Group Sequential Tests for Delayed Responses

Lisa Hampson
Department of Mathematics and Statistics,
Lancaster University, UK

Chris Jennison
Department of Mathematical Sciences,
University of Bath, UK

Read to the Royal Statistical Society
London, May 2012

Outline

Outline

Outline

Outline

Outline

Outline

Group sequential monitoring of clinical trials

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

Let the treatment effect, θ, be the difference in average response between the new treatment and control.

We can design a superiority trial to test

$$H_0\text{: } \theta \leq 0 \text{ against } \theta > 0$$

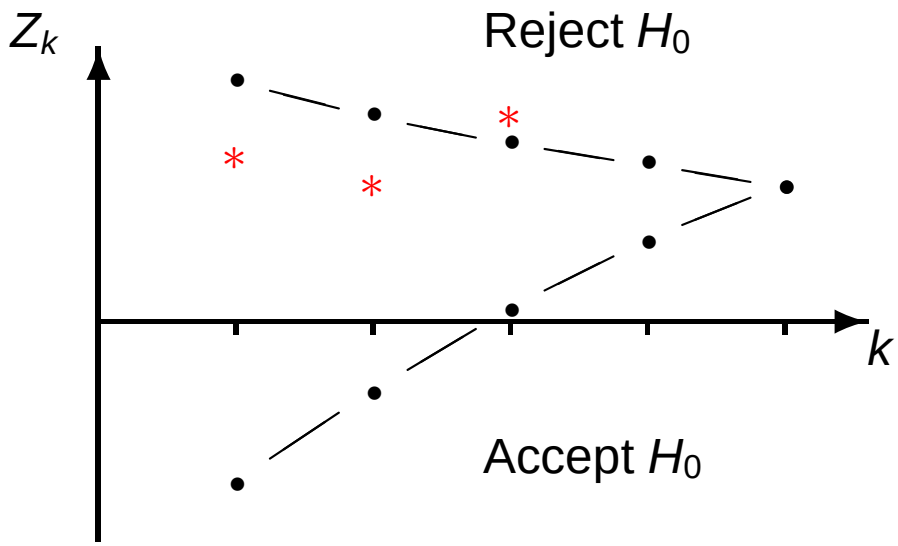
with one-sided type I error rate α and power $1 - \beta$ at $\theta = \delta$.

In a group sequential design, we monitor standardised test statistics Z_k at analyses $k = 1, 2, \ldots$.

The stopping rule allows an early decision to reject H_0 or accept H_0.

A group sequential test (GST)

A group sequential boundary for testing H_0: $\theta \leq 0$ vs $\theta > 0$



Here, the trial stops and rejects H_0 at the third of five analyses.

Sequential testing can reduce expected sample size to around 60% or 70% of that of a fixed sample size design.

The problem of delayed responses

Example A: Cholesterol reduction after 4 weeks of treatment

In this example, there is a delay of four weeks between the start of treatment and observation of the primary endpoint.

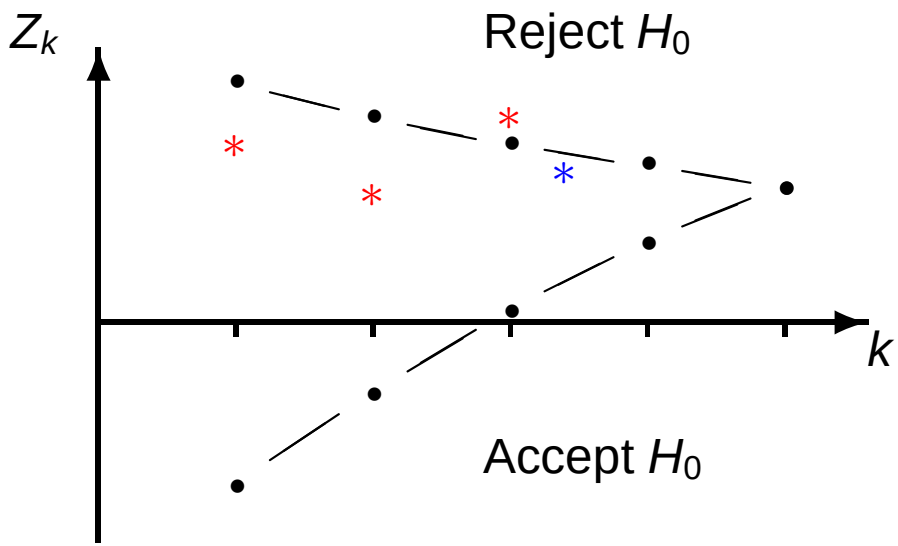
At each interim analysis we expect about 16 subjects to be “in the pipeline”, that is, to have started treatment but not yet provided a response.

