Group Sequential Designs and Sample Size Re-estimation

Christopher Jennison

Department of Mathematical Sciences,
University of Bath, UK
http://people.bath.ac.uk/mascj

Statistical Planning of Translational Studies

Göttingen

March, 2024

Motivation: Phase 3 clinical trials

Phase III trials are conducted as the last stage in the drug development process.

Regulators customarily require a hypothesis test to reach significance at the one-sided 2.5% level.

Studies may recruit hundreds, or even thousands, of subjects at a cost of as much as €10k to €50k per patient.

The time taken to reach a conclusion eats into the limited patent lifetime remaining to the company developing the drug.

Thus, there are strong incentives to reach an early conclusion for either a positive or negative decision.

However, the overall type I error rate must be protected.

Stopping a trial as soon as a conclusion can be reached is also desirable in early phase studies.

Group sequential tests (GSTs)

1.1. Introduction

1.2. Sequential distribution theory

1.3. Computations for group sequential tests

1.4. Benefits of group sequential testing

1.5. Error spending tests

2.1. Motivation for adaptive designs

2.2. Combination tests

2.3. Sample size re-estimation for a response variance

2.4. Sample size re-assessment in response to $\widehat{\theta}$

Critique of the “promising zone” approach

1.1 Group sequential tests: Introduction

Suppose a new treatment (Treatment A) is to be compared to a placebo or positive control (Treatment B) in a Phase III trial.

The treatment effect θ for the **primary endpoint** represents the advantage of Treatment A over Treatment B.

If $\theta > 0$, Treatment A is more effective.

We wish to test the null hypothesis H_0: $\theta \leq 0$ against $\theta > 0$ with

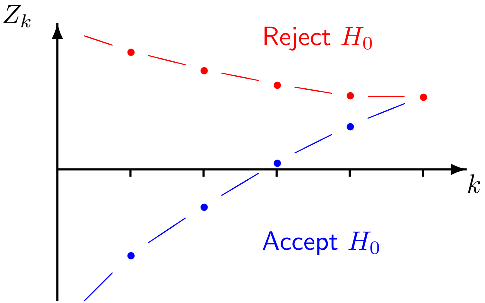
$$P_{\theta=0}\{\text{Reject } H_0\} = \alpha,$$

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

In a group sequential trial, data are examined on a number of occasions to see if an early decision may be possible.

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 in favour of $\theta > 0$.

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

Benefits of group sequential testing

Earlier decisions

Group sequential testing can speed up the process to introduce an effective new treatment.

Fewer patients recruited

Expected sample sizes for group sequential designs are, typically, around 60 to 70% of the fixed sample size for a trial with the same type I error rate and power.

Stopping failing trials early

Early stopping "for futility" can release resources to continue the development of other promising treatments.

1.2 Joint distribution of parameter estimates

Reference: Ch. 11 of *Group Sequential Methods with Applications to Clinical Trials*, Jennison & Turnbull, 2000 (hereafter, JT).

Let $\widehat{\theta}_k$ denote the estimate of θ based on data at analysis k.

The information for θ at analysis k is

$$\mathcal{I}_k = \{\mathsf{Var}(\widehat{\theta}_k)\}^{-1}, \quad k = 1, \ldots, K.$$

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

In many situations, $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ are approximately multivariate normal,

$$\widehat{\theta}_k \sim N(\theta, \mathcal{I}_k^{-1}), \quad k = 1, \ldots, K,$$

and

$$\mathsf{Cov}(\widehat{\theta}_{k_1}, \widehat{\theta}_{k_2}) = \mathsf{Var}(\widehat{\theta}_{k_2}) = \mathcal{I}_{k_2}^{-1} \quad \text{for } k_1 < k_2.$$

Sequential distribution theory

The joint distribution of $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ can be derived directly for:

θ a single normal mean,

$\theta = \mu_A - \mu_B,$ comparing two normal means.

