Monitoring clinical trial outcomes with delayed response: Incorporating "pipeline" data in group sequential and adaptive designs

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

Deming Conference, Atlantic City,

December 11, 2013

Outline of talk

Group sequential tests (GSTs)

1. Group sequential monitoring of clinical trials

Delayed responses

2. The problem of delayed responses
3. Defining a group sequential test with delayed responses
4. Optimising a Delayed Response GST
5. Efficiency loss when there is a delay in response
6. Error spending Delayed Response GSTs
7. Further topics

Outline of talk

Group sequential or "start small and ask for more"?

8. Adaptive and Group Sequential designs: Choosing the sample size of a clinical trial
9. Mehta & Pocock's example
10. The Mehta-Pocock (MP) design
11. Alternatives to the MP design
12. Deriving efficient sample size rules in the MP framework
13. Using the Conditional Probability of Rejection principle
14. Relation between MP designs and Delayed Response GSTs
15. Conclusions

1. Group sequential monitoring of clinical trials

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

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

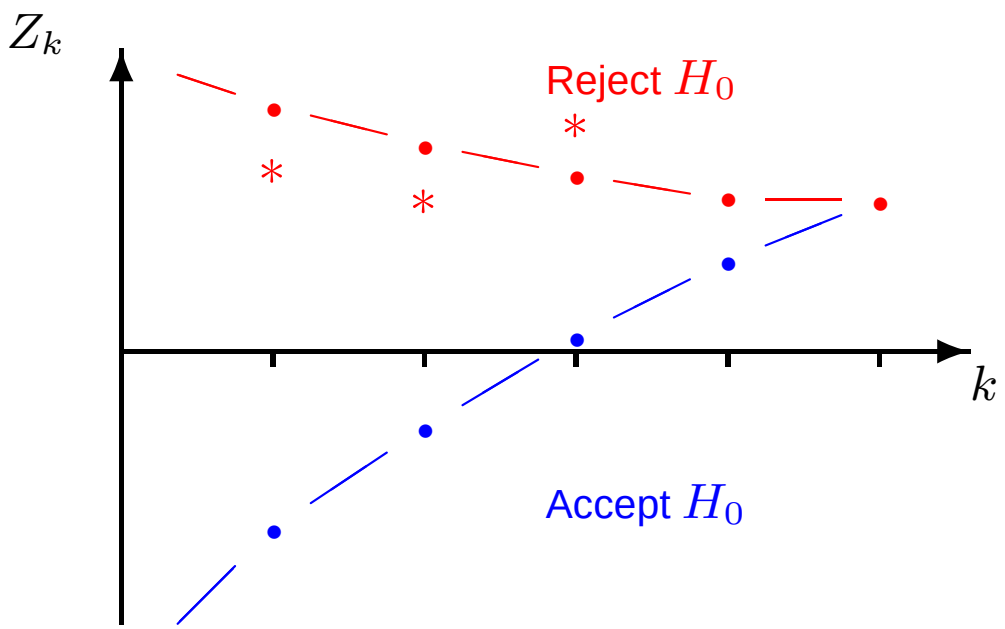
$$P_{\theta=0}\{\text{Reject } H_0\} = \alpha,$$

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

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 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.

Here, the trial stops to reject H_0 at the third of five analyses.

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.

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

A single normal mean

Suppose $X_1, X_2, \ldots$ are independent $N(\theta, \sigma^2)$ responses.

For $n_1 < n_2$, define

$$\widehat{\theta}_1 = \frac{X_1 + \ldots + X_{n_1}}{n_1} \quad \text{and} \quad \widehat{\theta}_2 = \frac{X_1 + \ldots + X_{n_1} + \ldots + X_{n_2}}{n_2}.$$

The joint distribution of $\widehat{\theta}_1$ and $\widehat{\theta}_2$ is bivariate normal.

Marginally

$$\widehat{\theta}_1 \sim N(\theta, \mathcal{I}_1{}^{-1}) \quad \text{and} \quad \widehat{\theta}_2 \sim N(\theta, \mathcal{I}_2{}^{-1}),$$

where

$$\mathcal{I}_1 = \frac{n_1}{\sigma^2} \quad \text{and} \quad \mathcal{I}_2 = \frac{n_2}{\sigma^2}.$$

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

It remains to check the covariance:

$$\begin{aligned}
\mathrm{Cov}(\widehat{\theta}_1, \widehat{\theta}_2) &= \mathrm{Cov}\left(\frac{X_1 + \ldots + X_{n_1}}{n_1}, \frac{X_1 + \ldots + X_{n_1} + \ldots + X_{n_2}}{n_2} \right) \\
&= \mathrm{Cov}\left(\frac{X_1 + \ldots + X_{n_1}}{n_1}, \frac{X_1 + \ldots + X_{n_1}}{n_2} \right) \\
&= \frac{1}{n_1\, n_2} \mathrm{Var}(X_1 + \ldots + X_{n_1}) \\
&= \frac{\sigma^2}{n_2} = \{\mathcal{I}_2\}^{-1} \\
&= \mathrm{Var}(\widehat{\theta}_2).
\end{aligned}$$

Canonical joint distribution of z-statistics

In testing H_0: $\theta = 0$, the *standardised statistic* at analysis k is

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

For this,

$(Z_1, \ldots, Z_K)$ is multivariate normal,

$Z_k \sim N(\theta\sqrt{\mathcal{I}_k}, 1), \quad k = 1, \ldots, K,$

$\mathsf{Cov}(Z_{k_1}, Z_{k_2}) = \sqrt{\mathcal{I}_{k_1}/\mathcal{I}_{k_2}}$ for $k_1 < k_2$.

Canonical joint distribution of score statistics

The *score statistics*, $S_k = Z_k\sqrt{\mathcal{I}_k}$, are also multivariate normal with

$$S_k \sim N(\theta\,\mathcal{I}_k,\, \mathcal{I}_k), \quad k = 1, \ldots, K.$$

The score statistics possess the "independent increments" property,

$$\text{Cov}\big(S_k - S_{k-1},\, S_{k'} - S_{k'-1}\big) = 0 \quad \text{for } k \neq k'.$$

It can be helpful to know that the score statistics behave as Brownian motion with drift θ observed at times $\mathcal{I}_1, \ldots, \mathcal{I}_K$.

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.

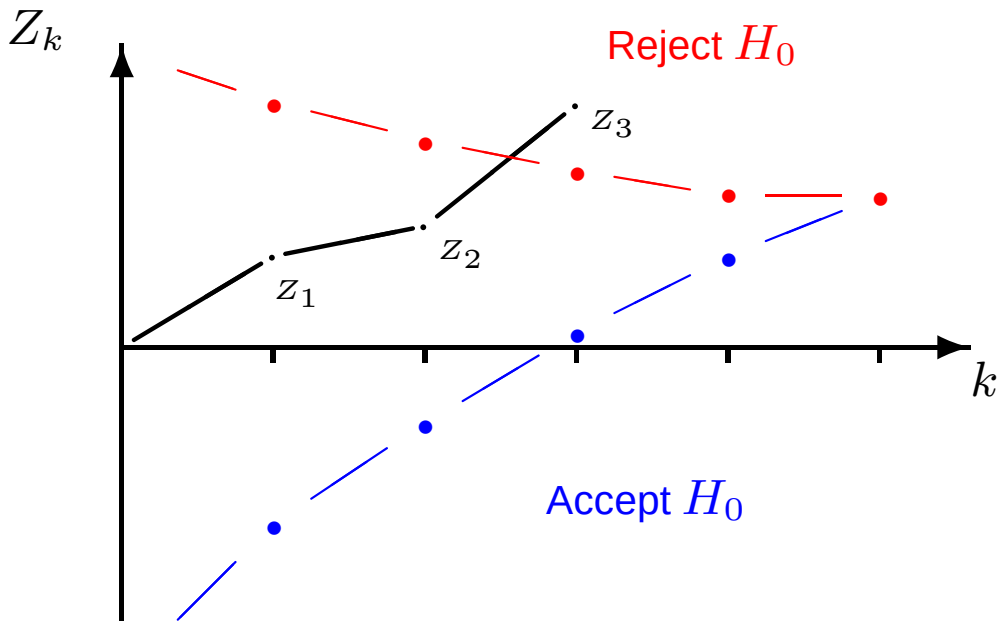
For survival data, observed information is roughly proportional to the number of failures.

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).

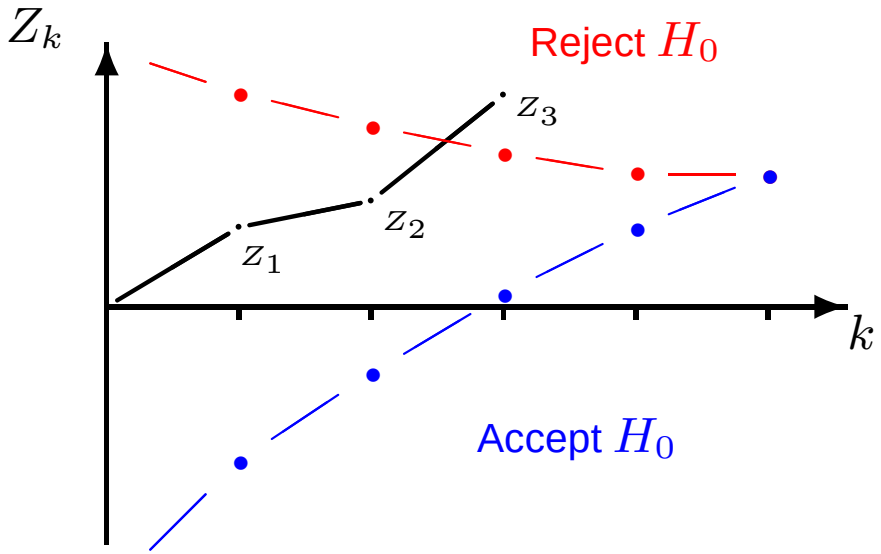
Computations for group sequential tests



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).

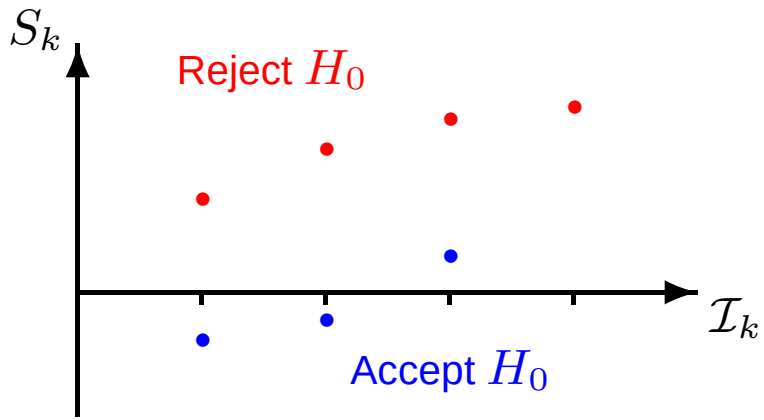
Combining such probabilities yields properties of a group sequential boundary.

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

Example of 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 $\theta = \delta$.



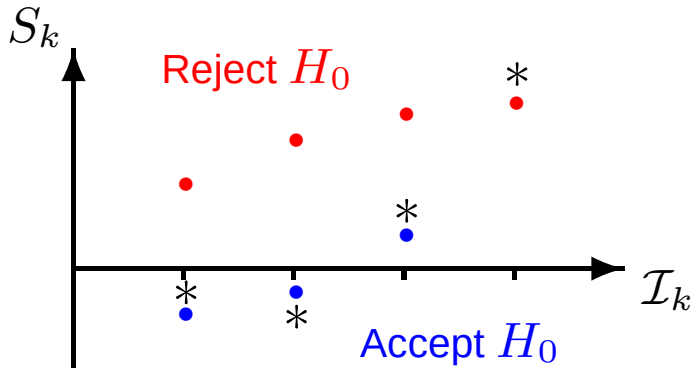
The computational methods just described can be used to define tests with parametric stopping boundaries meeting the design criteria.

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

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

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.

Benefits of group sequential testing

In order to test H_0: $\theta \leq 0$ against $\theta > 0$ with type I error probability α and power $1 - \beta$ at $\theta = \delta$, a fixed sample size test needs information

$$\mathcal{I}_{fix} = \frac{\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2}{\delta^2}.$$

Information is (roughly) proportional to sample size in many clinical trial settings.

A group sequential test with K analyses will need to be able to continue to a maximum information level $\mathcal{I}_K$ which is greater than $\mathcal{I}_{fix}$.

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

We term the ratio $R = \mathcal{I}_K / \mathcal{I}_{fix}$ the “inflation factor” for a group sequential design.

Benefits of group sequential testing

In specifying a group sequential test's boundary, one can aim to minimise the expected information $\mathbb{E}_\theta(\mathcal{I})$ under effect sizes of θ of most interest, subject to a fixed number of analyses K and inflation factor R.

Eales & Jennison (*Biometrika*, 1992) and Barber & Jennison (*Biometrika*, 2002) report on designs optimised for criteria of the form $\sum_i w_i \mathbb{E}_{\theta_i}(\mathcal{I})$ or

$$\int f(\theta)\, \mathbb{E}_\theta(\mathcal{I})\, d\theta,$$

where f is a normal density.