If a group sequential test reaches its conclusion at an interim analysis, one would still expect investigators to follow up the pipeline subjects and observe their responses.

How should these data be analysed?

The problem of delayed responses

A possible outcome for the cholesterol reduction trial of Example A



Suppose $Z_3 = 2.4$, exceeding the boundary value of 2.3.

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

Can the investigators claim significance at level α?

Using short term endpoints in Delayed Response GSTs

We shall show how to design a Delayed Response Group Sequential Test which makes the best possible use of “pipeline” data.

Nevertheless, we cannot achieve all of the reductions in expected sample size that are possible for an immediate response.

We therefore seek ways to recover some of this lost efficiency.

One way to do this is by fitting a joint model for the primary endpoint and a correlated response which is observed more rapidly.

Using short term endpoints in Delayed Response GSTs

Example D: Prevention of fracture in postmenopausal women

In this example, the primary endpoint is whether or not a fracture occurs within five years of entry to the study.

Changes in bone mineral density (BMD) are measured after one year.

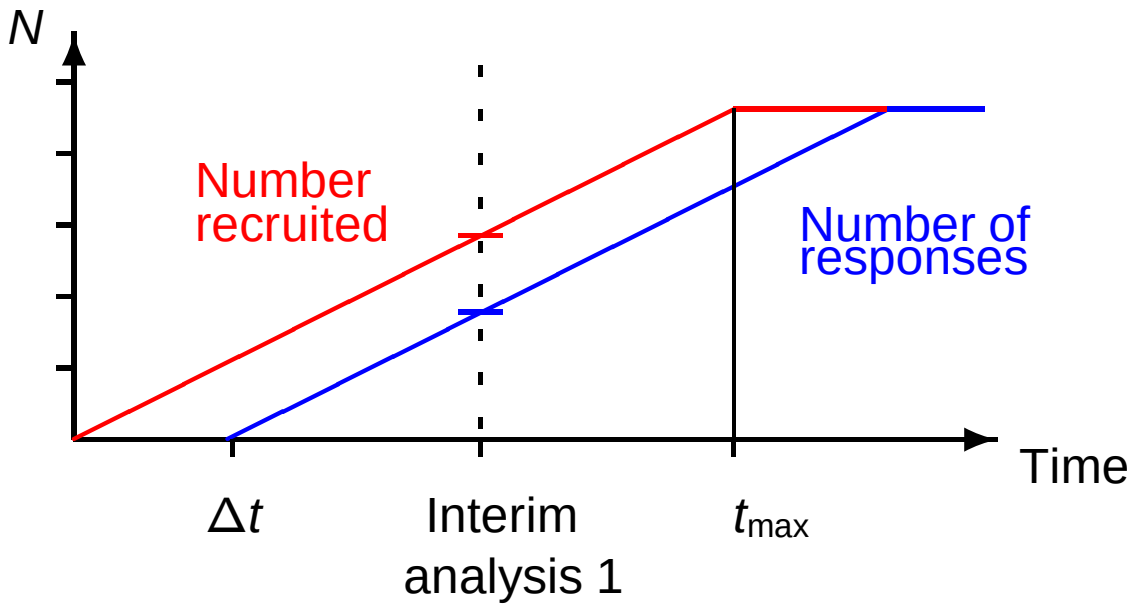
It is expected that these two variables are correlated.

How can we use the BMD data to gain information from subjects who have been followed for between one and five years?

Would fitting a Kaplan-Meier curve for time to first fracture also help — remembering that inference is about the binary outcome defined at five years?

Incorporating delayed responses into GSTs

Consider a trial where response is observed time Δ_t after treatment.



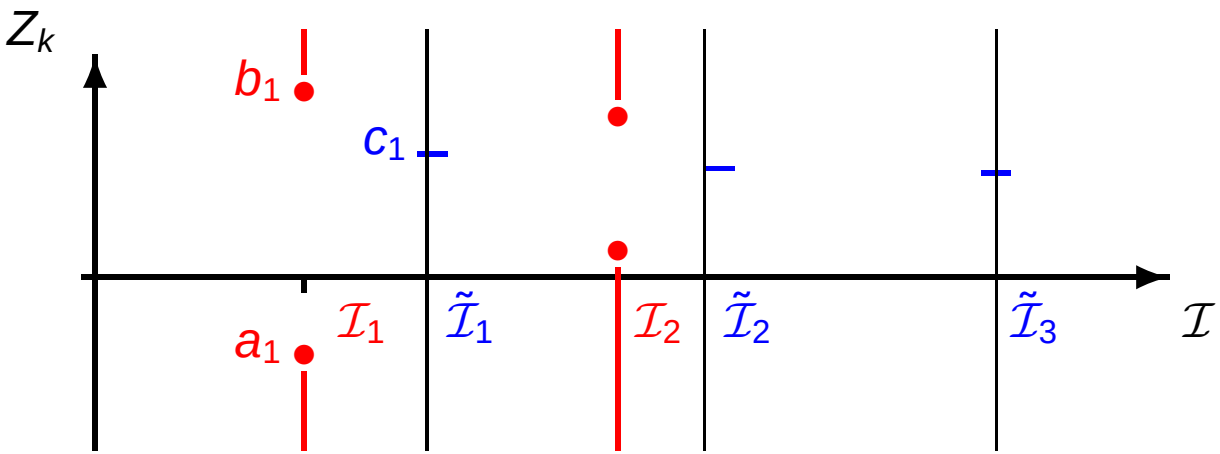
We assume information is proportional to the observed number of responses.

