Adaptive Sample Size Modification in Clinical Trials:

Start Small then Ask for More?

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

BASS XVIII, Savannah GA

November 2011

Planning to adaptively extend a trial

Issues may be:

1. Dispute over the effect size to use in setting power

   Anticipated effect size Δ_1,

   Minimum effect of interest Δ_2,

   with $\Delta_1 > \Delta_2$;

2. Uncertainty over the value of a nuisance parameter (e.g., response variance);

3. Co-primary endpoints with $\Delta_1 > \Delta_2$, e.g., PFS and OS in late stage cancer;

4. Testing for Superiority (with effect size Δ_1) or Non-inferiority (with margin Δ_2);

5. General population (effect size Δ_2) or targeted population (effect size Δ_1) — "Enrichment".

What's wrong with simply adding observations?

The false positive rate α is increased (selection bias).

FDA Guidances forbid this!

ICH E9 (September 1998),

Adaptive Design (February 2010).

Paired "cat" example.

Protection of Type I error (α)

- ICH E9 (p.25):

  "The procedure selected should always ensure that the overall probability of type I error is controlled."

- PhRMA White paper (2006, *J. Biopharmaceutical Statistics*):

  "The key issue in most contexts is preservation of the Type I error rate."

- Pocock and Hughes (1989, *Controlled Clinical Trials*, p. 211S):

  "Control of Type I error is a vital aid to prevent a flood of false positives into the medical literature."

Designing a trial with good power and sample size

Our topic is the design of a clinical trial that will

Protect the type I error rate and achieve sufficient power,

Using 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

We have previously compared proposals for group sequential tests (GSTs) and adaptive designs, and ***concluded in favour of GSTs***.

See, for example, our papers in

Statistics in Medicine (2003, 2006),

Biometrika (2006),

Biometrics (2006)

and Chapter 5 of

Handbook of Adaptive Designs in Pharmaceutical and Clinical Development (2011).

Two issues in the debate about adaptive designs are

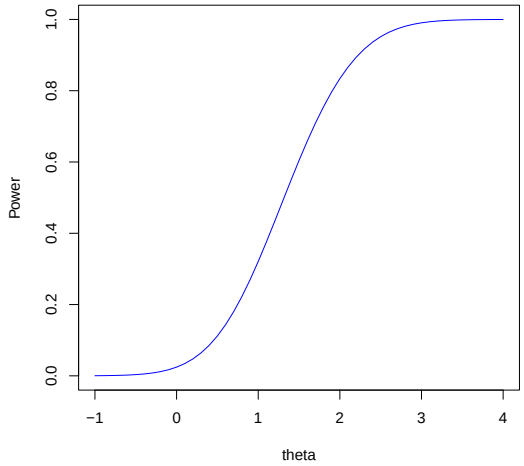
Emphasis on the *conditional* power of an adaptive design,

How to deal with uncertainty about the likely effect size.

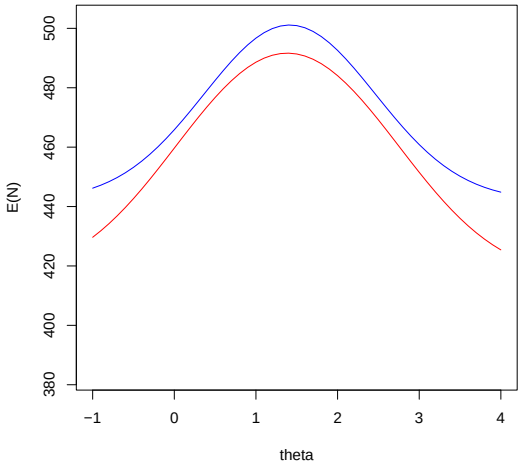
Designing a clinical trial

Let θ denote the treatment effect for experimental treatment versus control.

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 Designs and Group Sequential Tests

Suppose we specify objectives

- Type I error rate α,
- Power $1 - \beta$ at $\theta = \delta$,
- Values of θ at which low $E_\theta(N)$ is desired

and constraints

- At most K analyses,
- Maximum sample size R times the fixed sample size.

How can an adaptive or group sequential design best achieve these goals?

Since certain classes of designs are nested within each other, there is an ordering of optimal designs.

Adaptive Designs and Group Sequential Tests

Given α, $1-\beta$, δ, K, R and an $E_\theta(N)$ criterion, designs are nested:

$$\begin{array}{c} K\text{-group} \\ \text{GSTs} \end{array} \quad \subset \quad \begin{array}{c} K\text{-group} \\ \text{Adaptive GSTs} \end{array}$$

Here, an "Adaptive GST" is a group sequential test in which future group sizes are allowed to depend on the responses observed thus far.

Jennison & Turnbull (*Biometrika*, 2006) showed the efficiency gain of optimal Adaptive K-group GSTs over optimal K-group GSTs is small (around 2%).

Since optimal designs in the two classes are close in efficiency, for *any* K-group adaptive design, there is a (simpler) K-group GST of almost equal efficiency.

In our experience, many adaptive designs in the literature use sub-optimal sample size rules and are ***significantly less efficient*** than well-chosen GSTs.

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

Mehta & Pocock (*Statistics in Medicine*, 2011) recently published the paper

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

Their conclusions are counter to the findings we have reported.

In their example, response is measured some time after treatment, so many patients have been treated but are yet to produce a response at an interim analysis.

Delayed response is common — and not easily dealt with by standard GSTs.

Assessing the design proposed by Mehta & Pocock (MP) in their Example 1, we will

Test our conclusions about adaptive designs and GSTs in a new setting,

Suggest a new framework for deriving adaptive and group sequential designs,

Illuminate and address the issue of delayed response.

Outline of talk

1. Mehta & Pocock's Example 1
2. Mehta & Pocock's design for this example
3. Alternative fixed and group sequential designs
4. Deriving efficient designs in Mehta & Pocock's framework
5. Efficient designs using the conditional probability of rejection principle
6. Optimising in a general class of designs
7. Relation to proposed delayed response GSTs (Hampson & Jennison)
8. Conclusions
9. Connections to other research

1. 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.

Denote responses by

$Y_{Bi},\ i = 1, 2, \ldots,$ on the new treatment,

$Y_{Ai},\ i = 1, 2, \ldots,$ on the comparator treatment.

Responses are assumed to be normally distributed with variance 7.5^2, so each

$$Y_{Ai} \sim N(\mu_A, \sigma^2) \quad \text{and} \quad Y_{Bi} \sim N(\mu_B, \sigma^2),$$

where $\sigma^2 = 7.5^2$. The treatment effect is

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

Mehta & Pocock's Example

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

The final analysis will reject H_0: $\theta \leq 0$ if $Z_2 > 1.96$, where

$$Z_2 = \frac{\overline{Y}_B(n_2) - \overline{Y}_A(n_2)}{\sqrt{\{4\sigma^2/n_2\}}}$$

and $\overline{Y}_A(n_2)$ and $\overline{Y}_B(n_2)$ are treatment means from a total of n_2 observations.

This gives a test with one-sided 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. However, the sponsors will only increase sample size if interim results are "promising".

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

Due to uniform staggered accrual and the 26 week delay in obtaining a response, another 208 subjects will be treated by this time and await 26 weeks follow up.

Recruitment continues. The final data set will contain at least the original 442 subjects: with "promising" data, an increase up to 884 subjects is permitted.

Increasing the sample size

At the interim analysis we observe

$$\widehat{\theta}_1 = \overline{Y}_B(n_1) - \overline{Y}_A(n_1) \quad \text{and} \quad Z_1 = \frac{\widehat{\theta}_1}{\sqrt{\{4\sigma^2/n_1\}}}.$$

Define conditional power $CP_\theta(z_1)$ to be the probability the final test — with the original $n_2 = 442$ observations — rejects H_0, given $Z_1 = z_1$ and effect size θ, i.e.,

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

MP's adaptive design consider three cases at the interim analysis:

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* $n_2 = 442$.

It is crucial to protect α when increasing sample size in the promising zone.

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 a final test is carried out without adjustment for this adaptation.

