Group Sequential Tests for Delayed Responses

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

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

Lisa Hampson

Department of Mathematics and Statistics,

University of Lancaster, UK

BfArM/DIA Statistics Workshop

Bonn, October 2012

Outline of talk

1. Group sequential tests (GSTs)
2. Delayed responses
3. Group sequential designs for delayed responses
4. Optimal delayed response GSTs
5. Using short term endpoints to recover efficiency
6. Error spending designs
7. Further topics:

   Optimising for a variety of criteria

   Non-binding futility boundaries

   Unexpected over-running

   Inference on termination

   Adaptive choice of group sizes

1. Group sequential monitoring of clinical trials

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

Let the treatment effect θ represent the improvement in average response of the new treatment over the control.

We can design a superiority trial to test

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

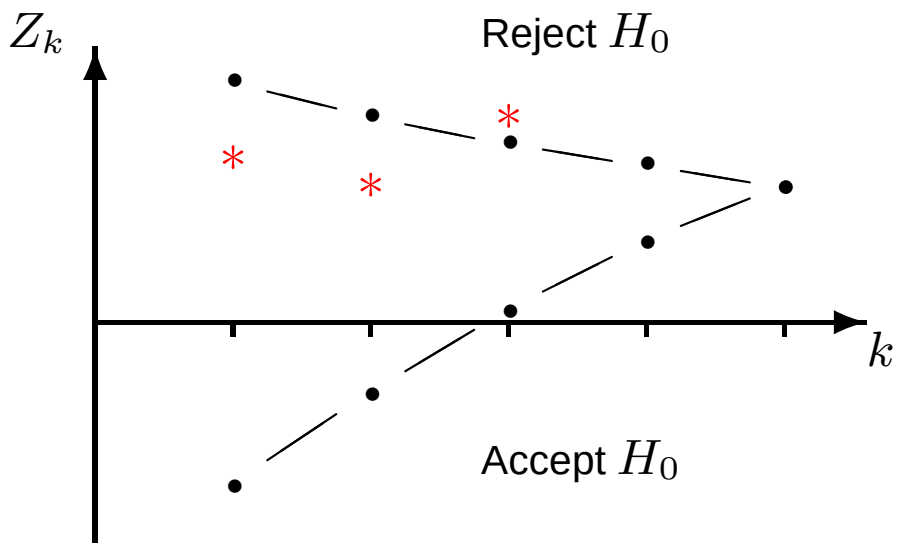
with one-sided type I error rate α and power $1 - \beta$ at $\theta = \delta$.

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 to 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 to reject 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.

2. The problem of delayed responses

Reference: Hampson & Jennison (HJ), (*JRSS B*, 2012)

Example: Cholesterol reduction after 4 weeks of treatment

In their Example A, HJ describe a trial where there is a delay of four weeks between the start of treatment and observation of the primary endpoint.

The recruitment rate is around 4 patients per week, so at each interim analysis we expect about 16 subjects to have started treatment but not yet given a response.

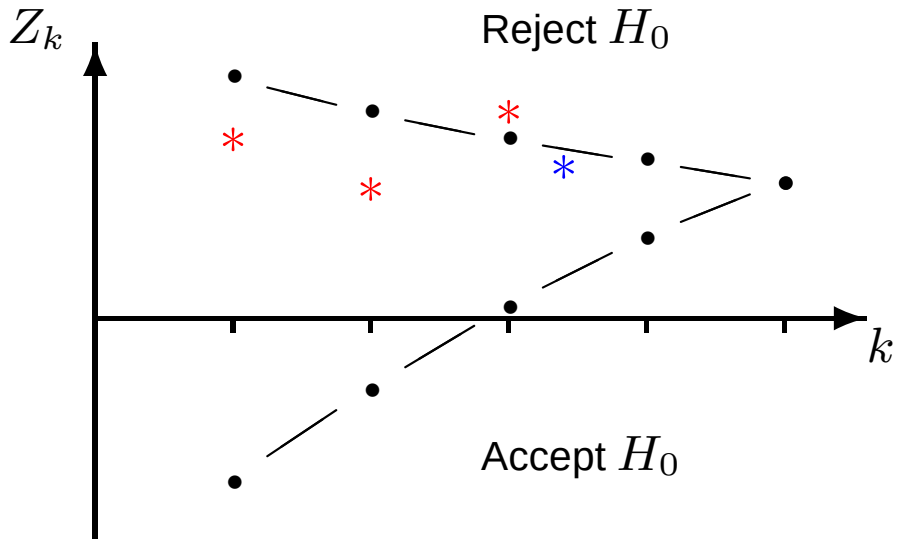
We refer to these as patients as being “in the pipeline”.

