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

Roche, Welwyn Garden City

May 2013

Choosing the sample size for a trial

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.

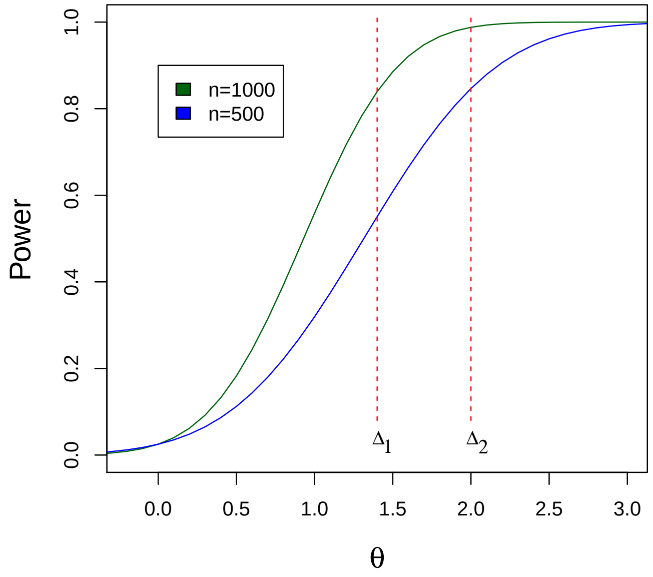
Should investigators use:

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 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:

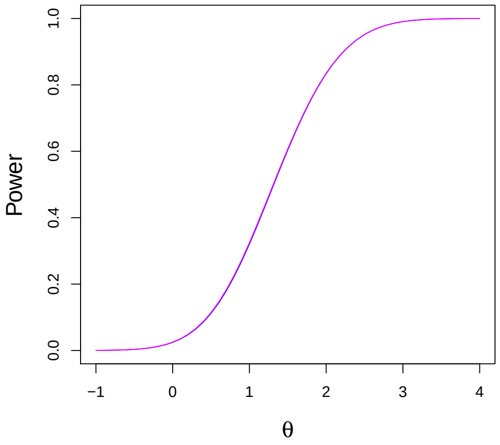
Specify the desired type I error and power function,

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

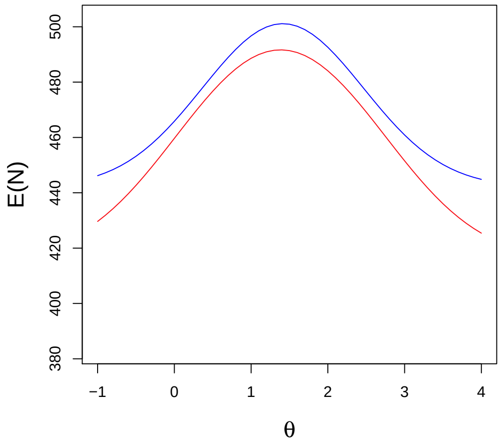
Stop the trial early if data support this.

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 — and not easily dealt with by standard GSTs.

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. Improving designs in Mehta & Pocock’s framework
5. Extending the framework
6. Relation to delayed response GSTs (Hampson & Jennison, *JRSS B*, 2013)
7. Conclusions

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.

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

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

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

Thus, H_0 is 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

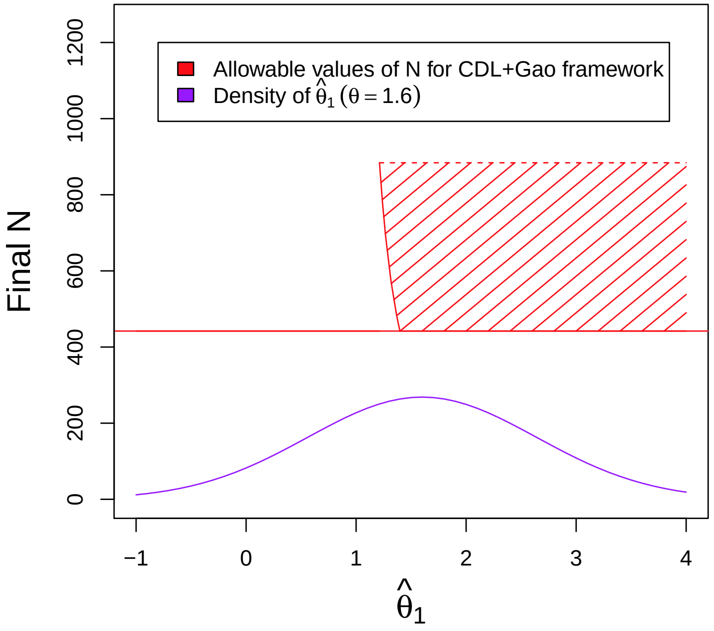
$$CP_{\hat{\theta}_1}(z_1) > 0.5,$$

then the type I error probability will not increase.

(In general, changes to sample size may increase or decrease the type I error rate.)

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^*.

