Analysing over-run data after a group sequential trial

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

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

and Lisa Hampson

Novartis,

Basel

Adaptive Designs Workshop

Heidelberg

July 1, 2022

Outline of talk

1. Group sequential tests
2. The problem of over-run data
3. Whitehead's proposal
4. Group sequential tests for a delayed response (Hampson & Jennison, 2013)
5. Error spending designs with a non-binding futility boundary and over-run data
6. Conclusions

(7. Details of Hampson & Jennison's optimised delayed response GSTs)

1. Group sequential tests

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

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

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

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

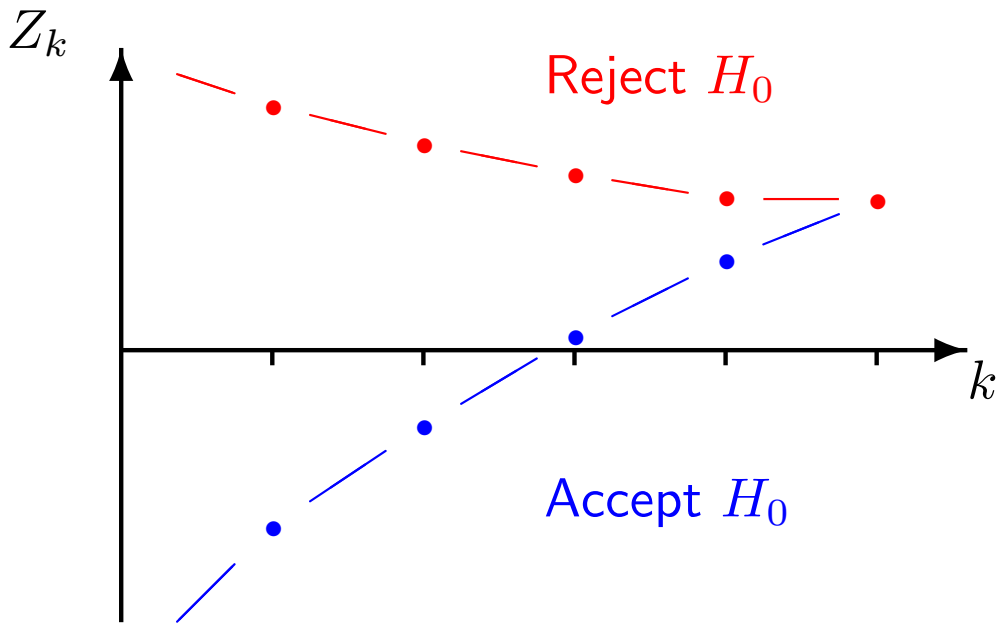
$$P_{\theta=0}\{\text{Reject } H_0\} = \alpha,$$

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

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

Group sequential tests

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



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

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

Benefits of group sequential testing

Earlier decisions

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

Fewer patients recruited

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

Stopping failing trials early

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

2. Over-run data

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

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

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

In practice, responses are observed some time after treatment.

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

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

How should the additional data be analysed?

Examples of over-run data

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

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

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

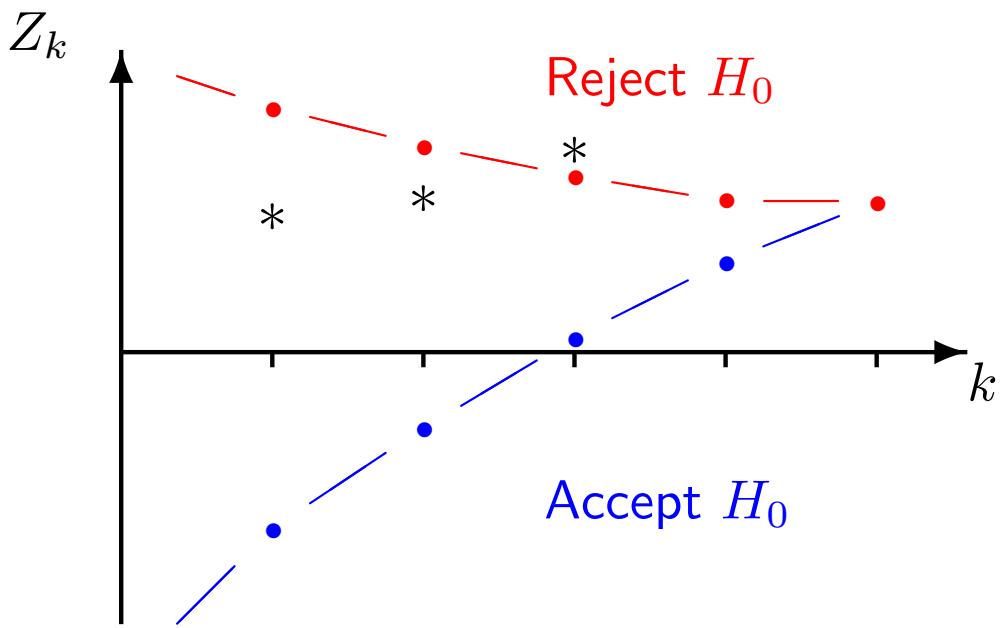
Example 2: Consider a clinical trial with a time-to-event endpoint.

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

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

Example 1: The cholesterol reduction trial

Suppose a standard group sequential test (GST) with type I error rate α is applied.