We will equally space interim analyses between times Δ_t and t_{max}.

T.W. Anderson (*JASA*, 1964) considers sequential tests for delayed responses. We follow this basic structure to construct GSTs.

Boundaries for a Delayed Response GST

At interim analysis k, Z_k is associated with information level $\mathcal{I}_k = \text{Var}(\hat{\theta}_k)$.



If $Z_k > b_k$ or $Z_k < a_k$, cease enrollment of future patients and follow-up all recruited subjects.

At the decision analysis, based on information $\tilde{\mathcal{I}}_k$, reject H_0 if $\tilde{Z}_k > c_k$.

Calculating properties of Delayed Response GSTs

Calculations of test properties (type I error rate, power, $\mathbb{E}_\theta(N)$) require the joint distributions of test statistic sequences:

- $\{Z_1, \ldots, Z_k, \tilde{Z}_k\}$, for $k = 1, \ldots, K-1$,
- $\{Z_1, \ldots, Z_{K-1}, \tilde{Z}_K\}$.

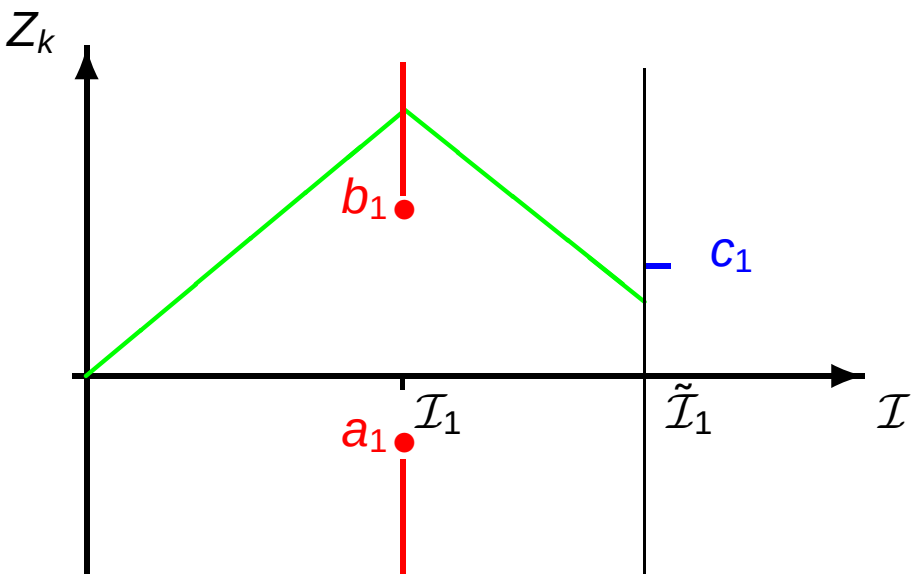
Each sequence is based on accumulating datasets.

Given $\{\mathcal{I}_1, \ldots, \mathcal{I}_k, \tilde{\mathcal{I}}_k\}$, the sequence $\{Z_1, \ldots, Z_k, \tilde{Z}_k\}$ follows the canonical distribution for statistics generated by a GST for immediate responses (Jennison & Turnbull, *JASA*, 1997).

Properties of Delayed Response GSTs can therefore be calculated using numerical routines devised for standard designs.