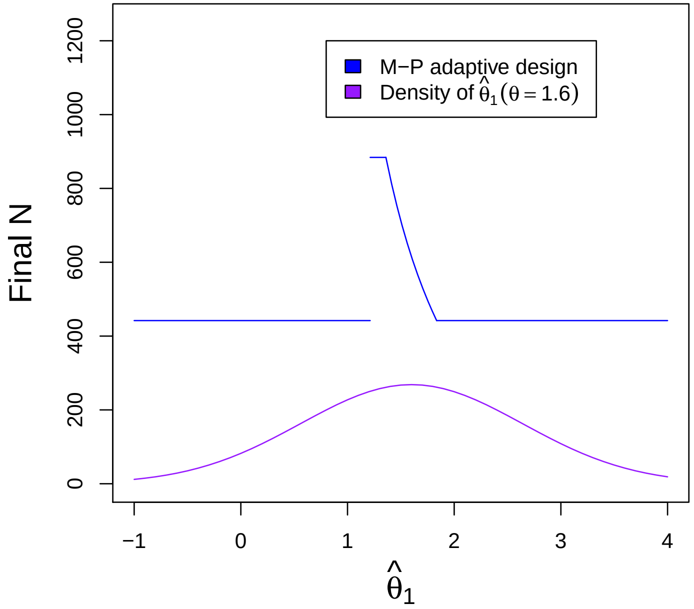


With an upper limit of $n_2^* = 884$, the final sample sizes permitted by the CDL+Gao approach are as shown in the figure.

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

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.

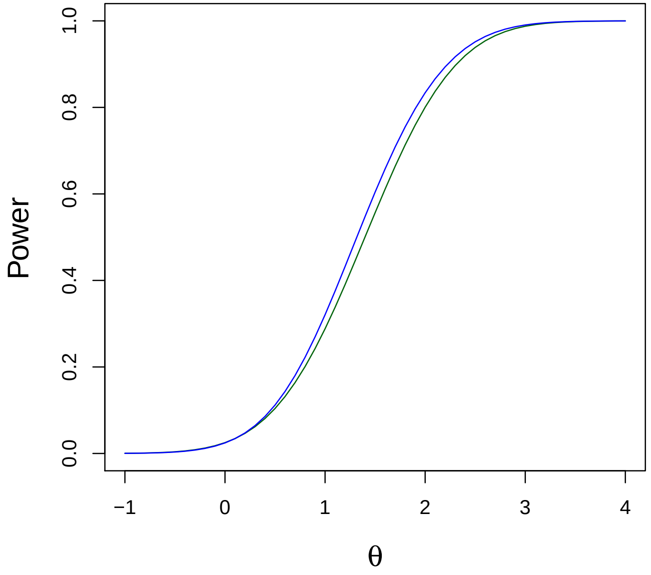


Comparison with the distribution of $\widehat{\theta}_1$ under $\theta = 1.6$ shows that increases in n_2 occur in a region of quite small probability.

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.

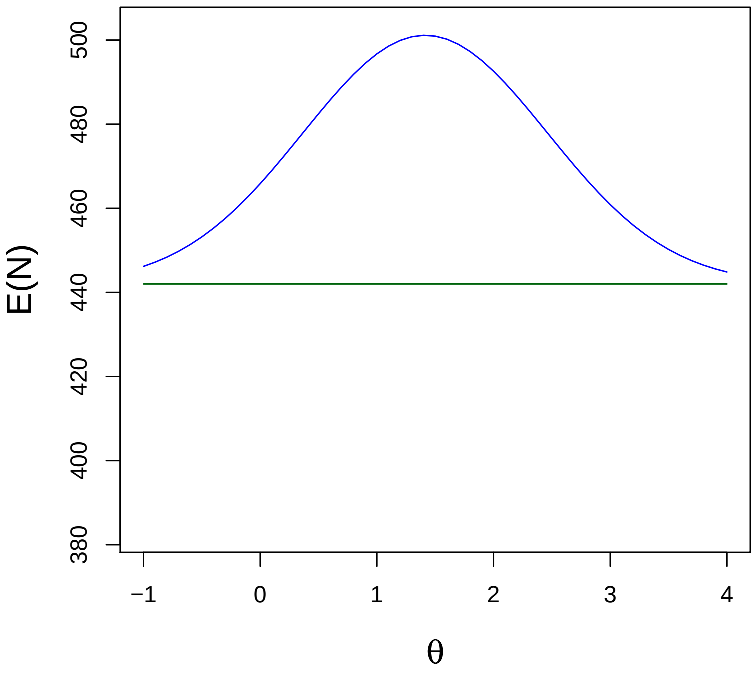


Given the limited range of values of $\widehat{\theta}_1$ for which n_2 is increased, only a small improvement in power can be expected.

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.



Aiming for higher conditional power under $\theta = \widehat{\theta}_1$ or raising the sample size beyond 884 gives small increases in power at the cost of large increases in $E(N)$.