If a group sequential test reaches its conclusion at an interim analysis, we still expect investigators to follow up 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



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?

Short term information on“pipeline” subjects

Example: Prevention of fracture in postmenopausal women

In their Example D, HJ consider a study where the primary endpoint is occurrence of a fracture within five years.

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

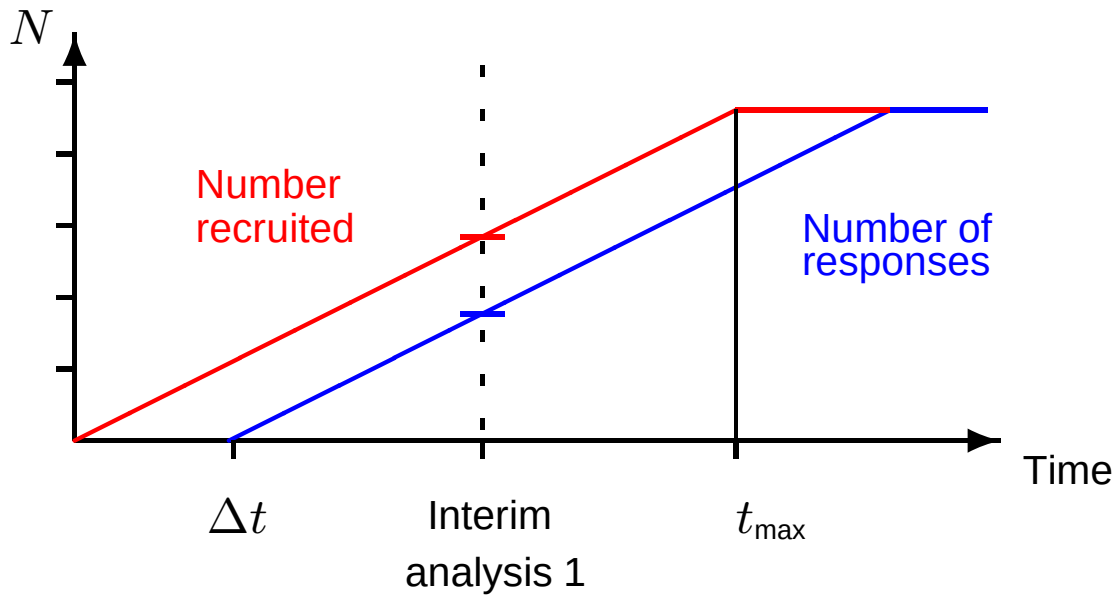
It is expected that these two variables are correlated.

How might 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 — remember that inference is about the binary outcome defined at five years?

3. Defining a group sequential test with delayed responses

Consider a trial where responses are observed time Δ_t after treatment.