These optimal group sequential designs can be used in their own right.

They also serve as benchmarks for other methods which may have additional useful features.

Computing optimal GSTs

In optimising a group sequential test, 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.

Benefits of group sequential testing

One-sided tests with binding futility boundaries, minimising $\{\mathbb{E}_0(\mathcal{I}) + \mathbb{E}_\delta(\mathcal{I})\}/2$ for equal group sizes, $\alpha = 0.025$, $1-\beta = 0.9$, K analyses, $\mathcal{I}_{max} = R\mathcal{I}_{fix}$.

Minimum values of $\{\mathbb{E}_0(\mathcal{I}) + \mathbb{E}_\delta(\mathcal{I})\}/2$, as a percentage of $\mathcal{I}_{fix}$

	R					
K	1.01	1.05	1.1	1.2	1.3	*Minimum over R*
2	80.8	74.7	73.2	73.7	75.8	73.0 at R=1.13
3	76.2	69.3	66.6	65.1	65.2	65.0 at R=1.23
5	72.2	65.2	62.2	59.8	59.0	58.8 at R=1.38
10	69.2	62.2	59.0	56.3	55.1	54.2 at R=1.6
20	67.8	60.6	57.5	54.6	53.3	51.7 at R=1.8

Note: $\mathbb{E}(\mathcal{I}) \searrow$ as $K \nearrow$ but with diminishing returns,

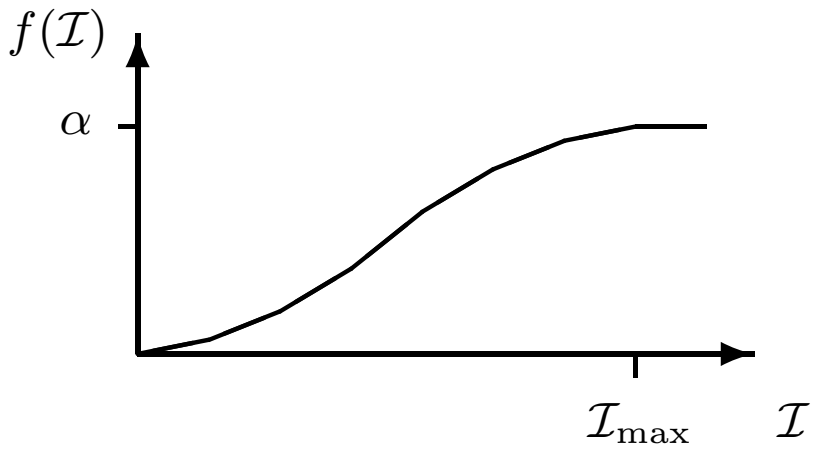
$\mathbb{E}(\mathcal{I}) \searrow$ as $R \nearrow$ up to a point.

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.

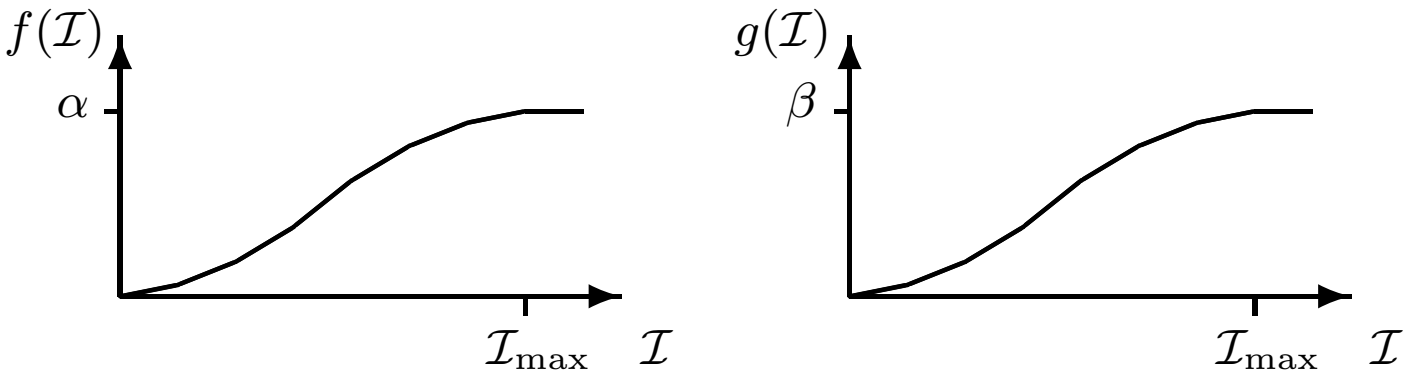
One-sided error spending tests

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:

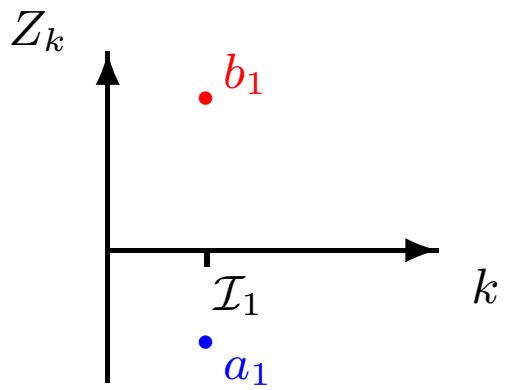
Observed information $\mathcal{I}_1$.

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 $\mathcal{I}_2$

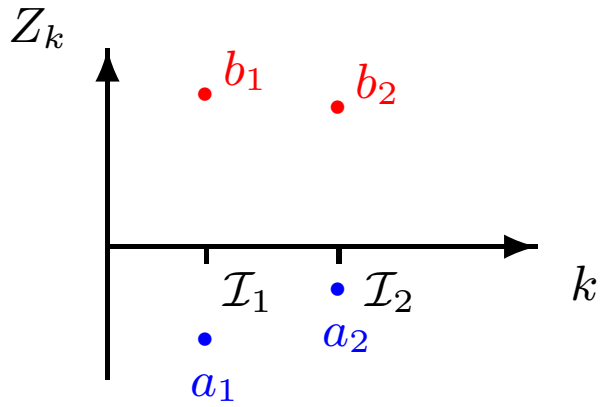
Reject H_0 if $Z_2 > b_2$, where

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

— 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

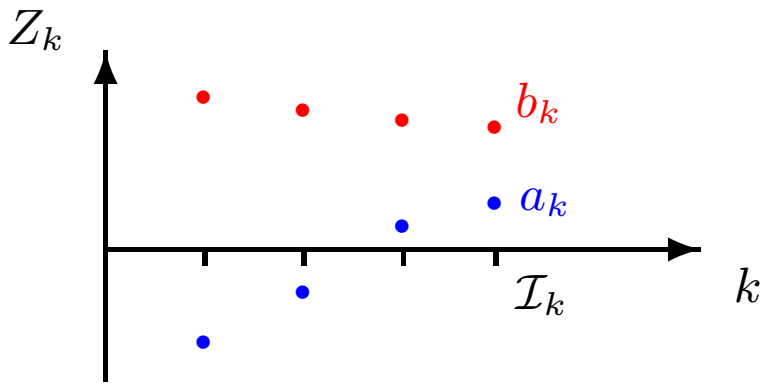
Analysis k: Observed information $\mathcal{I}_k$

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

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

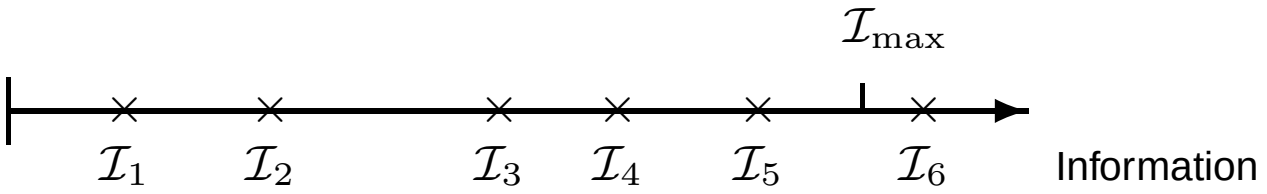


Remarks on error spending tests

1. Computation of $(a_k,\, b_k)$ does ***not*** depend on future information levels, $\mathcal{I}_{k+1},\, \mathcal{I}_{k+2},\, \ldots$.

2. A "maximum information design" continues until a boundary is crossed or an analysis with $\mathcal{I}_k \geq \mathcal{I}_{\max}$ is reached.

If necessary, patient accrual can be extended to reach $\mathcal{I}_{\max}$.



If a maximum of K analyses is specified, the study terminates at analysis K with $f(\mathcal{I}_K)$ defined to be α. Then, a_K is chosen to give cumulative type I error probability α and we set $b_K = a_K$.

Remarks on error spending tests

3. The value of $\mathcal{I}_{\max}$ can be chosen so that boundaries converge at the final analysis under a typical sequence of information levels, e.g.,

$$\mathcal{I}_k = (k/K)\,\mathcal{I}_{\max}, \quad k = 1, \ldots, K.$$

4. The ρ-family provides a convenient choice of error spending functions.

In the case of one-sided tests, type I error probability is spent as

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

and type II error probability as

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

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

Barber & Jennison (*Biometrika*, 2002) show ρ-family tests have excellent efficiency properties when compared with designs for the same number of analyses K and inflation factor R.

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

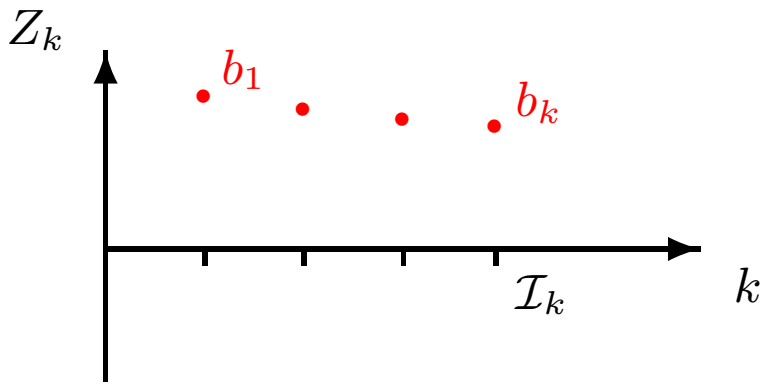
If the futility boundary is treated as non-binding, computation of the error-spending efficacy boundary only involves the type I error spending function $f(\mathcal{I})$.

Boundary values, $b_1, b_2, \ldots$, are calculated one by one as the trial proceeds.

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

Reject H_0 if $Z_k > b_k$, where

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



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

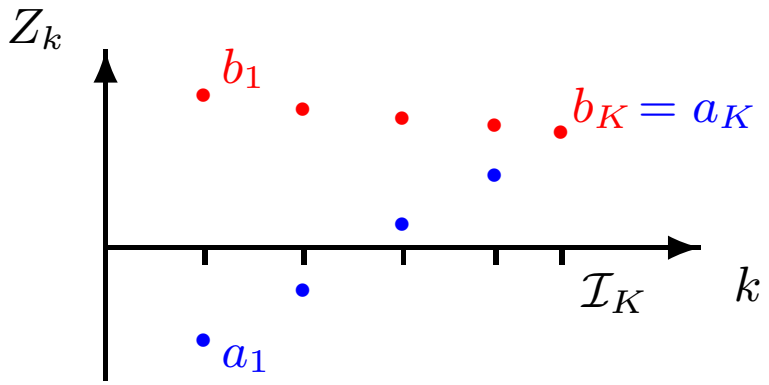
The futility boundary can be added through a type II error spending function $g(\mathcal{I})$.

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 $a_K = b_K$.



2. The problem of delayed responses

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

Example: Cholesterol reduction after 4 weeks of treatment

In their Example A, HJ describe a group sequential 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.

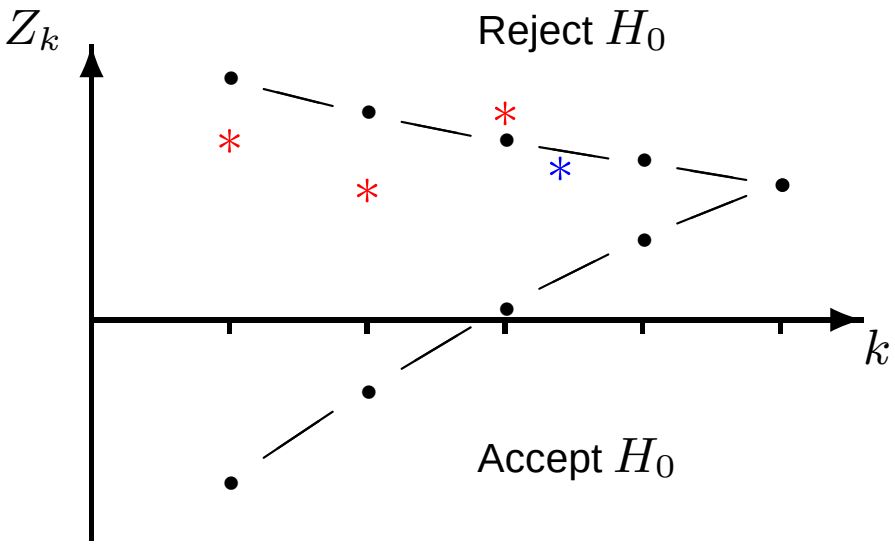
We refer to these as patients as being “in the pipeline”.

