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

Novartis, Basel
March 2018

Outline of talk

1. Choosing the sample size for a trial

   Group sequential and adaptive approaches

2. An example from Mehta & Pocock (*SiM*, 2011)

   (i)   Mehta & Pocock's "promising zone" design

   (ii) Alternative fixed and group sequential designs

   (iii) Improving designs in Mehta & Pocock's framework

   (iv) Extending the framework

   (v)   Relation to delayed response GSTs (Hampson & Jennison, *JRSS B*, 2013)

3. Conclusions

Research Article

Statistics
in Medicine

Adaptive sample size modification in clinical trials: start small then ask for more?

Christopher Jennison[a][*][†] and Bruce W. Turnbull[b]

We consider sample size re-estimation in a clinical trial, in particular when there is a significant delay before the measurement of patient response. Mehta and Pocock have proposed methods in which sample size is increased when interim results fall in a 'promising zone' where it is deemed worthwhile to increase conditional power by adding more subjects. Our analysis reveals potential pitfalls in applying this approach. Mehta and Pocock use results of Chen, DeMets and Lan to identify when increasing sample size, but applying a conventional level α significance test at the end of the trial does not inflate the type I error rate: we have found the greatest gains in power per additional observation are liable to lie outside the region defined by this method. Mehta and Pocock increase sample size to achieve a particular conditional power, calculated under the current estimate of treatment effect: this leads to high increases in sample size for a small range of interim outcomes, whereas we have found it more efficient to make moderate increases in sample size over a wider range of cases. If the aforementioned pitfalls are avoided, we believe the broad framework proposed by Mehta and Pocock is valuable for clinical trial design. Working in this framework, we propose sample size rules that apply explicitly the principle of adding observations when they are most beneficial. The resulting trial designs are closely related to efficient group sequential tests for a delayed response proposed by Hampson and Jennison. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: group sequential test; sample size re-estimation; adaptive design; clinical trial; optimal design; promising zone

1. Choosing the sample size for a clinical trial

You might think this would be a simple question . . .

Let θ denote the effect size of a new treatment, e.g., 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.
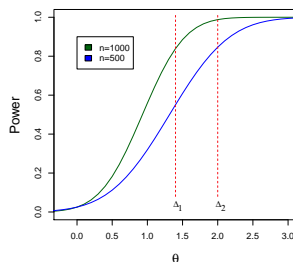
Dispute may arise over the choice of Δ.

Should investigators use:

   The minimum effect of interest Δ_1, or

   The anticipated effect size Δ_2 ?

Choosing the sample size for a clinical trial

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

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

With only 500 subjects, a high 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 high power and low sample size

In designing a clinical trial, we aim to

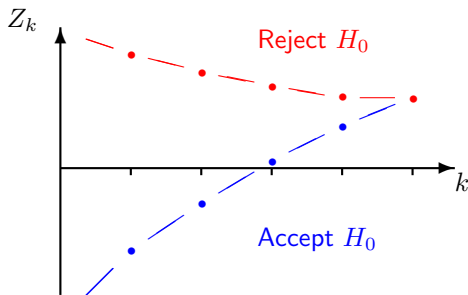*Protect the type I error rate,*

Achieve sufficient power,

Use as small a sample size as possible.

We can use **Group Sequential** or **Adaptive** designs to achieve desired error rates with a small average sample size.

(A) Group sequential tests

A typical boundary for a one-sided test, expressed in terms of standardised test statistics $Z_1, \ldots, Z_K$, has the form:

Crossing the upper boundary leads to early stopping for a positive outcome, rejecting H_0: $\theta \leq 0$ in favour of $\theta > 0$.

Crossing the lower boundary implies stopping for "futility" with acceptance of H_0.

Error spending group sequential tests

Let $\mathcal{I}_k = \text{Var}(\widehat{\theta}_k)$ denote the information for θ at analysis k.

When the sequence $\mathcal{I}_1$, $\mathcal{I}_2$, ... is unpredictable, a group sequential design must adapt to observed information levels.