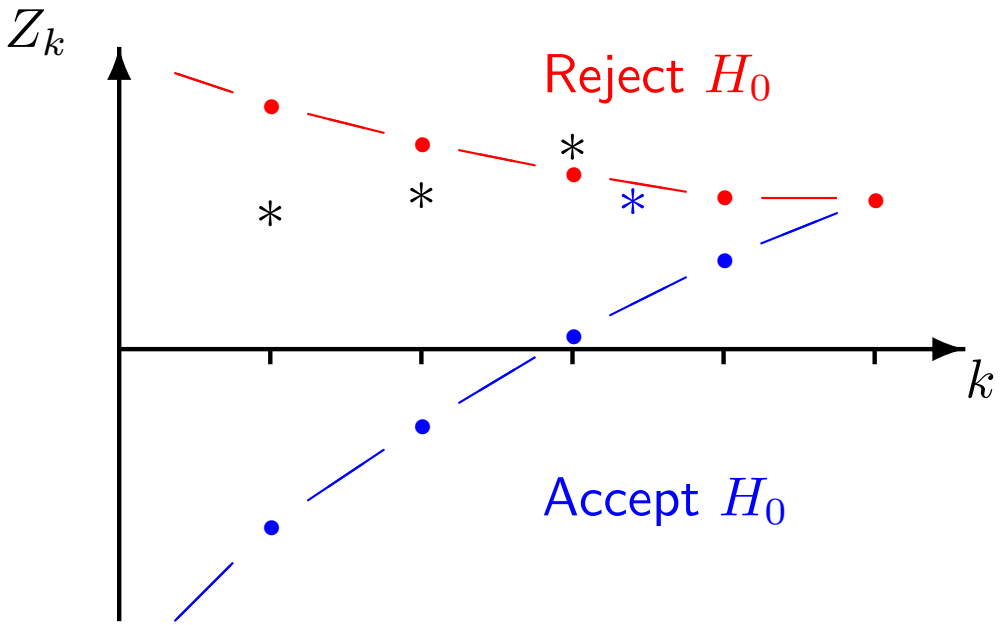
We observe $Z_3 = 2.4$, which exceeds the boundary value of 2.3.

The trial stops and the conclusion of the group sequential test is

"Reject H_0: $\theta \leq 0$".

Example 1: The cholesterol reduction trial

Now suppose that additional observations are observed for subjects who were "in the pipeline" at analysis 3.



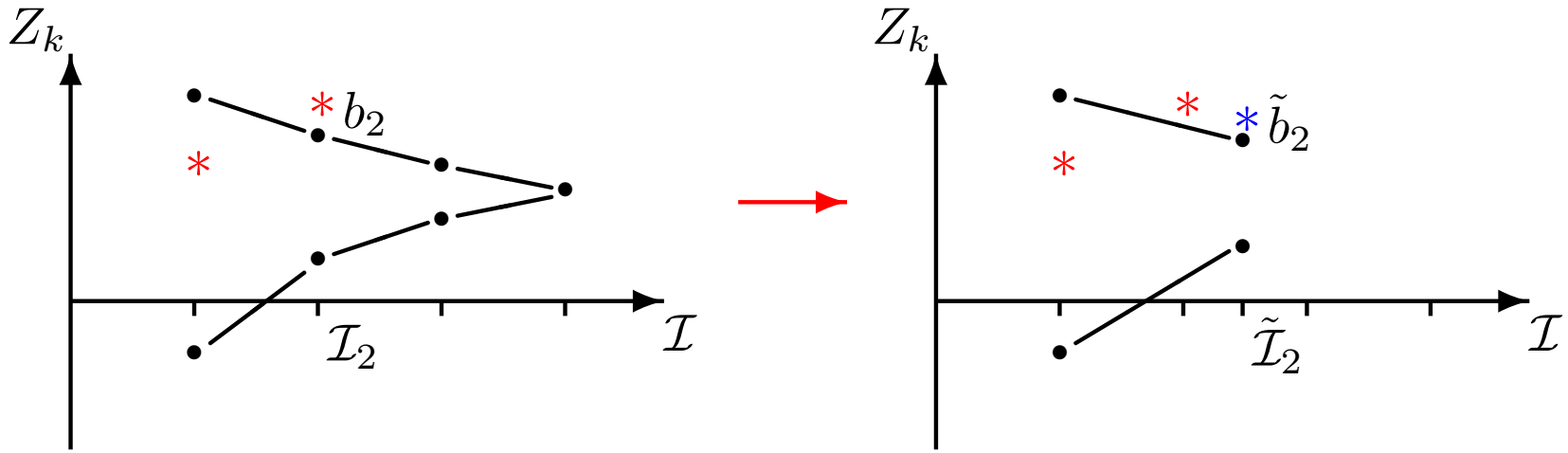
With the pipeline data included, we find $\tilde{Z}_3 = 2.1$.

Can the investigators claim significance at level α?

