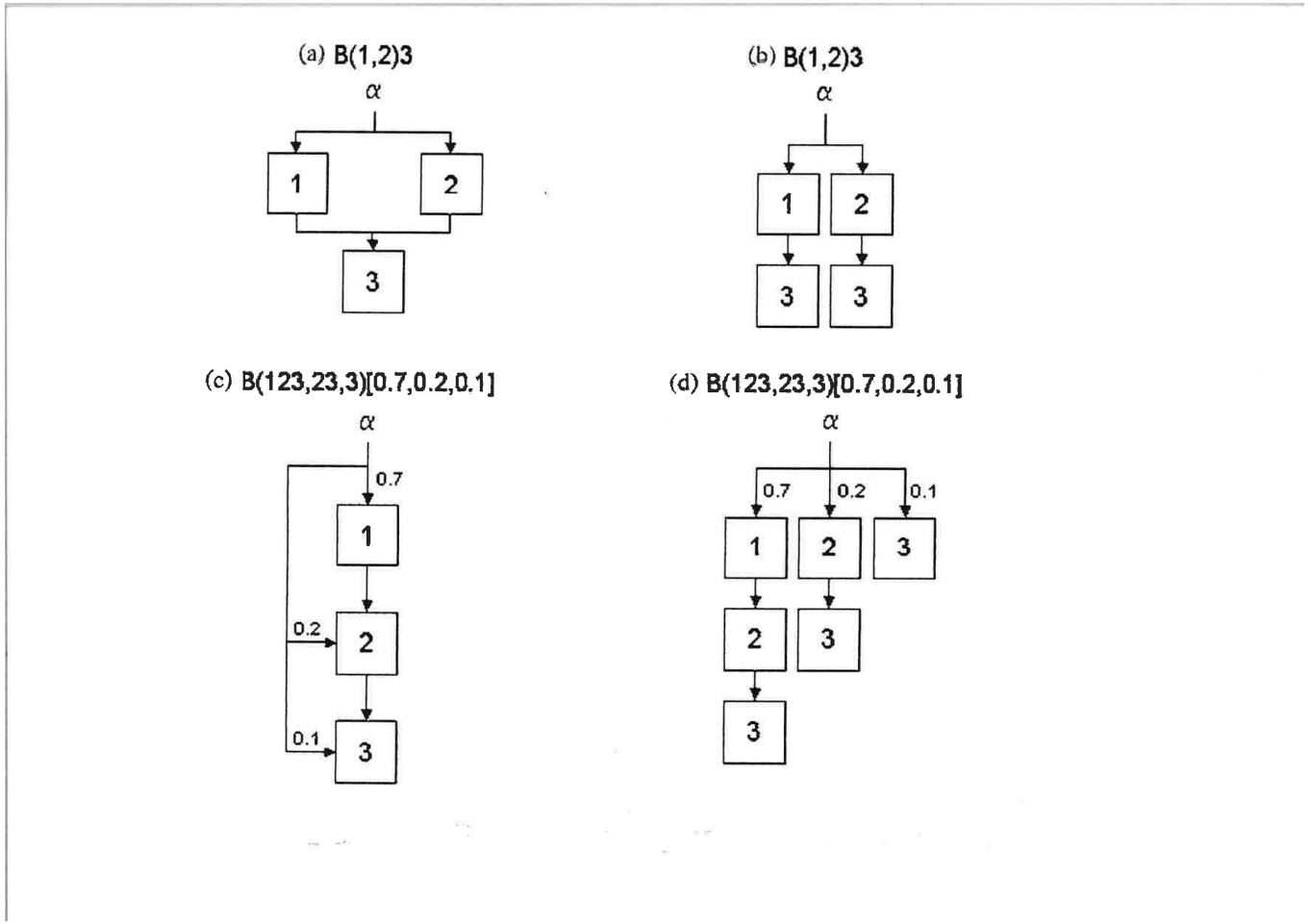
> "A graphical approach to sequentially rejective multiple test procedures" by Bretz, Maurer, Brannath and Posch, *Statistics in Medicine*, 2009.

The following diagrams give a flavour of what is possible and the graphical representations of multiple testing procedures used in the two papers .