The canonical distribution also applies when θ is a parameter in:

a general normal linear model,

a general model fitted by maximum likelihood (large sample theory).

Thus, theory supports general comparisons, including:

crossover studies,

analysis of longitudinal data,

comparisons adjusted for covariates.

Canonical joint distribution of Z-statistics

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

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

For these statistics,

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

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

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

Canonical joint distribution of score statistics

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

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

The score statistics possess the "independent increments" property,

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

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

Survival data

The canonical joint distributions also arise for

a) estimates of a parameter in Cox's proportional hazards regression model,

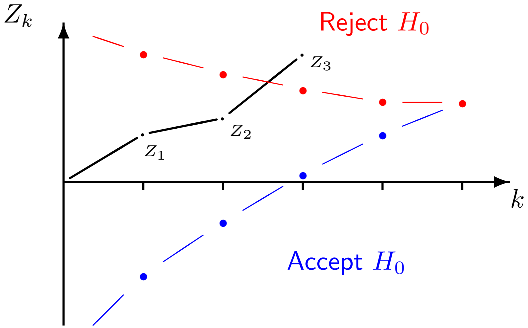
b) log-rank statistics for comparing two survival curves.

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

The "error spending" approach can be used to define group sequential tests that can handle unpredictable and unevenly spaced information levels.

Reference: "Group-sequential analysis incorporating covariate information", Jennison & Turnbull (*JASA*, 1997).

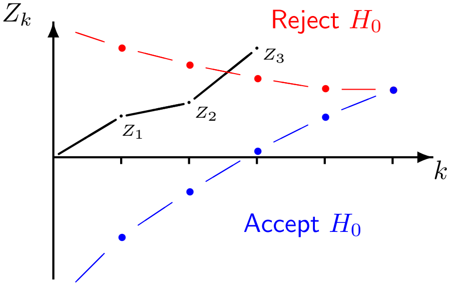
1.3 Computations for group sequential tests (GSTs)



In order to find $P_\theta\{\text{Reject } H_0\}$, etc., we need to calculate the probabilities of basic events such as

$$a_1 < Z_1 < b_1, \;\; a_2 < Z_2 < b_2, \;\; Z_3 > b_3.$$

Computations for group sequential tests



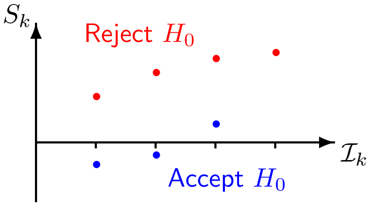
Probabilities such as $P_\theta\{a_1 < Z_1 < b_1,\ \ a_2 < Z_2 < b_2,\ \ Z_3 > b_3\}$ can be computed by repeated numerical integration (JT, Ch. 19).

Combining these probabilities yields type I error rate, power, expected sample size, etc., of a group sequential design.

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

One-sided tests: The Pampallona & Tsiatis family

To test H_0: $\theta \le 0$ against the *one-sided* alternative $\theta > 0$ with type I error probability α and power $1-\beta$ at $\theta = \delta$.



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

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

The computational methods described above can be used to find C_1, C_2 and $\mathcal{I}_K$ such that the test has the specified error rates.

Reference: Pampallona & Tsiatis (*JSPI*, 1994).

1.4 Benefits of group sequential testing

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

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

where Φ is the standard normal CDF.

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

A GST with K analyses will need to be able to continue to a maximum information level $\mathcal{I}_K$, greater than $\mathcal{I}_{fix}$.

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

We call $R = \mathcal{I}_K / \mathcal{I}_{fix}$ the *inflation factor* of a group sequential test.