3. Whitehead's method

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

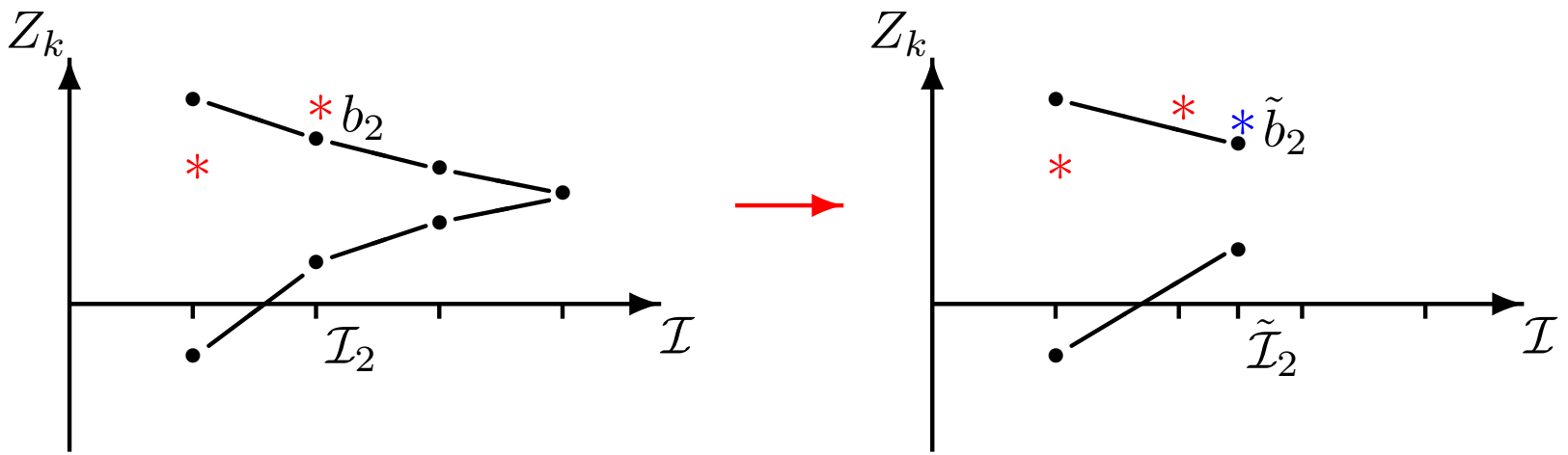
The analysis k at which termination occurs is deleted and one behaves as if analysis k had occurred with the information level $\tilde{\mathcal{I}}_k$ arising from the final set of responses.



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

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

Whitehead's method



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

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

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

4. Hampson & Jennison's method

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

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

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

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

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

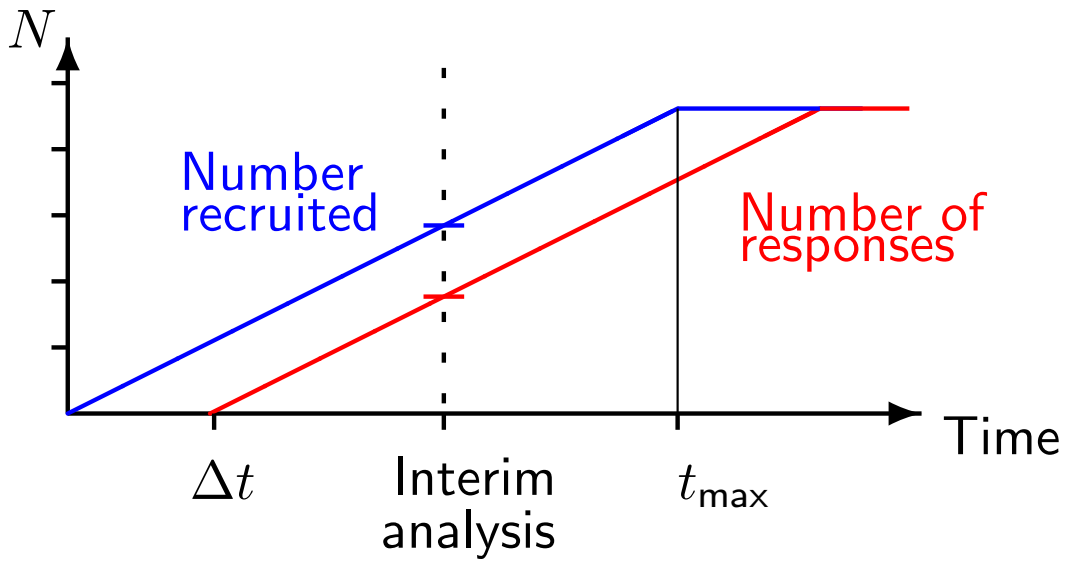
1. Stop recruitment of any more subjects,
2. After responses have been observed for all recruited subjects, make a decision to accept or reject H_0.

Defining a group sequential test with delayed responses

For now, we assume, as in Example 1:

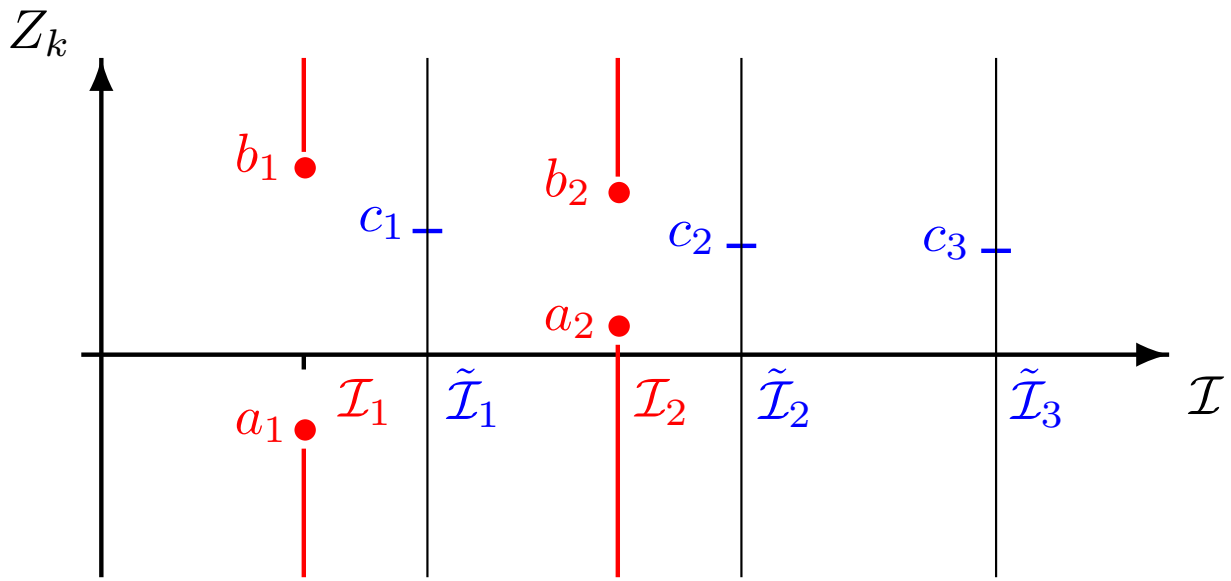
- The primary endpoint is measured a fixed time after treatment commences,
- The endpoint will be known (eventually) for all treated subjects,
- If recruitment is stopped, it cannot be re-started.