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?

Incorporating delayed observations after a GST terminates

1. Whitehead (*Controlled Clinical Trials*, 1992) proposed a "deletion" method.

The analysis k at which termination occurs is "deleted" and one behaves as if analysis k originally had information $\tilde{\mathcal{I}}_k$, appropriate to the final set of responses.

A boundary value $\tilde{b}_k$ is computed and H_0 rejected if the final statistic $\tilde{Z}_k \geq \tilde{b}_k$.

2. Hall & Ding (*Univ. Rochester, Technical Report*, 2002) applied a combination test (Bauer & Köhne, *Biometrics*, 1994) to the two sets of data obtained before and after the GST terminates.

Sorriyarachchi et al. (*Biometrics*, 2003) investigated these methods and found they perform poorly with respect to power:

> *The deletion method is conservative and can lead to lower power than a GST which ignores the additional data,*
>
> *With a moderate number of pipeline subjects, Hall & Ding's method leads to greater loss of power than the deletion method.*

Incorporating delayed observations after a GST terminates

The methods of Whitehead (1992) and Hall & Ding (2002) are based on applying a GST as if response were immediate, then trying to deal with 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 as 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.

3. Defining a group sequential test with delayed responses

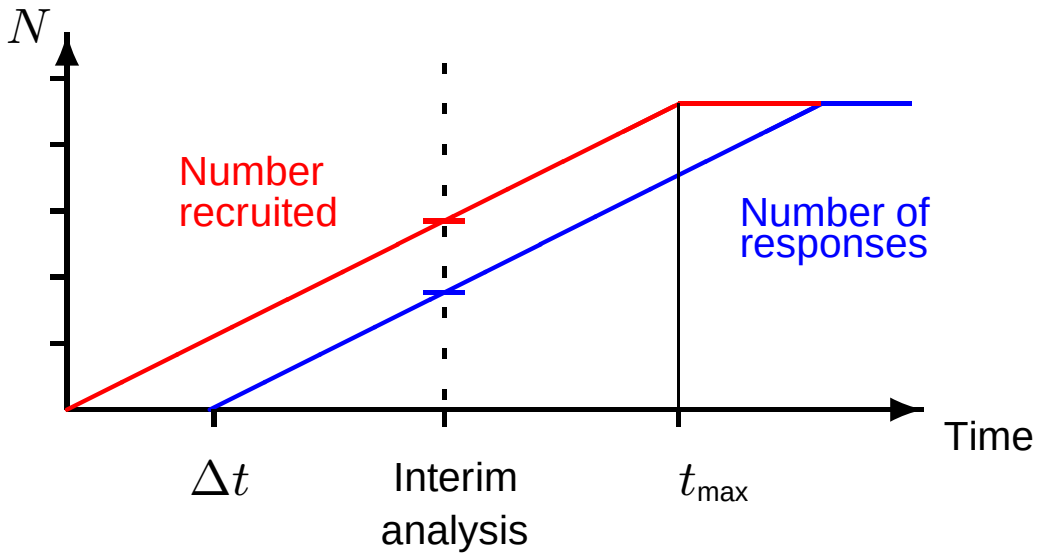
We assume:

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.

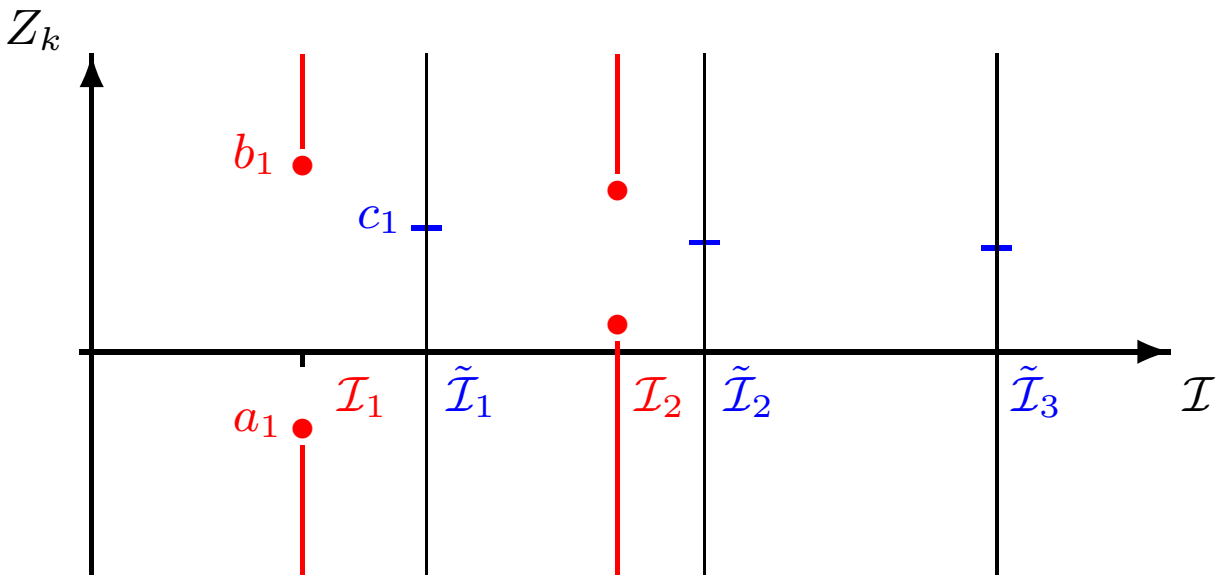
Consider a trial where responses are observed time Δ_t after treatment.



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

Boundaries for a Delayed Response GST

At interim analysis k, the observed information level 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.

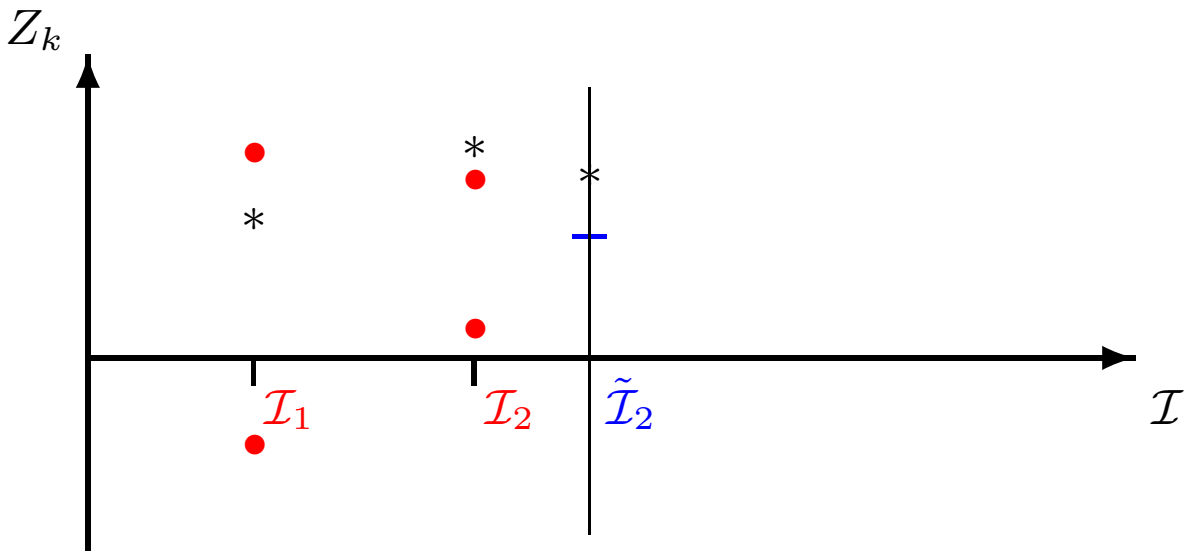
At the subsequent decision analysis, denote the observed information by $\tilde{\mathcal{I}}_k$ and reject H_0 if $\tilde{Z}_k > c_k$.

Delayed Response Group Sequential Tests (DR 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 extra "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 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 distribution we saw earlier for the sequence of Z-statistics in a GST with immediate responses (JT, Ch. 11).

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

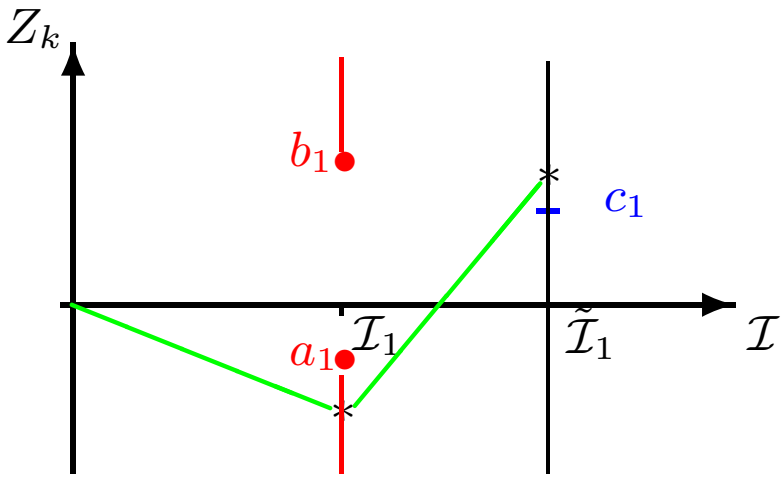
The value of information from pipeline subjects

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 use in this decision.

The pipeline data will occasionally produce a “reversal”, with the final decision differing from that anticipated when recruitment was terminated.

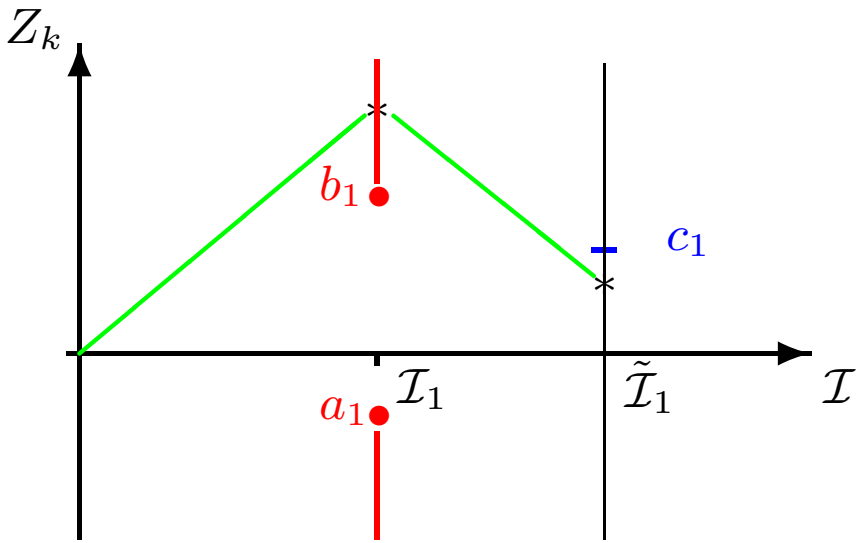
We could observe:



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 that there is no option of "banking" the good evidence at analysis 1 — we are assuming all pipeline subjects will eventually be observed.

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

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

4. Optimising a Delayed Response GST

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

We set 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:

Given α, β, δ, n_{max}, K and r, we 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

We follow the same approach as for optimising a GST with immediate response.

We create a Bayes sequential decision problem, placing a prior on θ and defining costs for sampling and for making incorrect decisions.

This 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.

Again, the Bayes decision problem is introduced as a computational device, but the derivation demonstrates the relationship between admissible frequentist designs and Bayes procedures.

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,

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:

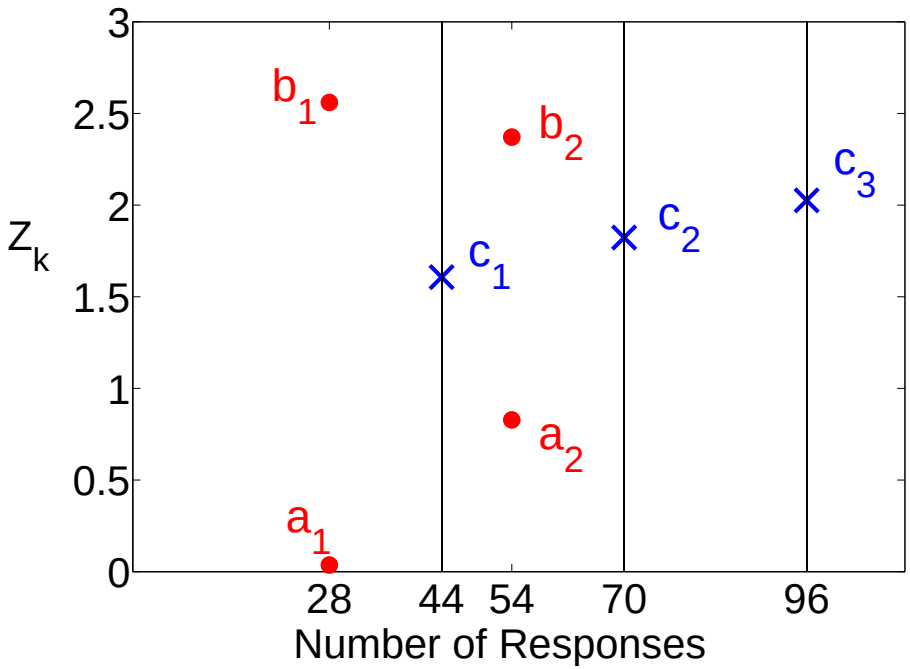
$\tilde{n}_1 = 44$ responses if recruitment stops at interim analysis 1,

$\tilde{n}_2 = 70$ 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.