Reversals of anticipated final decisions

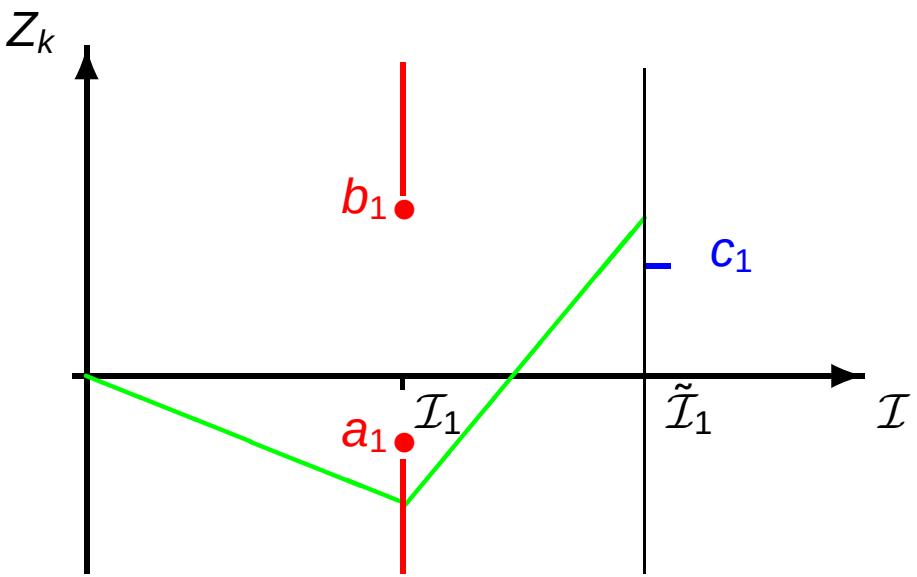
Stopping with $Z_k > b_k$ or $Z_k < a_k$ indicates our *likely* final decision but there there may be a reversal. We could observe



We optimise our designs to maximise the value of the additional pipeline responses for increasing the test's power.

Reversals of anticipated final decisions

Stopping with $Z_k > b_k$ or $Z_k < a_k$ indicates our *likely* final decision but there there may be a reversal. We could observe



We optimise our designs to maximise the value of the additional pipeline responses for increasing the test's power.

Optimal Delayed Response GSTs

Let N represent the total number of subjects recruited.

Let r be the fraction of a test's maximum sample size in the pipeline at each interim analysis.

Objective: For a given r, maximum sample size n_{max}, stages K and analysis schedule, we find the Delayed Response GST minimising

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

with type I error rate α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$. Here $f(\theta)$ is the density of a $N(\delta/2, (\delta/2)^2)$ distribution.

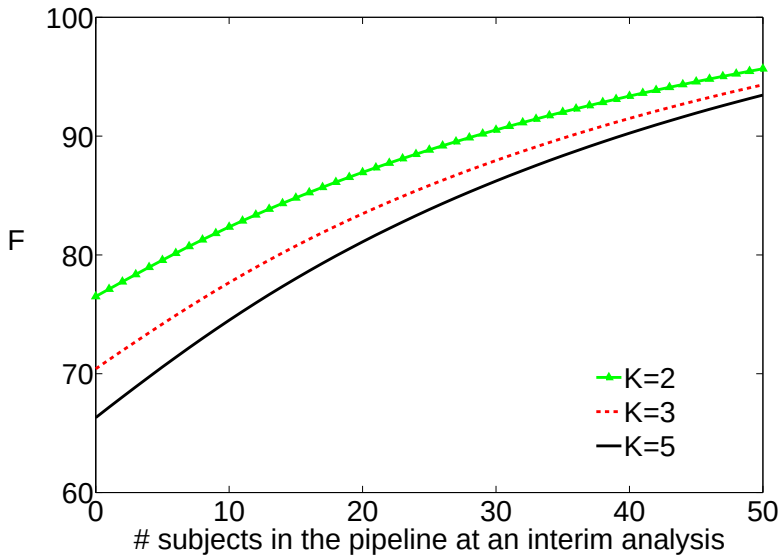
We create an unconstrained Bayes problem by adding a prior on θ and costs for sampling and for making incorrect decisions. We search for the combination of prior and costs which gives a solution with frequentist error rates α and β.

Efficiency loss when there is a delay in response

It is required to test $H_0 : \theta \leq 0$ against $\theta > 0$ with $\alpha = 0.025$ and $\beta = 0.1$.

Suppose (δ, σ^2) are such that the fixed sample test needs $n_{fix} = 100$ subjects and set $n_{max} = 1.1\, n_{fix}$.

The figure shows the minima of F attained by optimised versions of our designs.



Revisiting Example A

Example A: Cholesterol reduction after 4 weeks of treatment

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

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

- type I error rate $\alpha = 0.025$ at $\theta = 0$,
- power $1 - \beta = 0.9$ at $\theta = \delta = 1.0$.

The fixed sample test needs $n_{fix} = 86$ subjects divided between the two treatments.

We consider designs with a maximum sample size of 96, assuming a recruitment rate of 4 per week, giving $4 \times 4 = 16$ pipeline subjects at each interim analysis.