Thus, H_0 is rejected if

$$Z_2(n_2^*) \;=\; \frac{\overline{Y}_B(n_2^*) - \overline{Y}_A(n_2^*)}{\sqrt{\{4\sigma^2/n_2^*\}}} \;>\; 1.96.$$

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

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

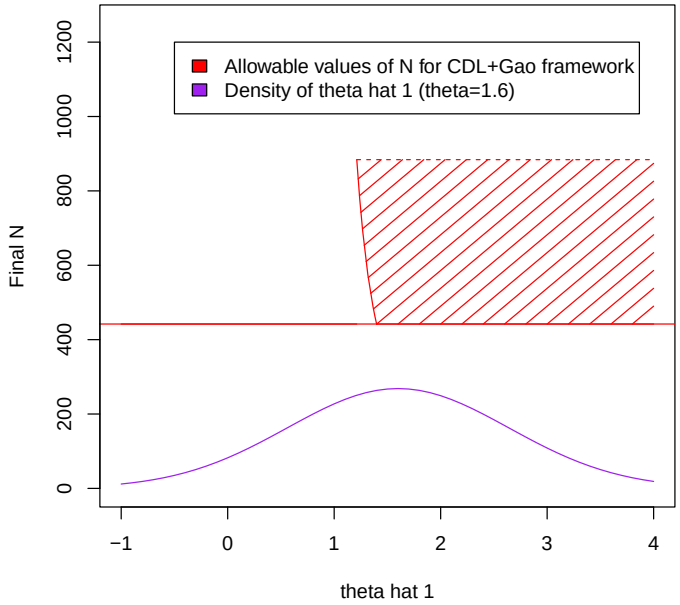
then the type I error probability will not increase.

(In general, adaptive changes to sample size are liable to increase type I error rate.)

Gao's extension of the CDL method

Gao et al. extend the CDL result to lower values of $\widehat{\theta}_1$.

They show the type I error rate does not increase as long as a sufficiently high value is chosen for n_2^*.

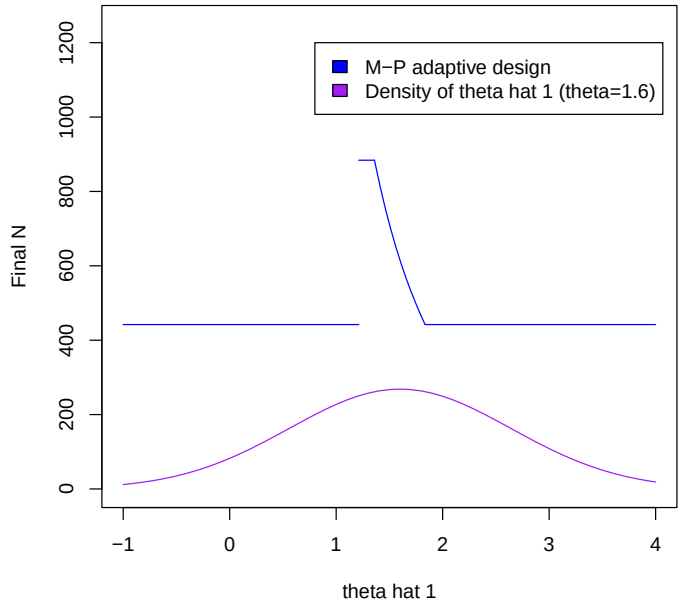


With an upper limit of $n_2^* = 884$, the final sample sizes permitted when using the CDL+Gao approach are as shown in the figure.

Now, n_2 can be increased for cases down as far as $CP_{\hat{\theta}_1}(z_1) = 0.365$.

2. The MP 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.

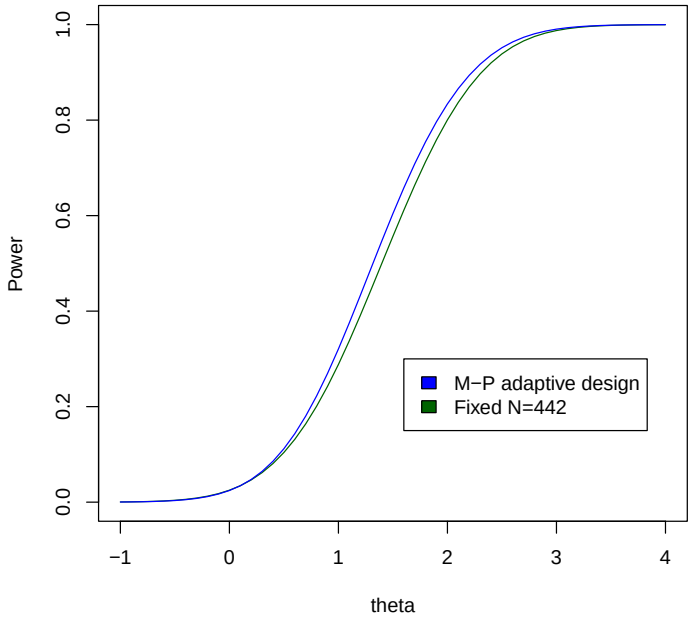


Comparison with the distribution of $\widehat{\theta}_1$ under $\theta = 1.6$ shows that increases in n_2 occur in a region of quite small probability.

(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.

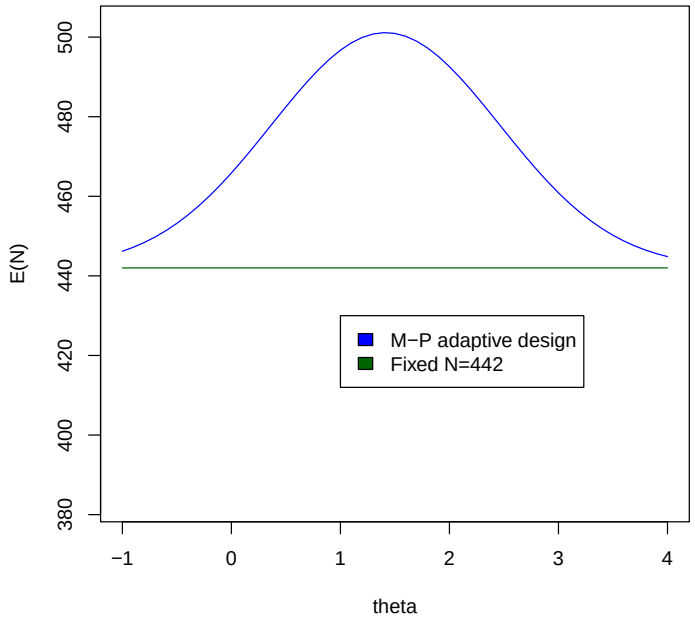


Given the limited range of values of $\widehat{\theta}_1$ for which n_2 is increased, only a small improvement in power can be expected.

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 cost of higher power is an increase in expected sample size.



The MP design could be modified by:

- Aiming for higher conditional power under $\theta = \widehat{\theta}_1$, or
- Raising the maximum for n_2 above 884.

However, the resulting gains in power are small for the increases in $E_\theta(N)$.

3. Alternatives to the MP design

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

Could the same power have been achieved more efficiently by another design?

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 — fewer than the expected sample size of the MP design for effect sizes θ between 0.8 and 2.0.

A group sequential test

Despite the delayed response, we can still consider a group sequential design with an interim analysis after 208 responses.

If the trial stops to reject H_0 or accept H_0 at the first analysis, the actual sample size, including all subjects treated thus far, must be counted as 416.

A group sequential test

We consider an error spending design for a one-sided test using a ρ-family error spending function with $\rho = 2$ (Jennison & Turnbull, 2000, Ch. 7).

This design has an interim analysis after 208 responses and a final analysis after 514 responses. The stopping rule and decision rule are:

At analysis 1

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

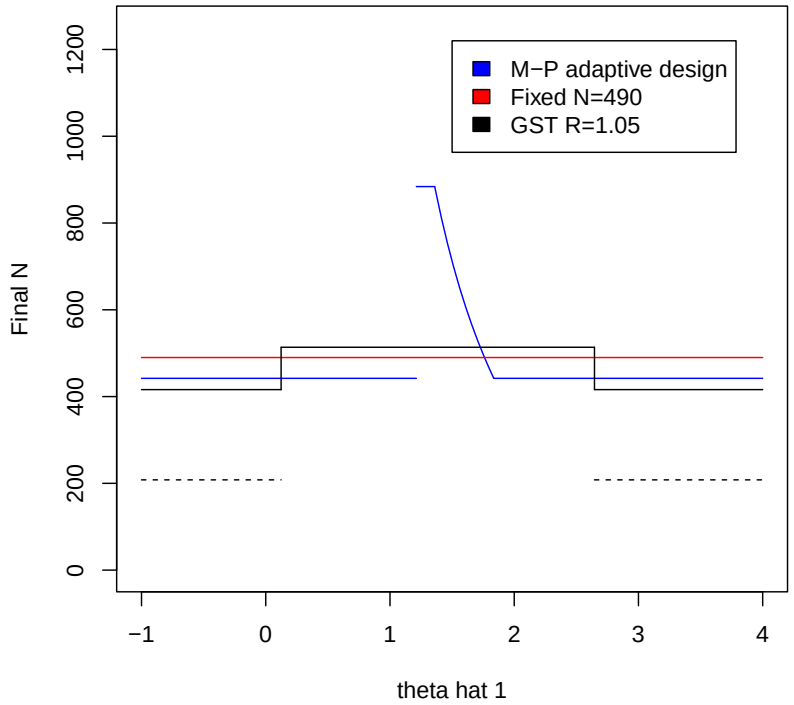
At analysis 2

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

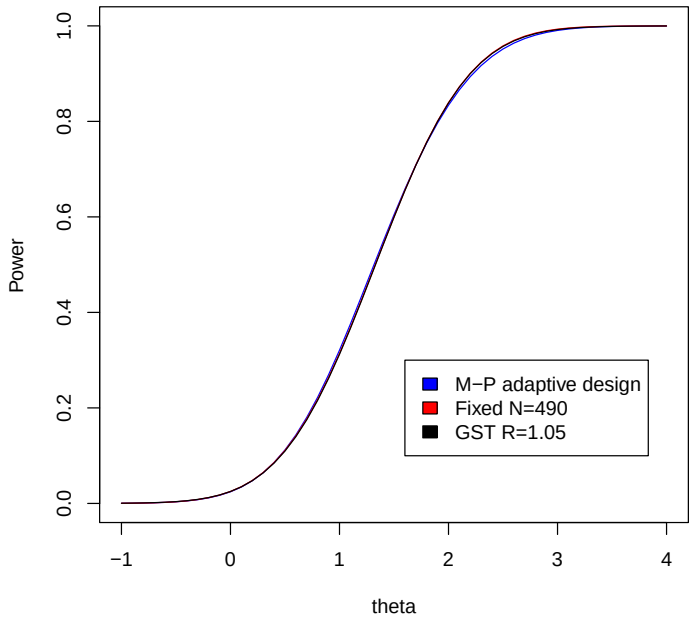
For the group sequential test (GST), a sample size of 416 is charged on stopping at the first analysis with 208 observed responses.

The GST's maximum sample size is a factor $R = 1.05$ times the 490 needed to achieve the same power in a fixed sample design.

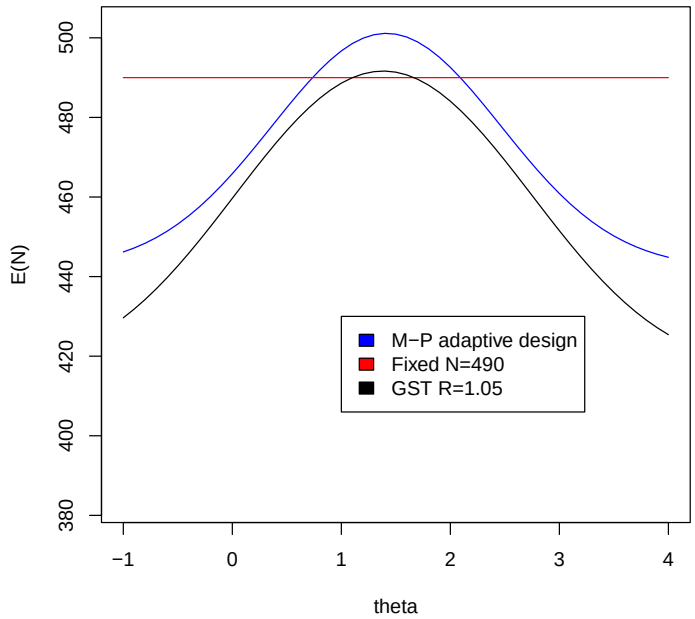


Comparison of designs

Power curves



$E_\theta(N)$ curves



All three designs have essentially the same power curve.

Clearly, it is quite possible to improve on the efficiency of the MP design.

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

Questions

Improved designs in the MP framework

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

Adding observations when they will do the most good seems a reasonable idea. Can we work out how to do this efficiently?

Group sequential tests for a delayed response

Although the GST does well, it does not make use of the data eventually obtained from subjects “in the pipeline” at analysis 1.

This is inefficient. There may also be a problem if later data do not support the conclusion at the interim analysis.

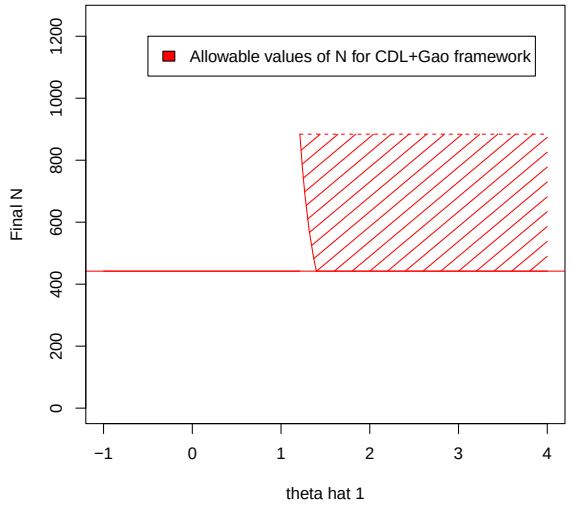
Can GSTs be extended to give a proper treatment of “pipeline” data and what is the most efficient way to do this?

4. 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 specify a sample size rule by matching conditional power against sample size.

Suppose we observe $Z_1 = z_1$ and choose final sample size n_2^*. Let

$$Z_2(n_2^*) \;=\; \frac{\overline{Y}_B(n_2^*) - \overline{Y}_A(n_2^*)}{\sqrt{\{4\sigma^2/n_2^*\}}}.$$

Denote conditional power under $\theta = \tilde{\theta}$, given $Z_1 = z_1$ and sample size n_2^*, by

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

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

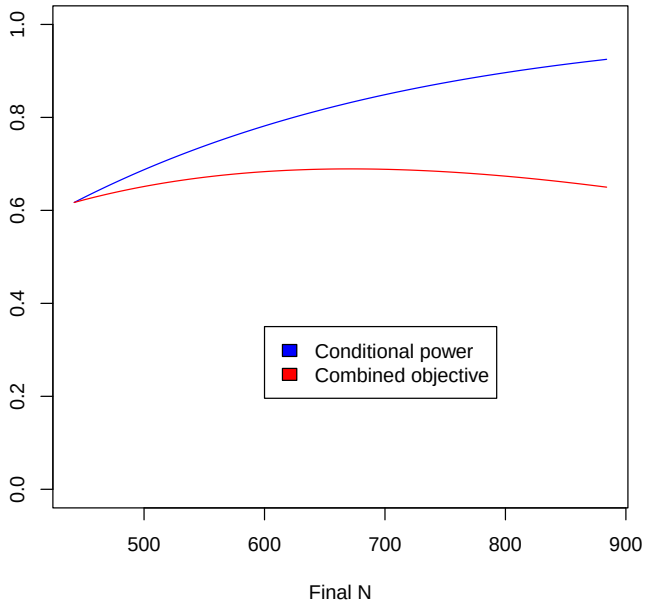
Choose n_2^* to optimise a combined objective

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

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

Plots of $CP_{\tilde{\theta}}(z_1, n_2^*)$ and $CP_{\tilde{\theta}}(z_1, n_2^*) - \gamma(n_2^* - 442)$

For the case $\tilde{\theta} = 1.6$, $\gamma = 0.14/(4\,\sigma^2)$ ***and*** $\widehat{\theta}_1 = 1.5$



We shall see that using $\gamma = 0.140/(4\sigma^2)$ gives similar power to the MP design.

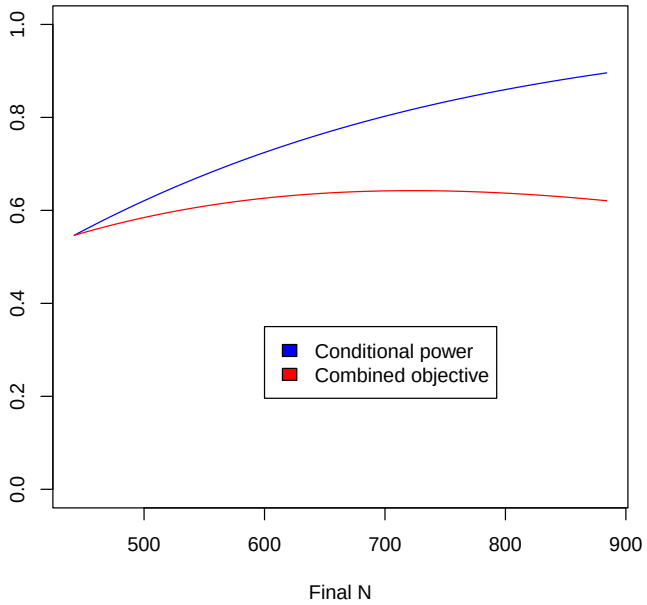
The objective $CP_{\tilde{\theta}}(z_1, n_2^*) - \gamma(n_2^* - 442)$ has a maximum at $n_2^* = 654$.

At the optimal value of n_2^*, the slope of the conditional power curve is γ.

The outcome is similar to MP's choice of n_2^*.