Benefits of group sequential testing

One-sided GSTs with binding futility boundaries, minimising $\{\mathbb{E}_0(\mathcal{I}) + \mathbb{E}_\delta(\mathcal{I})\}/2$ for K equally sized groups, $\alpha = 0.025$, $1 - \beta = 0.9$ 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			*Minimum*
K	1.01	1.05	1.1	1.2	1.3	*over R*
2	80.8	74.7	**73.2**	73.7	75.8	73.0 at R=1.13
3	76.2	69.3	66.6	**65.1**	65.2	65.0 at R=1.23
5	72.2	65.2	62.2	59.8	**59.0**	58.8 at R=1.38
10	69.2	62.2	59.0	56.3	**55.1**	54.2 at R=1.6
20	67.8	60.6	57.5	54.6	**53.3**	51.7 at R=1.8

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

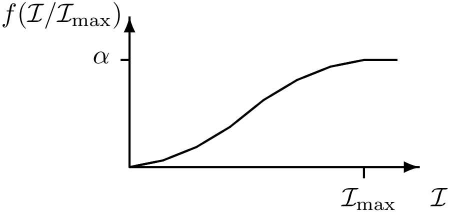
$\mathbb{E}(\mathcal{I}) \searrow$ as $R \nearrow$ up to a point.

1.5 Error spending tests (JT Ch. 7)

When the sequence $\mathcal{I}_1$, $\mathcal{I}_2$, $\dots$ 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 spending function $f(\mathcal{I}/\mathcal{I}_{\max})$



The boundary at analysis k is set to give cumulative type I error probability $f(\mathcal{I}_k/\mathcal{I}_{\max})$.

If $\mathcal{I}_{\max}$ is reached without rejecting H_0, then H_0, is accepted.

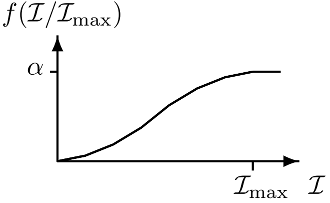
One-sided error spending tests

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 $f(\mathcal{I}/\mathcal{I}_{\max})$, and type II error probability β according to $g(\mathcal{I}/\mathcal{I}_{\max})$.

One-sided error-spending tests

Analysis 1:

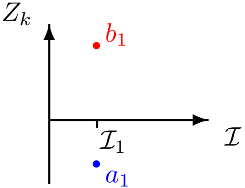
Observed information $\mathcal{I}_1$.

Reject H_0 if $Z_1 > b_1$, where

$$P_{\theta=0}\{Z_1 > b_1\} = f(\mathcal{I}_1/\mathcal{I}_{\max}).$$

Accept H_0 if $Z_1 < a_1$, where

$$P_{\theta=\delta}\{Z_1 < a_1\} = g(\mathcal{I}_1/\mathcal{I}_{\max}).$$



One-sided error-spending tests

Analysis 2: Observed information $\mathcal{I}_2$

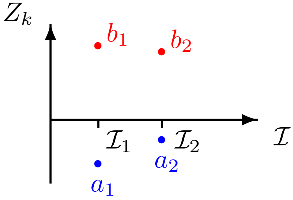
Reject H_0 if $Z_2 > b_2$, where

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

— *note that, for now, we assume the futility boundary is binding.*

Accept H_0 if $Z_2 < a_2$, where

$$P_{\theta=\delta}\{a_1 < Z_1 < b_1, Z_2 < a_2\} = g(\mathcal{I}_2/\mathcal{I}_{\max}) - g(\mathcal{I}_1/\mathcal{I}_{\max}).$$



One-sided error-spending tests

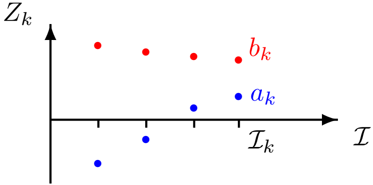
Analysis k: Observed information $\mathcal{I}_k$

Find a_k and b_k to satisfy

$$P_{\theta=0}\{a_1 < Z_1 < b_1, \ldots, a_{k-1} < Z_{k-1} < b_{k-1}, Z_k > b_k\}$$
$$= f(\mathcal{I}_k/\mathcal{I}_{\max}) - f(\mathcal{I}_{k-1}/\mathcal{I}_{\max}),$$

and

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

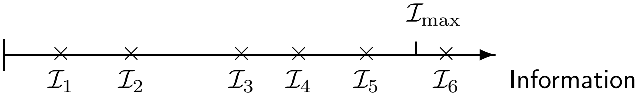


Remarks on error spending tests

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

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

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



3. If a maximum of K analyses is specified, the study terminates at analysis K with $f(\mathcal{I}_K/\mathcal{I}_{\max})$ defined to be α.