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.

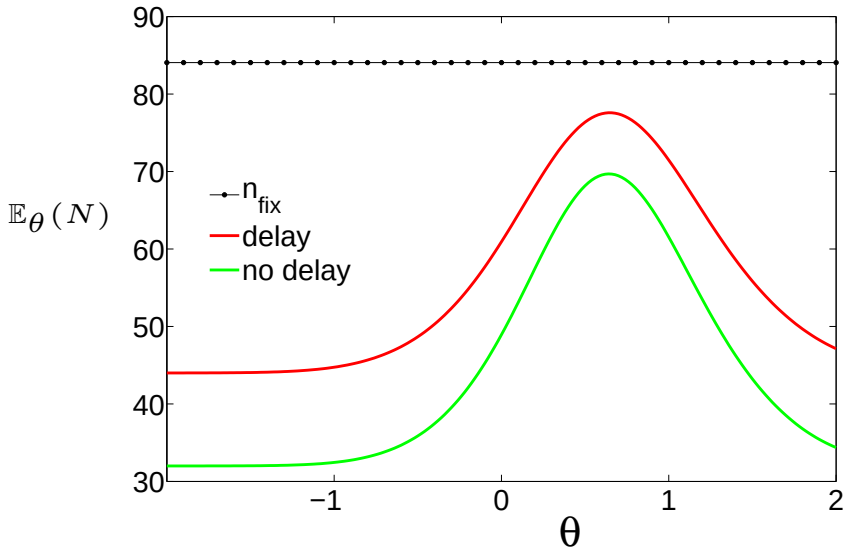
An optimal design for the cholesterol treatment example

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.



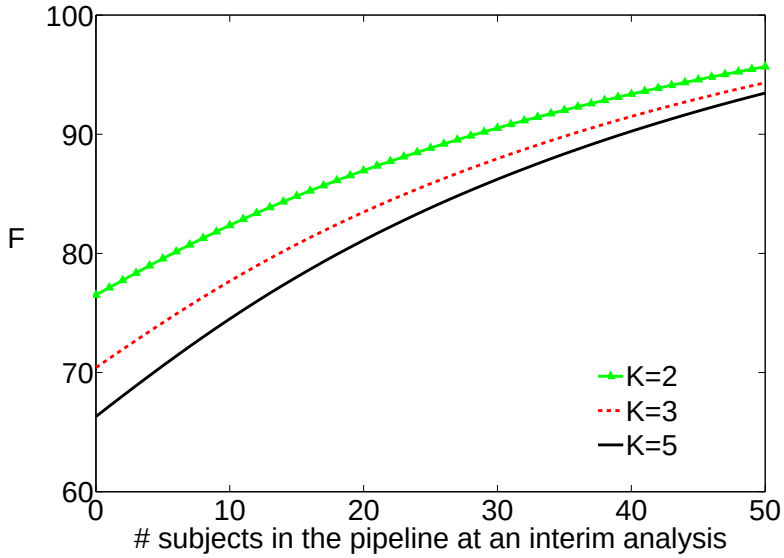
5. Efficiency loss when there is a delay in response

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 for a range of "pipeline" sizes.



The reduction in average $\mathbb{E}_\theta(N)$ when the pipeline size is 25 patients is around half the reduction achieved by a GST when response is observed immediately.

Using a short term endpoint to improve 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 improve 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

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

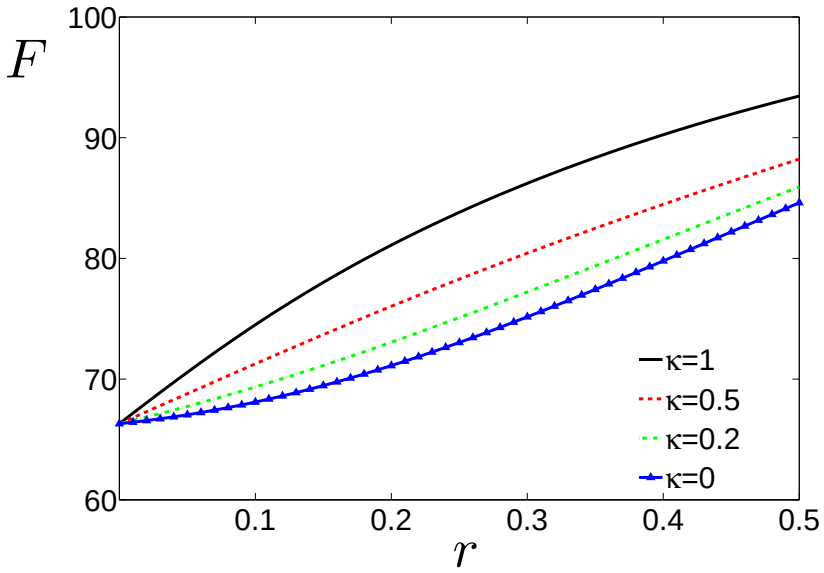
Including the short term endpoint in the model increases the information, $\mathcal{I}_k$, for the long term endpoint.

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 improve 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 improve 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.

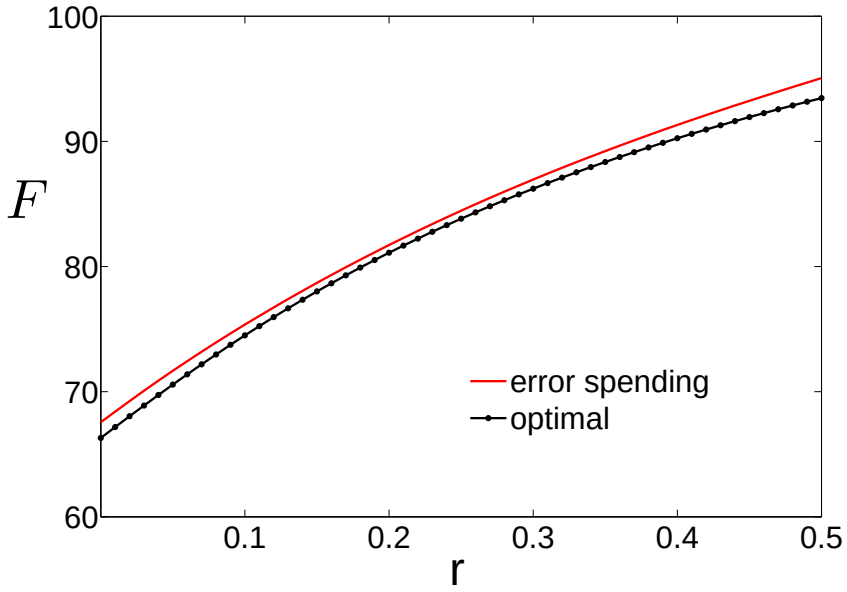
The ρ-family of error spending functions

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.

With an immediate response, these designs can be regarded as GSTs with the added feature that the size of each group is data-dependent.

HJ derived optimal “adaptive” versions of 2-group Delayed Response GSTs designs. They found only minor benefits were achieved by adapting group sizes in response to treatment effect estimates.

Faldum & Hommel (*J. Biopharm. Statistics*, 2007) and Mehta & Pocock (*Statistics in Medicine*, 2011) present 2-group designs with sample size re-estimation and a delayed response: we shall explore how the Mehta-Pocock designs compare to HJ’s Delayed Response GSTs, both non-adaptive and adaptive.

Summary of Delayed Response GSTs

We have described group sequential tests for a delayed response (DR GSTs).

These designs offer (nearly) all the usual feature of GSTs for an immediate response.

We can design DR GSTs to be as efficient as possible, subject to the specified constraints.

Understanding the impact of a delayed response, we can take steps to improve efficiency, for example, by using short term end-points to capture interim information from pipeline subjects.

The methods are ready to be considered for application — which will, no doubt, raise further challenges.

8. Adaptive and Group Sequential trial designs: Choosing the sample size of a clinical trial

We now return to a fundamental issue in clinical trial design — the sample size.

Let θ denote the effect size of a new treatment, i.e., the difference in mean response between the new treatment and the control.

Sample size is determined by:

- Type I error rate α, and
- Treatment effect size $\theta = \Delta$ at which power $1 - \beta$ is to be achieved.

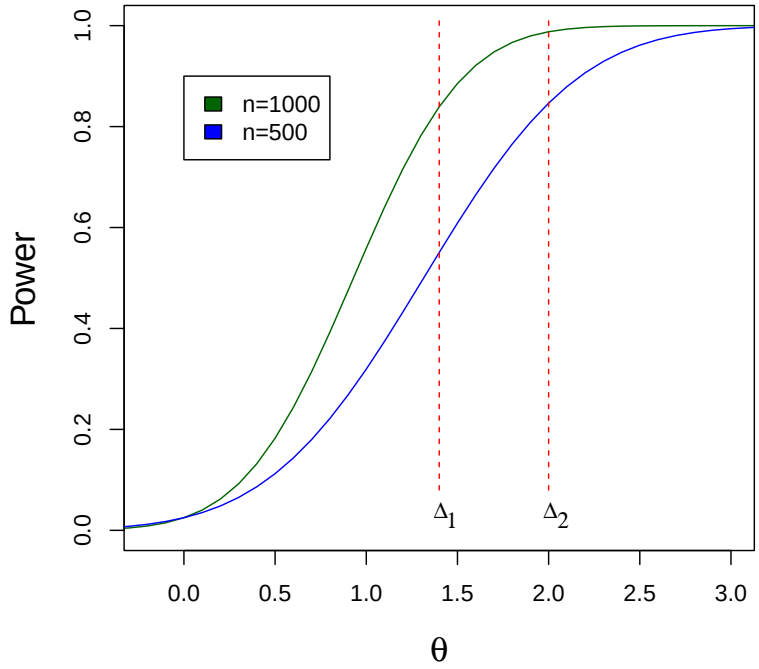
Dispute may arise over the choice of Δ.

Should investigators use:

- The minimum effect of interest Δ_1, or
- The anticipated effect size Δ_2 ?

Choosing the sample size for a trial

Power curves for designs with fixed sample sizes of 500 and 1000.



With 1000 subjects, there is good power at the minimum clinical effect, Δ_1.

With only 500 subjects, good power is achieved at the more optimistic Δ_2.

If $\theta = \Delta_2$, a sample size of 1000 is unnecessarily high.

Designing a trial with good power and sample size

In designing a clinical trial, we aim to

Protect the type I error rate,

Achieve sufficient power,

Use as small a sample size as possible.

Adaptive designs in this context often have the form:

Start with a fixed sample size design,

Examine interim data,

Add observations to improve power where most appropriate.

In contrast, ***Group Sequential*** designs require one to:

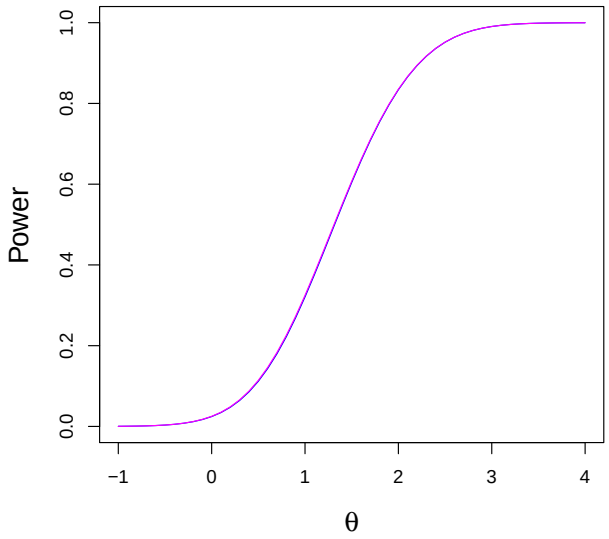
Specify the desired type I error and power function,

Set maximum sample size a little higher than the fixed sample size,

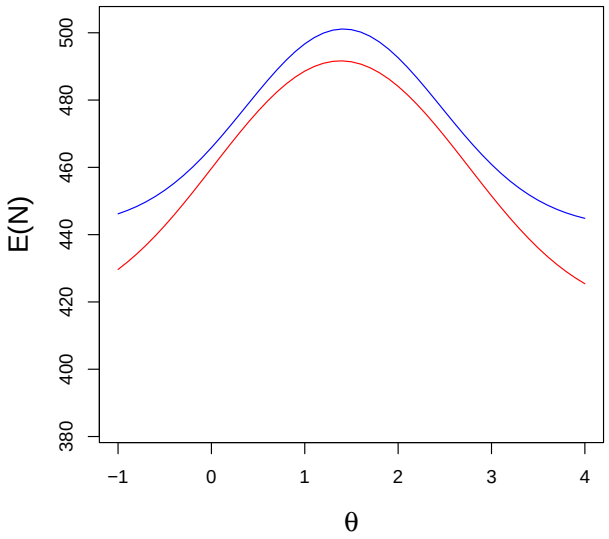
Stop the trial early if data support this.