Lan & DeMets (*Biometrika*, 1983) introduced "error spending" tests of H_0: $\theta = 0$ against $\theta \neq 0$.

Maximum information design with error spending function $f(\mathcal{I})$
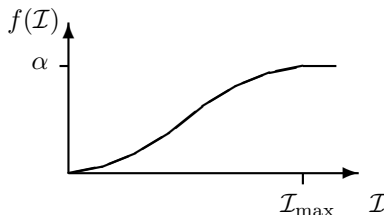
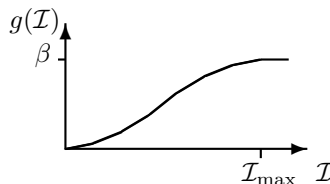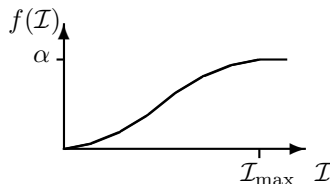The boundary at analysis k is set to give cumulative type I error probability $f(\mathcal{I}_k)$.

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

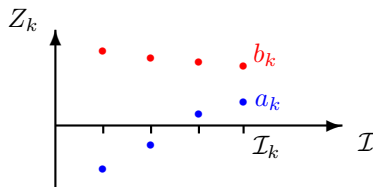*At analysis k:* With observed information $\mathcal{I}_k$

We 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

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

Benefits of group sequential testing

Properties of one-sided tests with binding futility boundaries, minimising $\{\mathbb{E}_0(\mathcal{I}) + \mathbb{E}_\delta(\mathcal{I})\}/2$ for K equally sized groups, $\alpha = 0.025$, power $1 - \beta = 0.9$ at $\theta = \delta$, and $\mathcal{I}_{max} = R\mathcal{I}_{fix}$.

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
2	80.8	74.7	**73.2**	73.7	75.8
3	76.2	69.3	66.6	**65.1**	65.2
5	72.2	65.2	62.2	59.8	**59.0**
10	69.2	62.2	59.0	56.3	**55.1**
20	67.8	60.6	57.5	54.6	**53.3**

Here, $\mathcal{I}_{fix}$ is the information required for a fixed sample size test to achieve type I error rate α and power $1 - \beta$ at $\theta = \delta$.

(B) Adaptive trial designs with sample size re-estimation

Suppose we run a clinical trial adaptively in two stages:

Set the design of Stage 1,

Conduct Stage 1,

Analyse results from Stage 1,

Set the design of Stage 2, choosing the sample size based on Stage 1 results,

Conduct Stage 2,

Analyse the results from Stage 2.

We need a way to test the null hypothesis H_0: $\theta \leq 0$ that properly protects the type I error rate — a combination test can do this.

Combination tests (Bauer & Köhne, *Biometrics*, 1994)

Before the trial commences, define the null hypothesis.

Let θ denote the treatment effect *vs* control for a specified form of the treatment, patient population and endpoint.

We test H_0: $\theta \leq 0$ against $\theta > 0$, with type I error rate α at $\theta = 0$.

Define one-sided P-values P_1 and P_2 from hypothesis tests of H_0 based on Stage 1 and Stage 2 data, respectively.

Under $\theta = 0$

$P_1 \sim U(0, 1)$.

Conditionally on all Stage 1 data and the Stage 2 design, $P_2 \sim U(0, 1)$.

Hence, P_1 and P_2 are independent $U(0, 1)$ variates.

The inverse normal combination test

Initial design

> Specify the **inverse normal test** for null hypothesis H_0, with weights w_1 and w_2 where $w_1^2 + w_2^2 = 1$.

> Design Stage 1, fixing sample size and test statistic.

Stage 1

> Observe the one-sided P-value, P_1, based on Stage 1 data.

> Compute $Z_1 = \Phi^{-1}(1 - P_1)$.

> Design Stage 2 in the light of Stage 1 data.

Stage 2

> Observe the P-value, P_2, based **only** on Stage 2 data.

> Compute $Z_2 = \Phi^{-1}(1 - P_2)$.

NB Under $\theta = 0$, $Z_1 \sim N(0,1)$, $Z_2 \sim N(0,1)$, independent.