Plots of $CP_{\tilde{\theta}}(z_1, n_2^*)$ and $CP_{\tilde{\theta}}(z_1, n_2^*) - \gamma(n_2^* - 442)$

For the case $\tilde{\theta} = 1.6$, $\gamma = 0.14/(4\,\sigma^2)$ ***and*** $\widehat{\theta}_1 = 1.3$



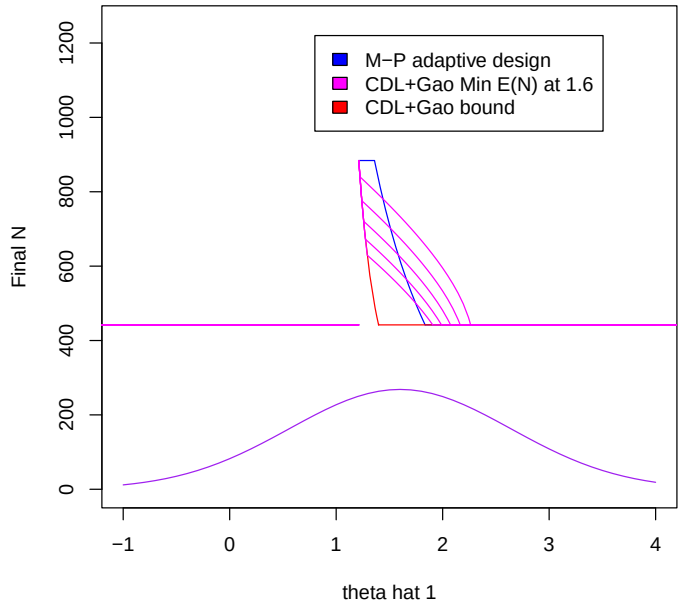
Here, the slope of the conditional power curve is higher and the optimum, where the derivative is γ, occurs later.

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 $n_2^* = 884$.

Efficient sample size rules in the MP framework

Sample size rules to optimise $CP_{\tilde{\theta}}(z_1, n_2^*) - \gamma(n_2^* - 442)$ ***for*** $\tilde{\theta} = 1.6$ ***and*** $4\sigma^2\gamma = 0.10, 0.12, \ldots, 0.18$

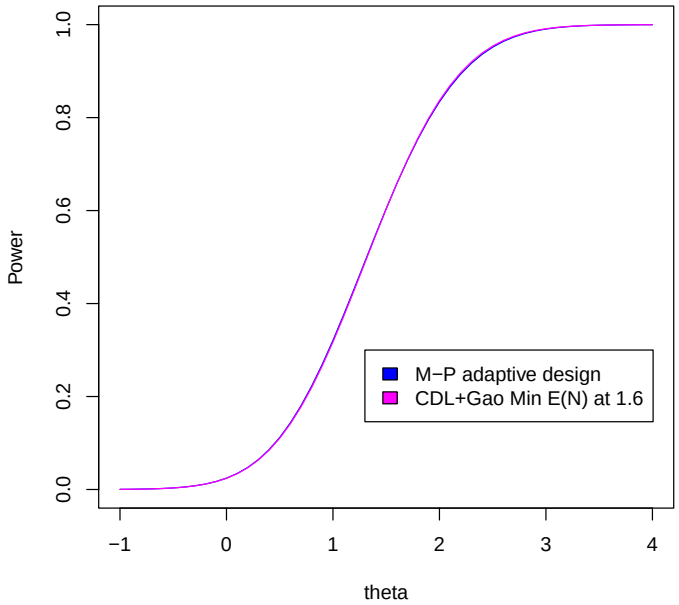


The sample size rules for various values of γ have similar — but different — shapes from that of the MP design.

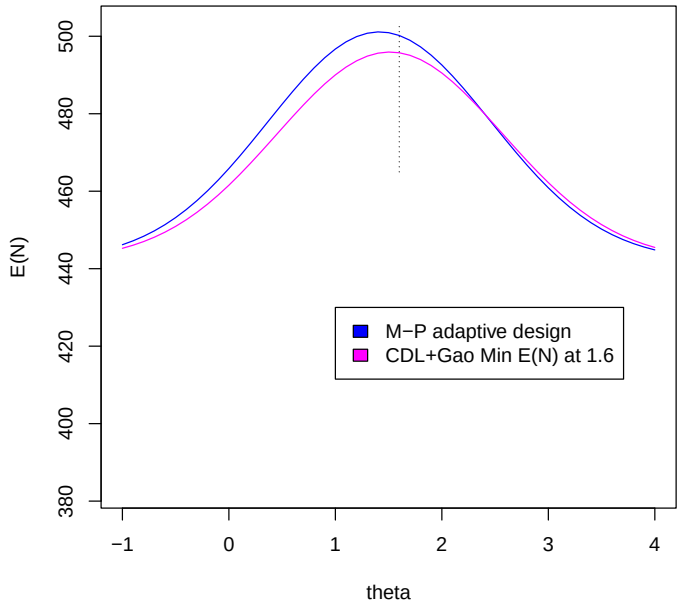
The rule for $\gamma = 0.140/(4\sigma^2)$ gives power 0.658 at $\theta = 1.6$, the same as the MP design.

Efficient sample size rules in the MP framework

Power curves



$E_\theta(N)$ curves



There is, essentially, a one parameter family of power curves for procedures with type I error 0.025 at $\theta = 0$. Thus matching the MP design's power at one value of θ implies matching its whole power curve.

Our new design has the same power curve as the MP design and lower $E_\theta(N)$.

An overall optimality property

Given a sample size rule for choosing n_2^* as a function of z_1,

$$(\text{ Power at } \theta = \tilde{\theta}\,) \;-\; \gamma E_{\tilde{\theta}}(N) \;=$$

$$\int \{CP_{\tilde{\theta}}(z_1, n_2^*(z_1)) - \gamma n_2^*(z_1)\}\; f_{\tilde{\theta}}(z_1)\, dz_1, \qquad (1)$$

where $f_{\tilde{\theta}}(z_1)$ denotes the density of Z_1 under $\theta = \tilde{\theta}$.

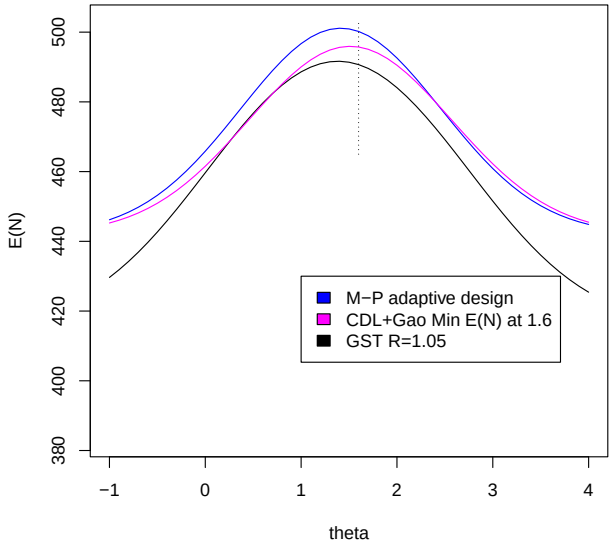
A sample size rule that maximises $CP_{\tilde{\theta}}(z_1, n_2^*(z_1)) - \gamma n_2^*(z_1)$ for every z_1 must also maximise (1).

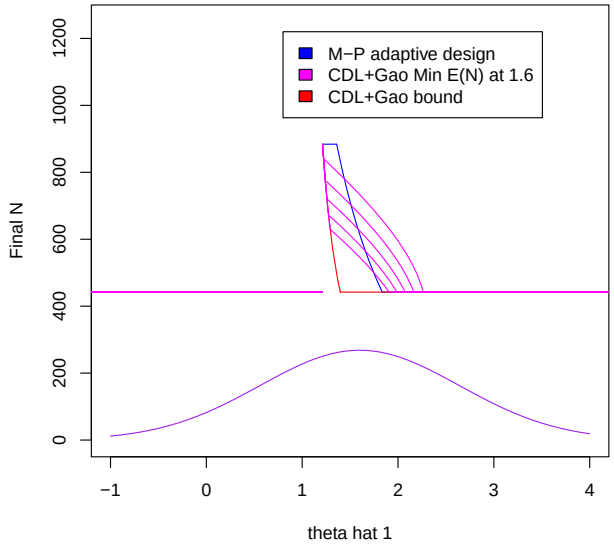
It follows that such a rule has minimum $E_{\tilde{\theta}}(N)$ among all rules that achieve the same power under $\theta = \tilde{\theta}$.

Hence, our rule for $\gamma = 0.14$ has the lowest possible $E_{\theta=1.6}(N)$ among all rules following the CDL+Gao framework that achieve power 0.658 at $\theta = 1.6$.

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.





The conservatism of the CDL construction is not the root of the problem.

Shapes of the 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 does allow such adaptations.

5. Using the Conditional Probability of Rejection principle

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

The initial fixed sample size design with $n_2 = 442$ has type I error probability α. Thus

$$\int CP_{\theta=0}(z_1)\ f_{\theta=0}(z_1)\, dz_1\ =\ \alpha.$$

If we define a new procedure which preserves the conditional probability of rejecting H_0 under $\theta = 0$ (the CPR) for every value of z_1, then the overall type I error rate will remain the same.