3. 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 can still consider a 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

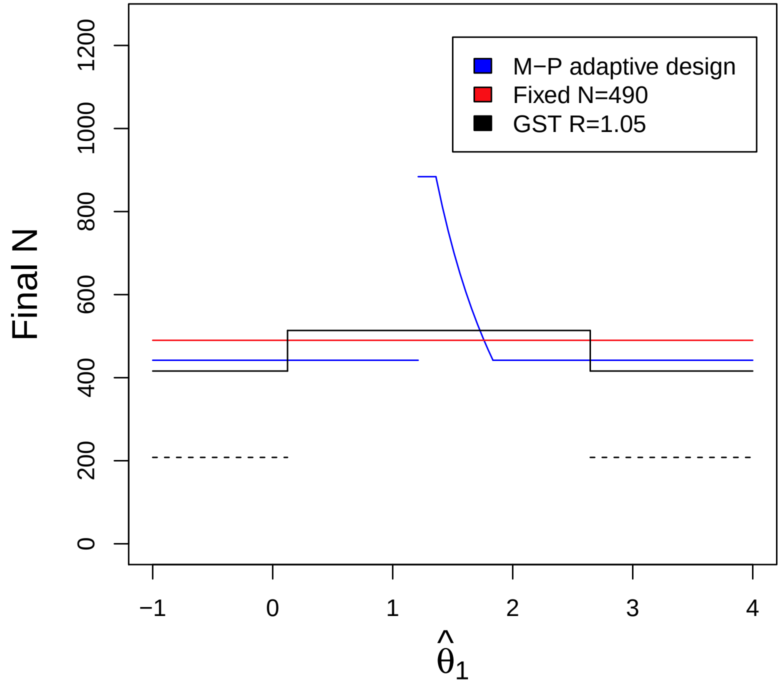
Sample size rules for MP, fixed and group sequential designs

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

The fixed sample size design has 490 observations.

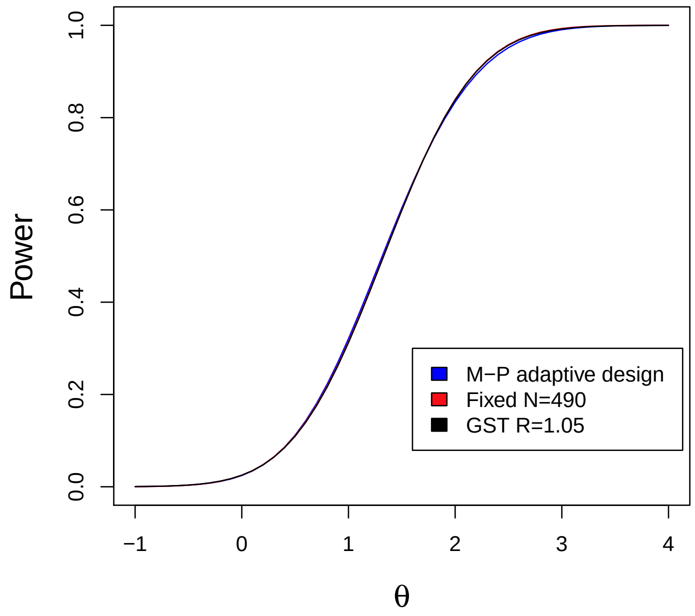
The group sequential test can stop with a sample size of 416 or 514.