A figure from Burman et al. (2009)

(a) and (b) A parallel gatekeeping procedure (equivalent versions)

(c) and (d) A fallback procedure (equivalent versions)



A figure from Bretz et al. (2009)

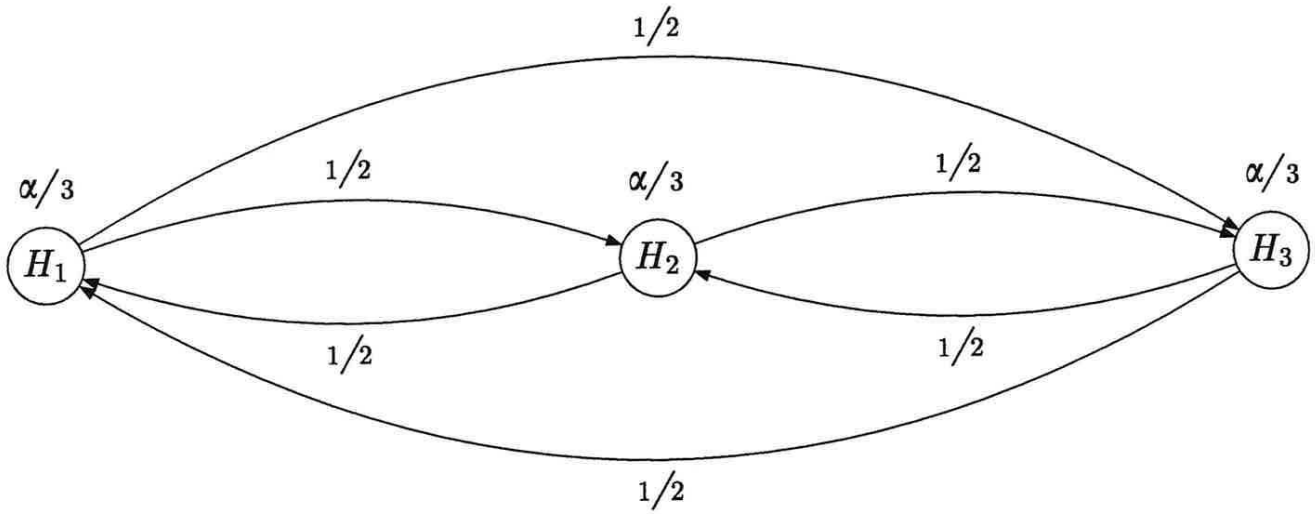


Figure 3. Graphical illustration of the Bonferroni–Holm procedure with $m=3$ hypotheses and initial allocation $\boldsymbol{\alpha}=(\alpha/3, \alpha/3, \alpha/3)$.

Multiple testing procedures and group sequential designs

We have just described multiple testing procedures in the context of a fixed sample size trial design.

Here, the sample space is simple and it is straightforward to define a Z-statistic or determine a P-value for each null hypothesis.

We can follow the same principles to test multiple hypotheses when a study is conducted group sequentially — but we shall need to define any P-values with proper attention to the sequential sampling frame.

In particular, the definition of a P-value should not change in response to observed data, either on the endpoint in question or other, correlated endpoints.

These considerations underlie discussion in the paper

> "Statistical considerations for testing multiple endpoints in group sequential or adaptive clinical trials" by Hung, Wang and O'Neill, *J. Biopharmaceutical Statistics*, 2007.

Testing a secondary endpoint after a group sequential test

In our earlier example, a trial compares two treatments with regard to their effect on a primary endpoint, then a secondary endpoint is analysed if a positive result is obtained for the primary endpoint.

Denoting the treatment effect on the primary endpoint by θ_1, a group sequential test is conducted of H_1: $\theta_1 \leq 0$ vs $\theta_1 > 0$.

If H_1 is rejected by the group sequential test, the secondary endpoint, with treatment effect θ_2, is analysed.

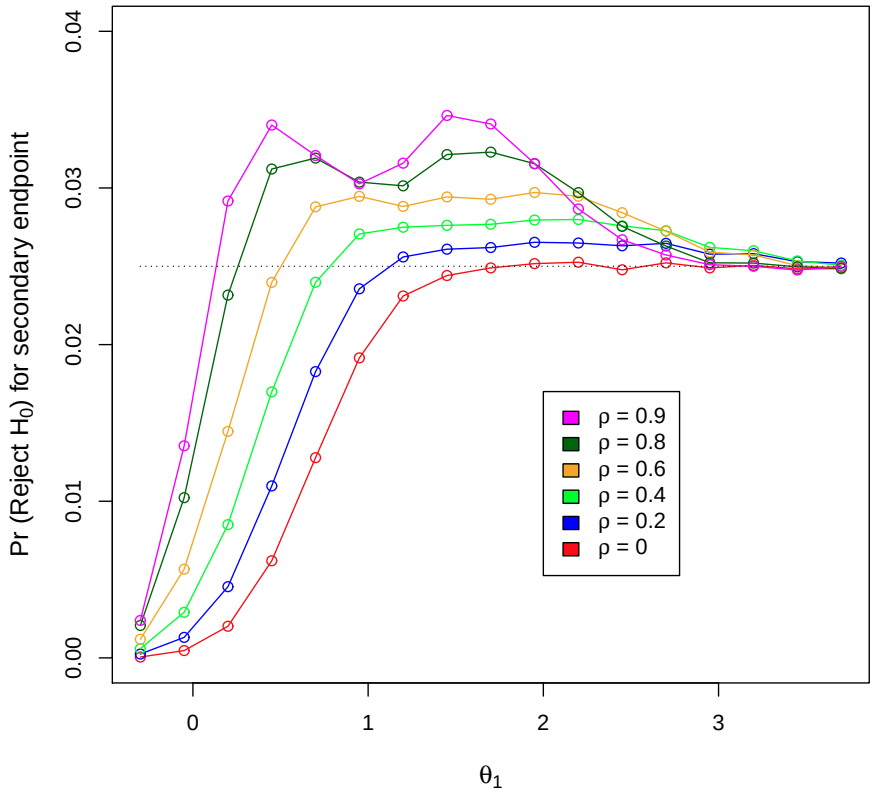
We supposed that investigators chose to conduct a fixed sample size, level α test of H_2: $\theta_2 \leq 0$ against $\theta_2 > 0$ using the data available for the second endpoint.