(This can also be regarded as a "weighted inverse normal combination test".)

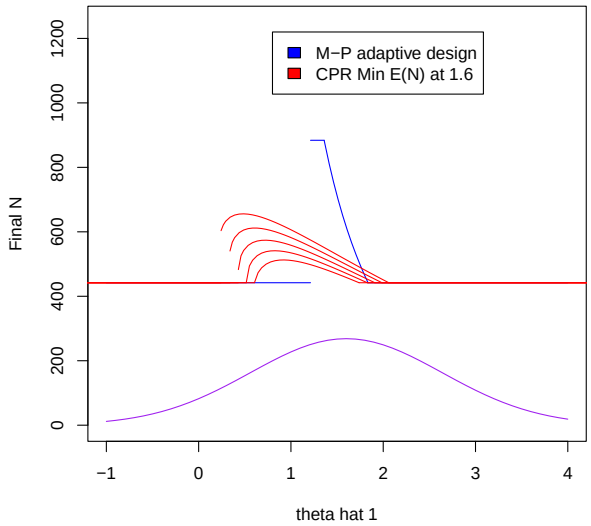
Thus, we can choose a new final sample size n_2^* for each z_1 and set a critical value for $Z_2(n_2^*)$ at the final analysis to maintain the CPR.

Again, we can set n_2^* to maximise $CP_{\tilde{\theta}}(z_1, n_2^*) - \gamma(n_2^* - 442)$, where $\tilde{\theta} = 1.6$.

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

Efficient sample size rules in the CPR framework

Sample size rules optimising $CP_{\tilde{\theta}}(z_1, n_2^*) - \gamma(n_2^* - 442)$ ***for*** $\tilde{\theta} = 1.6$ ***and*** $4\sigma^2\gamma = 0.21, 0.23, \ldots, 0.29$



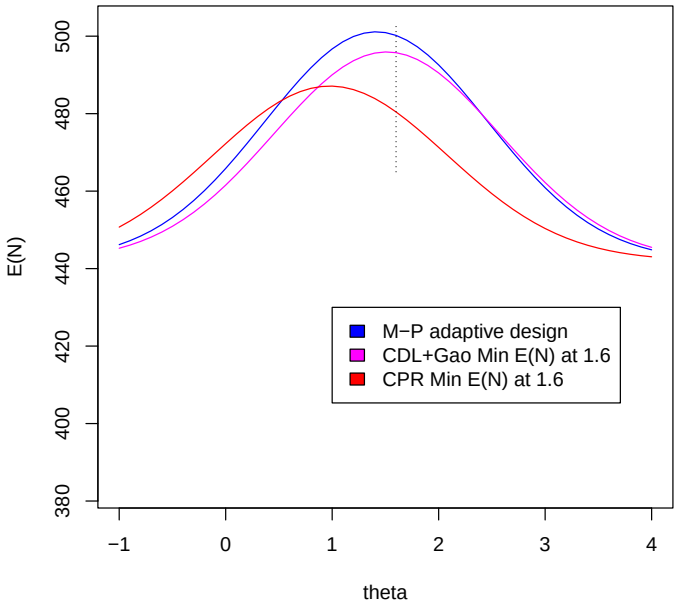
The middle rule, with $\gamma = 0.250/(4\sigma^2)$), matches the MP design's power of 0.658 at $\theta = 1.6$.

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

The most productive opportunities for investing additional resource are *not* in the "promising zone" identified by Mehta & Pocock.

Efficient sample size rules in the CPR framework

$E_\theta(N)$ curves



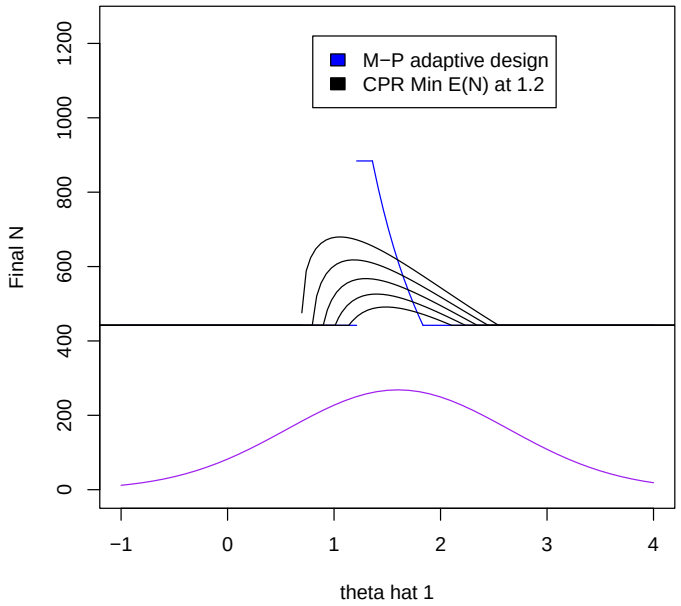
The CPR design provides a further reduction in $E_\theta(N)$ at $\theta = 1.6$.

Using the CPR principle for higher values of $\widehat{\theta}_1$ helps a little — but most of the improvement in $E_\theta(N)$ comes from being permitted to increase sample size for values of $\widehat{\theta}_1$ below the MP design's "promising zone".

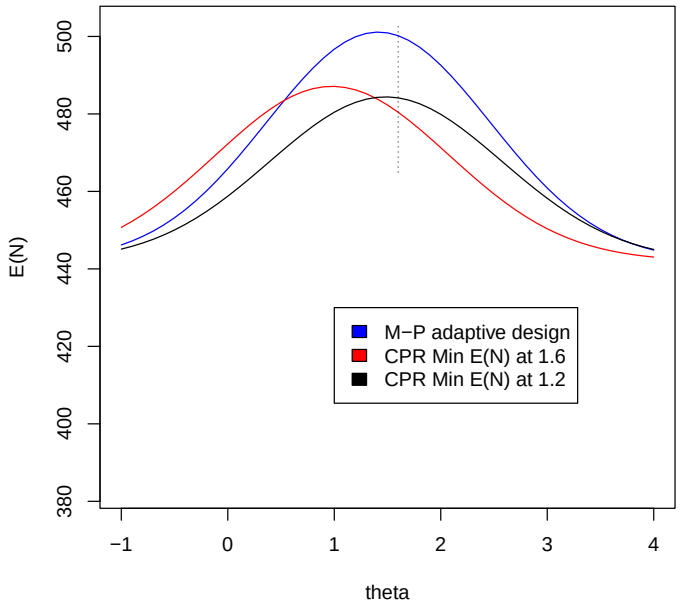
Optimising $CP_{\tilde{\theta}}(z_1, n_2^*) - \gamma(n_2^* - 442)$ with $\tilde{\theta} = 1.2$

A closer match to the shape of the MP design's $E_\theta(N)$ curve is obtained by maximising $CP_{\tilde{\theta}}(z_1, n_2^*) - \gamma(n_2^* - 442)$ for $\tilde{\theta} = 1.2$.

Sample size rules



$E_\theta(N)$ curves



The optimal CPR design for $\tilde{\theta} = 1.2$ that has power 0.658 at $\theta = 1.6$ achieves lower $E_\theta(N)$ than the MP design over the whole range of θ values.

6. Further generalisations

1. We can allow a general final decision rule, rather than the CPR rule.

To do this, we need to add a cost for a type I error to our optimality criterion, then tune the cost parameter so the type I error probability is $\alpha = 0.025$.

This will give the best possible sampling and decision rules, of any kind, with $n_1 = 208$ and n_2 in the range 442 to 884.

Rules are functions of the sufficient statistic for θ (unlike those of CPR procedures).

2. We can allow recruitment to be terminated at the interim analysis, so the minimum final sample size is set at $n_2 = 416$, rather than 442 (assuming this will provide sufficient data to evaluate safety).

The same optimisation over sample size rules and decision rules can be performed to give the most efficient design with a given type I error rate and specified power.

This would appear a natural way to deal with the volume of “pipeline” subjects when there is a delayed response.