Revisiting Example A

Once the trial is underway, data start to accrue after 4 weeks. Recruitment will close after 24 weeks.

Interim analyses are planned after $n_1 = 28$ and $n_2 = 54$ observed responses.

A decision analysis will be 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 in the absence of early stopping.

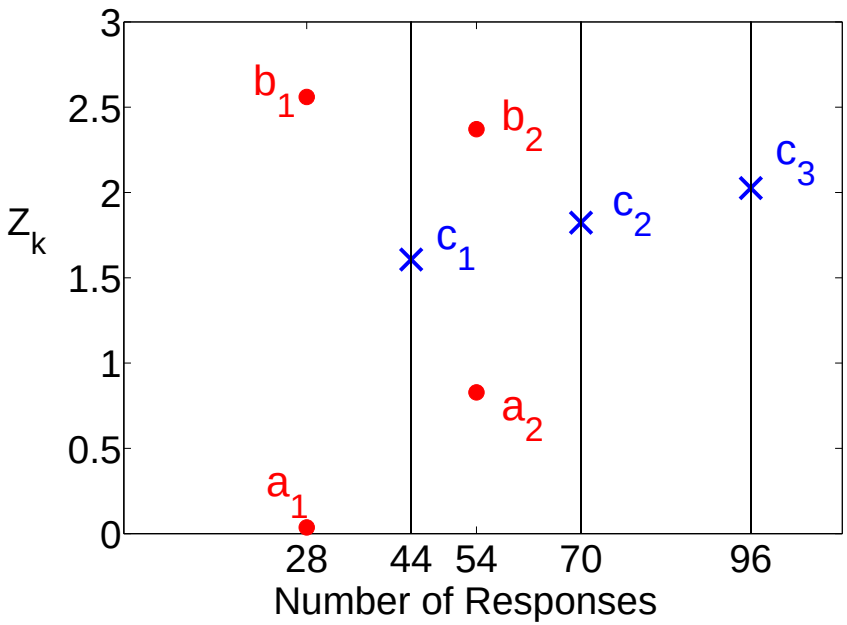
We derive a Delayed Response GST minimising

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

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

Revisiting Example A

Critical values for the optimised Delayed Response GST are shown below.



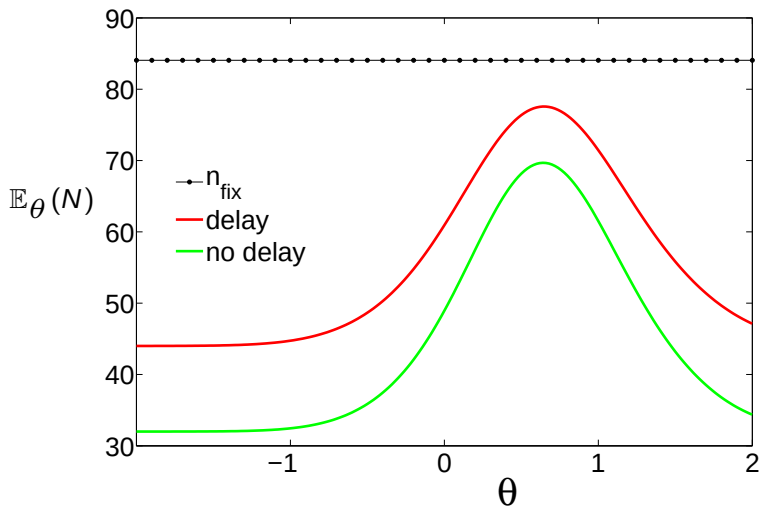
Critical values c_1 and c_2 are well below b_1 and b_2, so the probability of a reversal is small.

Both c_1 and c_2 are less than 1.96. If desired, these can be raised to 1.96 with little change to the design's power curve.

Revisiting Example A

The figure shows expected sample size curves for

- the fixed sample test with $n_{fix} = 85$ patients,
- the Delayed Response GST minimising F,
- the GST for immediate responses with analyses after 32, 64 and 96 responses, also minimising F.