The investigators claim type I error probability α is passed from the test of H_1 to the test of H_2, just as in a "gatekeeping" procedure.

Does general theory ensure the familywise type I error rate is protected?

Testing a secondary endpoint after a group sequential test

We have already seen plots of the overall probability of rejecting H_2: $\theta_2 \leq 0$ when $\theta_2 = 0$ which show that familywise error rate is not protected at level $\alpha = 0.025$.



For $\rho > 0$, $Pr\{\text{Reject } H_2\} > \alpha$ for sufficiently high values of θ_1.

Why the "gatekeeping" argument does not apply

In the proposed design, H_2 is tested if H_1 is rejected.

The test of H_2 is based on the set of measurements of the secondary endpoint at analysis $j = 1, 2, 3,$ or 4, depending on when H_1 is rejected.

Each analysis $j = 1, \ldots, 4$ will give a different value for P_2, $P_2(j)$ say.

The plan is to reject H_2 if $P_2(j) \leq \alpha$ when H_1 is rejected at analysis j.

If a single value of j were specified and the trial always continued to analysis j (so we learn the value of $P_2(j)$), we would have

$$P_2(j) \sim U(0,1) \quad \text{under } \theta_2 = 0.$$

Then, rejecting H_2 when $P_2(j) \leq \alpha$ would give a level α test.

Instead, the proposal is to reject H_2 when $P_2(J) \leq \alpha$ where J is the random variable denoting the analysis at which the trial terminates.

Why the "gatekeeping" argument does not apply

We have defined the random variable J as the analysis at which the trial stops.

The value taken by J depends on observations on the primary endpoint.

These observations are correlated with those on the secondary endpoint, so there is dependence between J and the values $P_2(1)$, $P_2(2)$, $P_2(3)$ and $P_2(4)$.

The danger is that $P_2(J)$ is more likely to be one of the smaller values in the set $\{P_2(1), P_2(2), P_2(3), P_2(4)\}$, increasing the probability that $P_2(J) \leq \alpha$, and H_2 is rejected, above α.

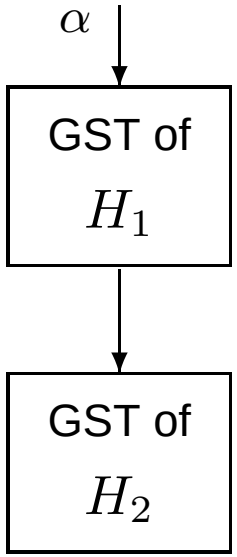
The simulation results for our example show this does indeed happen.

Solution:

We must test H_2 in a way which ***does not change in response to observed data, either on the endpoint in question or other, correlated endpoints.***

A correct gatekeeping procedure

We need to specify a single test of H_2 which can 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

The group sequential test of H_2 provides a critical value at each analysis.

If the first test rejects H_1 at analysis J, we compare data on the secondary endpoint to the critical value given by the GST of H_2 at analysis J.

A correct gatekeeping procedure

Let $\{Z_{1,1}, \ldots, Z_{1,K}\}$ denote Z-statistics for testing H_1: $\theta_1 \leq 0$ at analyses $1, \ldots, K$ when information for θ_1 is $\mathcal{I}_{1,1}, \ldots, \mathcal{I}_{1,K}$.

Similarly, let $\{Z_{2,1}, \ldots, Z_{2,K}\}$ be Z-statistics for testing H_2: $\theta_2 \leq 0$ at analyses $1, \ldots, K$ when information for θ_2 is $\mathcal{I}_{2,1}, \ldots, \mathcal{I}_{2,K}$.