Further generalisations

3. We could choose other expected sample size criteria.

We can minimise a weighted sum

$$\sum_i w_i \, E_{\theta_i}(N)$$

or an integral

$$\int w(\theta) \, E_{\theta_i}(N) \, d\theta.$$

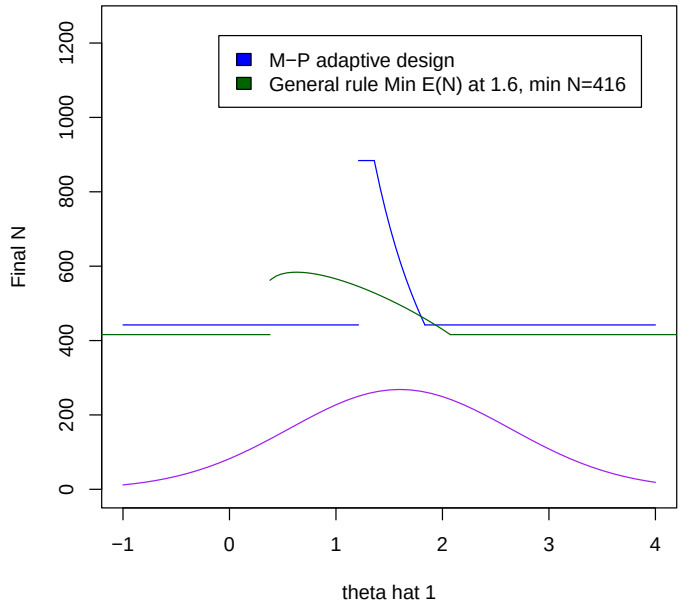
Then, our approach will lead to minimisation of this function of $E_\theta(N)$ subject to a related average power property.

As before, we can match this average power to the same property of (say) the MP design and this will give a close match to the whole power curve.

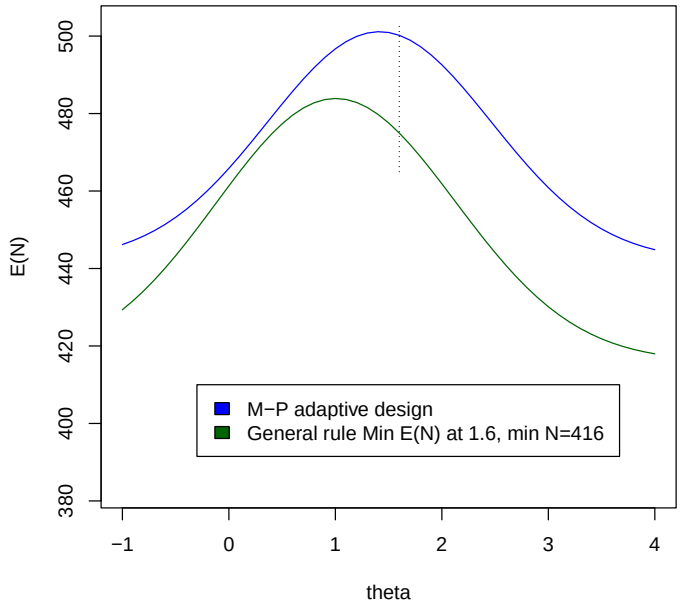
General sampling rule and early termination of recruitment

Following generalisations (1) and (2) above gives the sample size rule shown below.

Sample size rule



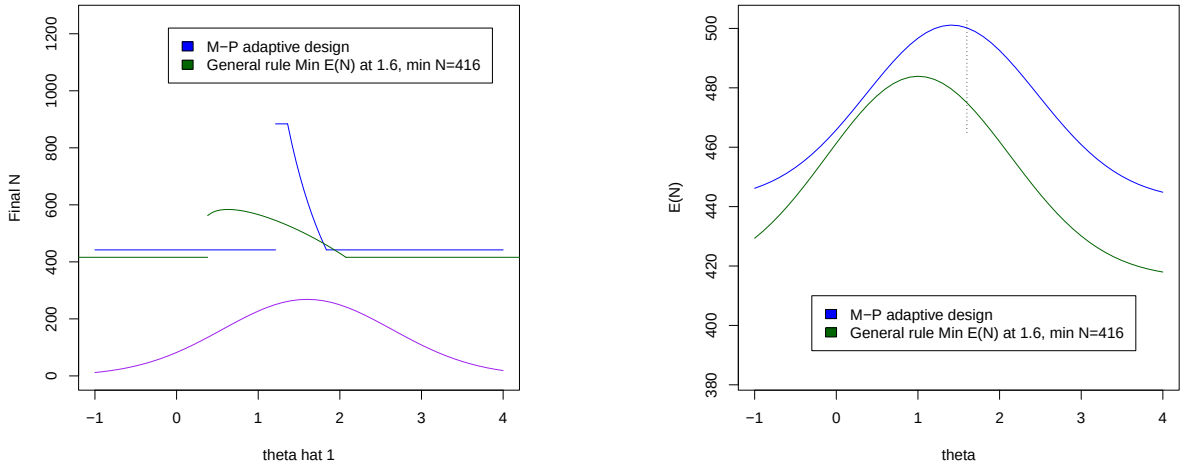
$E_\theta(N)$ curves



We achieve further, useful reductions in $E_\theta(N)$. Most of the benefit comes from generalisation (2), which allows the maximum sample size to be limited to 416.

As for the optimal CPR rule, the highest final sample size is chosen at values of $\widehat{\theta}_1$ below Mehta & Pocock's "promising zone".

7. Relation to proposals for Delayed Response GSTs



The methods we have developed are related to derivations of optimal GSTs (see references in Section 9).

We can ask the question:

Could we have achieved something similar to this efficient design by extending GSTs to deal with a delayed response?

A GST would have a discrete set of values for the final sample size — but perhaps the above sample size function can be approximated by a simple step function.

Hampson & Jennison's Delayed Response GSTs

Reference: "Group sequential tests for delayed response" *(Submitted)*.

Hampson & Jennison (HJ) formulate group sequential designs for delayed response in which the 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.

In a design with up to K analyses, let Z_k denote the standardized test statistic for testing H_0: $\theta \leq 0$ at interim analysis k.

Recruitment can terminate at interim analysis k

For a high values of Z_k, suggesting a positive treatment effect,

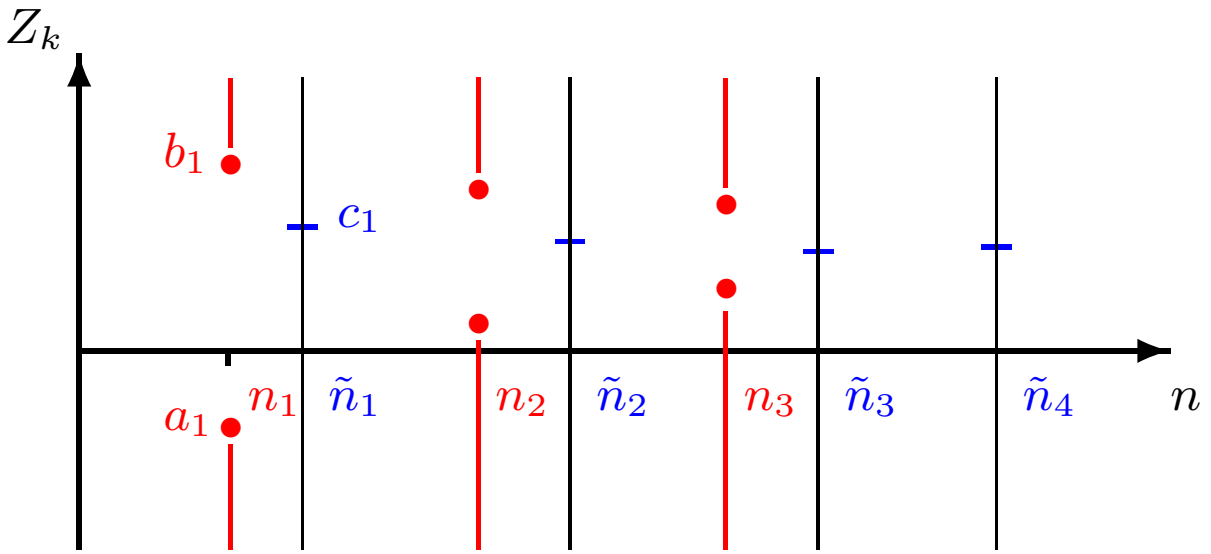
For a low value of Z_k, suggesting no treatment effect.

However, the decision to accept or reject H_0 is not taken until the subsequent "decision analysis" when responses from pipeline subjects are available.

Hampson & Jennison's Delayed Response GSTs

At interim analysis k, with n_k observations, compare Z_k to critical values a_k, b_k.

If $Z_k < a_k$ or $Z_k > b_k$, cease recruitment of new patients.



Now wait for responses from "pipeline" subjects who have been treated but have no response at interim analysis k.