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

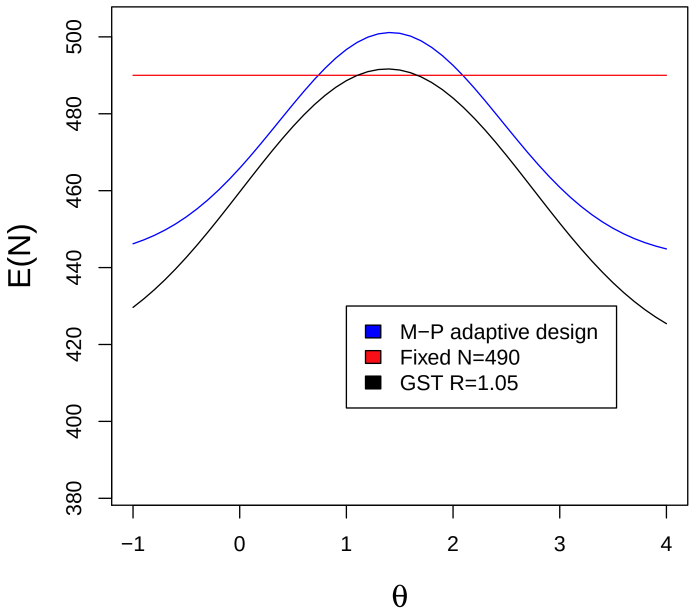


Comparison of designs

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,

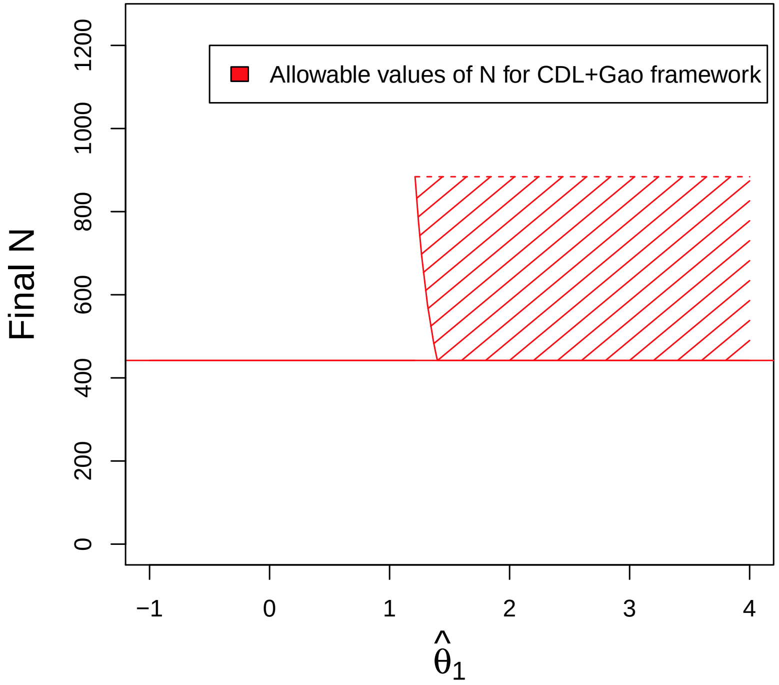
Can we work out how to do this effectively?

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 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.

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,

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

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

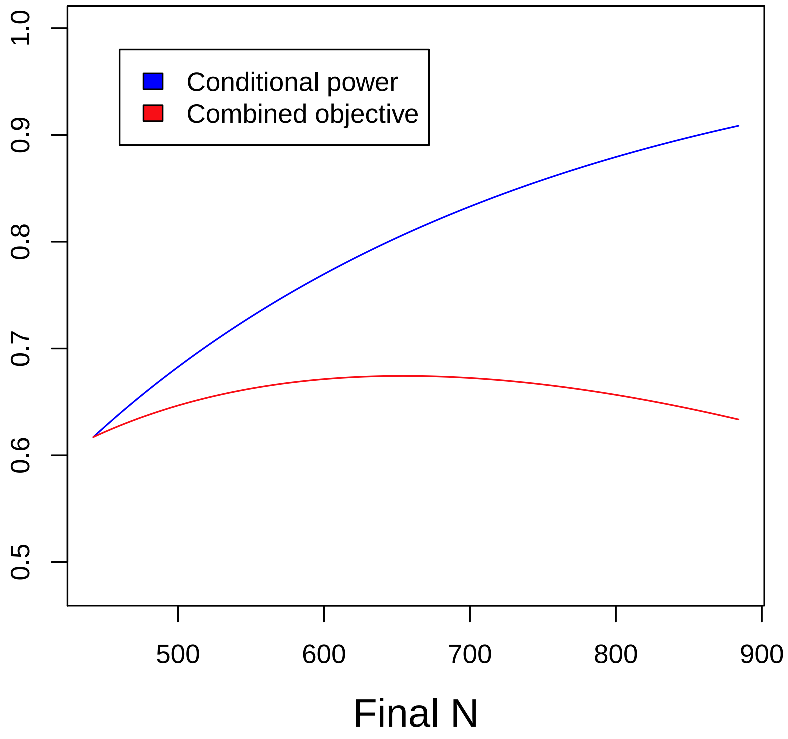
$$\int \{CP_{\tilde{\theta}}(z_1, n_2^*(z_1)) - \gamma n_2^*(z_1)\}\, f_{\tilde{\theta}}(z_1)\, dz_1,$$

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 $\theta = 1.6$.

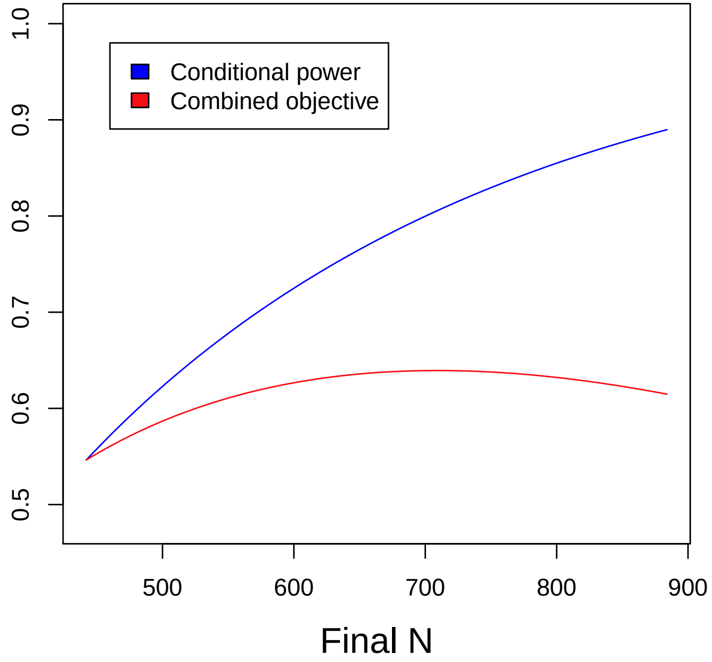
Plots for $\tilde{\theta} = 1.6$, $\gamma = 0.14/(4\,\sigma^2)$ and $\widehat{\theta}_1 = 1.5$