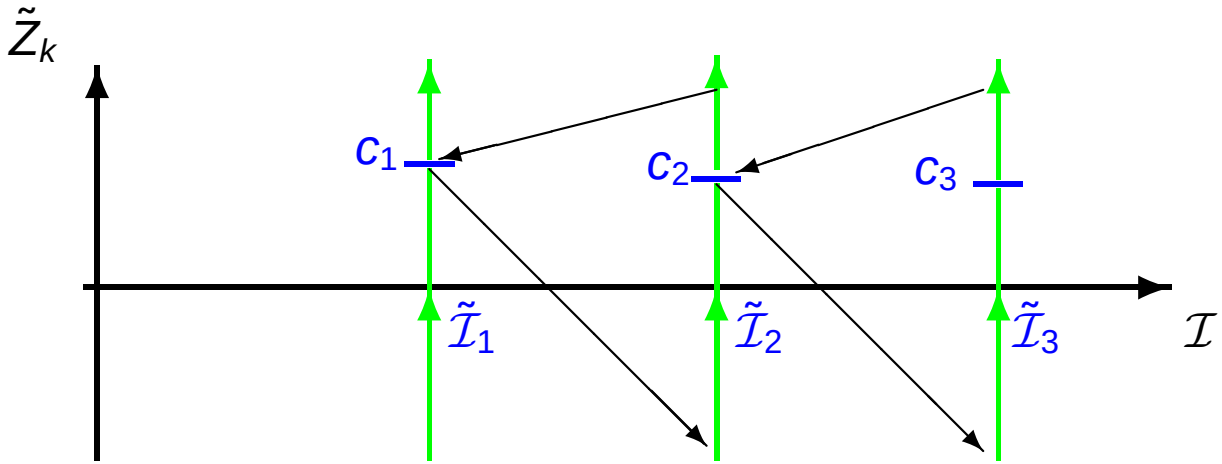


The delay in response means savings in $\mathbb{E}_\theta(N)$ are smaller than they would be if response were immediate.

Making inferences on termination

How can we calculate a p-value for $H_0 : \theta \leq 0$ and a CI for θ?

On termination of the test at stage T, $(\tilde{\mathcal{I}}_T, \tilde{Z}_T)$ is a sufficient statistic for θ. We base inferences on a "stage-wise" ordering of the test's sample space for this pair.



The sample space at $\tilde{\mathcal{I}}_T = \tilde{\mathcal{I}}_k$ is partitioned by c_k into "high" and "low" sets.

This ordering ensures p-value calculations do not depend on future, possibly *unpredictable*, information levels.

Error spending Delayed Response GSTs

We design error spending Delayed Response GSTs which

- reach a target information level $\tilde{\mathcal{I}}_{max}$ in absence of early stopping,
- spend error probabilities as a function of $\mathcal{I}/\mathcal{I}_{max}$.

Let π_k and γ_k be cumulative type I and II error rates to be spent by stage k.

Choosing c_k to balance reversal probabilities under $\theta = 0$ implies we may choose (a_k, b_k) to satisfy

$$\mathbb{P}_{\theta=0}\{Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1}, Z_k \geq b_k\} = \pi_k - \pi_{k-1}$$
$$\mathbb{P}_{\theta=\delta}\{Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1}, Z_k \leq a_k\} = \gamma_k - \gamma_{k-1},$$

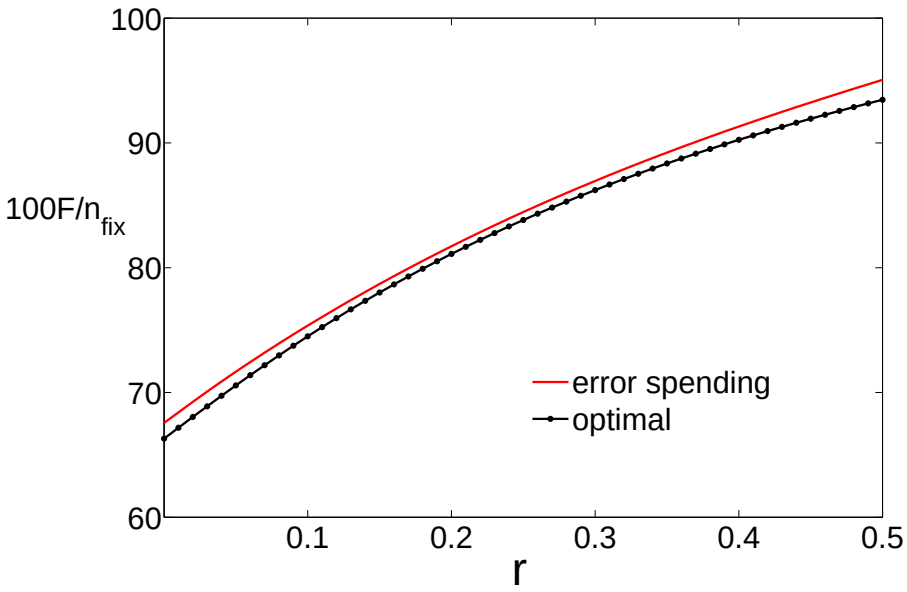
and control the type I error rate at level α, and the type II error rate at a level just below β.