The inverse normal combination test

Under $\theta = 0$, Z_1 and Z_2 are independent $N(0,1)$, so

$$w_1 Z_1 + w_2 Z_2 \sim N(0,1).$$

Hence, for an overall one-sided test with type I error rate α, we reject H_0 if

$$w_1 Z_1 + w_2 Z_2 > \Phi^{-1}(1-\alpha).$$

If $\theta < 0$, then Z_1 and Z_2 are stochastically smaller than $N(0,1)$ random variables and the type I error rate is less than α.

If w_1 and w_2 are proportional to the square roots of the Stage 1 and Stage 2 sample sizes then $w_1 Z_1 + w_2 Z_2$ is the standard Z-statistic based on the data at the end of Stage 2.

However, it is crucial that initially specified weights, w_1 and w_2, are used in the final test.

Adaptive sample size modification

The general adaptive framework allows investigators to choose a moderately large sample size initially, then

If interim data are in line with original assumptions, the trial continues as planned,

If interim data suggest the treatment effect is lower than expected, the final sample size can be increased.

In a two stage design, as outlined in previous slides, a pre-specified combination test protects the type I error rate at level α.

Multi-stage adaptive designs are also possible:

Lehmacher & Wassmer (*Biometrics*, 1999) describe a multi-stage version of the combination test.

Cui, Hung & Wang (*Biometrics*, 1999) show how to increase group sizes in a group sequential design.

Designing a trial with good power and sample size

In designing a clinical trial, with a given type I error rate, adequate power, and as small a sample size as possible:

Adaptive designs have the form:

Start with a fixed sample size design,

Examine interim data,

Add observations to improve power where most appropriate.

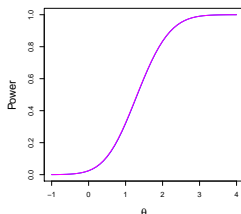**Group Sequential** designs require one to:

Specify the desired type I error and power function,
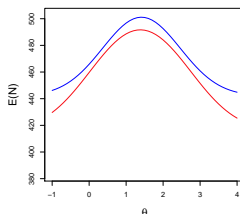
Set maximum sample size a little more than the fixed sample size,

Stop the trial early if data support this.

Designing a trial with good power and sample size

Power curve

$E_\theta(N)$ curves

All designs, *including adaptive designs*, 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.
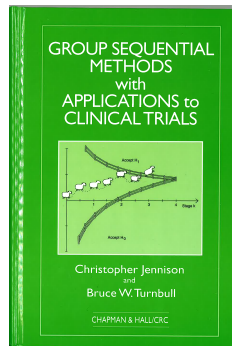
Adaptive or group sequential designs?

Jennison & Turnbull have studied optimal versions of adaptive and non-adaptive sequential designs (e.g., *Statist. in Med.*, 2003 and 2006; *Biometrika*, 2006).

The set of group sequential tests (GSTs) is a subset of the set of adaptive designs (which can adapt group sizes to observed responses)

Adaptive designs are, at best, a little more efficient than GSTs with the same number of analyses, reducing average sample size by 1% or 2% for the same power,

Many published adaptive designs are considerably less efficient than a well chosen GST.

And advice is available on how to create good group sequential designs:

GROUP SEQUENTIAL
METHODS
with
APPLICATIONS to
CLINICAL TRIALS

Christopher Jennison
and
Bruce W. Turnbull

CHAPMAN & HALL/CRC

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.

2. Mehta & Pocock's Example

MP's Example 1 concerns a Phase 3 trial of a new treatment for schizophrenia, comparing the new drug to an active control.

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

Responses are

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

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

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

The treatment effect is

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

and we estimate θ by

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

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 $\theta > 0$ at the final analysis, we reject H_0 if

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

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 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.