4. If the trial ends with $\mathcal{I}_K > \mathcal{I}_{\max}$, we set $f(\mathcal{I}_K/\mathcal{I}_{\max}) = \alpha$.

Then, b_K is chosen to give cumulative type I error probability α and we set $a_K = b_K$.

Remarks on error spending tests

5. The value of $\mathcal{I}_{\max}$ can be chosen so that boundaries converge at the final analysis when, say,

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

6. In a one-sided test with ρ-family error spending function, type I error probability is spent as

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

and type II error probability as

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

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

Barber & Jennison (*Biometrika*, 2002) show the ρ-family provides tests with excellent efficiency for a given number of analyses K and inflation factor R.

2.1 Adaptive clinical trials: Motivation

Wall Street Journal, July 2006:

FDA Signals it's Open to Drug Trials that Shift Midcourse

Adaptive designs may allow trials to be adjusted:

- Route more patients to the treatment that seems to work best
- Drop treatments that don't seem to be effective
- Add more of the type of patients ... reacting best to a particular treatment
- Merge two different phases of drug development into one trial

With views from:

Bob O'Neill , FDA	Michael Krams, Wyeth
Paul Gallo, Novartis	Don Berry, M. D. Anderson Cancer Center
Tom Fleming, Univ. Washington	Bruce Turnbull, Cornell University

2.2 Combination tests

Suppose we run a clinical trial adaptively in two stages:

Set the design of Stage 1, then conduct this part of the trial,

Analyse results from Stage 1,

Consider external information, if appropriate.

Set the design of Stage 2, informed by Stage 1 results and external information,

Conduct Stage 2,

Analyse the results from Stage 2.

How can we test a null hypothesis with proper protection of the type I error rate?

Combination tests

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 data 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, if $\theta = 0$, $P^{(1)}$ and $P^{(2)}$ are independent $U(0, 1)$ variates.

The inverse χ^2 combination test

Reference Bauer & Köhne (*Biometrics*, 1994).

Initial design

Define H_0 and specify the **inverse $\boldsymbol{\chi^2}$ combination test.**

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

Stage 1

Observe the one-sided P-value, $P^{(1)}$, based on Stage 1 data.

Design Stage 2 in the light of Stage 1 data.

Stage 2

Observe the P-value, $P^{(2)}$, based on **only** Stage 2 data.

NB: Under $\theta = 0$, $P^{(1)} \sim U(0,1)$, $P^{(2)} \sim U(0,1)$, independent.

Bauer & Köhne's inverse χ^2 combination test

Bauer & Köhne's test rejects H_0 for low values of $P^{(1)}\,P^{(2)}$.

If $P \sim U(0,1)$, then

$$-\ln(P) \,\sim\, \mathsf{Exp}\,(1) = \frac{1}{2}\,\chi_2^2.$$

Thus, under $\theta = 0$,

$$-\ln(P^{(1)}\,P^{(2)}) \,\sim\, \frac{1}{2}\,\chi_4^2.$$

Combining the two P-values in an overall test, we reject H_0 if

$$-\ln(P^{(1)}\,P^{(2)}) \,>\, \frac{1}{2}\,\chi^2_{4,\,1-\alpha}.$$

If $\theta < 0$, then $P^{(1)}$ and $P^{(2)}$ are stochastically larger than $U(0,1)$ random variables and the type I error rate is less than α.