The GST of H_1 stops at analysis k to

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

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

Boundary values are set to control type I error at level α under $\theta_1 = 0$, i.e.,

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

A correct gatekeeping procedure

The GST of H_2 rejects H_2 at analysis k if $Z_{2,k} \geq c_k$, where

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

Since the stopping rule for the trial is based on the primary endpoint, the test of H_2 does not need a futility boundary, which would imply early acceptance of H_2.

In the overall procedure, if the GST of H_1 stops to reject H_1 at analysis k^*, then we also reject H_2 if

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

A gatekeeping procedure using all of $\{Z_{2,1}, \ldots, Z_{2,K}\}$ *could* reject H_2 if

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

Hence, our overall procedure protects the familywise type I error rate conservatively.

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

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

This suggests 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 the most recent data (and the sufficient statistic for θ_2) would cast doubt on the credibility of this decision.

2. Continue the trial in the hope that $Z_{2,k} \geq c_k$ at some future analysis k.
   However, if the primary endpoint is also observed for future subjects, is there a risk of "losing" the positive result on the primary endpoint ?

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

Example: Testing primary and secondary endpoints

A trial compares two treatments with normally distributed responses.

The treatment effect is θ_1 for the primary and θ_2 for the secondary endpoint.

The trial is designed group sequentially with a Pampallona & Tsiatis test of the primary endpoint using $\Delta = 0$, 4 analyses, $\alpha = 0.025$ and power 0.8 at $\theta_1 = 1$.

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

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

Case A (Pocock):

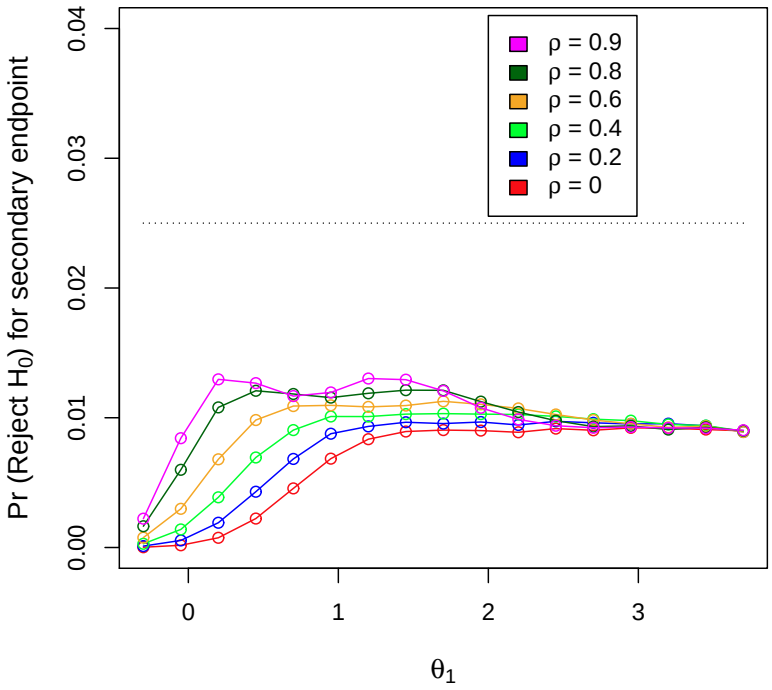
$$c_k \ = \ 2.361, \quad k = 1, \ldots, 4.$$

Case B (O'Brien & Fleming):

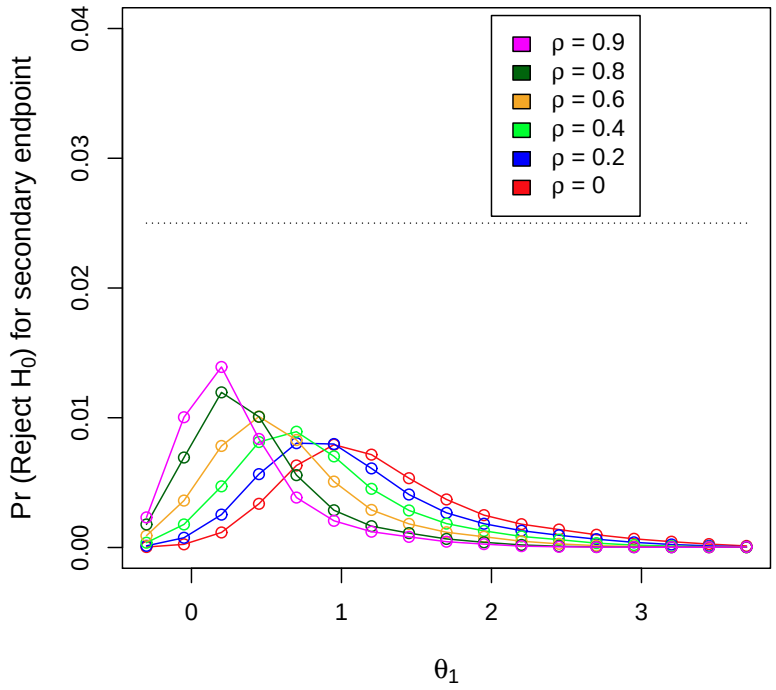
$$c_k \ = \ 2.024 \sqrt{\frac{4}{k}}, \quad k = 1, \ldots, 4.$$