Designing a clinical trial

Power curve



$E_\theta(N)$ curves



All designs, *including adaptive procedures*, have overall power curves.

Designs with similar power curves can be compared in terms of their average sample size functions, $E_\theta(N)$.

Even if there is uncertainty about the likely treatment effect, investigators should be able to specify the values of θ under which early stopping is most desirable.

Adaptive design or GST?

Jennison & Turnbull (JT) have compared group sequential tests (GSTs) and adaptive designs. See, for example, papers in

Statistics in Medicine (2003, 2006), *Biometrika* (2006), *Biometrics* (2006)

JT conclude that:

GSTs are excellent

They do what is required with low expected sample sizes,

Error spending versions handle unpredictable group sizes, etc.

Adaptive designs can be as good as GSTs

However, many published adaptive designs require higher expected sample sizes to achieve the same power as good GSTs.

Re-visiting the *Group Sequential* vs *Adaptive* question

The paper by Mehta & Pocock (*Statistics in Medicine*, 2011)

"Adaptive increase in sample size when interim results are promising:
A practical guide with examples"

has re-opened this question.

Conclusions of Mehta & Pocock (MP) are counter to the findings we have reported.

An important feature:

In MP's first example, response is measured some time after treatment.

Thus, at an interim analysis, many patients have been treated but are yet to produce a response.

Delayed responses are common — yet, prior to Hampson & Jennison (2013), they received little attention in the GST literature.

Re-visiting the *Group Sequential* vs *Adaptive* question

We shall consider the first example presented by Mehta & Pocock and describe their proposed trial design.

We shall describe alternative fixed and group sequential designs which could be used for this example: these designs achieve the same power curves with smaller expected sample sizes.

We shall discuss how one can improve Mehta & Pocock's design while working in their overall framework.

We then extend this framework to obtain a wider class of designs.

We relate the 2-group designs obtained at the end of this development to the Delayed Response GSTs of Hampson & Jennison (2013).

9. Mehta & Pocock's example

MP's Example 1 concerns a Phase 3 trial of a new treatment for schizophrenia in which a new drug is to be compared to an active comparator.

The efficacy endpoint is improvement in the Negative Symptoms Assessment score from baseline to week 26.

Responses are

$$Y_{Bi} \sim N(\mu_B, \sigma^2), \ i = 1, 2, \ldots, \text{ on the new treatment,}$$

$$Y_{Ai} \sim N(\mu_A, \sigma^2), \ i = 1, 2, \ldots, \text{ on the comparator treatment.}$$

where $\sigma^2 = 7.5^2$.

The treatment effect is

$$\theta = \mu_B - \mu_A.$$

and we estimate θ by

$$\hat{\theta} = \hat{\mu}_B - \hat{\mu}_A = \overline{Y}_B - \overline{Y}_A.$$

Mehta & Pocock's Example

The initial plan is for a total of $n_2 = 442$ patients, 221 on each treatment.

In testing H_0: $\theta \leq 0$ vs $\theta > 0$, the final analysis will reject H_0 if Z_2

$$Z_2 = \frac{\hat{\theta}(n_2)}{\sqrt{\{4\sigma^2/n_2\}}} > 1.96,$$

where $\hat{\theta}(n_2)$ is the standard estimate of θ at the final analysis.

This design and analysis gives type I error rate 0.025 and power 0.8 at $\theta = 2$.

Higher power, e.g., power 0.8 at $\theta = 1.6$, would be desirable.

But, the sponsors will only increase sample size if interim results are "promising".

An interim analysis is planned after observing $n_1 = 208$ responses.

Increasing the sample size

At the interim analysis with $n_1 = 208$ observed responses, the estimated treatment effect is

$$\widehat{\theta}_1(n_1) = \overline{Y}_B(n_1) - \overline{Y}_A(n_1)$$

and

$$Z_1 = \frac{\widehat{\theta}_1(n_1)}{\sqrt{\{4\sigma^2/n_1\}}}.$$

At the time of this analysis, a further 208 subjects will have been treated for less than 26 weeks. Their responses will be observed in due course.

As recruitment continues, we use the value of Z_1 in choosing a new total sample size between the original figure of 442 and a maximum of 884.

In deciding whether to increase the sample size, MP consider conditional power of the original test (with $n_2 = 442$ observations), given the observed value of Z_1.

Increasing the sample size

Definition

The conditional power $CP_\theta(z_1)$ is the probability the final test, with $n_2 = 442$ observations, rejects H_0, given $Z_1 = z_1$ and effect size θ,

$$CP_\theta(z_1) \;=\; P_\theta\{Z_2 > 1.96 \,|\, Z_1 = z_1\}.$$

MP's adaptive design is based on conditional power under $\theta = \hat{\theta}_1$.

They divide the range of z_1 into three regions:

Favourable	$CP_{\hat{\theta}_1}(z_1) \geq 0.8$	*Continue to* $n_2 = 442$,
Promising	$0.365 \leq CP_{\hat{\theta}_1}(z_1) < 0.8$	*Increase* n_2,
Unfavourable	$CP_{\hat{\theta}_1}(z_1) < 0.365$	*Continue to* $n_2 = 442$.

When increasing sample size in the promising zone, the final test of H_0 must protect the type I error rate at level α.

The Chen, DeMets & Lan method

References:

Chen, DeMets & Lan, *Statistics in Medicine* (2004),

Gao, Ware & Mehta, *J. Biopharmaceutical Statistics* (2008).

Suppose at interim analysis 1, the final sample size is increased to $n_2^* > n_2$ and one wishes to carry out a final test without adjustment for this adaptation.

Thus, H_0 will be rejected if

$$Z_2(n_2^*) = \frac{\widehat{\theta}(n_2^*)}{\sqrt{\{4\sigma^2/n_2^*\}}} > 1.96.$$

Chen, DeMets & Lan (CDL) show that if n_2 is only increased when

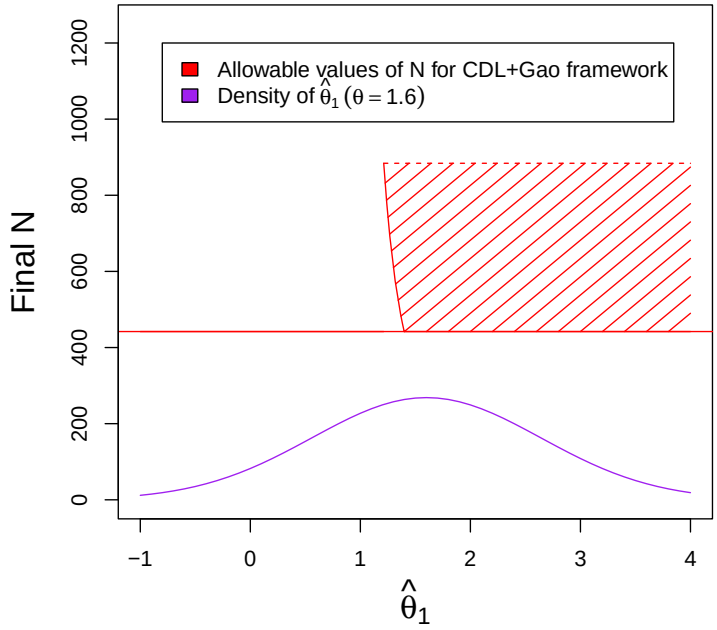
$$CP_{\hat{\theta}_1}(z_1) > 0.5, \tag{1}$$

then the type I error probability will not increase.

So, the "standard" test of H_0 can be used if n_2 is only increased when (1) holds.

Gao's extension of the CDL method

Gao et al. extended the CDL method to lower values of $\widehat{\theta}_1$, as long as a sufficiently high value is chosen for the final sample size, n_2^*.

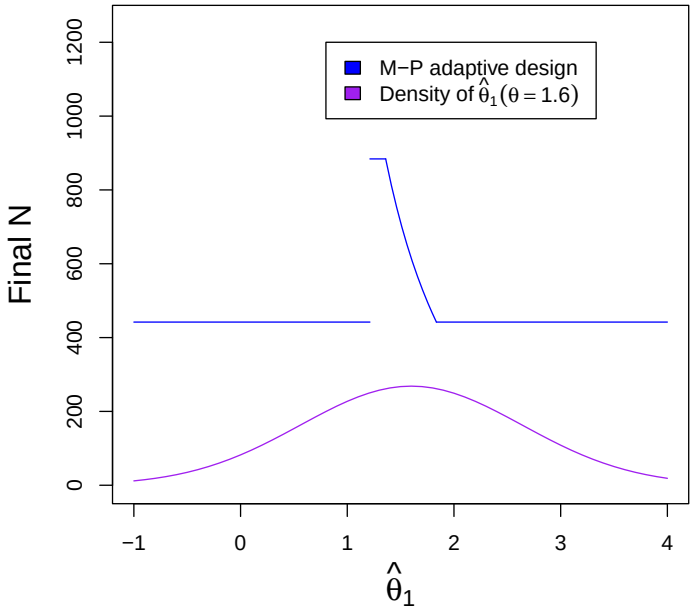


With an upper limit of $n_2^* = 884$, the final sample sizes permitted by the CDL+Gao approach are as shown in the figure.

Now, n_2 can be increased when $CP_{\hat{\theta}_1}(z_1)$ is as low as 0.365.

10. The Mehta-Pocock design

In their "promising zone", MP increase n_2 to achieve conditional power 0.8 under $\theta = \widehat{\theta}_1$, truncating this value to 884 if it is larger than that.

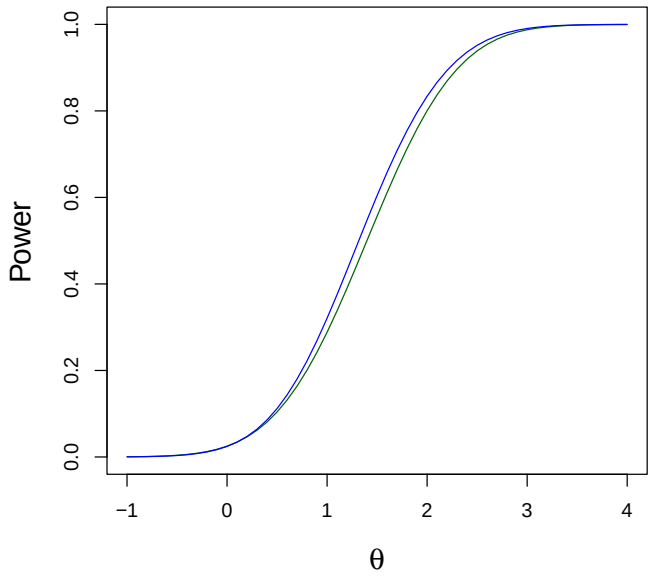


Comparison with the distribution of $\widehat{\theta}_1$ under $\theta = 1.6$ shows that increases in n_2 occur in a region of quite small probability.

The distribution of $\widehat{\theta}_1$ under other values of θ is shifted but has the same variance.

Properties of the MP design

The increase in n_2 in the "promising zone" has increased the power curve a little.

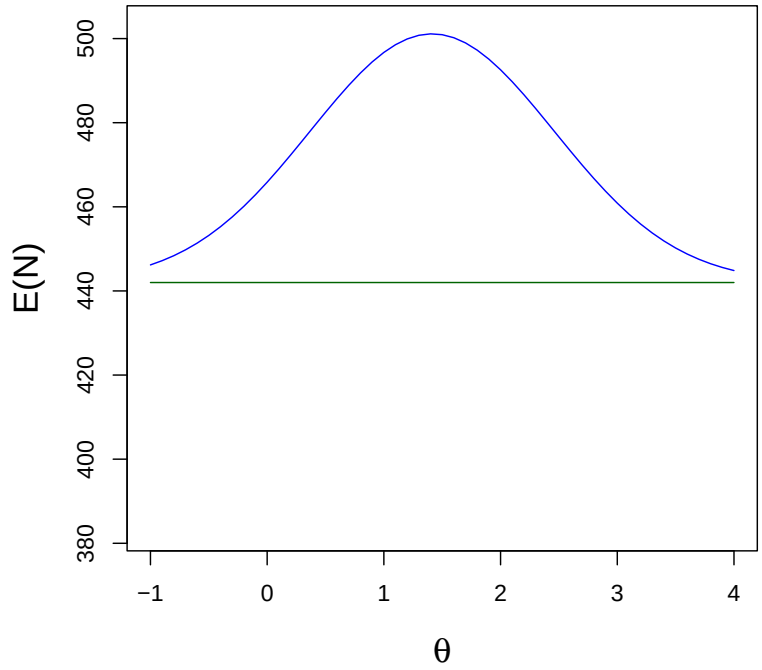


Given the limited range of values of $\widehat{\theta}_1$ for which n_2 is increased, only a small improvement in power can be expected.

Although it was stated that power 0.8 at $\theta = 1.6$ would be desirable, power at this effect size has only risen from 0.61 to 0.66.

Properties of the MP design

The MP design has $E_\theta(N)$ close to 500 at $\theta = 1.5$, compared to the original design's $E_\theta(N) = 442$ at all values of θ.