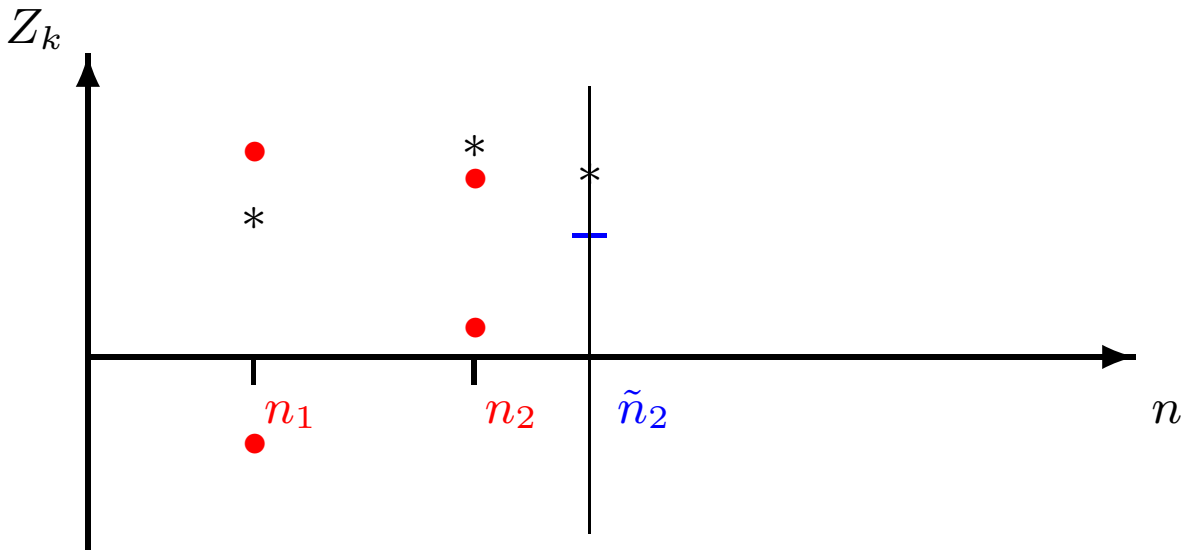
At the final decision analysis, with $\tilde{n}_k$ observations, reject H_0 if $\tilde{Z}_k > c_k$.

Delayed Response GSTs

For a particular sequence of observed responses, we apply boundary points at a sequence of sample sizes of the form

$$n_1, \ldots, n_k, \ \tilde{n}_k.$$

In the example below, recruitment ceases at the second interim analysis and the final decision is made with extra “pipeline” data bringing the information up to $\tilde{n}_2$.



Delayed Response GSTs

Computations for Delayed Response GSTs

It is not difficult to compute properties of a given Delayed Response Group Sequential Test (DR GST).

The sequence of estimates $\{\widehat{\theta}_1, \ldots, \widehat{\theta}_k, \tilde{\theta}_k\}$ and the related sequence of Z-statistics $\{Z_1, \ldots, Z_k, Z_k\}$ have the same forms of joint distribution seen for standard GSTs.

Hence, the same methods can be used to compute properties of a DR GST.

Optimising a Delayed Response GST

Suppose we stipulate type I error rate α, power $1 - \beta$ at $\theta = \delta$, and K interim analyses and decision analyses at specified times.

Then it is possible to optimise a DR GST with respect to a measure of expected sample size on termination.

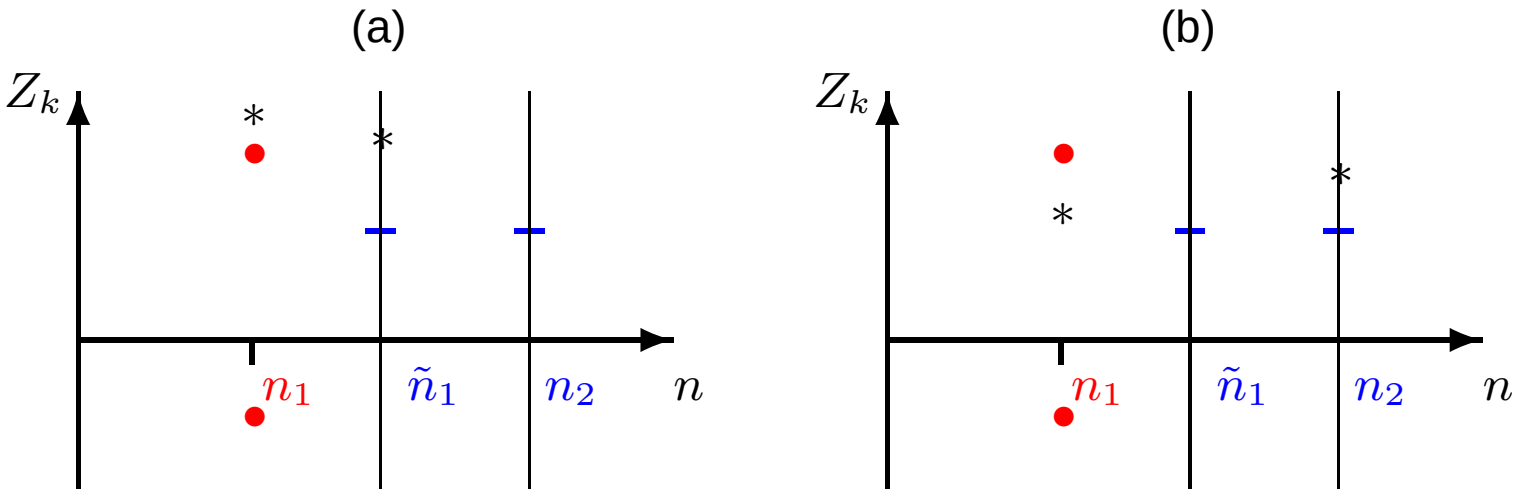
Relation between DR GSTs and Mehta & Pocock's design

With $K = 2$, the DR GST is similar to the MP design.

On the basis of data at interim analysis 1, a decision is made

(a) To cease recruitment and wait for pipeline subjects to respond, or

(b) Continue recruitment, then wait for responses from all subjects.



So, interim data guide the choice made between a final group size of $\tilde{n}_1$ and n_2.

In the MP design, final sample size is chosen from a *continuous* range of values.

Delayed Response GST for the MP example

For Mehta & Pocock's example we can define a DR GST with:

Type I error rate $\alpha = 0.025$,

Power 0.658 at $\theta = 1.6$,

First analysis with $n_1 = 208$ observed responses,

Second analysis with

either $\tilde{n}_1 = 416$ responses (pipeline subjects only),

or $n_2 = 518$ responses (recruiting 102 more patients).

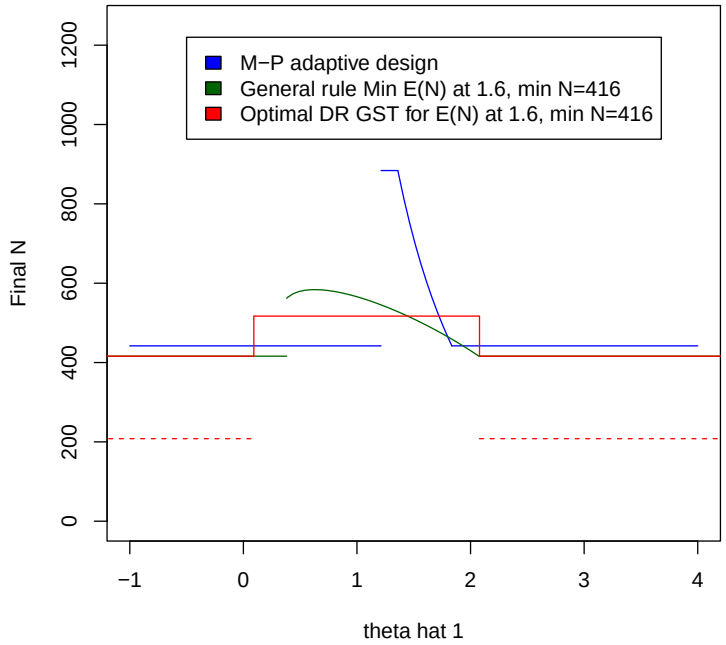
We have found the DR GST minimising $E_{\theta=1.6}(N)$ subject to these constraints.

An "Adaptive DR GST" would allow the final sample size to be chosen as any value greater than the minimum $\tilde{n}_1 = 416$ arising from the pipeline subjects — but these are the defining properties of the general optimal rule in Section 6

Delayed Response GST for the MP example

Optimising a DR GST gives the sample size rule shown in the figure below.

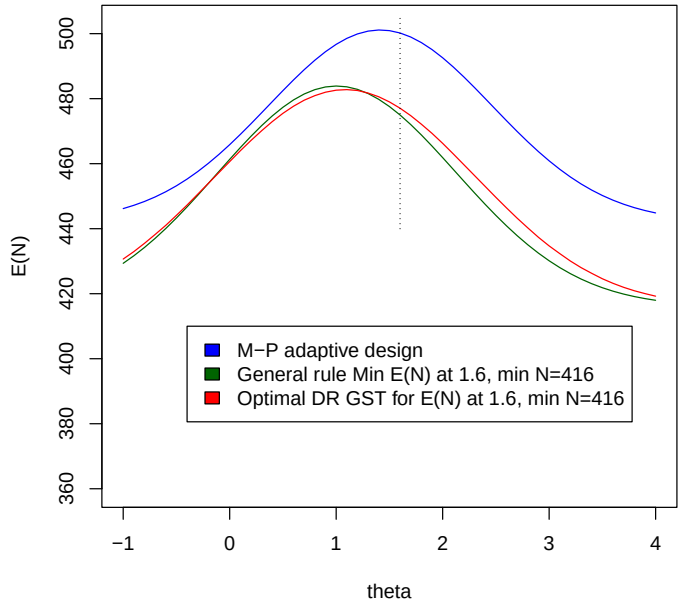
This DR GST has type I error rate 0.025 and power 0.658 at $\theta = 1.6$, and is optimised for $E_{\theta=1.6}(N)$.