Consider a trial with responses observed time Δ_t after treatment.



Boundaries for a Delayed Response GST

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



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

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

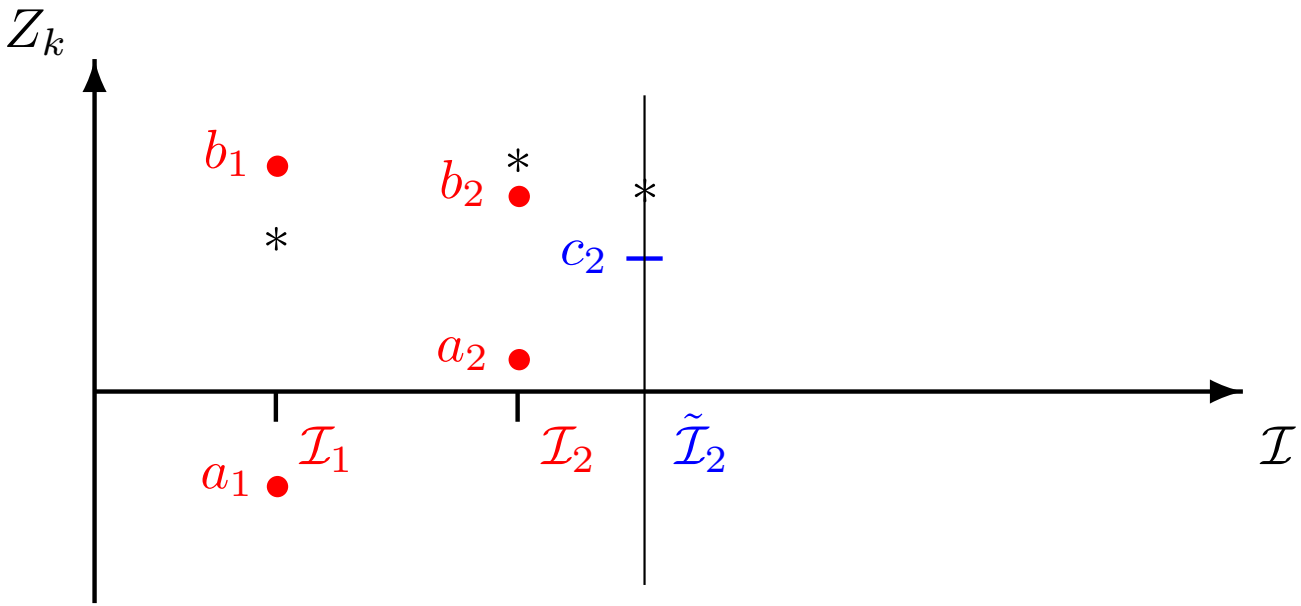
If we reach the **final** analysis K, we reject H_0 if $\tilde{Z}_K > c_K$.

Delayed Response GSTs

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

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

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



Calculations for a Delayed Response GST

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

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

and

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

Each sequence is based on accumulating data sets.

Given $\{\mathcal{I}_1, \ldots, \mathcal{I}_k, \tilde{\mathcal{I}}_k\}$, the sequence $\{Z_1, \ldots, Z_k, \tilde{Z}_k\}$ follows the "canonical joint distribution" for the sequence of Z-statistics observed in a GST with immediate response (Jennison & Turnbull, 2000, Ch. 11).

Calculations for a Delayed Response GST

Specifically, with

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

we have:

$(Z_1, \ldots, Z_K)$ 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 $k_1 < k_2$.

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

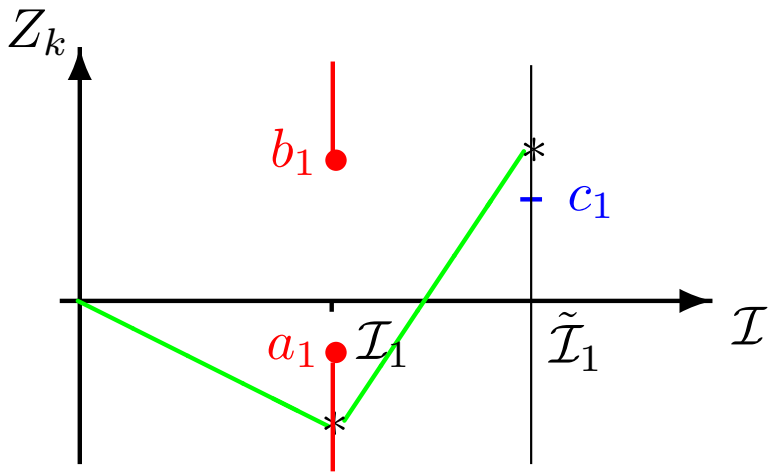
The value of information from pipeline subjects