We could increase the sample size to give higher conditional power under $\theta = \widehat{\theta}_1$ or raise the maximum sample size beyond 884 — but such modifications give small additional power at the cost of a large increase in $E(N)$.

11. Alternatives to the MP design

Suppose we are satisfied with the overall power function attained by MP's design.

The same power curve can be achieved by other designs.

A fixed sample design

Emerson, Levin & Emerson (*Statistics in Medicine*, 2011) note that the same power is achieved by a fixed sample size study with 490 subjects.

This looks like an attractive option since, for effect sizes θ between 0.8 and 2.0, the expected sample size of the MP design is greater than 490.

There is more to the sample size distribution than $E_\theta(N)$

High variance in N is usually regarded as undesirable, so the wide variation in N for the MP design is a negative feature.

Perhaps variation in N is viewed more positively when investors in a small bio-tech company are thinking of adding resource to a study when it is most helpful?

A group sequential test

Despite the delayed response, we could apply a "standard" group sequential design.

Suppose an interim analysis takes place after 208 observed responses.

If the trial stops at this analysis, the sample size is taken as 416, counting all subjects treated thus far, even though only 208 have provided a response.

We consider an error spending design in the ρ-family (JT 2000, Ch. 7):

At analysis 1 *after 208 responses*

If $Z_1 \geq 2.54$	Stop, reject H_0
If $Z_1 \leq 0.12$	Stop, accept H_0
If $0.12 < Z_1 < 2.54$	Continue

At analysis 2 *after 514 responses*

If $Z_2 \geq 2.00$	Reject H_0
If $Z_2 < 2.00$	Accept H_0

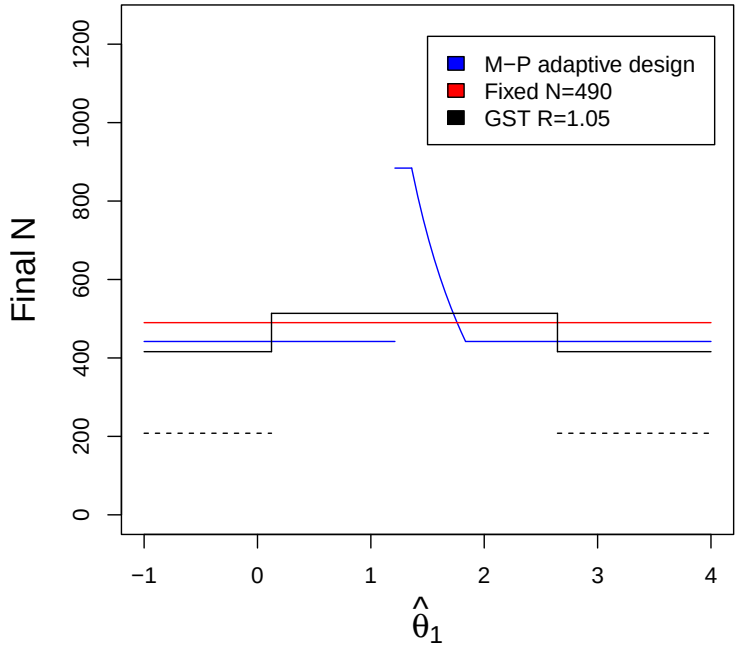
Sample size rules for MP, fixed and group sequential designs

Sample size for the MP design varies between 442 and 884.

The fixed sample size design has 490 observations.

The group sequential test can stop with a sample size of 416 or 514.

Since $514 = 490 \times 1.05$, it has an "inflation factor" of $R = 1.05$.

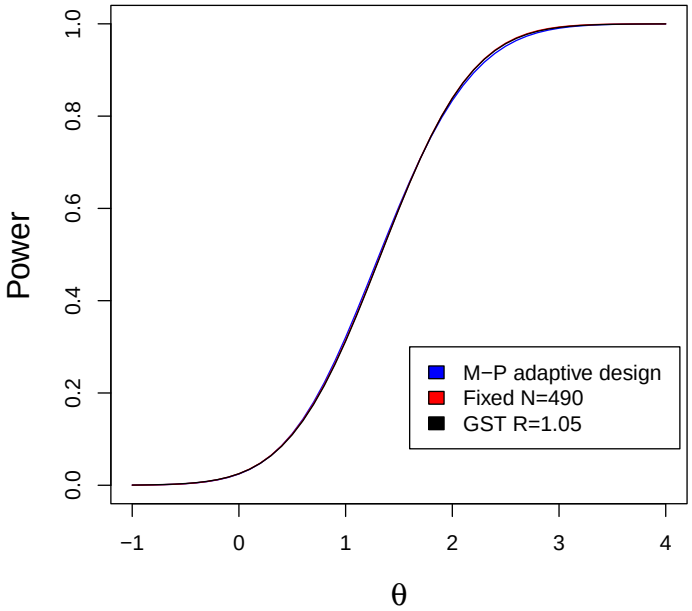


For the GST, the dashed line shows the 208 responses observed at the first analysis.

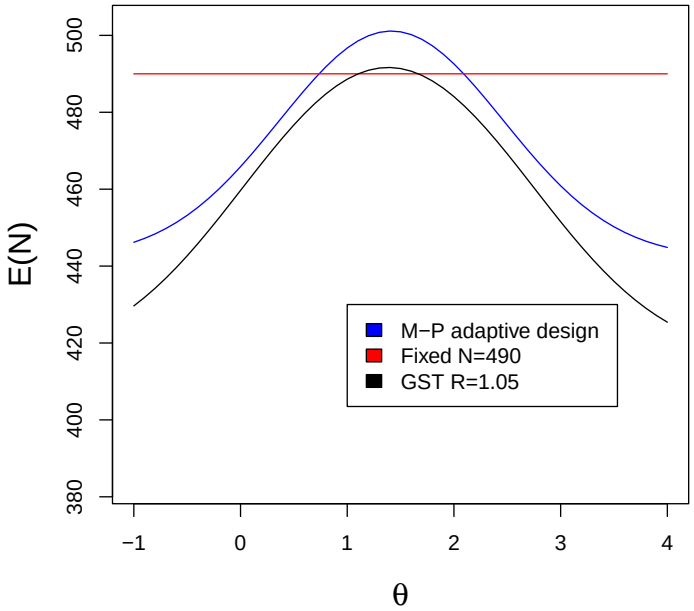
When a decision is made based on these 208 responses, the sample size is counted as $N = 416$, the number of subjects enrolled.

Comparison of designs

Power curves



$E_\theta(N)$ curves



All three designs have essentially the same power curve.

Clearly, it is quite possible to improve on the $E_\theta(N)$ curve of the MP design.

NB, Mehta & Pocock discuss two-stage group sequential designs but they only present an example with much higher power (and, thus, higher sample size).

Can we improve the trial design within the MP framework?

Why does the MP design have high $E_\theta(N)$ for its achieved power?

Mehta & Pocock describe their method as adding observations in situations where they will do the most good:

This seems a good idea, but the results are not so great,

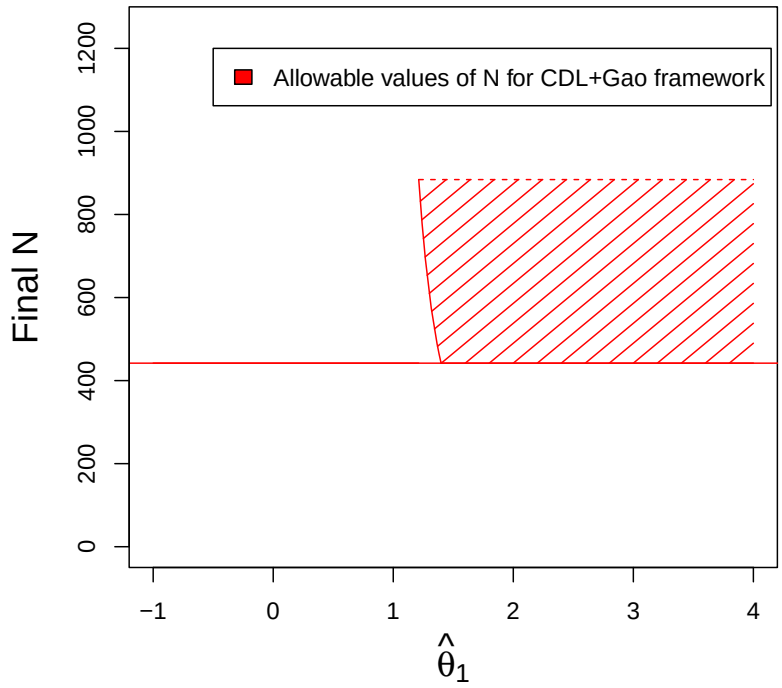
Can we work out how to do this effectively?

12. Deriving efficient sample size rules in the MP framework

We stay with MP's example and retain the basic elements of their design.

The interim analysis takes place after 208 observed responses.

A final sample size n_2^* is chosen based on $\widehat{\theta}_1$ (or equivalently Z_1).



Values of $n_2^* \in [442, 884]$ that satisfy the CDL+Gao conditions are allowed.

At the final analysis, we reject H_0 if $Z_2 > 1.96$, where Z_2 is calculated without adjustment for adaptation.

Efficient sample size rules in the MP framework

We shall assess the conditional power that an increase in sample size achieves.

Suppose $Z_1 = z_1$ and we are considering a final sample size n_2^* with

$$Z_2(n_2^*) \;=\; \frac{\hat{\theta}(n_2)}{\sqrt{\{4\sigma^2/n_2\}}}.$$

and conditional power under $\theta = \tilde{\theta}$

$$CP_{\tilde{\theta}}(z_1, n_2^*) = P_{\tilde{\theta}}\{Z_2(n_2^*) > 1.96 \,|\, Z_1 = z_1\}.$$

Setting γ as a "rate of exchange" between sample size and power, we shall:

Choose n_2^* to optimise a combined objective

$$\boldsymbol{CP_{\tilde{\theta}}(z_1, n_2^*) - \gamma(n_2^* - 442).}$$

We shall do this with $\tilde{\theta} = 1.6$, a value where we wish to "buy" additional power.

For consistency, we use the same γ when considering different values of z_1.

An overall optimality property

The rule that maximises $CP_{\tilde{\theta}}(z_1, n_2^*(z_1)) - \gamma n_2^*(z_1)$ for every z_1 also maximises, unconditionally,

$$P_{\theta=\tilde{\theta}}\left(\text{Reject } H_0\right) \; - \; \gamma E_{\tilde{\theta}}(N).$$

This can be seen by writing $P_{\theta=\tilde{\theta}}\left(\text{Reject } H_0\right) - \gamma E_{\tilde{\theta}}(N)$ as

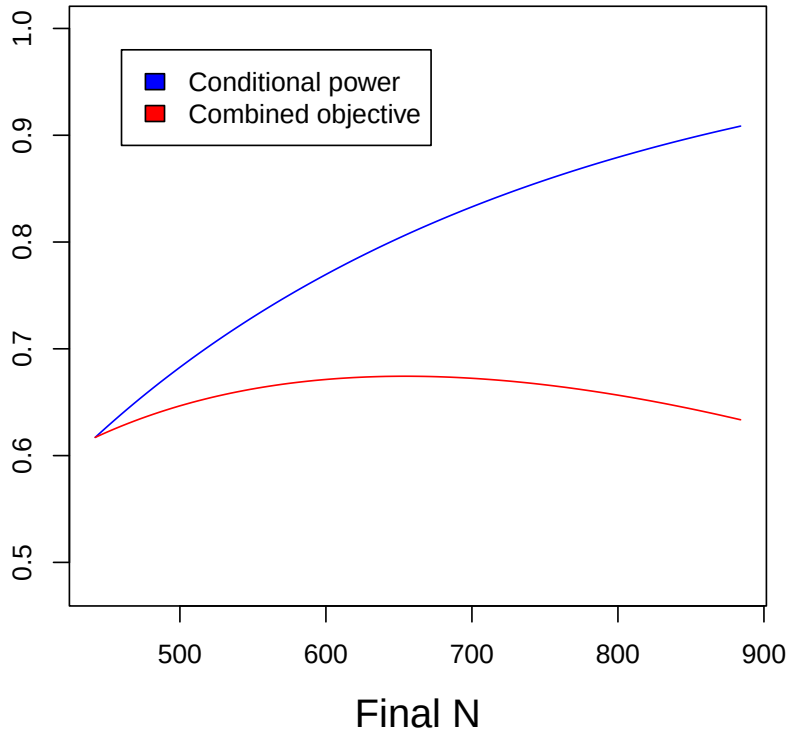
$$\int \left\{CP_{\tilde{\theta}}(z_1, n_2^*(z_1)) - \gamma n_2^*(z_1)\right\} f_{\tilde{\theta}}(z_1)\, dz_1,$$

where $f_{\tilde{\theta}}(z_1)$ denotes the density of Z_1 under $\theta = \tilde{\theta}$, and noting that we have minimised the integrand for each z_1.

We shall set $\gamma = 0.14/(4\,\sigma^2)$ to achieve the same power curve as the MP design.

So, the resulting procedure will have minimum possible $E_{\theta=1.6}(N)$ among all designs following the CDL+Gao framework that achieve power 0.658 at $\theta = 1.6$.