The sampling rule approximates that of the optimal adaptive DR GST — which is the same as the general optimal rule found in Section 6.

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

There is very little difference in $E_\theta(N)$ between the optimal DR GST and the optimal *adaptive* DR GST, which is allowed to vary the second group size.



As Jennison & Turnbull (*Biometrika*, 2006) found for an immediate response, there is minimal benefit from tuning the final sample size in response to interim data.

HJ find a similarly low return from adaptation in other examples with $K = 2$.

Hampson & Jennison's Delayed Response GSTs

HJ's DR GSTs are defined for a general number of analyses $K \geq 2$.

Efficient forms of these designs are known.

Their performance can be compared for different choices of

Number of analyses of K *and*

Maximum overall sample size

to find the most suitable design.

HJ define error spending versions of their DR GSTs, which can deal with departures in group sizes from their planned values.

HJ also present methods for inference (P-values and confidence intervals) on termination of a DR GST.

Efficiency implications of a delayed response

We are used to GSTs giving reductions in $E_\theta(N)$ from a fixed sample size design.

Suppose the pipeline subjects at each interim analysis comprise a fraction r of the overall maximum sample size.

The reductions in $E_\theta(N)$ that can be achieved by interim monitoring and early termination of recruitment decrease as r increases.

HJ show that the loss of efficiency is small for values of $r < 0.1$.

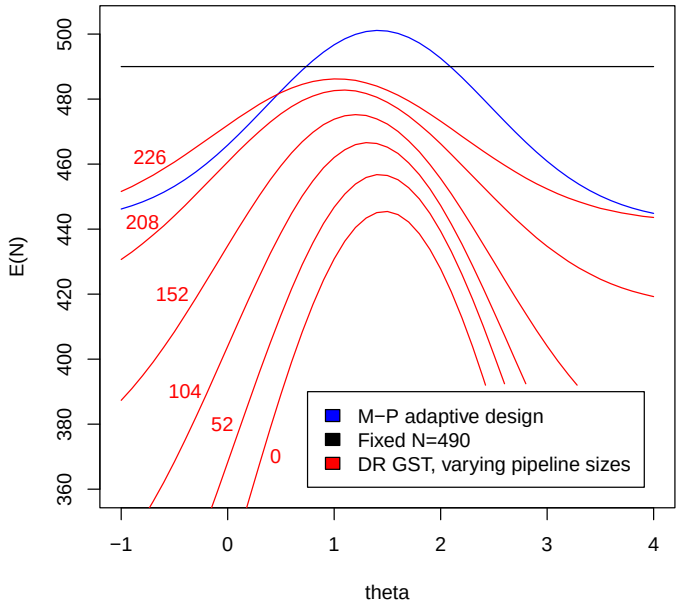
However, as r rises to 0.25, we lose about half the savings in expected sample size that a GST for immediate response would deliver.

In the MP example, the value of r is much higher than this! The two-stage GST we considered in Section 3 had a maximum of 514 subjects, so

$$r \;=\; \frac{208}{514} \;=\; 0.40.$$

Efficiency implications of a delayed response

We have computed DR GSTs for the MP example under a range of scenarios with delay in response leading to different numbers of pipeline subjects.



The number of pipeline subjects in the example as described by MP is 208.

MP's adaptive design has at least 226 new responses observed between analyses.

For lower numbers of pipeline subjects, DR GSTs produce lower $E_\theta(N)$ curves, coinciding with the GST for immediate response when this number is zero.

Ameliorating the effects of a delayed response

Once we understand how a delayed response affects what can be achieved by group sequential monitoring, steps can be taken to address this problem.

Recall that the MP example concerns a Phase 3 trial of a new treatment for schizophrenia. The primary endpoint is improvement in the Negative Symptoms Assessment (NSA) score from baseline to week 26.

Slower recruitment

Suppose the costs of running the trial are the limiting factor and investigators do not object to taking a little longer to reach a conclusion.

Then recruiting patients more slowly will reduce the number in the pipeline at an interim analysis — and we can achieve one of the lower $E_\theta(N)$ curves in the previous figure.

Ameliorating the effects of a delayed response

Using data on a short term endpoint

DR GSTs can incorporate data on a short term response, correlated with the long-term endpoint, still making inferences on the desired long-term endpoint.

In the MP example, there is an opportunity to measure the NSA score at one or more intermediate times between baseline and week 26.

Fitting a longitudinal model to these data at an interim analysis yields an improved estimate of the treatment effect on the primary endpoint.

The increase in information about the primary endpoint at the interim analysis has the same effect as a reduction in the "pipeline size". So, this will lead to a more efficient DR GST.

With a good correlation between short-term and long-term endpoints, e.g., $\rho = 0.7$, this approach recoups much of the efficiency loss due to the delay in response.

8. Conclusions

1. Mehta & Pocock describe a problem for sequential monitoring of a clinical trial with a delayed response: this poses problems for "conventional" GSTs.

2. MP use the Chen, DeMets & Lan (2004) approach with sample size set to attain conditional power 0.8 if $\theta = \widehat{\theta}_1$. This does not yield a particularly efficient design.

3. We have pursued MP's idea of spending resource where it will have the greatest benefit, and found efficient adaptive designs for their problem.

4. The solution to our most general formulation of this problem is also an optimal "Adaptive Delayed Response GST", as proposed by Hampson & Jennison (2011). However, a non-adaptive version of this design — the natural extension of a GST to a delayed response — is almost as efficient and likely to be simpler to implement.

5. Understanding how a delay in response affects the monitoring process can help address this problem. In MP's example, either slower accrual or taking interim measurements of the NSA score could help reduce expected sample size.

9. Connections to other work

Optimal non-adaptive GSTs for an immediate response

Eales & Jennison (1992) *Biometrika*,
Barber & Jennison (2002) *Biometrika*

These papers present derivations of optimal GSTs with specified type I error rate and power. Designs minimise a weighted average or integral of $E_{\theta(N)}$ values. Derivations are similar to those required to implement generalisations (1) to (3) listed in Section 6 of this talk — but with $K \geq 2$ groups of fixed size.

Optimal adaptive GSTs for an immediate response

Posch, Bauer & Brannath (2003), *Statistics in Medicine*

In Section 3.3.3 of this paper, the authors consider a form of combination test and find an optimised sample size rule by searching over a 4-parameter family of functions. Extending this approach to the delayed response setting would give a similar result to our optimised CPR method.

Connections to other work

Optimal adaptive GSTs for an immediate response

Jennison & Turnbull (2006) *Biometrika*,

Banerjee & Tsiatis (2006) *Statistics in Medicine*,

Lokhnygina & Tsiatis (2008) *JSPI*

These authors derive optimal adaptive GSTs with given type I error rate and power. Derivations are essentially those needed to carry out generalisations (1) to (3) listed in Section 6 of this talk.

Jennison & Turnbull (2006) implement the methods for designs with $K \geq 2$ groups.

Optimal adaptive and non-adaptive GSTs for a delayed response

Hampson & Jennison (2011) "Group sequential tests for delayed response"
Submitted for publication

Connections to other work

Conditional probability of rejection (CPR)

Lan, Simon & Halperin (1982) *Communications in Statistics*,

Lan & Wittes (1988) *Biometrics*,

Proschan & Hunsberger (1995) *Biometrics*,

Denne (2001) *Statistics in Medicine*,

Müller & Schäfer (2001) *Biometrics Medicine*,

Müller & Schäfer (2004) *Statistics in Medicine*,

Jennison & Turnbull (2003) *Statistics in Medicine*

There is a long history of making use of the conditional probability, under the null hypothesis, that a test will ultimately reject H_0. Lan, Simon & Halperin used this quantity in defining “stochastic curtailment” procedures; Lan & Wittes discussed more general usage in data monitoring.

Proschan & Hunsberger made use of CPR, showing that preserving this quantity when a design is adapted will protect the type I error rate.

Connections to other work

Conditional probability of rejection (CPR)

Denne (2001) and Müller & Schäfer (2001, 2004) used similar constructions, with a greater emphasis on flexible (i.e., not fully pre-specified) procedures.

Jennison & Turnbull (2003) showed formally that such preservation of the conditional type I error rate is essential in flexible adaptations in order to avoid the possibility of inflating the overall type I error rate.