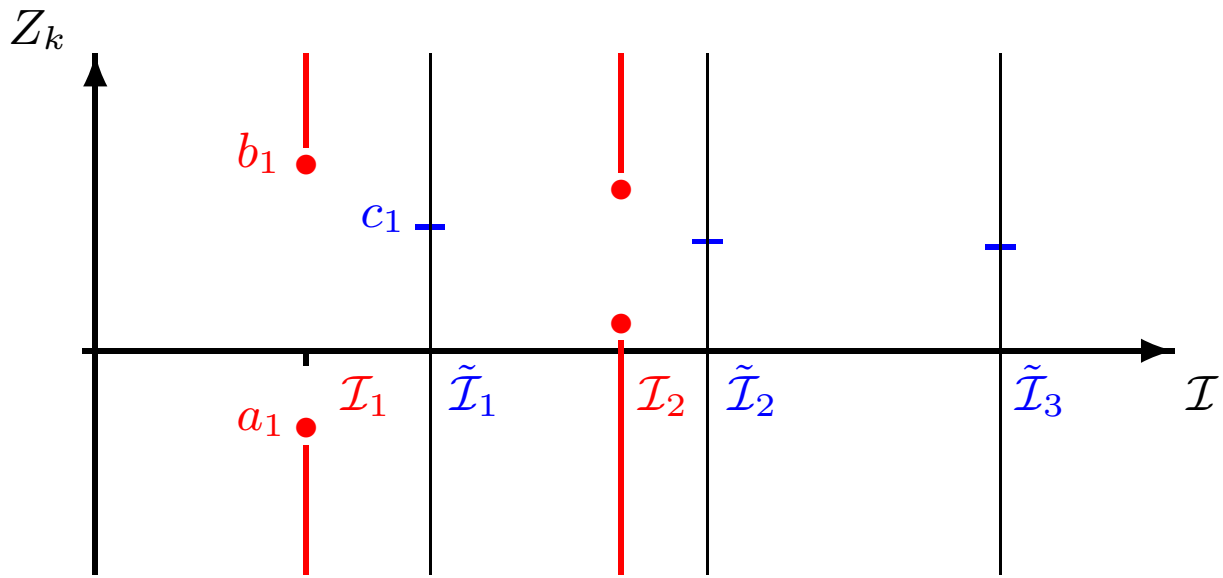
At each analysis, patients arriving in the last Δ_t units of time are “in the pipeline”.

T.W. Anderson (*JASA*, 1964) proposed a way to accommodate delayed responses in a Sequential Probability Ratio Test.

We follow his basic structure to construct our GSTs for delayed responses.

Boundaries for a Delayed Response GST

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



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

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

Calculations for a Delayed Response GST

The type I error rate, power and expected sample size of a Delayed Response GST depend on 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 same canonical distribution as the sequence of Z-statistics in a GST with immediate responses (Jennison & Turnbull, *JASA*, 1997).

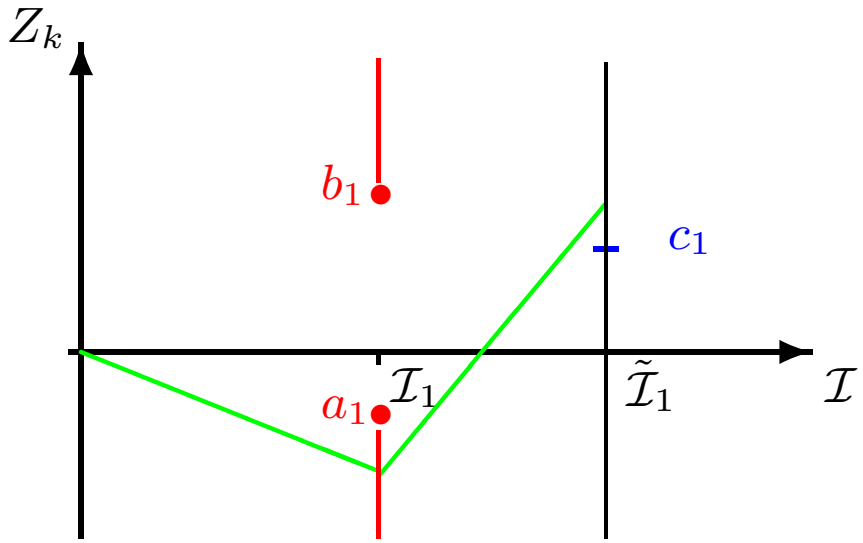
Thus, properties of Delayed Response GSTs can be calculated using numerical routines devised for standard group sequential designs.