This χ^2 test was originally proposed for combining results of several studies by R. A. Fisher in 1932.

The inverse normal combination test

Initial design

Specify the **inverse normal combination 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

The combination test is based on the statistic $w_1 Z^{(1)} + w_2 Z^{(2)}$.

Under $\theta = 0$, $Z^{(1)}$ and $Z^{(2)}$ are independent $N(0,1)$ so, with $w_1^2 + w_2^2 = 1$,

$$w_1 Z^{(1)} + w_2 Z^{(2)} \sim N(0,1).$$

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 essential that w_1 and w_2 are pre-specified and not changed in response to observed data.

2.3 Sample size re-estimation for a response variance

A combination test can be used to protect the type I error rate when a trial's sample size is changed.

Consider a two-treatment comparison in which observations on Treatments A and B, respectively, are distributed as

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

Objective

It is desired to test H_0: $\theta = \mu_A - \mu_B \leq 0$ against $\theta > 0$ with type I error rate α and power $1 - \beta$ at $\theta = \delta$.

In the case of known variance, the sample size formula

$$n = \frac{\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2 \, 2\,\sigma^2}{\delta^2} \tag{1}$$

gives the required value of n, the sample size per treatment.

However, in practice, only an estimate of σ^2 is usually available.

Sample size re-estimation for a response variance

We can follow an adaptive approach, using an estimate of σ^2 from early trial data to modify the initial choice of sample size.

Initial design

Specify a two-stage adaptive design, using the inverse χ^2 combination rule to test H_0: $\theta \leq 0$ against $\theta > 0$.

Use an initial estimate σ_0^2 in the sample size formula (1) to obtain a sample size of n_0 per treatment.

Stage 1

Conduct Stage 1 with $n_1 = n_0/2$ subjects per treatment.

Observe estimates $\widehat{\theta}_1$, $\hat{\sigma}_1^2$ and the t-statistic t_1 for testing H_0.

Convert t_1 to a one-sided P-value, $P^{(1)} = P_{\theta=0}\{T_{2n_1-2} > t_1\}$.

Sample size re-estimation for a response variance

After Stage 1

Calculate a new Stage 2 sample size of n_2 per treatment arm.

Here, n_2 may be obtained simply by using the new variance estimate $\hat{\sigma}_1^2$ in the original sample size formula.

Or, n_2 might be chosen to give conditional power $1-\beta$ given $P^{(1)}$, assuming $\theta = \widehat{\theta}_1$ and $\sigma^2 = \hat{\sigma}_1^2$.

Stage 2

Calculate the t-statistic t_2 for testing H_0 based on Stage 2 data alone, and obtain the P-value $P^{(2)} = P_{\theta=0}\{T_{2n_2-2} > t_2\}$.

The inverse χ^2 combination test, which rejects H_0 if

$$-\ln(P^{(1)}\,P^{(2)}) > \frac{1}{2}\,\chi^2_{4,\,1-\alpha}$$

has type I error rate exactly equal to α.

Sample size re-estimation for a response variance

The above approach adds to the variety of methods for dealing with an unknown parameter, ϕ, that affects sample size.

Internal pilot:

Wittes & Brittain (*Statistics in Medicine*, 1990) proposed a simple "plug in" of the current estimate $\hat{\phi}$ to update sample size. Bias in the final $\hat{\phi}$ tends to cause a small inflation of the type I error rate.

Blinded variance estimation:

Friede & Miller (*Applied Statistics*, 2012) show that, for normally distributed data, sample size modification based on a blinded estimate of σ^2 leads to almost zero type I error rate inflation.

Information monitoring:

Mehta & Tsiatis (*Drug Information Journal*, 2001) "plug in" estimated information in an error spending group sequential design. Typically, this leads to a small inflation of the type I error rate.