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 $\widehat{\theta}_1 = 1.3$

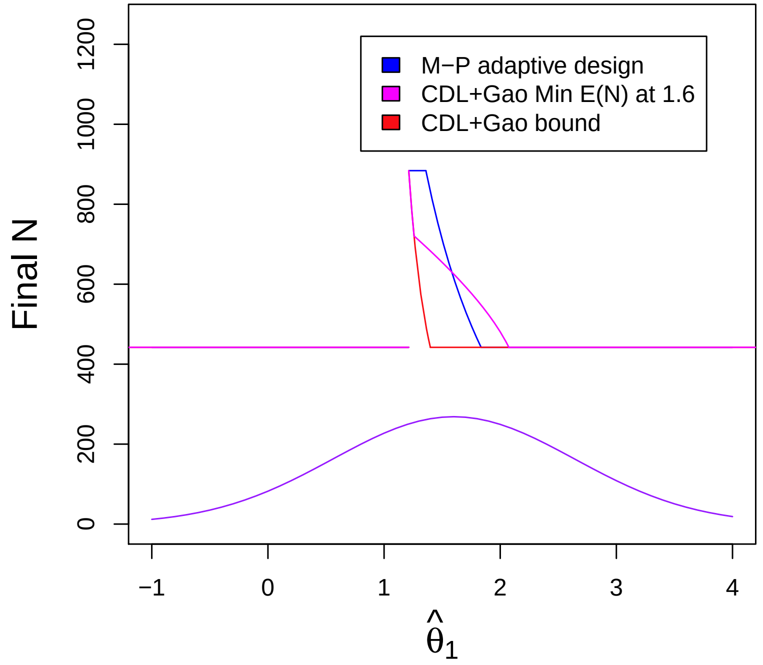


The conditional power curve is steeper and the optimum occurs at a higher n_2^*.

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 $\gamma = 0.14/(4\,\sigma^2)$



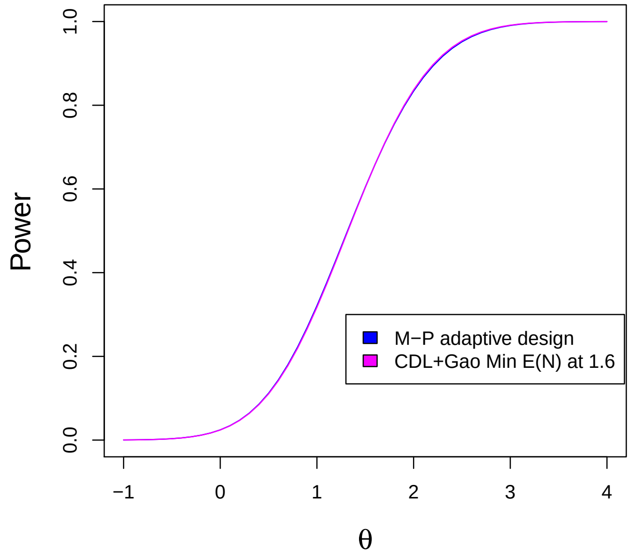
This rule gives power 0.658 at $\theta = 1.6$, the same as the MP design.

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

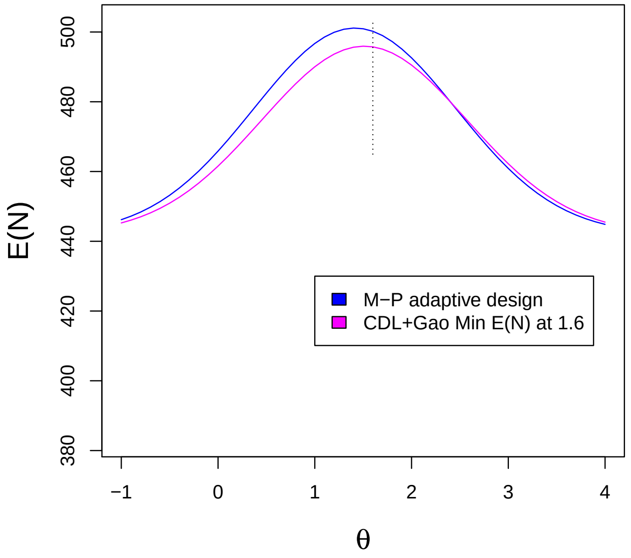
As $\widehat{\theta}_1$ decreases, sample size increases less steeply than for the MP design.

Efficient sample size rules in the MP framework

Power curves



$E_\theta(N)$ curves



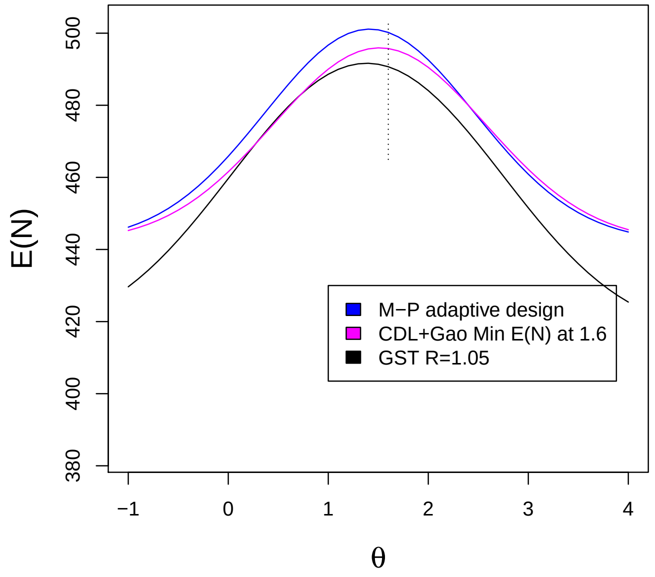
With the type I error rate at $\theta = 0$ fixed at 0.025, matching the MP design's power at one value of θ implies matching the whole power curve.