Definition

The conditional power $CP_\theta(z_1)$ is the probability the final test, with 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 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, naively, H_0 is rejected if

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

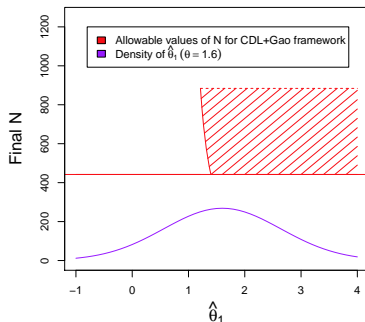Chen, DeMets & Lan (CDL) show that if n_2 is only increased when

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

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.

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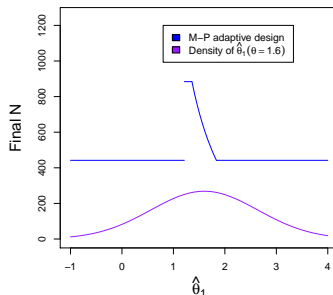
Comparison with the distribution of $\widehat{\theta}_1$ under $\theta = 1.6$ shows that increases in n_2 occur in a region of quite small probability.

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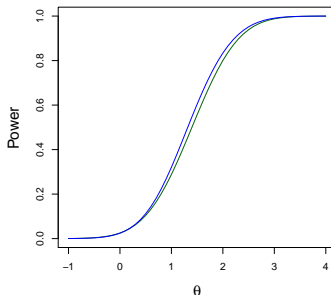
Given the limited range of values of $\widehat{\theta}_1$ for which n_2 is increased, only a small improvement in power can be expected.

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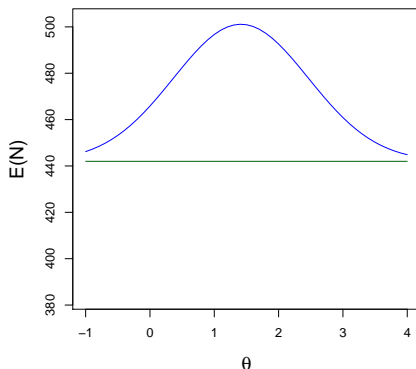
Aiming for higher conditional power under $\theta = \widehat{\theta}_1$ or raising the sample size beyond 884 would give a small increase in power at the cost of a large increase in $E(N)$.

Alternatives to the MP design

Suppose we are satisfied with the overall power function attained by MP's design: the same power can be achieved by other designs.

A fixed sample design

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

This is an attractive option since $E_\theta(N)$ for the MP design is greater than 490 for effect sizes θ between 0.8 and 2.0.

NB: There is more to 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, we take the sample size as 416, counting all subjects treated thus far — even though only 208 have provided a response.

We apply an error spending design in the ρ-family (JT, 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 seq. designs

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

The fixed sample size design has 490 observations.

The group sequential test stops with a sample size of 416 or 514.
Since $514 = 490 \times 1.05$, it has an "inflation factor" of $R = 1.05$.
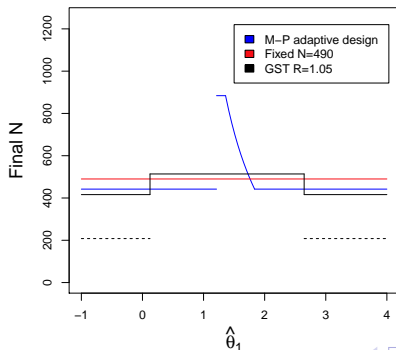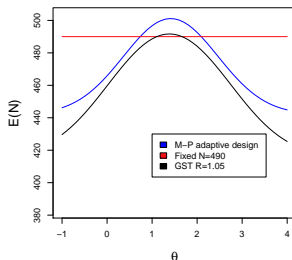
Comparison of designs

Power curves
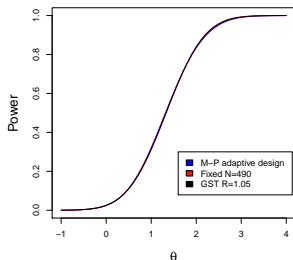
$E_\theta(N)$ curves

All three designs have essentially the same power curve.

It is clearly possible to improve on the MP design's $E_\theta(N)$ curve.

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

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?