2.4 Sample size re-assessment in response to $\widehat{\theta}$

In the early 2000s, the possibility of adaptive design prompted interest in procedures that increase sample size in response to a low interim estimate of the treatment effect.

The objective here is to increase power, recognising that the effect size used in the original power calculation was over-optimistic.

The resulting procedures have an overall maximum possible sample size but, depending on the observed data, the actual sample size can be smaller than this — just like a GST.

Such designs can be viewed as group sequential tests in which group sizes are chosen based on the data observed so far.

The decision rule to reject or accept H_0 must take account of the data-dependent group sizes in order to protect the type I error rate.

Norbert Schmitz (1993) had proposed similar designs, which he called "Sequentially planned sequential decision procedures".

Sample size re-assessment in response to $\widehat{\theta}$

JT (*Statistics in Medicine*, 2003 and 2006) discussed proposals for adaptive designs that incorporate sample size re-assessment.

They claimed one could usually do better by

- Thinking carefully about power when designing the trial,
- Planning a maximum sample size to achieve this power,
- Using a GST to stop early when possible.

Proposals for adaptive designs with sample size re-assessment continued to appear.

JT (*Biometrika*, 2006) computed properties of optimal adaptive GSTs ("Schmitz" designs) and optimal GSTs.

They showed that familiar group sequential designs provide almost all the available efficiency gains — whereas some proposals for adaptive designs can be quite inefficient.

Mehta & Pocock's "Promising zone" design

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

> *"Adaptive increase in sample size when interim results are promising: a practical guide with examples".*

Mehta & Pocock (MP) refer to the work of JT, saying

> *"These results are of great theoretical interest but of limited practical value for sponsors of industry trials."*

Jennison & Turnbull (*Statistics in Medicine*, 2015) discussed the MP designs and demonstrated various inefficiencies.

They proposed new, decision theoretic rules for sample size re-assessment and a different final hypothesis test.

Hsiao, Liu & Mehta (*Biometrical Journal*, 2019) wrote the paper "*Optimal promising zone designs*". They refer to JT's proposals as the "gold standard" and propose variations on these methods.

Mehta & Pocock's "Promising zone" design

The following slides summarise the content of JT (*SiM*, 2015)

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

JT compare **adaptive trial designs** that

Start with a fixed sample size design,

Examine interim data,

Add observations to increase power when appropriate

with **group sequential designs** that

Specify the desired power function,

Set the maximum sample size appropriately,

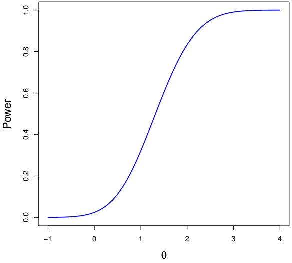
Stop at an interim analysis if data support an early conclusion.

Mehta & Pocock's "Promising zone" design

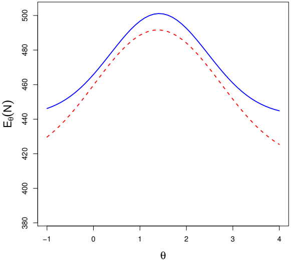
JT note that the "unconditional" properties of a trial design are important.

Suppose the treatment effect is denoted by θ and a trial is conducted to test H_0: $\theta \leq 0$ vs $\theta > 0$.

Common power curve



$E_\theta(N)$ curves



If two designs have the same power curve, then the design with the lower expected sample size function is to be preferred.

MP's Example 1: Phase 3 trial of schizophrenia treatments

A new drug is to be tested against an active comparator.

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

The initial plan is for a total of 442 patients (221 per treatment) which gives power 0.61 if $\theta = 1.6$ — but power 0.8 is desired.

An interim analysis is planned after observing $n_1 = 208$ responses (104 on each treatment).

With staggered accrual and the 26-week time to response, another 208 pipeline subjects will have been randomised but followed up for less than 26 weeks by the interim analysis time.

At the interim analysis, the total sample size will be revised to a value n_2, where n_2 is in the range 442 to 884.