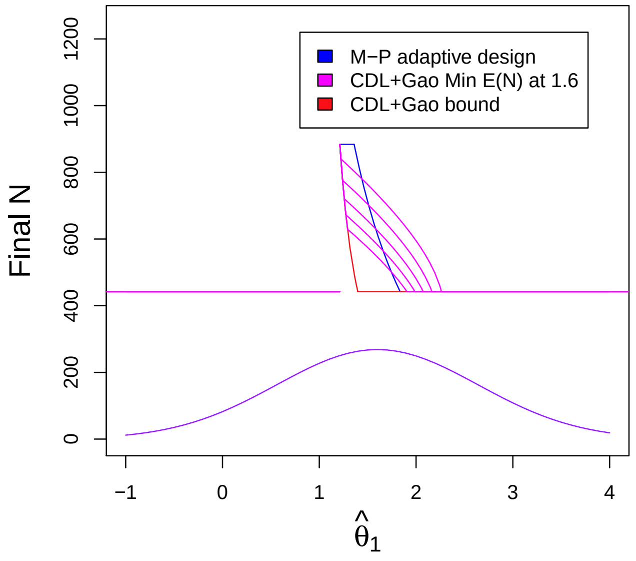
Our optimised design has the same power curve as the MP design and lower $E_\theta(N)$ (just about) at all θ values.

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.

5. 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.

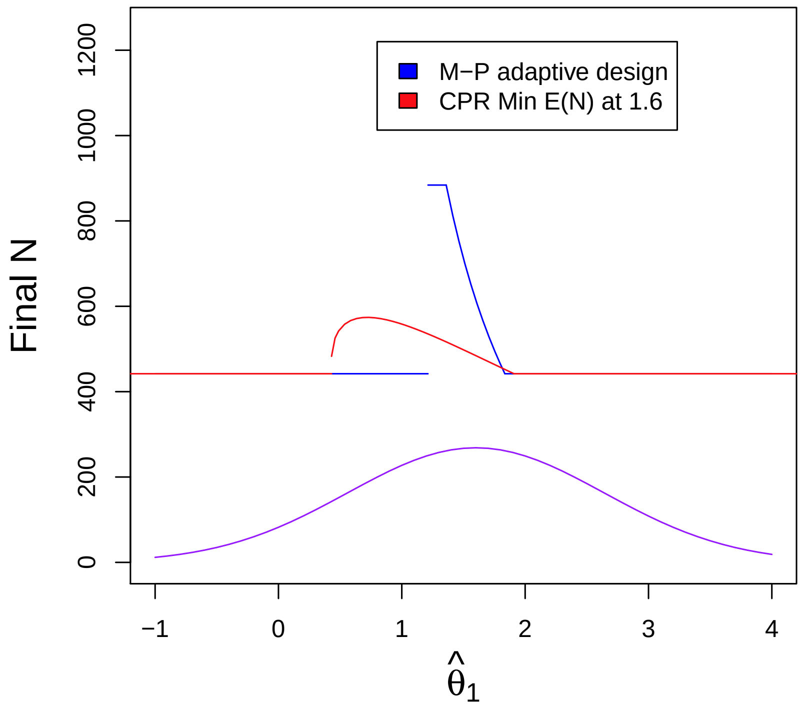
The overall type I error rate is the integral of the conditional type I error rate, and this remains the same.

This type of adaptation can also be regarded as a “weighted inverse normal combination test” Bauer & Köhne (1994).

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 $\tilde{\theta} = 1.6$



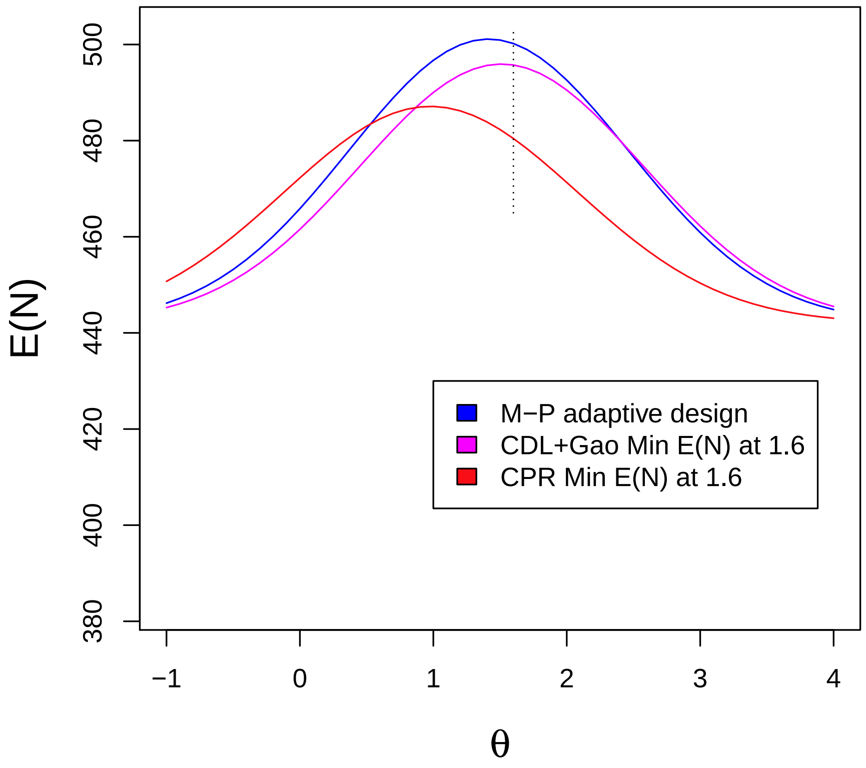
The rule with $\gamma = 0.25/(4\,\sigma^2)$ matches the MP test's power of 0.658 at $\theta = 1.6$.