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

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

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

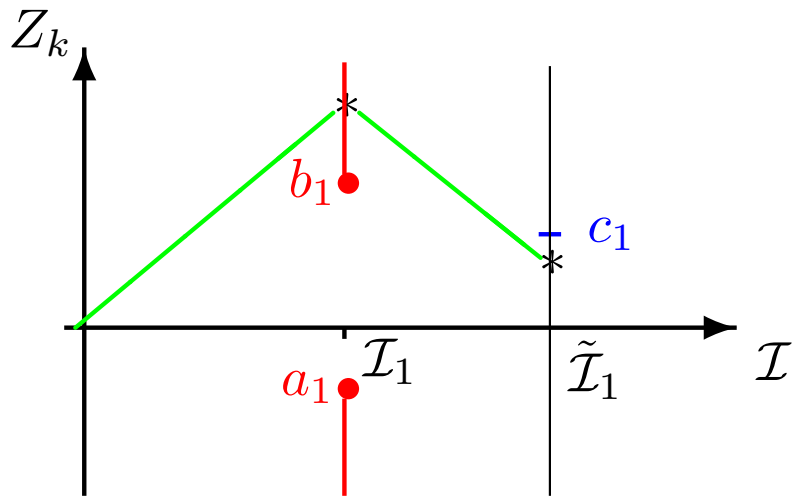
We could, for example, observe:



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

The value of information from pipeline subjects

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



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

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

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

Optimising a Delayed Response GST

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

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

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

Let N denote the total number of subjects recruited.

Objective:

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

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

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

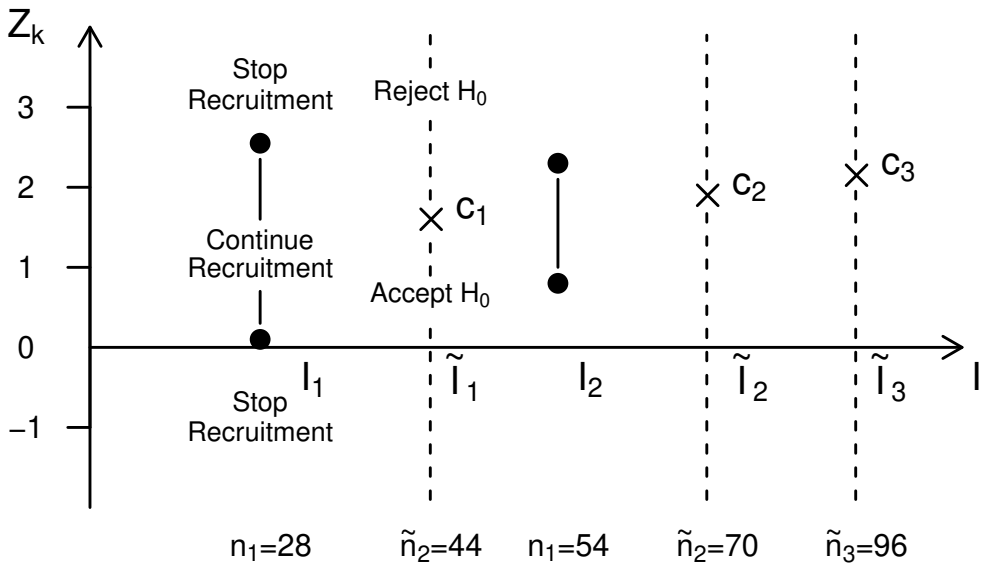
Other weighted combinations of $\mathbb{E}_\theta(N)$ can also be used.

An optimal design for Example 1: Cholesterol treatment

For normally responses with $\sigma^2 = 2$, $\alpha = 0.025$, power 0.9 required at $\theta = 1$, and given group sizes, the following design minimises

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

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



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

4. Error spending Delayed Response GSTs

HJ show how to construct Error Spending Delayed Response GSTs.

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

The test is defined through two error spending functions:

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

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

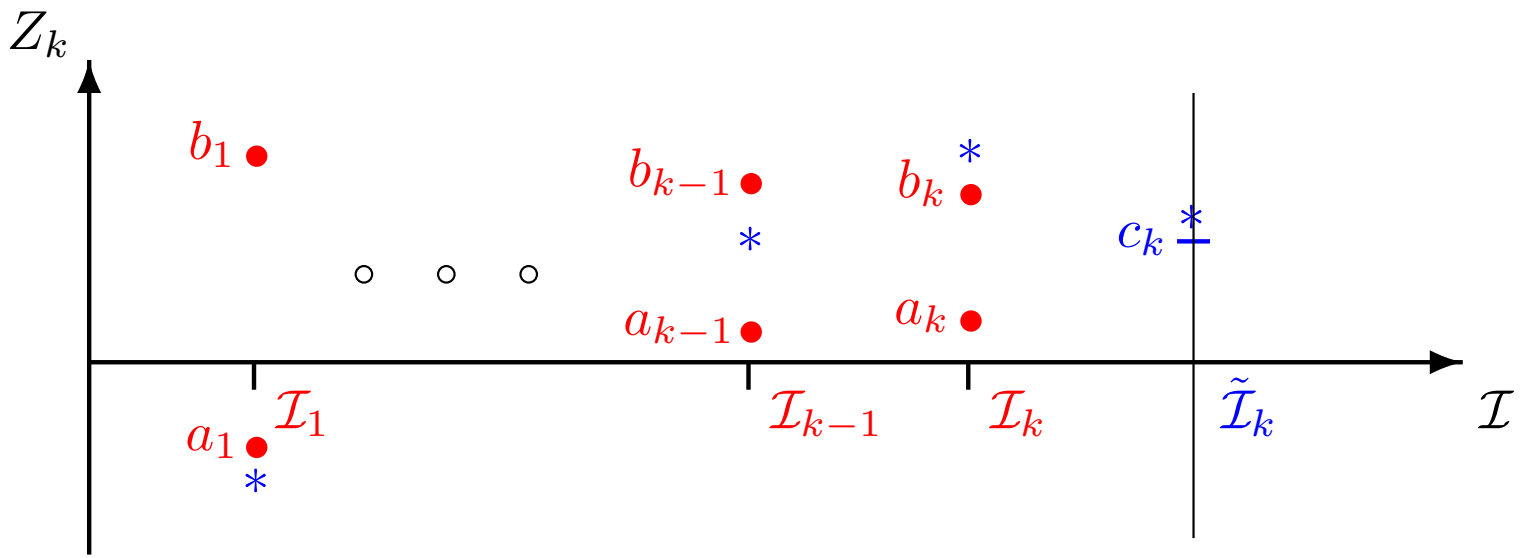
Recruitment stops when the target information $\mathcal{I}_{max}$ is reached (or will be reached with the responses from pipeline subjects).

After analysis k and its subsequent decision analysis:

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

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

Error spending Delayed Response GSTs

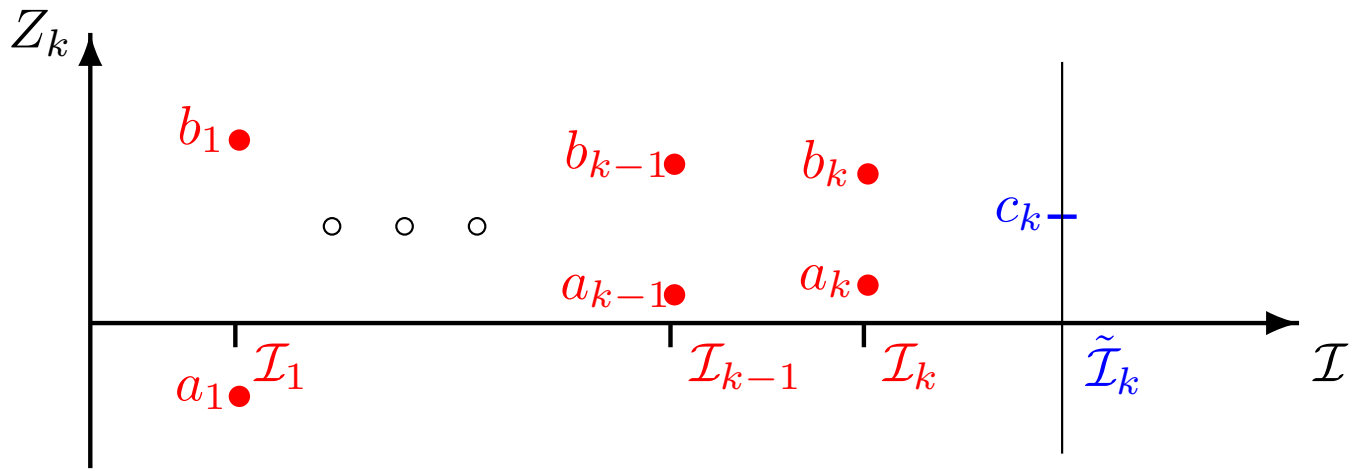


Non-binding futility boundary:

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

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

Computing an error spending Delayed Response GST



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

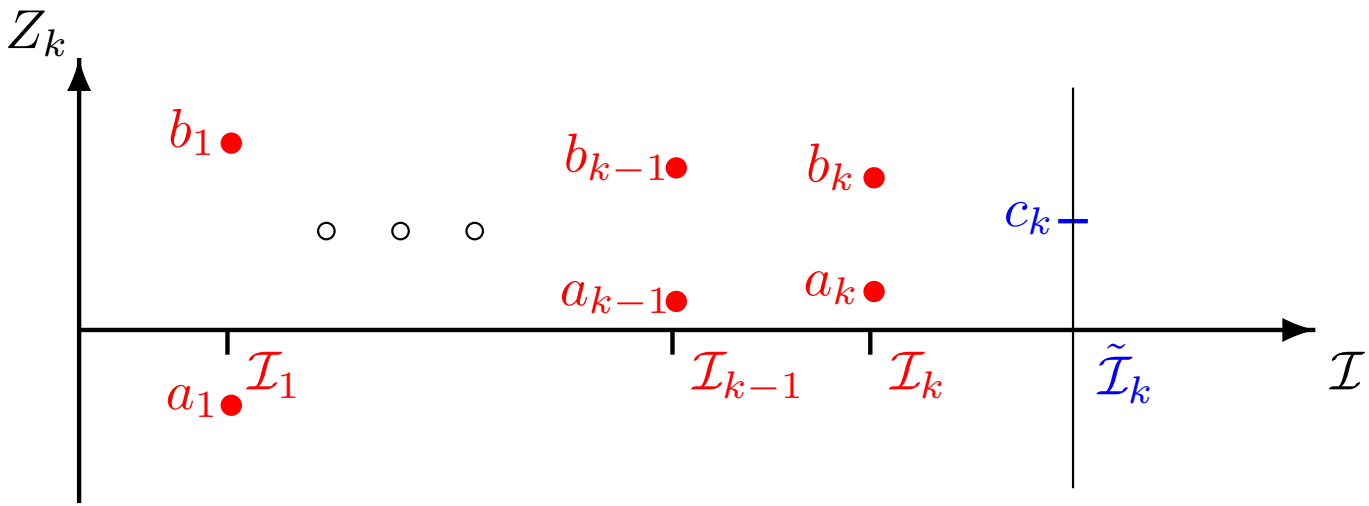
$$P_{\theta=0}\{Z_1 < b_1, \ldots, Z_{k-1} < b_{k-1}, Z_k > b_k, \tilde{Z}_k > c_k\}$$
$$= f(\mathcal{I}_k/\mathcal{I}_{\max}) - f(\mathcal{I}_{k-1}/\mathcal{I}_{\max}),$$