Type I error probability for testing H_2

A: Pocock boundary for H_2



B: OBF boundary for H_2

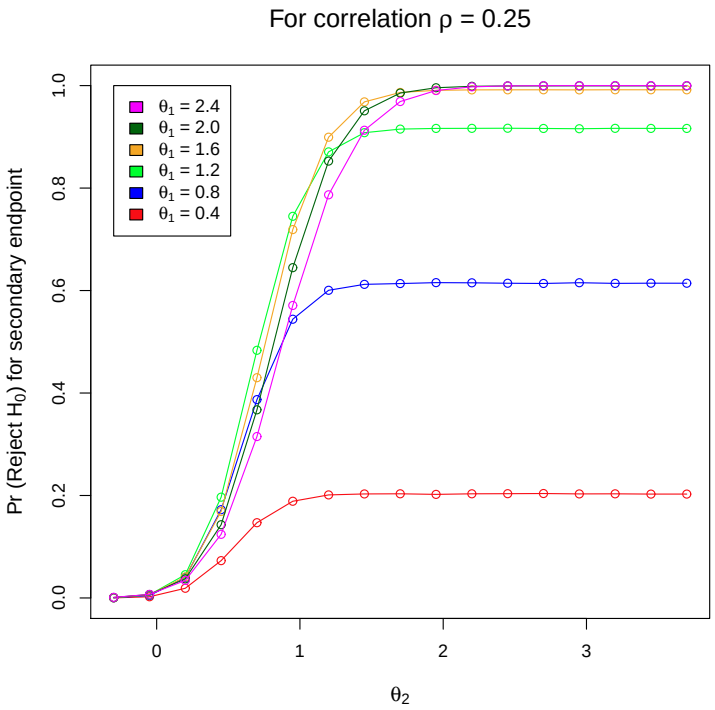


Type I error probabilities are calculated under $\theta_2 = 0$, but they also depend on θ_1 and the correlation, ρ, between the primary and secondary endpoints.

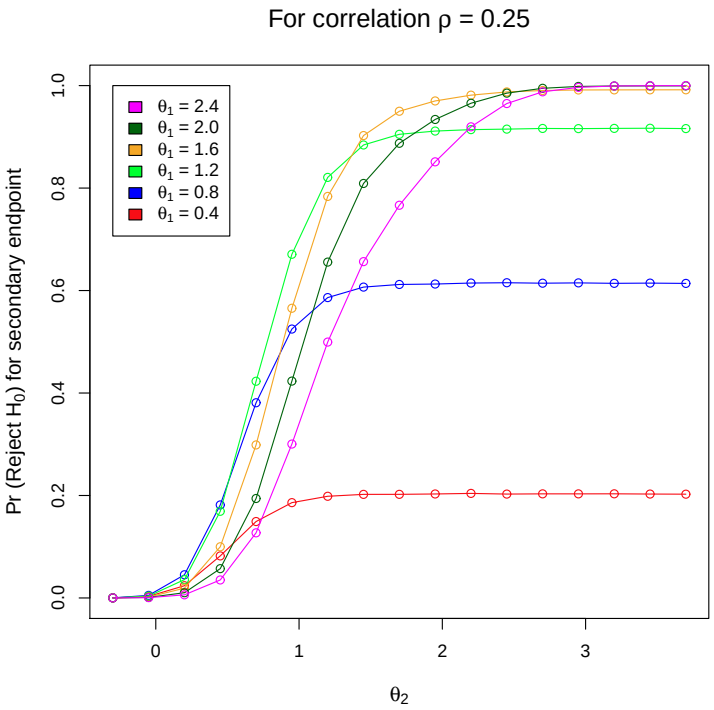
The test of H_2 is particularly conservative under large values of θ_1.

Power for testing H_2, $\rho = 0.25$

A: Pocock boundary for H_2



B: OBF boundary for H_2



We have supposed (without real loss of generality) that $\mathcal{I}_{2,k} = 2\,\mathcal{I}_{1,k}$.