The choice of n_2 will be based on the conditional power under the current estimate of θ.

MP's Example 1: Protecting the type I error probability

Chen, DeMets & Lan (*Statistics in Medicine*, 2004) consider a trial testing H_0: $\theta \leq 0$ vs $\theta > 0$ with type I error probability α.

Suppose an interim analysis is performed after a certain fraction of the total sample size has been observed.

Define the conditional power under treatment effect θ as

$$CP(\theta) = Pr_\theta\{H_0 \text{ will be rejected} \,|\, \text{Interim estimate} = \widehat{\theta}_1\}.$$

Chen, DeMets & Lan (CDL) show that if

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

The sample size is increased,

A standard fixed sample analysis is carried out,

then the type I error probability will be no greater than α.

MP's Example 1: The "Promising zone" design

MP control the type I error rate using an extension of the CDL result, due to Gao, Ware & Mehta (*Biopharm. Statistics*, 2008).

The total sample size, n_2, is a function of $CP(\widehat{\theta}_1)$, the conditional power under the current treatment effect estimate .

In MP's Example 1, the sample size rule is

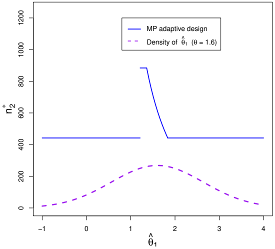
Favorable	$CP(\widehat{\theta}_1) \geq 0.8$	Continue to $n_2 = 442$,
Promising	$0.365 \leq CP(\widehat{\theta}_1) \leq 0.8$	Increase n_2,
Unfavorable	$CP(\widehat{\theta}_1) < 0.365$	Continue to $n_2 = 442$.

The "Promising zone" is simply the region where $CP(\widehat{\theta}_1) \leq 0.8$ and the Gao et al. extension of the CDL result can be employed.

In this zone, n_2 is increased to give $CP(\widehat{\theta}_1) = 0.8$, subject to a cap of $n_2 = 884$.

MP's Example 1: The "Promising zone" design

The sample size re-assessment rule for MP's "Promising zone" design is shown below.



Note that the region in which changes to the sample size occur is small when compared to the distribution of $\widehat{\theta}_1$.

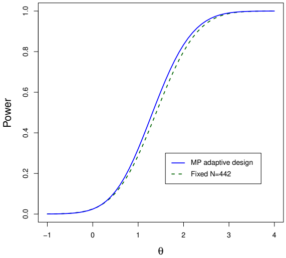
The estimate $\widehat{\theta}_1$ has a double role in $CP(\widehat{\theta}_1)$: it is both the current data and the value of θ at which conditional power is calculated.

This helps explain the high sensitivity of n_2 to $\widehat{\theta}_1$.

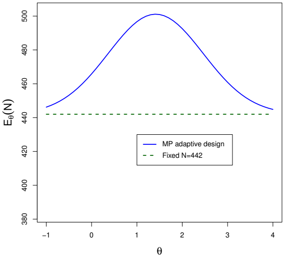
MP's Example 1: The "Promising zone" design

We can compare the overall, "unconditional" properties of the MP design with those of the fixed sample size design with $N = 442$.

Power curves



$E_\theta(N)$ curves



Although it is 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.

The cost of this increase in power is a considerably higher expected sample size function.

MP's Example 1: What might we do instead?

We could opt for:

(i) A fixed sample design with $N = 490$.

(ii) A group sequential design with $n_1 = 208$ and stopping rule:

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 to $n_2 = 514$

At analysis 2

If $Z_2 \geq 2.00$ Reject H_0

If $Z_2 < 2.00$ Accept H_0.

Error spending GST,
Rho-family, $\rho = 2$,
Power 0.8 at $\theta = 1.9$.

If the trial stops at analysis 1, "pipeline subjects" are not used in hypothesis testing, but they will contribute to $E_\theta(N)$.