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.

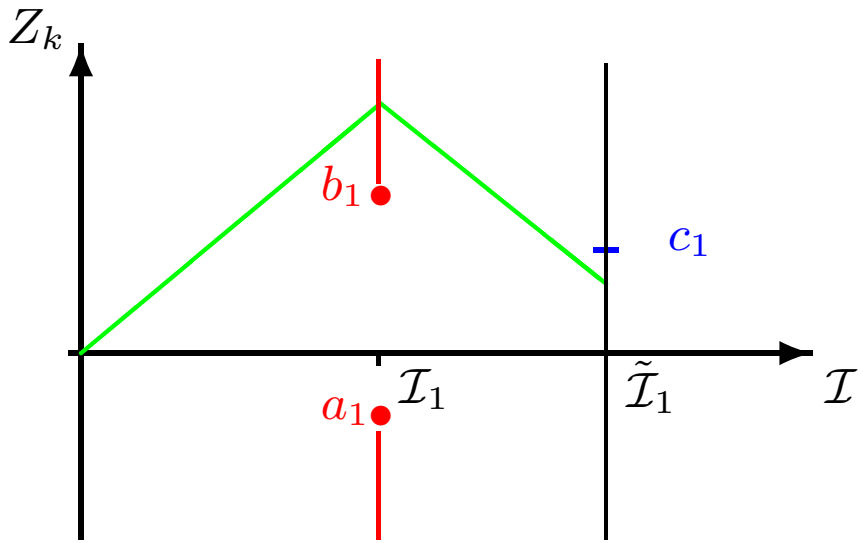
However, the pipeline data provide further information to be used in this decision.

We could observe:



The value of information from pipeline subjects

Or, we might see:



We can optimise the placement of boundary points in a Delayed Response GST design to achieve high power with low expected sample size.

These optimised designs will occasionally produce a “reversal”, with the final decision differing from that anticipated when recruitment was terminated.

4. Optimising a Delayed Response GST

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

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

Let r be the fraction of n_{max} in the pipeline at each interim analysis.

Let N denote the total number of subjects recruited.

Objective:

For 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

In solving this optimisation problem, we create a Bayes sequential decision problem, placing a prior on θ and defining costs for sampling and for making incorrect decisions.

Such a problem can be solved rapidly by dynamic programming.

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

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

Note: Although the Bayes decision problem is introduced as a computational device, this derivation demonstrates that an efficient frequentist procedure should also be good from a Bayesian perspective.

An optimal design for the cholesterol treatment example

In the cholesterol treatment trial, the primary endpoint is reduction in serum cholesterol after 4 weeks of treatment.

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

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

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

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

Type I error rate $\alpha = 0.025$,

Power 0.9 at $\theta = 1$.

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

An optimal design for the cholesterol treatment example

We consider designs with a maximum sample size of 96.

We assume a recruitment rate of 4 per week:

Data start to accrue after 4 weeks,

At each interim analysis, there will be $4 \times 4 = 16$ pipeline subjects,

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:

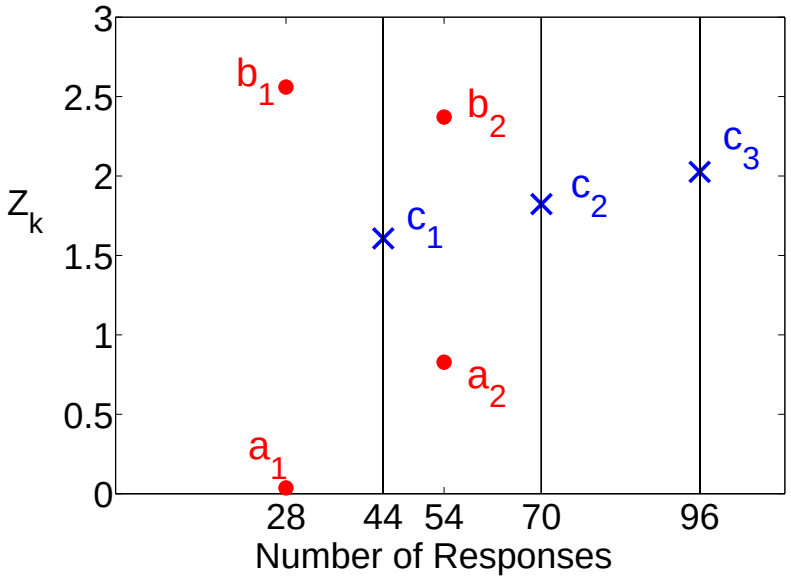
$\tilde{n}_1 = 44$ responses if recruitment stops at interim analysis 1,

$\tilde{n}_2 = 70$ 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.



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.

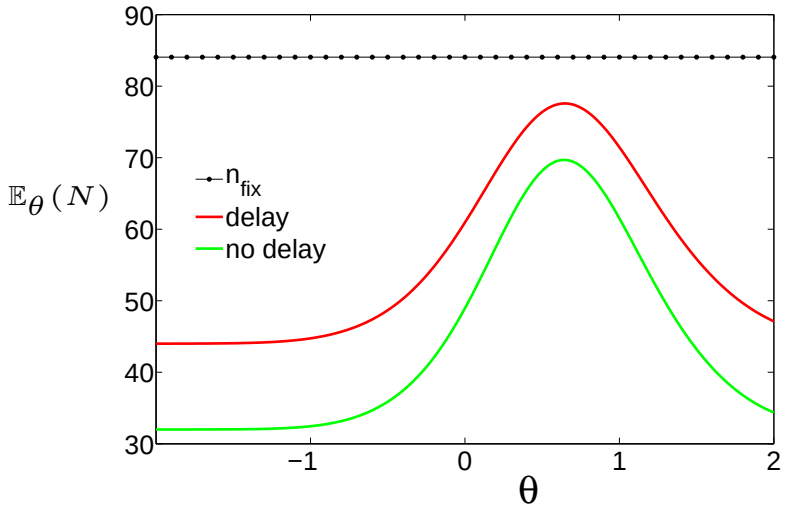
An optimal design for the cholesterol treatment example

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.



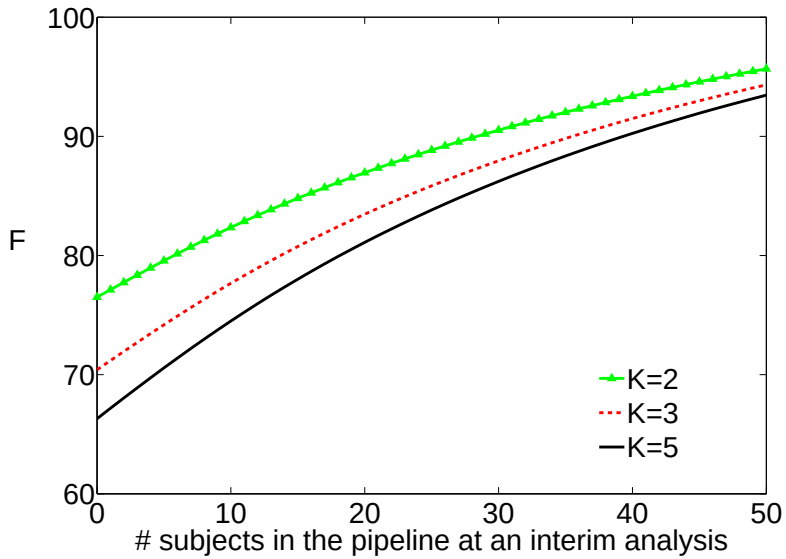
Efficiency loss when there is a delay in response

In general, a delay in response erodes the benefits of sequential testing.