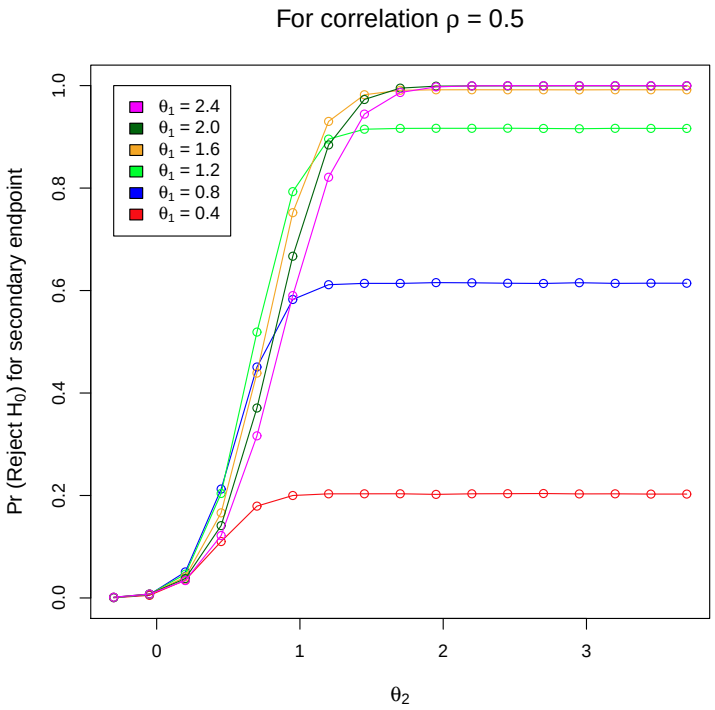
Power is shown as a function of θ_2 for selected values of θ_1 and ρ.

The value of θ_1 has a large effect on the analysis at which a test of H_2 may occur.

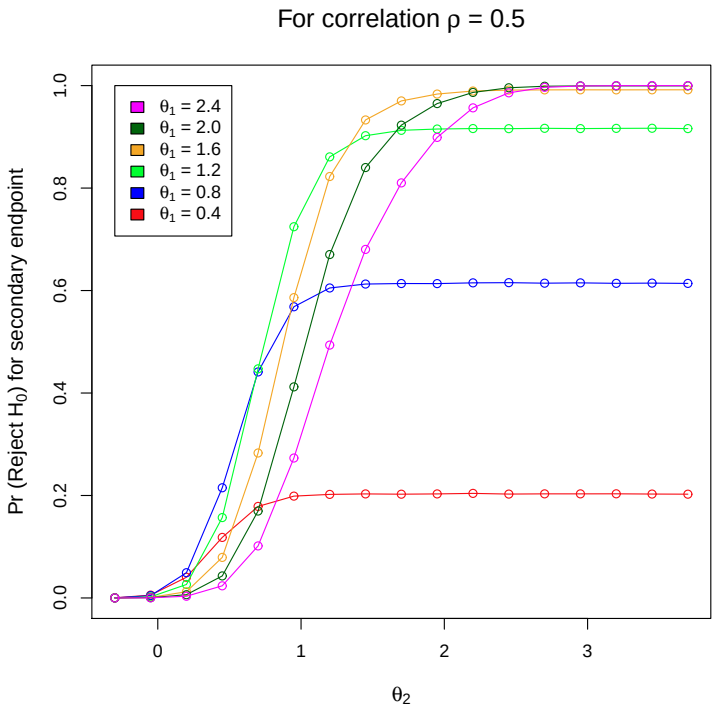
The Pocock boundary for H_2 deals better with the trial's uncertain termination time.

Power for testing H_2, $\rho = 0.5$

A: Pocock boundary for H_2



B: OBF boundary for H_2



We have supposed (without real loss of generality) that $\mathcal{I}_{2,k} = 2\,\mathcal{I}_{1,k}$.

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

The value of θ_1 has a large effect on the analysis at which a test of H_2 may occur.

The Pocock boundary for H_2 deals better with the trial's uncertain termination time.

GSTs and multiple hypothesis testing

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

2. Closed testing procedures encompass a variety of useful types of multiple hypothesis test.

3. Graphical representations (SiM papers, 2009) can help investigators to select — and understand — an appropriate procedure.

4. There are many options to choose from. A 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.

5. When testing multiple hypotheses in a group sequential trial design, the key point is to use GSTs as the “testing rules” in the multiple testing scheme: if this is not done correctly, FWER may be too high.

GSTs and multiple hypothesis testing: further reading

Tang & Geller (*Biometrics*, 1999) Closed testing procedures for group sequential clinical trials with multiple endpoints.

One treatment vs control with multiple endpoints, or multiple treatments vs control with a single endpoint.

In the closed testing procedure, each intersection hypothesis has its own GST.

Intersection hypotheses are tested systematically, starting with the intersection of all k hypotheses, then intersections of $(k-1)$ hypotheses, etc.

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

GMB consider hierarchical testing of a secondary endpoint in a group-sequential clinical trial that is mainly driven by a primary endpoint.

The "secondary" endpoint may actually be of prime interest and the primary endpoint only a surrogate to indicate when to test the secondary endpoint.