Under this construction, the stage k stopping rule can be set without knowledge of $\tilde{\mathcal{I}}_k$.

Efficiency of error spending tests

In the figure below, error spending tests are designed using the ρ-family of error spending functions.

Values of F are attained by tests designed and conducted with $K = 5$, $n_{max} = 1.1\, n_{fix}$, $\alpha = 0.025$ and $\beta = 0.1$.



Error spending Delayed Response GSTs are flexible and closely match the optimal tests for savings in $\mathbb{E}_\theta(N)$.

Dealing with unexpected overrunning

Suppose a standard GST designed with $\mathcal{I}_k$ and boundaries (a_k, b_k) stops at analysis $k^\star < K$ with $Z_k > b_k$ or $Z_k < a_k$.

Question: If additional data are observed, how can these be incorporated into the final analysis while preserving the type I error rate?

Solution: We partition the sample space at $\tilde{\mathcal{I}}_{k^\star}$ such that

- if $\tilde{Z}_{k^\star} \geq c_{k^\star}$, reject H_0,
- if $\tilde{Z}_{k^\star} \leq c_{k^\star}$, accept H_0.

Requiring $c_{k^\star}$ to balance the probabilities of reversing decisions under $\theta = 0$ at stage $k^\star$ preserves the test's overall type I error rate.

In addition, p-value calculations do not depend on $\tilde{\mathcal{I}}_1, \ldots, \tilde{\mathcal{I}}_{k^\star - 1}$, nor on information levels beyond stage $k^\star$.

Efficiency loss when there is a delay in response

In Section 3 of the paper, we consider tests with type I error rate 0.025 and power 0.9 at $\theta = \delta$, minimising F, the integral of $\mathbb{E}_\theta(N)$ with respect to a $N(\delta/2, (\delta/2)^2)$ distribution for θ.

Maximum sample size, $n_{\max}$, is 1.1 times the fixed sample size.

There are $r\, n_{\max}$ subjects "in the pipeline" at each interim analysis.

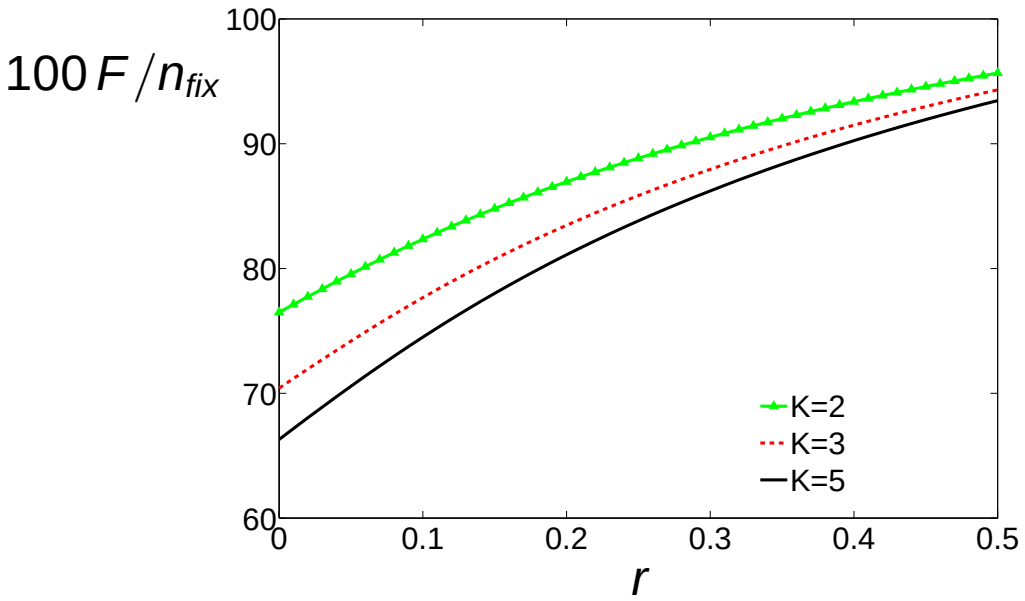
We find optimal Delayed Response GSTs with $K = 2$, 3 and 5 analyses when the number of observed responses at analysis k is

$$n_k = \frac{k}{K}(1 - r)n_{\max}, \quad k = 1, \ldots, K - 1.$$

The effect of delayed response on the efficiency of a Delayed Response GST increases with the "pipeline fraction" r.

Efficiency loss when there is a delay in response

Minima of average $\mathbb{E}_\theta(N)$, F, as a function of the pipeline fraction r



Substantial savings in $\mathbb{E}_\theta(N)$ are still present for small values of r, when response is rapidly observed.