Deriving efficient sample size rules in the MP framework

Continuing with MP's example, we retain the basic elements of the MP 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).
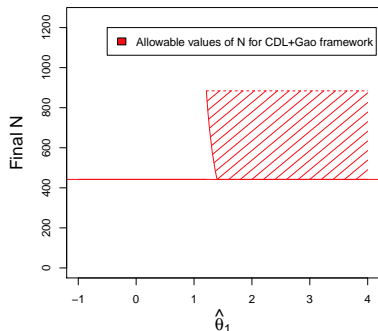
Values of $n_2^* \in [442, 884]$ that satisfy the CDL+Gao conditions are allowed.

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 value of an increase in sample size in terms of the conditional power that it 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 choose n_2^* to optimise a combined objective

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

We shall do this taking $\tilde{\theta} = 1.6$, 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 set $\gamma = 0.14/(4\,\sigma^2)$ to achieve the power of 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 of conditional power and combined objective function

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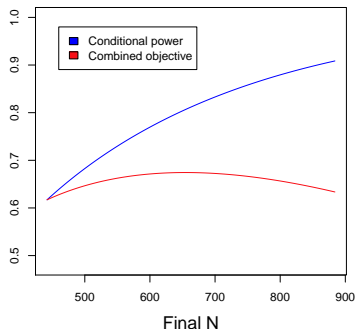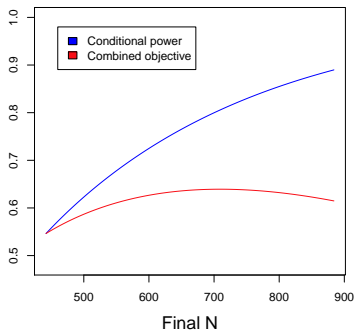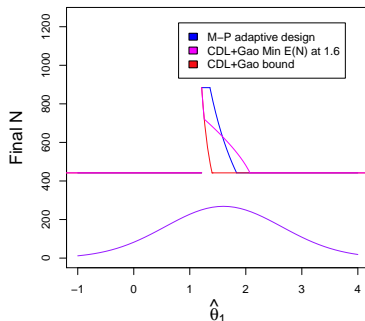

The objective $CP_{\tilde{\theta}}(z_1, n_2^*) - \gamma(n_2^* - 442)$ has a maximum at $n_2^* = 654$.

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 $\widehat{\theta}_1 = 1.3$

In this case, the conditional power curve is steeper and the optimum occurs at a higher n_2^*.

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

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

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

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

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

As $\hat{\theta}_1$ 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
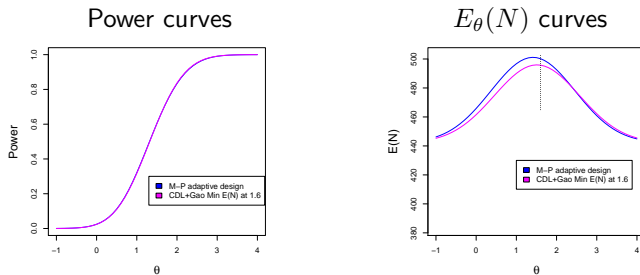
With the type I error rate fixed at 0.025, matching the MP design's power at one value of θ will match the whole power curve.

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.
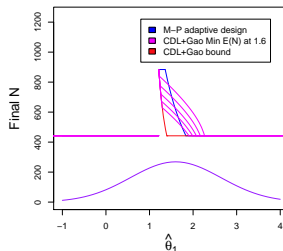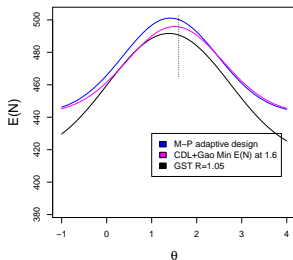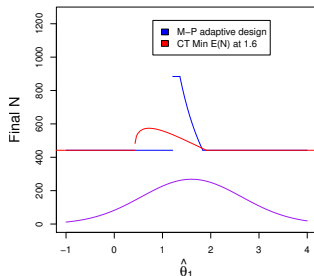

Shapes of optimised sample size rules suggest it would help to increase n_2^* at lower values of $\widehat{\theta}_1$ — but this is not allowed in the CDL+Gao framework.