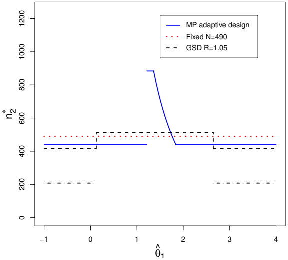
MP's Example 1: What might we do instead?

Sample size rules for

Mehta & Pocock design (MP),

Fixed sample size trial (N = 490),

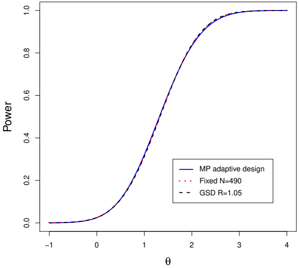
Group sequential design (GSD).



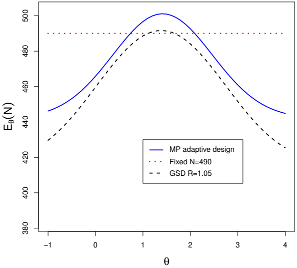
Two black lines show $n_1 = 208$, the number of observations, and $n_1^* = 416$, the number of patients enrolled at analysis 1.

MP’s Example 1: Properties of three trial designs

Power curves



$E_\theta(N)$ curves



By construction, the fixed sample design and GSD have power curves matching that of the MP design.

The fixed sample size design is more efficient than the MP design at important values of θ between 1 and 2.

Despite the “pipeline” patients, the GSD is more efficient than the MP design at all values of θ.

MP's Example 1: Refining the "Promising zone" design

Jennison & Turnbull (2015) discuss how to modify Mehta & Pocock's design to improve its efficiency.

They recommend dropping the CDL methodology.

Instead, they propose using an inverse normal combination test to combine data from before and after sample size re-assessment.

JT derive a sample size rule in which n_2 is chosen to maximise a combined objective

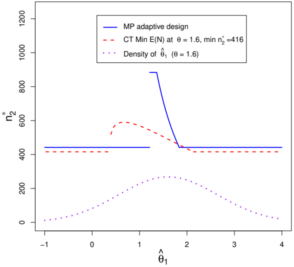
$$Pr_{\theta=1.6}\{\text{Reject } H_0 \,|\, \widehat{\theta}_1, n_2\} - \gamma(n_2 - 442).$$

Here, the tuning parameter $\gamma > 0$ represents a "rate of exchange" between conditional power and sample size.

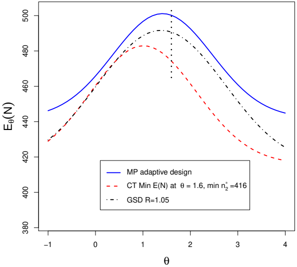
JT prove that this sample size rule achieves the minimum possible $E_{\theta=1.6}(N)$ among rules that achieve the same power when $\theta = 1.6$.

MP's Example 1: Refining the "Promising zone" design

Sample size rules



$E_\theta(N)$ curves



JT's refined design has the same power curve as the MP design.

It increases n_2 by a smaller amount of a wider range of $\widehat{\theta}_1$.

In the refined design, there is no stopping at the interim analysis.

Responses of "pipeline" patients are included in the final analysis.

The refined design reduces $E_\theta(N)$ over a wide range of θ values.

Sample size re-assessment in response to $\widehat{\theta}$: Conclusions

Group sequential designs

GSDs offer an excellent way to adapt sample size to observed data.

Efficient error spending designs are available.

Designs with sample size re-assessment

Designs with SSR offer a little extra efficiency, but many proposed SSR designs do not achieve this efficiency.

If you want an SSR design, consider the approach of JT (2015).

Can you start small and ask for more?

Both GSDs and SSR designs require a commitment to obtain their implied maximum sample size, should this be needed.

Dealing with pipeline data:

See"Group sequential tests for delayed responses" by Hampson & Jennison (*JRSS, B*, 2013).