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.

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

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

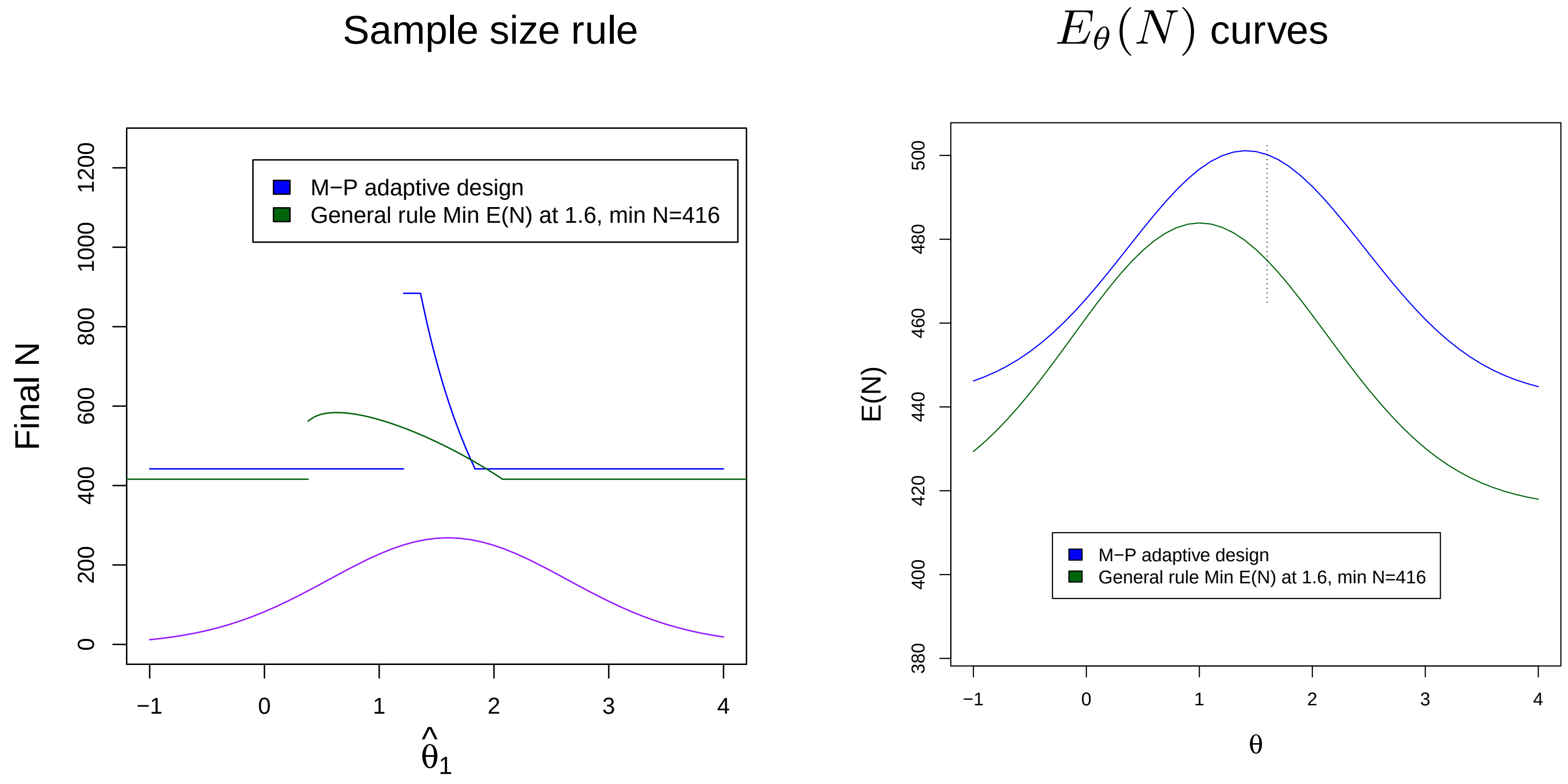
The resulting designs deal neatly with the “pipeline” subjects arising when there is a delayed response.

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”.