and

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

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

Computing an error spending Delayed Response GST



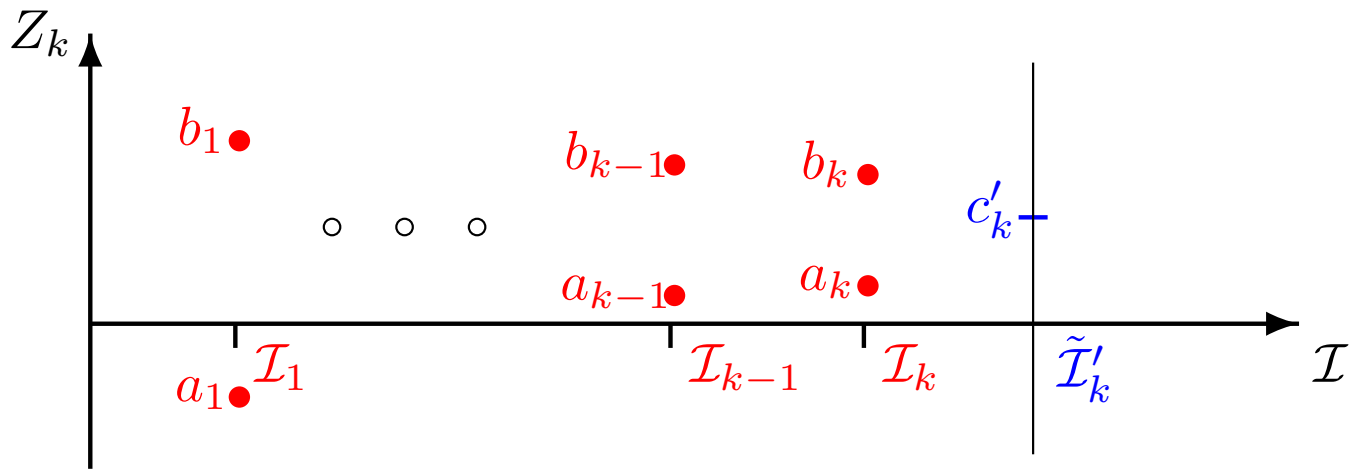
HJ noted their optimal Delayed Response GSTs with $\alpha = 0.025$ often had values of $c_1, \ldots, c_{K-1}$ less than $\Phi^{-1}(1-\alpha) = 1.96$.

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

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

Updating c_k on observing $\tilde{\mathcal{I}}_k$ — preserving type I error

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



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

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

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

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

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

The ρ-family of error spending functions

HJ considered ρ-family error spending functions of the form

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

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

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

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

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

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

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

Example 1: A ρ-family error spending GST

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

In the cholesterol trial, the anticipated sample sizes

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

lead to

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

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

In this case, the boundary values are

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

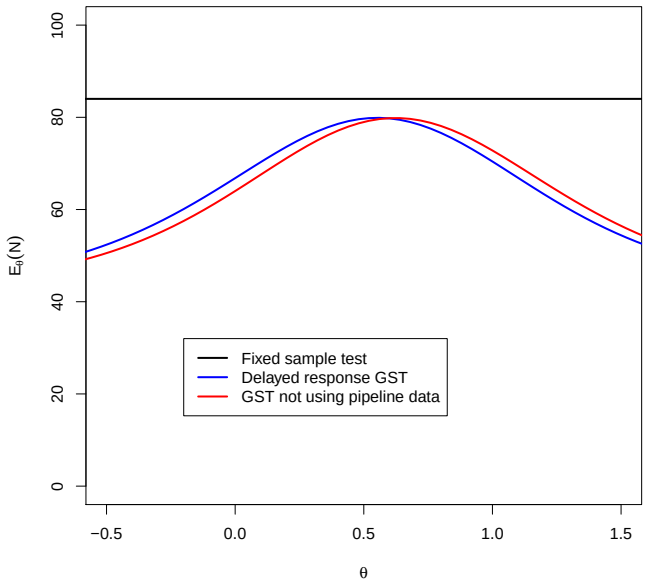
$$a_2 = 0.664, b_2 = 2.244, c_2 = 1.960;$$