If we use a **Combination Test,** such adaptations are permissible.

# Optimal sample size rule for a combination test design

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

The design with this sample size rule for $\gamma = 0.25/(4\,\sigma^2)$ 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".

Optimal sample size rule for a combination test design

$E_\theta(N)$ curves
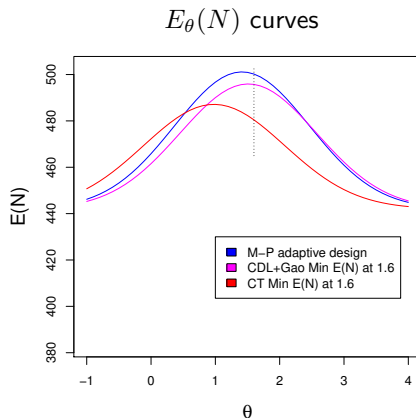
When using a combination test, we can increase sample size for $\widehat{\theta}_1$ below the CDL+Gao region.

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

Further extensions

1. We can allow recruitment to terminate at the interim analysis, so the minimum sample size is $n_2 = 416$, rather than 442.

2. We can use a general combination test or, equivalently, a general conditional type I error function (Proschan & Hunsberger, *Biometrics*, 1995).

3. We can minimise other average 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 two-stage designs deal neatly with the "pipeline" subjects arising when there is a delayed response.

For the chosen criteria, they will give the best possible sampling and decision rules with $n_1 = 208$ and n_2 in the range 416 to 884.
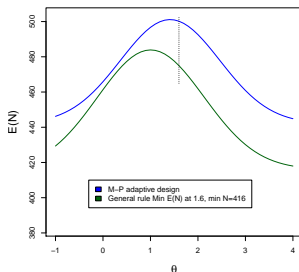
General sampling rule, early termination of recruitment

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

Sample size rule
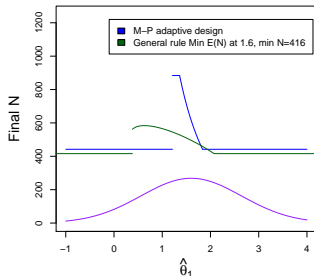
$E_\theta(N)$ curves

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".

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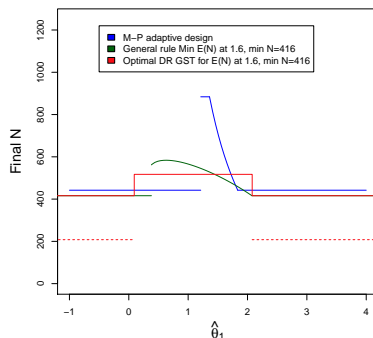
The sampling rule approximates that of the general adaptive method, but with a step function rather than a continuous sample size function.

Properties of the optimal DR GST

The optimised DR GST has an almost identical $E_\theta(N)$ curve to the general rule that uses the continuum of possible sample sizes.
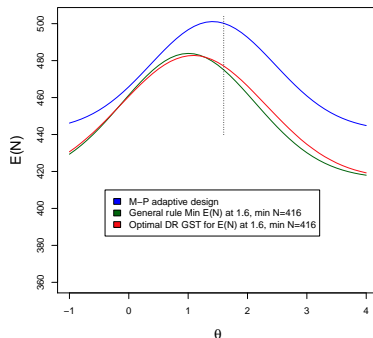
Thus, as Jennison & Turnbull (*Biometrika*, 2006) found for an immediate response, there is minimal benefit from fine-tuning the total sample size in response to interim data.

3. Conclusions

Although JT had shown that adaptive designs offer at most a slight improvement on GSTs, it is appropriate to consider 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.

2. We have developed MP's idea of spending resources where they have the greatest benefit — and we have found adaptive designs with greater efficiency.

3. Our most general solution is very similar to a "Delayed Response GST", as proposed by Hampson & Jennison (2013). Such a design offers the benefits of established group sequential methods and extensions, e.g., error spending tests.

4. The adaptive approach (start small, then ask for more) *can* give good trial designs — but there are pitfalls to avoid!