6. Relation to proposals for Delayed Response GSTs

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

Hampson & Jennison have extended methodology for group sequential tests to handle a delayed response.

Their "Delayed Response GSTs" allow any number of interim analyses and can be optimised for specified criteria.

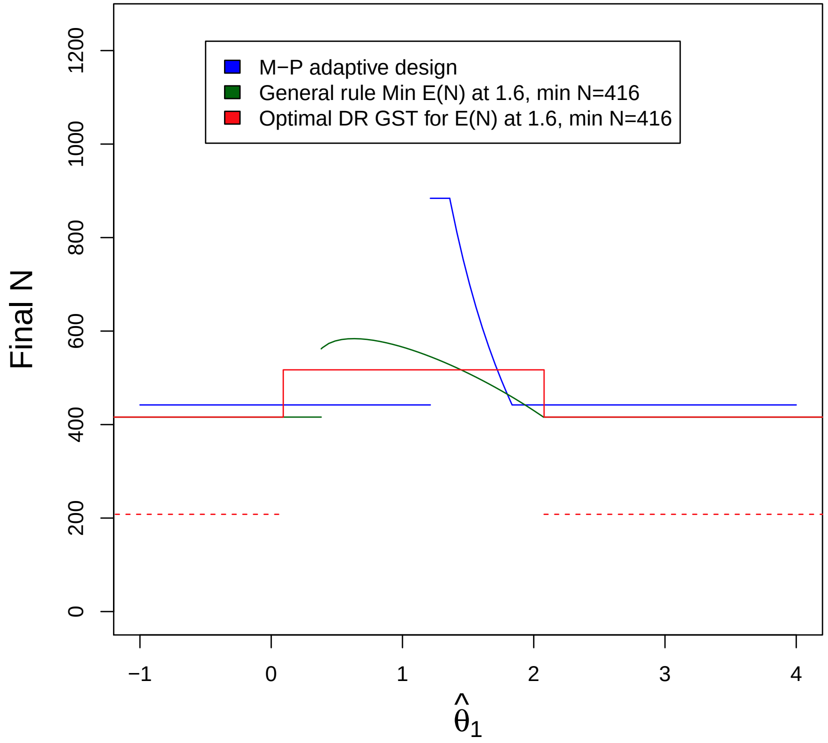
Applying this approach in the case of just 2 analyses:

> 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.

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

Delayed Response GST for the MP example

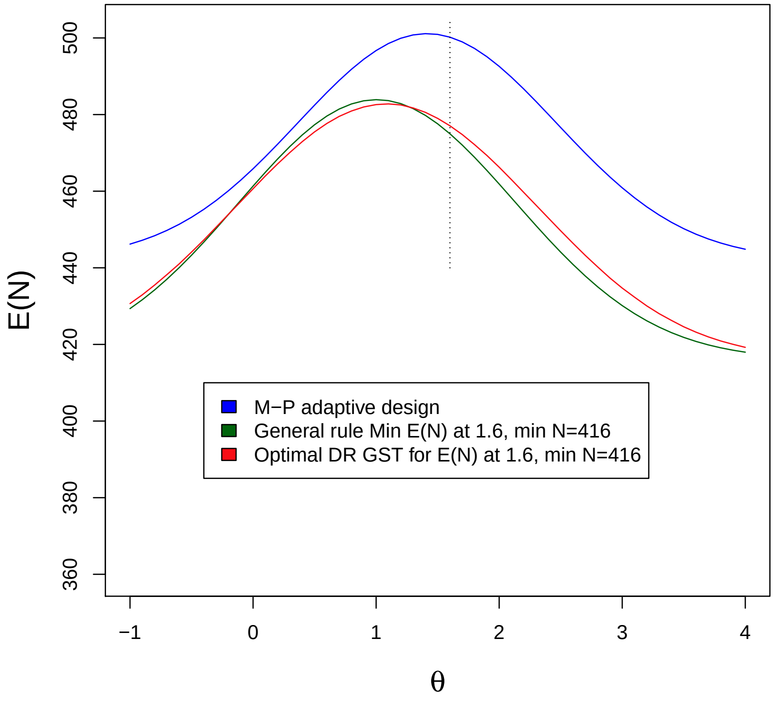
Optimising a DR GST to minimise $E_{\theta=1.6}(N)$ while matching the power of the MP design gives the sample size rule shown below.



The sampling rule approximates that of the general adaptive method, but with a step function rather than a continuous sample size function.

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

The optimised DR GST has an almost identical $E_\theta(N)$ curve to the general rule using the continuum of possible sample sizes.



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

7. Conclusions

Although JT had previously shown that adaptive designs offer at most a slight improvement on GSTs, it is appropriate to re-visit this issue for the case of a delayed response, as in Mehta & Pocock's example.

1. MP use the Chen, DeMets & Lan (2004) approach, choosing sample size by a conditional power rule. This does not yield a particularly efficient design.

2. We have developed MP's idea of spending resources where they have the greatest benefit — and found efficient adaptive designs for this problem.

3. The solution to our most general version of the problem is very similar to a "Delayed Response GST", as proposed by Hampson & Jennison (2013).

   Following this approach offers the benefits of established group sequential methodology and its extensions, e.g., error spending tests.

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