$$c_3 = 2.069.$$

Example 1: A ρ-family error spending GST

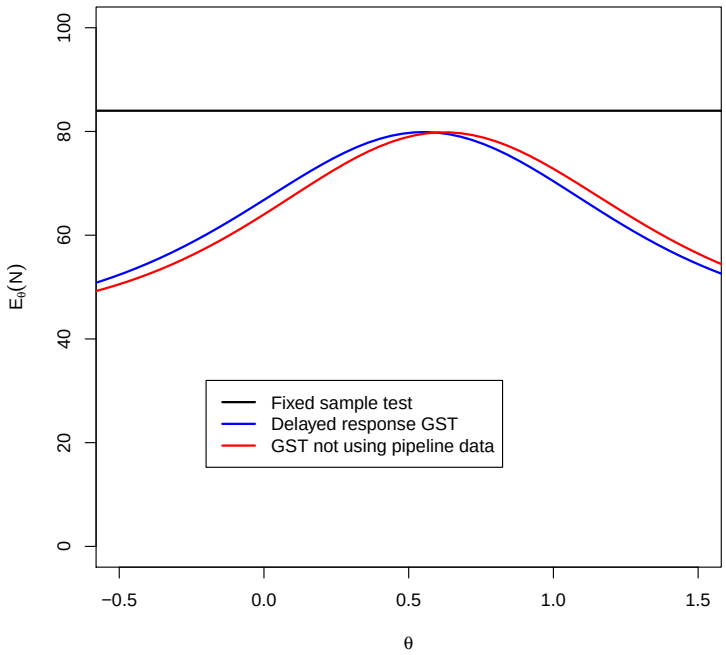
The figure shows $\mathbb{E}_\theta(N)$ for:

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



Both GSTs have non-binding futility boundaries.

Example 1: A ρ-family error spending GST



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

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

In contrast, if we apply Whitehead's deletion method, starting from the ρ-family error spending GST for immediate response, power at $\theta = 1$ falls from 0.9 to 0.872.

A 10% increase in overall sample size would be needed to recover this loss of power.

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

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

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

If events require adjudication, a further increase may follow.

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

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

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

- If $\theta = 0$, using these data may help avoid a type I error;
- If $\theta = \delta$, pipeline data are unlikely to "reverse a positive result".

Detailed calculations for Example 1 show this is possible!

6. Conclusions

HJ (2013) showed how GSTs for a delayed response can be optimised for criteria involving both the number of subjects recruited and the time to a final decision.

These optimised designs provide a template for error spending GSTs that can accommodate over-run data.

Our error spending GSTs reduce the possibility (compared to Whitehead's method) that over-run data will produce a "reversal" where a high Z_k causes recruitment to stop but the final $\tilde{Z}_k$ is below the boundary — even though there is still quite strong evidence against H_0.

In "reversals" where the final test statistic $\tilde{Z}_k$ falls below 1.96, we have the consolation that a simple fixed sample test based on the same data would not have produced a significant result.

7. Additional slides: Optimising a Delayed Response GST

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

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

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

Let N denote the total number of subjects recruited.

Objective:

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

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

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

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

Computing optimal Delayed Response GSTs

HJ (*JRSS B*, 2013) explain how to derive an optimal delayed response GST.

They create a Bayes sequential decision problem, placing a prior on θ and defining costs for sample size and the time taken to reach a decision, plus a penalty for incorrect decisions.

They solve this problem by dynamic programming.

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

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

An optimal design for the cholesterol treatment example

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

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

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

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

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

Type I error rate $\alpha = 0.025$,

Power 0.9 at $\theta = 1$.

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

An optimal design for the cholesterol treatment example

Consider designs with a maximum sample size of 96.

Assuming a recruitment rate of 4 per week:

Data start to accrue after 4 weeks,

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

Recruitment will close after 24 weeks.

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

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

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

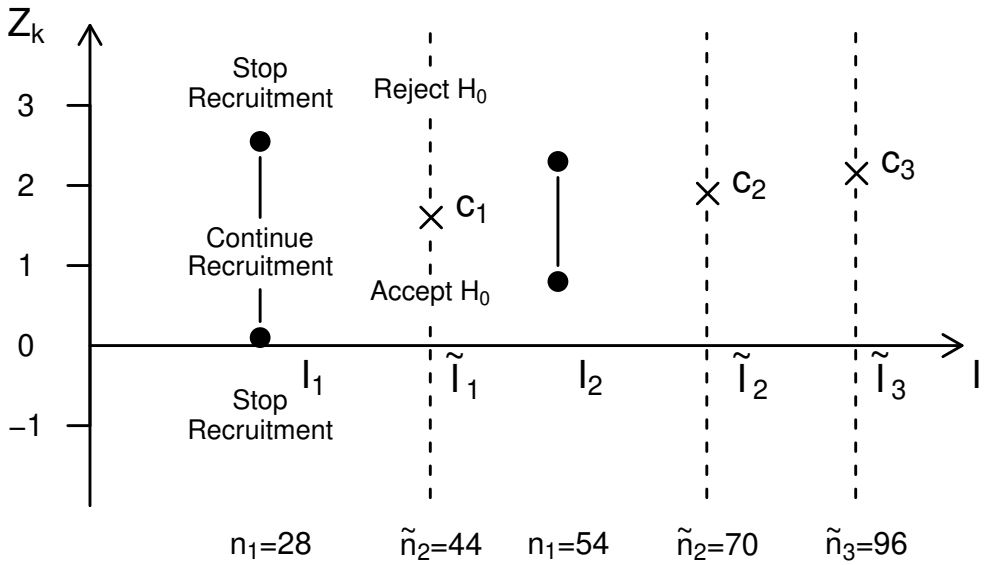
$\tilde{n}_3 = 96$ responses if there is no early stopping.

An optimal design for the cholesterol treatment example

The following Delayed Response GST minimises

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

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



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