GSTs and multiple hypothesis testing: further reading

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

TML reduce the conservatism in Tang & Geller's method for the case of known correlation, ρ, between endpoints.

For given GSTs of H_1 and H_2 and a known value of ρ, they calculate the overall type I error rate for H_2. They then calibrate the GST for the secondary endpoint so the maximum overall type I error rate for H_2, over all values of θ_1, is α.

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.

TWM obtain an upper confidence bound, r, for the correlation ρ. They proceed on the basis that $\rho \leq r$, allocating fractions of α to (i) type 1 error for testing H_2 assuming $\rho \leq r$ and (ii) the probability that $\rho > r$.

GSTs and multiple hypothesis testing: further reading

Ye, Liu & Yao (*Statist. in Med.*, 2012) A group sequential Holm procedure with multiple primary endpoints.

In applying the Holm procedure to test h multiple hypotheses, one starts by dividing the familywise type I error probability α between the h hypotheses.

If a hypothesis is rejected, its error probability is re-distributed to the others.

This process continues until no more hypotheses can be rejected.

YLY follow this approach with GSTs for each hypothesis — at the appropriate collection of type I error rates.

Maurer & Bretz (*Statist. in Biopharm. Research*, 2013) Multiple testing in group sequential trials using graphical approaches.

M&B apply GSTs in multiple testing procedures with a graphical representation.

They give a thorough account of the details of this methodology, including the issue of "concordance" and when a set of GSTs has this property.

Co-primary endpoints

Earlier, we mentioned an example of a trial comparing a new treatment against control with respect to two endpoints,

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

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

One possibility is that approval for the new treatment could be sought based on a positive outcome on at least one endpoint.

In this case, the previously described methods are appropriate.

Suppose instead that a positive outcome is required on *both* endpoints in order for a New Drug Application to be possible.

What are the multiple testing implications?

What can a group sequential design offer in this case?

Co-primary endpoints

Suppose it is required to show a new treatment is effective on both endpoints.

Denote the treatment effects on the two endpoints by θ_1 and θ_2.

We wish to demonstrate that $\theta_1 > 0$ and $\theta_2 > 0$.

Formally we test the null hypothesis

$$H_0\text{: } \theta_1 \leq 0 \text{ or } \theta_2 \leq 0$$

against the alternative

$$H_A\text{: } \theta_1 > 0 \text{ and } \theta_2 > 0.$$

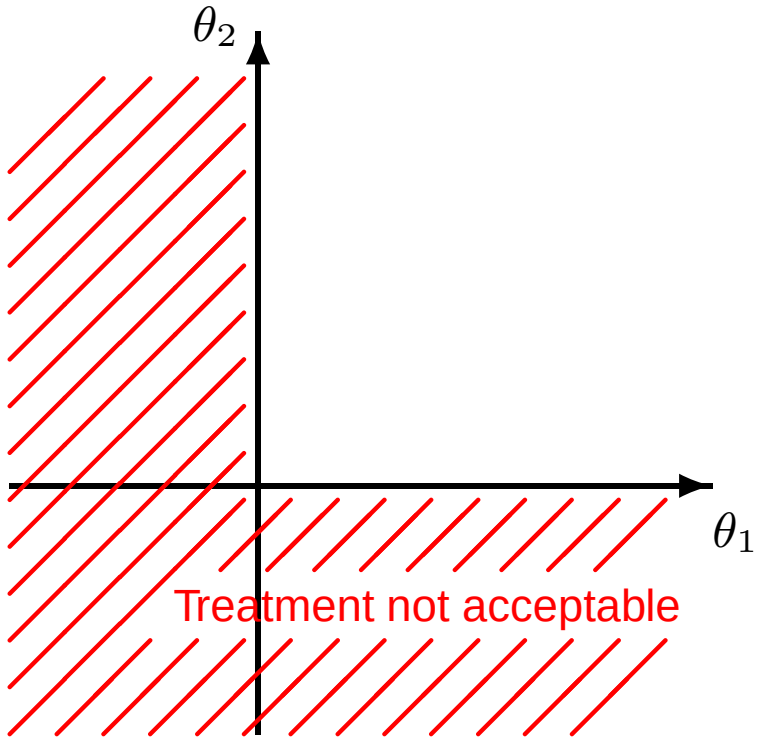
A type I error occurs if the new treatment is claimed to be effective,

i.e., if both $H_1\text{: } \theta_1 \leq 0$ and $H_1\text{: } \theta_2 \leq 0$ are rejected,

when *either* $\theta_1 \leq 0$ *or* $\theta_2 \leq 0$.

Co-primary endpoints

The type I error probability must be controlled over all values (θ_1, θ_2) in the null hypothesis H_0: $\theta_1 \leq 0$ or $\theta_2 \leq 0$, as shown below.