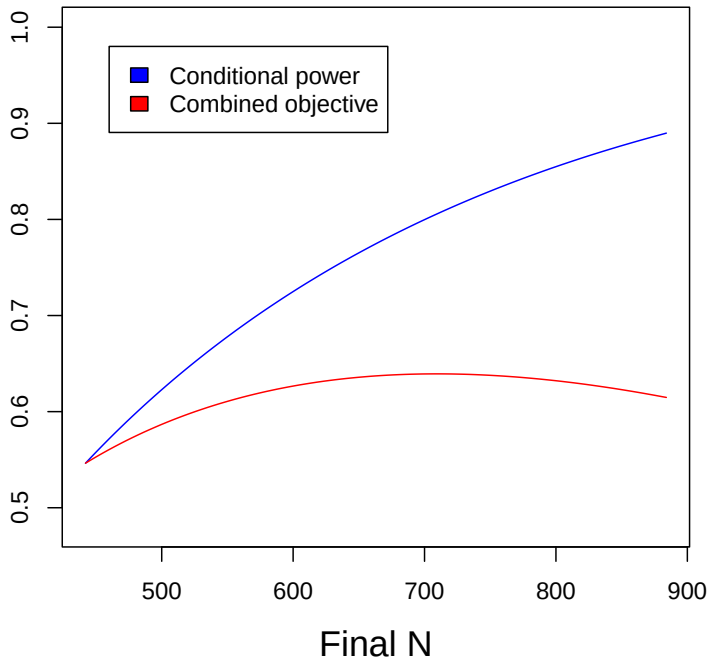
Plots for $\tilde{\theta} = 1.6,\ \gamma = 0.14/(4\,\sigma^2)$ and $\widehat{\theta}_1 = 1.5$



The objective $CP_{\tilde{\theta}}(z_1, n_2^*) - \gamma(n_2^* - 442)$ has a maximum at $n_2^* = 654$.

This value is similar to MP's choice of n_2^* when $\widehat{\theta}_1 = 1.5$.

Plots for $\tilde{\theta} = 1.6,\ \gamma = 0.14/(4\,\sigma^2)$ and $\widehat{\theta}_1 = 1.3$

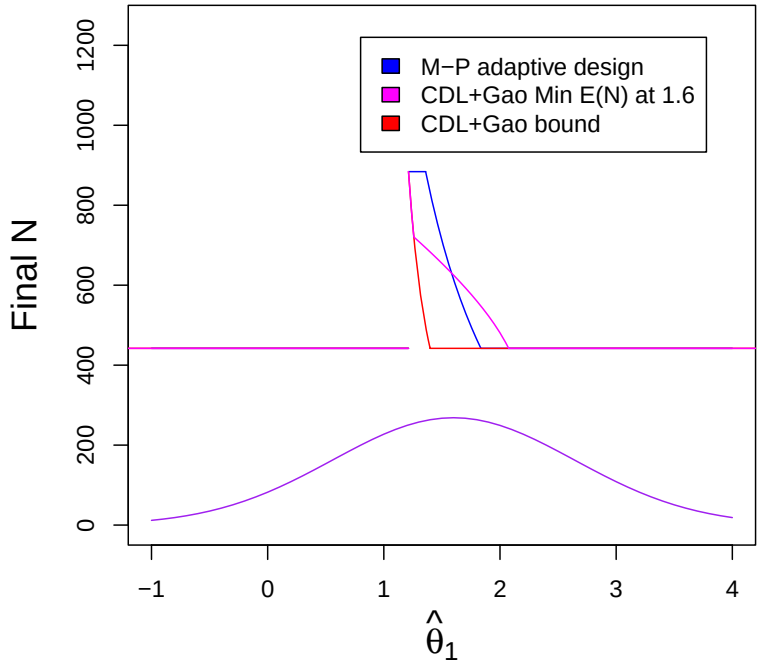


The conditional power curve is steeper and the optimum occurs at a higher n_2^*.

The objective $CP_{\tilde{\theta}}(z_1, n_2^*) - \gamma(n_2^* - 442)$ is maximised at $n_2^* = 707$.

In this case, MP's design takes the maximum permitted value of $n_2^* = 884$.

Optimal sample size rule for $\tilde{\theta} = 1.6$ and $\gamma = 0.14/(4\,\sigma^2)$



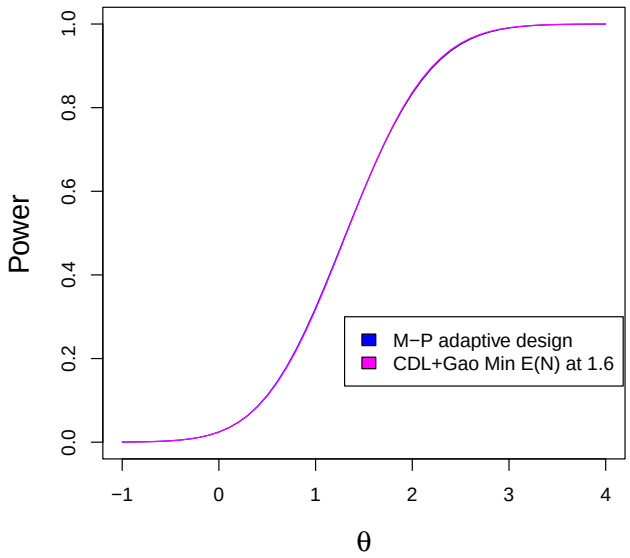
This rule gives power 0.658 at $\theta = 1.6$, the same as the MP design.

Decisions about the final sample size are based on a consistent comparison of the value of higher power and the cost of additional observations.

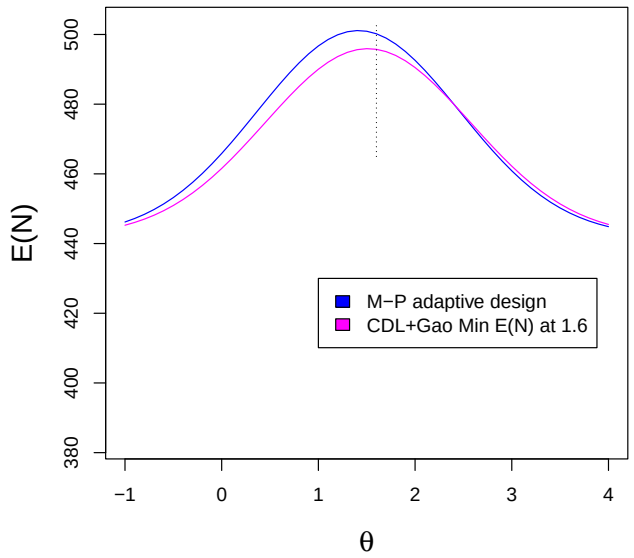
As $\widehat{\theta}_1$ decreases, sample size increases less steeply than for the MP design.

Efficient sample size rules in the MP framework

Power curves



$E_\theta(N)$ curves



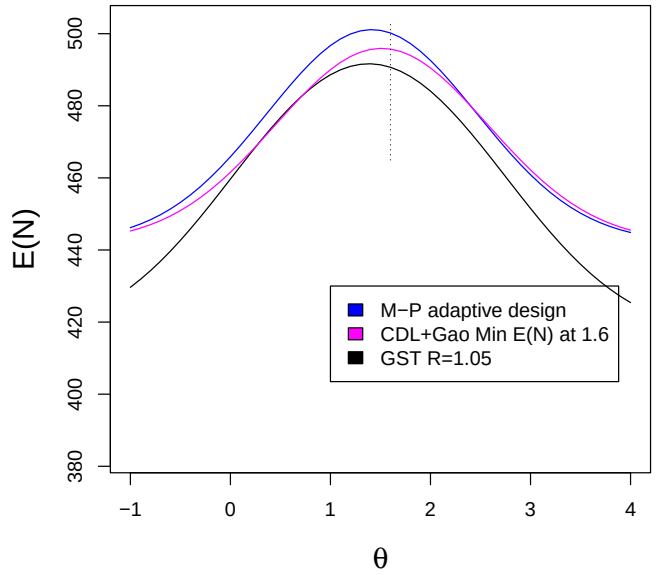
With the type I error rate at $\theta = 0$ fixed at 0.025, matching the MP design's power at one value of θ implies matching the whole power curve.

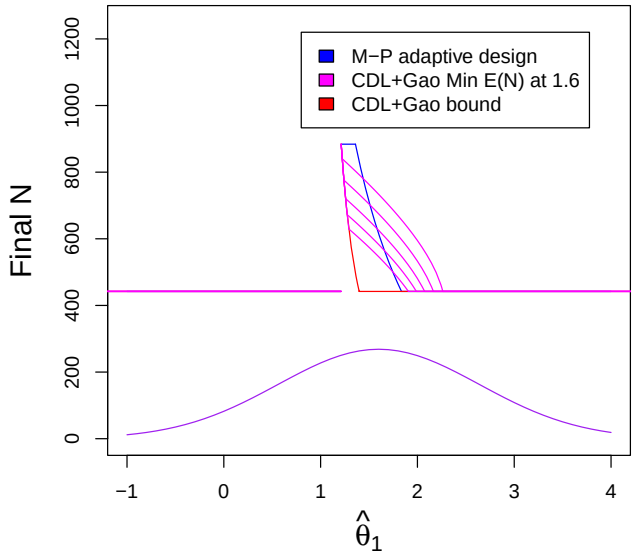
Our optimised design has the same power curve as the MP design and lower $E_\theta(N)$ (just about) at all θ values.

The reductions in $E_\theta(N)$ are modest — but given the optimality property of the sampling rule in the Mehta & Pocock framework, this is as good as it gets.

Further efficiency gains

Our new, optimised procedure still has higher $E_\theta(N)$ than the two-stage GST that ignores (but is charged for) pipeline data.





Shapes of optimised sample size rules suggest it would help to increase n_2^* at lower values of $\widehat{\theta}_1$ — but this is not permitted in the CDL+Gao framework.

The ***Conditional Probability of Rejection*** principle, or equivalently using a Bauer & Köhne (*Biometrics*, 1994) ***Combination Test*** does allow such adaptations.

13. Using the Conditional Probability of Rejection principle

Reference: Proschan & Hunsberger, (*Biometrics*, 1995)

On observing $\widehat{\theta}_1$, choose a new final sample size n_2^*.

Then, set the critical value for $Z_2(n_2^*)$ at the final analysis to maintain the Conditional Probability of Rejection (CPR) under $\theta = 0$ in the original design.

The overall type I error rate is the integral of the conditional type I error rate, and this remains the same.

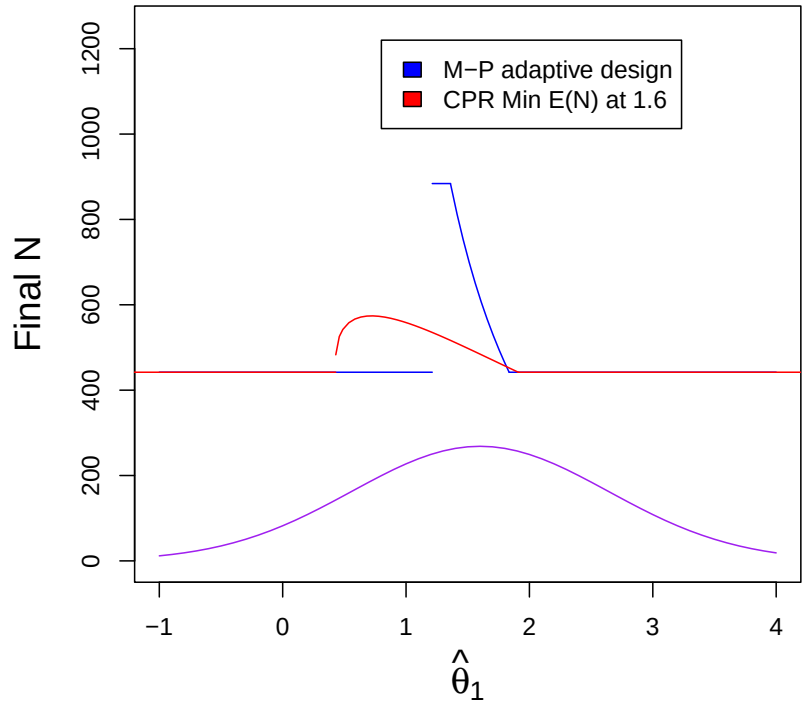
This can also be regarded as a "combination test" Bauer & Köhne (1994):

We reject H_0 if $w_1 Z_1 + w_2 \tilde{Z}_2 > 1.96$, where Z_1 is as before, $\tilde{Z}_2$ is based on the new data in Stage 2, w_1 and w_2 are pre-specified, and $w_1^2 + w_2^2 = 1$.

We can follow our previous strategy in this new framework and set n_2^* to maximise $CP_{\tilde{\theta}}(z_1, n_2^*) - \gamma(n_2^* - 442)$. Again, we shall use $\tilde{\theta} = 1.6$.

The resulting design has the minimum value of $E_{\tilde{\theta}}(N)$ among all designs in this larger class that achieve the same power under $\theta = \tilde{\theta}$.