About half the benefits of group sequential testing are lost as r increases to 0.25.

Using a short term endpoint to recover efficiency

Suppose a second endpoint, correlated with the primary endpoint, is available soon after treatment.

For patient i on treatment $T = A$ or B, let

$$Y_{T,i} = \textit{The short term endpoint},$$

$$X_{T,i} = \textit{The long term endpoint}.$$

Assume we have a parametric model for the joint distribution of $(Y_{T,i}, X_{T,i})$ in which

$$\mathbb{E}(X_{A,i}) = \mu_{A,2}, \quad \mathbb{E}(X_{B,i}) = \mu_{B,2} \quad \text{and} \quad \theta = \mu_{A,2} - \mu_{B,2}.$$

We analyse all the available data at each interim analysis.

Using a short term endpoint to recover efficiency

At interim analysis k, subjects are

- *Unobserved*,
- *Partially observed (just $Y_{T,i}$ available), or*
- *Fully observed (both $Y_{T,i}$ and $X_{T,i}$ available).*

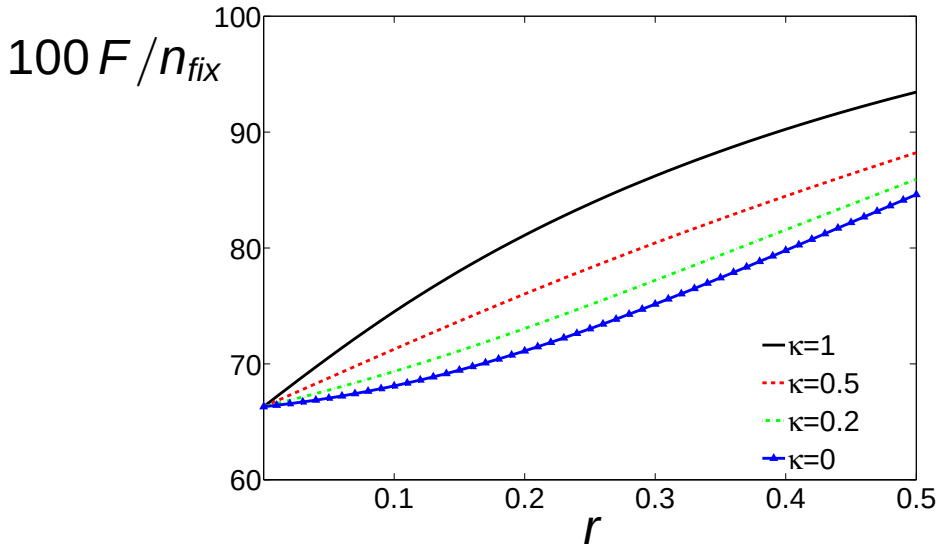
We use maximum likelihood estimation to fit the full model to all the data available at analysis k, then extract $\widehat{\theta}_k$ and $\mathcal{I}_k = \{\text{Var}(\widehat{\theta}_k)\}^{-1}$.

The sequence of estimates $\{\widehat{\theta}_k\}$ follows the standard joint distribution for a group sequential trial with observed information levels $\{\mathcal{I}_k\}$.

Thus, these estimates can be used to design a Delayed Response GST in the usual way.

Using a short term endpoint to recover efficiency

Values of F achieved using a second, short-term endpoint



Results are for the previous testing problem with $K = 5$ analyses.

We assume $Y_{T,i}$ and $X_{T,i}$ are bivariate normal with correlation 0.9.

The ratio of time to short-term and long-term endpoints is κ.

The solid line for $\kappa = 1$ is also the case of no short-term endpoint.

Using a short term endpoint to recover efficiency

Although a short-term endpoint may be of clinical interest, we still make inferences about the primary endpoint alone.

It is straightforward to extend this approach to use repeated measurements as follow-up continues for each patient.

In Example D, we can fit a joint model for bone mineral density measured at one year and incidence of fracture within five years — thereby increasing information about the latter.

In an extended model, we could also use censored information on the fracture endpoint for subjects with less than five years of follow-up.

Nuisance parameters, such as variances and correlation between short-term and long-term endpoints, can be estimated within the trial.

Summary

In this presentation, we have discussed

- Formulation of a Delayed Response GST
- Optimisation of a Delayed Response GST
- P-values and confidence intervals on termination
- Error spending versions of these tests
- Unexpected overrunning
- Using a short-term endpoint to improve efficiency

Additional topics covered in the paper

The paper also addresses

- Existence and uniqueness of optimal Delayed Response GSTs
- Computation of optimal Delayed Response GSTs
- Optimising designs for an objective combining expected sample size and time to a conclusion
- Adaptive choice of group sizes in a Delayed Response GST