Consider tests with $\alpha = 0.025$, power 0.9 and response variance, σ^2, such that the fixed sample test needs $n_{fix} = 100$ subjects.

Suppose a group sequential design has $n_{max} = 1.1\, n_{fix} = 110$.

The figure shows the minima of $F = \int \mathbb{E}_\theta(N)\, f(\theta)\, d\theta$, attained by optimal Delayed Response GSTs with K analyses.



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

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

We analyse all the available data at each interim analysis.

Using a short term endpoint to recover efficiency

At an interim analysis, subjects are

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

We fit the full model to all the data available at analysis k, then extract

$$\widehat{\theta}_k \quad \text{and} \quad \mathcal{I}_k = \{\text{Var}(\widehat{\theta}_k)\}^{-1}.$$

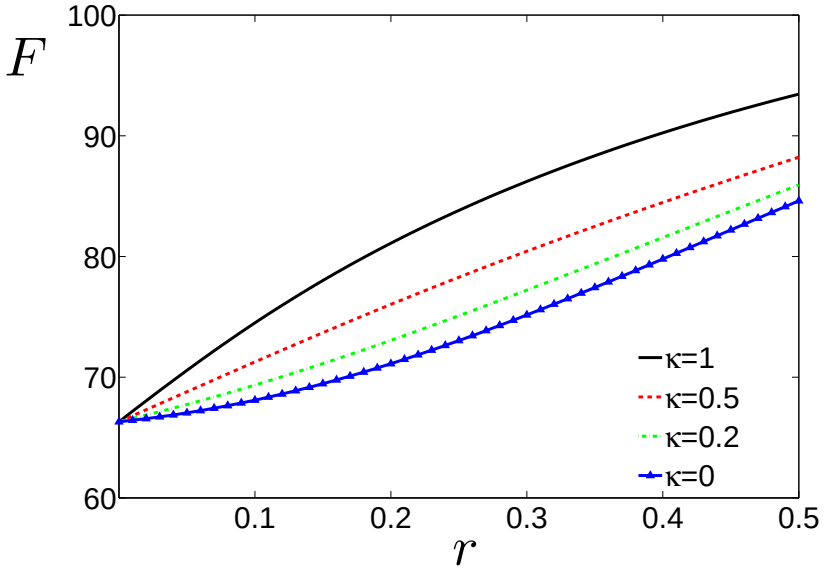
Including the short term endpoint in the model increases the information, $\mathcal{I}_k$, for the long term endpoint.

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, we can 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.

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

The parameter κ is the ratio of time to recording the short-term and long-term endpoints, so $\kappa = 1$ equates to having no short-term endpoint.

Using a short term endpoint to recover efficiency

Note: Although the short-term endpoint may itself be of clinical interest, the final inference is about the primary endpoint alone.

The same approach can be used with repeated measurements as follow-up continues for each patient.

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

In HJ's Example D, prevention of fracture in postmenopausal women, we could:

> *Fit a joint model for bone mineral density measured at one year and incidence of fracture within five years,*
>
> *Use censored time-to-event data on the fracture endpoint for subjects with less than five years of follow-up.*

6. Error spending Delayed Response GSTs

In practice, information levels at interim analyses and decision analyses are unpredictable.

In the error spending approach, the type I error probability to be spent by stage k is defined through a function $f(\mathcal{I}_k)$.

Similarly, the type II probability to be spent by stage k is specified as $g(\mathcal{I}_k)$.

A target information level $\mathcal{I}_{max}$ is defined and recruitment stops when this is reached (or will be reached with the responses from pipeline subjects).

HJ show how to construct error spending Delayed Response GSTs that protect type I error rate exactly.

The attained power is close to its specified level as long as the information levels take values similar to those assumed in planning the trial.