Optimal sample size rule for a CPR design with $\tilde{\theta} = 1.6$



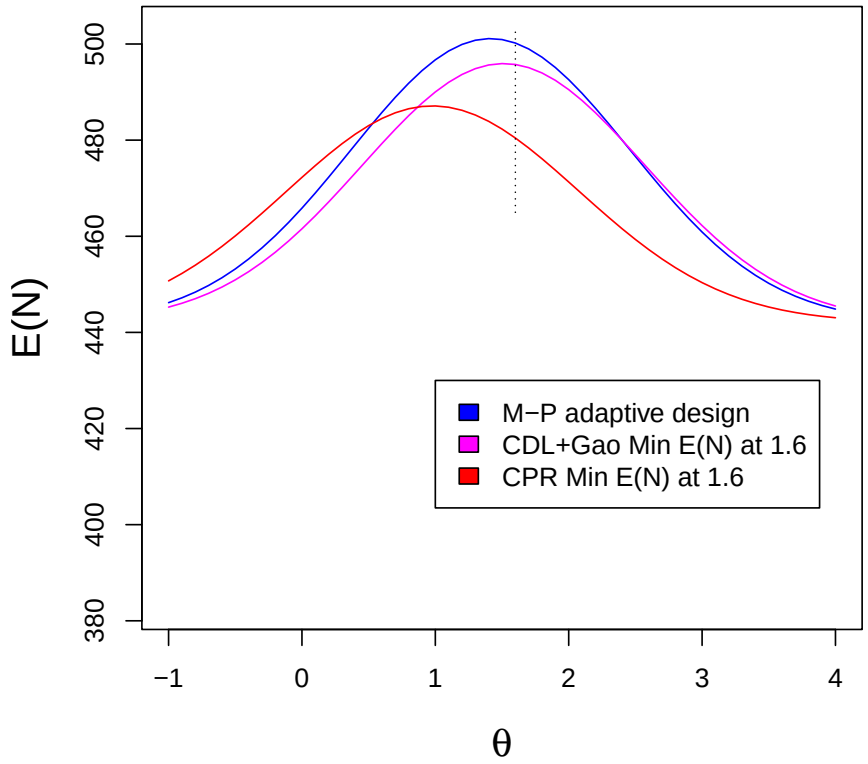
The rule with $\gamma = 0.25/(4\,\sigma^2)$ matches the MP test's power of 0.658 at $\theta = 1.6$.

Shapes of optimised sample size rules are ***very different*** from the MP design.

The best opportunities for investing additional resource are ***not*** in Mehta & Pocock's "promising zone".

Efficient sample size rules in the CPR framework

$E_\theta(N)$ curves



The CPR principle allows sample size increases for $\widehat{\theta}_1$ below the CDL+Gao region.

This leads to a useful reduction in $E_\theta(N)$ at $\theta = 1.6$.

Further extensions

1. We can allow recruitment to be terminated at the interim analysis, so the minimum final sample size is $n_2 = 416$, rather than 442. (Such a reduction in final sample size is *not* allowed in the CDL approach.)

2. We can use a general conditional type I error function (Proschan & Hunsberger, 1995) or, equivalently, a general Bauer & Köhne (1994) combination rule.

3. We can minimise other sample size criteria, such as a weighted sum or integral

$$\sum_i w_i \, E_{\theta_i}(N) \quad \text{or} \quad \int w(\theta) \, E_{\theta_i}(N) \, d\theta.$$

The resulting designs deal neatly with the "pipeline" subjects arising when there is a delayed response.

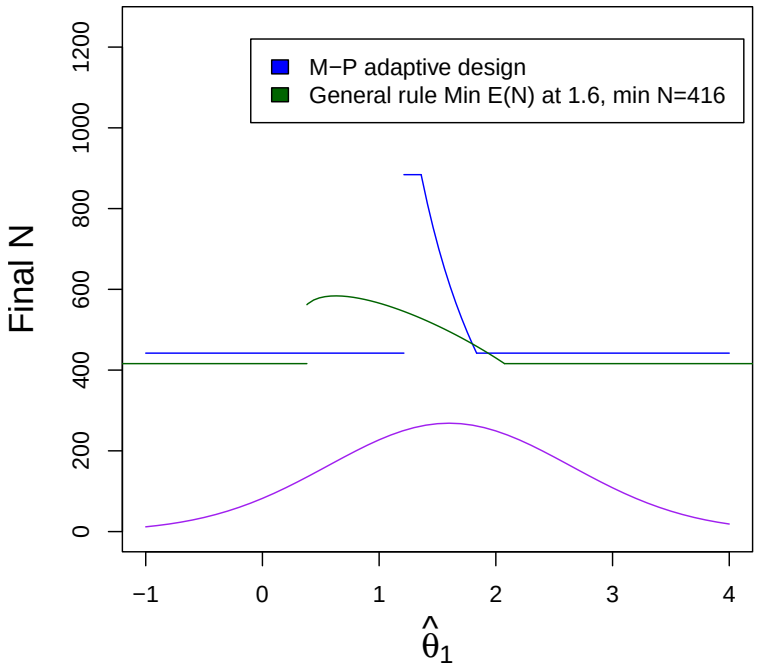
They will give the best possible sampling and decision rules with $n_1 = 208$ and n_2 in the range 416 to 884.

We could also aim for higher power, now we have a good way to achieve this.

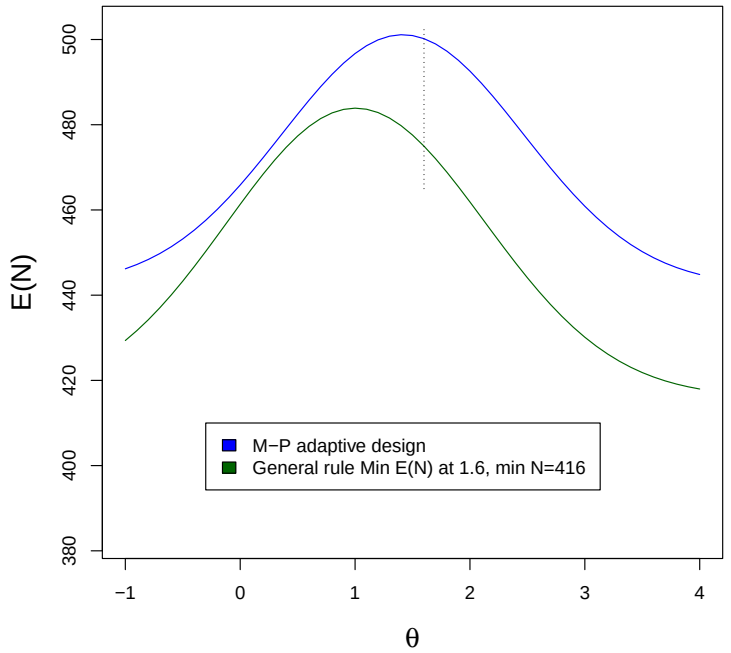
A general sampling rule with early termination of recruitment

We have followed (1) and (2) above in minimising $E_{\theta=1.6}(N)$.

Sample size rule



$E_\theta(N)$ curves



Reductions in $E_\theta(N)$ are mostly due to (1), which allows n_2 to be limited to 416.

The highest final sample sizes arise at values of $\widehat{\theta}_1$ below MP's "promising zone".

14. Relation between MP designs and Delayed Response GSTs

Consider using Hampson & Jennison's (2013) "Delayed Response GSTs" when there are just two analyses.

Either recruitment stops at analysis 1 and the final analysis occurs when all pipeline subjects have been observed,

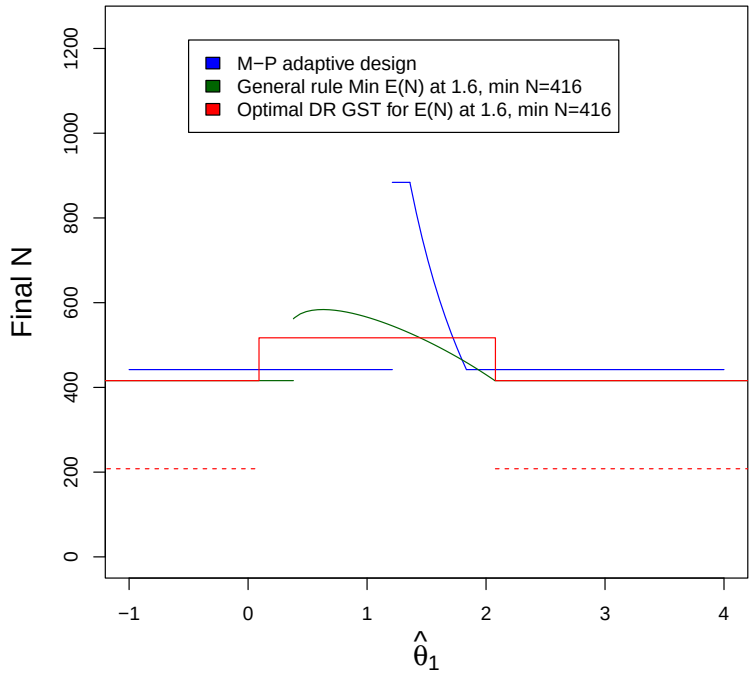
Or, an additional group of subjects is recruited and the final analysis has pipeline subjects plus these new subjects.

This is a special case of the designs we have been developing where only two values of n_2 are possible.

HJ optimised their DR GSTs to minimise criteria such as $E_{\theta=1.6}(N)$.

HJ also derived optimal adaptive DR GSTs which allow n_2 to take any value, up to a specified maximum: here, the class of possible designs is exactly the same as in our extension of the MP framework.

Sample size rules for designs minimising $E_{\theta=1.6}(N)$



The green line is for the optimal design in our extension of the MP framework. The trial can stop with 416 patients; Stage 1 and Stage 2 data can be combined in any way that protects type I error rate α.

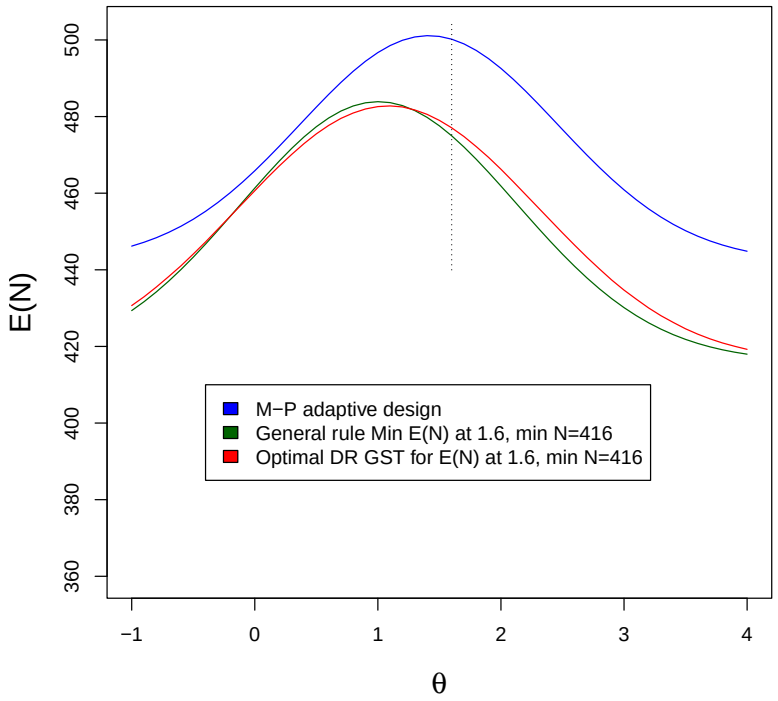
The optimal Adaptive DR GST has the same sample size function — and the same decision rule.

The red line is the sample size function for the non-adaptive DR GST which has two possible final sample sizes, 416 and 517, and minimises $E_{\theta=1.6}(N)$.

The red line is seen to be a step function approximation to the continuous function defined by the green line.

Plot of $E_\theta(N)$ for the optimal DR GST

The optimised non-adaptive DR GST has an almost identical $E_\theta(N)$ curve to the optimised adaptive design which uses the continuum of possible sample sizes.



As JT (*Biometrika*, 2006) found for an immediate response, there is minimal benefit from fine-tuning the total sample size in response to interim data.

Relation between MP designs and Delayed Response GSTs

For a trial with 2 stages, there is little difference in the operating characteristics of

A well-chosen design from our extension of the MP framework,

An optimised adaptive Delayed Response GST,

An optimised non-adaptive Delayed Response GST.

In practice, the fixed group sizes of the non-adaptive Delayed Response GST may make it easier to plan and manage a trial.

In presenting their methods, Mehta & Pocock

Talk in terms of the appropriate sample size to complete the trial,

Avoid the notion of “reversing” a provisional decision,

Introduce the attractive terminology “the promising zone”.

15. Conclusions

1. MP use the Chen, DeMets & Lan (2004) approach, choosing sample size by a conditional power rule: this does not yield very efficient designs.

We have developed MP's idea of spending resources where they have the greatest benefit — and obtained efficient adaptive designs.

If used well, the adaptive approach (start small, then ask for more) can give good trial designs — but there are pitfalls to be avoided!

2. The optimal design in our most general extension of MP's framework is very similar to a "Delayed Response GST" (Hampson & Jennison, 2013).

Using a Delayed Response GST offers the benefits of established group sequential methodology and its extensions, e.g., more than two analyses, error spending designs.