The type I error is largest at $(0, \infty)$ or $(\infty, 0)$.

Hence, one can define separate level α tests of $H_{0,1}$: $\theta_1 \leq 0$ and $H_{0,2}$: $\theta_2 \leq 0$ and claim the new treatment is effective if both null hypotheses are rejected.

Co-primary endpoints

Suppose a clinical trial is conducted in the hope of showing a treatment effect on both of two co-primary endpoints.

The trial will test

$$H_0\text{: } \theta_1 \leq 0 \text{ or } \theta_2 \leq 0$$

against the alternative

$$H_A\text{: } \theta_1 > 0 \text{ and } \theta_2 > 0.$$

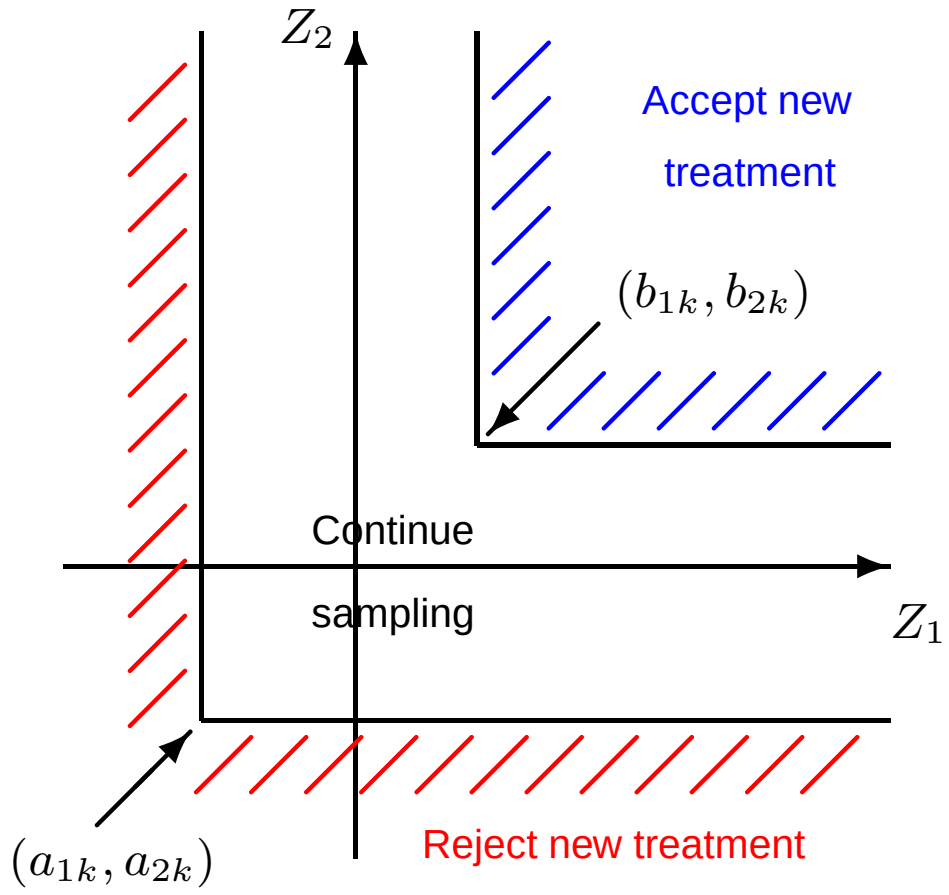
A group sequential design is possible — but this should only stop early for a positive outcome when there is evidence of a treatment effect for both endpoints.

Jennison & Turnbull (*Biometrics*, 1993) proposed such a group sequential design for a trial with efficacy and safety endpoints. (They used a non-inferiority criterion for safety, and so had $\theta_2 \leq -\delta$ rather than $\theta_2 \leq 0$ in their null hypothesis.)

Co-primary endpoints

Jennison & Turnbull's (1993) group sequential designs for a bivariate response have L-shaped boundaries at each analysis k.

The design is set up to achieve power at a specific pair of positive treatment effects.



Recapitulation: Group sequential tests and multiple hypotheses

- It is natural to monitor clinical trials with a view to possible early stopping.
- Distribution theory and computation support a variety of group sequential designs, including error spending tests, which control the type I error rate.
- Inference on termination can be conducted to give point estimates, p-values and confidence intervals with the usual frequentist properties.
- Such inferences can be extended to secondary endpoints — and adjustment for the stopping rule can be just as important for these inferences.
- When a trial is designed to test multiple endpoints:

  Care needs to be taken when combining multiple testing procedures (set up to protect FWER) with group sequential stopping rules.

  A safe approach is (i) to describe the mutiple testing procedure graphically, as per Burman et al. or Bretz et al. (2009), then (ii) specify group sequential tests that can be applied at each node of the graph.