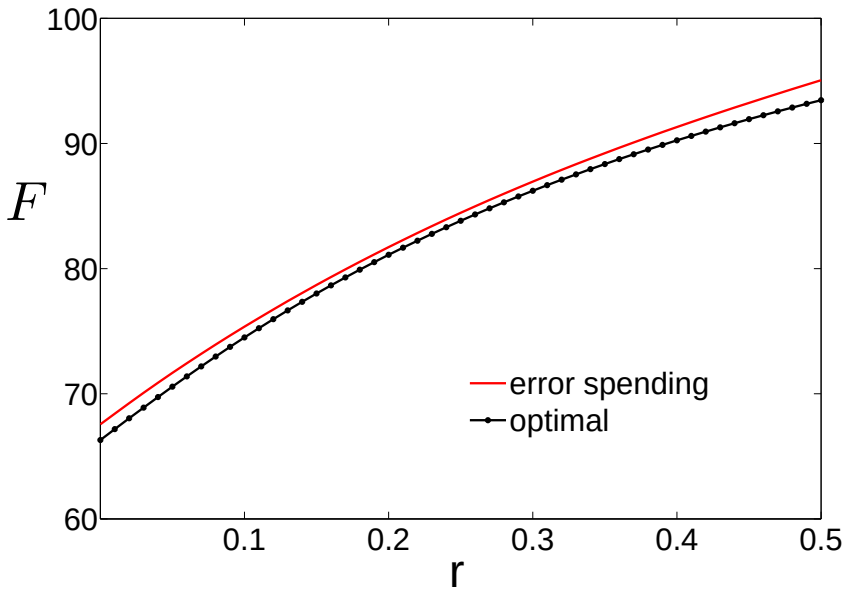
The ρ-family of error spending functions

HJ recommend error spending functions of the form

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

The efficiency of the resulting designs can be seen in our example with $\alpha = 0.025$, power 0.9, $K = 5$ stages, $n_{fix} = 100$ and $n_{max} = 110$.

Values of F achieved by ρ-family error spending designs



7. Further topics

A variety of optimality criteria

HJ show how designs can be optimised for criteria involving both the number of subjects recruited and the time to a final decision.

The nature of a specific clinical trial will determine which approaches may be possible, depending on whether:

All pipeline subjects must be followed to the response time,

Investigators may decide whether to wait and observe pipeline subjects,

Data from (some) pipeline subjects will not be “valid” and cannot be used.

Discussants of the HJ paper commented on the nature of “pipeline” data and HJ categorised possible types of situation in their response.

Further topics

Inference on termination

HJ explain how to construct p-values and confidence intervals, with the usual frequentist properties, on termination of a Delayed Response GST.

These methods can also provide median unbiased point estimates.

The bias of maximum likelihood estimates can be reduced following the approach which Whitehead (*Biometrika*, 1986) introduced for standard GSTs.

Non-binding futility boundaries

It is commonly required that a group sequential design should protect the type I error rate, even if the trial may continue after crossing the “futility” boundary.

We are currently working to extend our error spending methods to the “non-binding” case.

Further topics

Adaptive choice of group sizes in a Delayed Response GST

There have been many proposals for “sample size re-estimation” in response to interim treatment effect estimates.

In the case of an immediate response, the resulting methods can be regarded as group sequential tests with the added feature that the size of each group is data-dependent.

The papers of Faldum & Hommel (*J. Biopharm. Statistics*, 2007) and Mehta & Pocock (*Statistics in Medicine*, 2011) present examples of sample size re-estimation with a delayed response.

HJ show that, when designs are optimised, there is little to be gained from such adaptations — in agreement with the findings of Jennison & Turnbull (*Biometrika*, 2006) for the case of immediate response.

Further topics

Unexpected over-running

HJ describe how their methods can be used to handle data that arrive after the conclusion of a “standard” group sequential test.

The basic requirement for the approach to be valid is an understanding that this form of adjustment will be used when over-run data arise unexpectedly.