Optimising Group Sequential and Adaptive Designs:

Where Frequentist meets Bayes

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

http://people.bath.ac.uk/mascj

Lancaster University

December 2019

Group Sequential and Adaptive Designs

Group sequential and adaptive clinical trial designs have been proposed for a number of important applications:

Early stopping for efficacy or futility,

Sample size modification,

Treatment selection and testing (seamless Phase 2/3 trials),

Population selection and testing (enrichment designs).

There are usually options to choose from within such a design.

How should one make such choices and assess the end result?

Choosing a group sequential or adaptive design

A Phase 3 trial must protect the type I error rate.

This can be a complex problem when testing multiple null hypotheses — type I error rate must be controlled over a high-dimensional region.

We wish to be efficient, gaining high power with low sample size.

How should we make decisions:

At interim analyses?

At the final analysis?

Type I error rate is a frequentist property.

But Bayesian methods have advantages when optimising a design.

Outline of talk

1. Monitoring clinical trials

   Group sequential stopping rules

   Optimising the stopping boundary

2. Sample size re-estimation

   Optimising the Mehta-Pocock "Promising zone" design

3. Seamless Phase 2/3 designs

   Designs that protect family-wise error rate,

   Optimising decision rules and sample size allocation.

4. Enrichment designs

   Adaptive enrichment in response to interim data.

   Optimising the decision rule for when to enrich.

1. A group sequential clinical trial

Consider a Phase 3 clinical trial comparing a new treatment against a standard.

Let θ denote the "effect size", a measure of the improvement in the new treatment over the standard.

We shall test the null hypothesis H_0: $\theta \leq 0$ against $\theta > 0$.

Rejecting H_0 allows us to conclude the new treatment is superior.

We allow type I error probability α for rejecting H_0 when it is true.

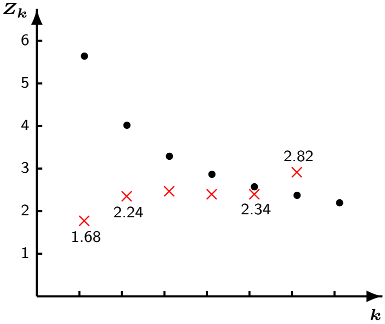
We specify power $1 - \beta$ as the probability of rejecting H_0 when $\theta = \delta$. Here δ is, typically, the minimal clinically significant treatment difference.

The trial design, including the method of analysis and stopping rule, must be set up to attain these error rates.

An early example: The BHAT trial

DeMets et al. (*Cont. Clin. Trials*, 1984) report on the Beta-Blocker Heart Attack Trial, that compared propanolol with placebo.

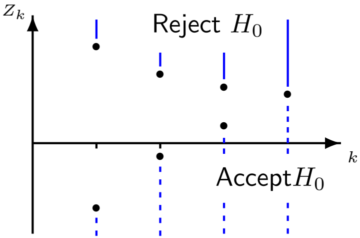
An "O'Brien and Fleming" stopping boundary was defined with overall type I error probability 0.025.



The trial stopped after the 6th of 7 planned analyses.

Group sequential tests: Stopping for futility

Adding a lower boundary allows stopping when there is little chance of a positive conclusion.



Rosner & Tsiatis (*Statistics in Medicine*, 1989) carried out retrospective analyses of 72 cancer studies of the U.S. Eastern Co-operative Oncology Group.

Had group sequential stopping rules been applied, early stopping (mostly to accept H_0) would have occurred in ∼80% of cases.

Requirements for clinical trial designs

We seek designs which:

Achieve specified type I error rate and power,

Stop early, on average, under key parameter values,

Can be applied to a variety of response types.

We shall present distribution theory which shows that a common set of methods can be applied to many data types.

To define efficient tests, we shall formulate and solve an optimal stopping problem.

Sequential distribution theory

Let $\widehat{\theta}_k$ denote the estimate of the treatment effect θ at analysis k.

Information for θ at analysis k is $\mathcal{I}_k = \{\mathsf{Var}(\widehat{\theta}_k)\}^{-1},\ k = 1, \ldots, K.$

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

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

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

References:

Jennison & Turnbull, *JASA*, 1997,

Scharfstein et al, *JASA*, 1997.

An optimal stopping problem

Consider a trial designed to test H_0: $\theta \leq 0$ vs $\theta > 0$, with:

Type I error rate α,

Power $1 - \beta$ at $\theta = \delta$,

Up to K analyses.

A fixed sample test needs information

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

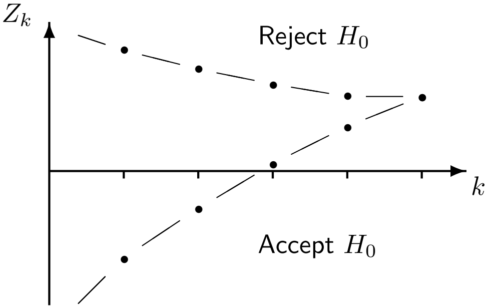
We set the maximum information to be

$$\mathcal{I}_{max} = R\,\mathcal{I}_{fix},$$

where $R > 1$, with equal increments between analyses.

Optimal group sequential tests

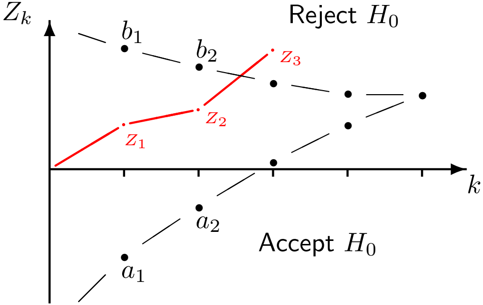
The error rates impose two constraints on the $2K - 1$ boundary points — leaving a high dimensional space of possible boundaries.



We shall look for a boundary that minimises

$$\{E_0(\mathcal{I}) + E_\delta(\mathcal{I})\}/2.$$

Computations for group sequential tests



We need to be able to calculate the probabilities of basic events such as

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

Combining such probabilities gives key properties, such as $Pr_\theta\{$Reject $H_0\}$ and $E_\theta(\mathcal{I})$.

Numerical integration

We can write probabilities as nested integrals, e.g.,

$$Pr\{a_1 < Z_1 < b_1,\, a_2 < Z_2 < b_2,\, Z_3 > b_3\} =$$

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} \int_{b_3}^{\infty} f_1(z_1)\; f_2(z_2|z_1)\; f_3(z_3|z_2)\; dz_3\; dz_2\; dz_1.$$

Applying numerical integration, we replace each integral by a sum of the form

$$\int_a^b f(z)\, dz \;=\; \sum_{i=1}^{n} w(i)\, f(z(i)),$$

where $z(1), \ldots, z(n)$ is a grid of points from a to b.

Numerical integration

Thus, we have

$$Pr\{a_1 < Z_1 < b_1,\, a_2 < Z_2 < b_2,\, Z_3 > b_3\} \;\approx$$

$$\sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \sum_{i_3=1}^{n_3} w_1(i_1)\, f_1(z_1(i_1))\, w_2(i_2) f_2(z_2(i_2)|z_1(i_1))$$

$$w_3(i_3)\, f_3(z_3(i_3)|z_2(i_2)).$$

Multiple integrations and summations will arise, e.g., for an outcome at analysis k,

$$\sum_{i_1=1}^{n_1} \ldots \sum_{i_k=1}^{n_k} w_1(i_1)\, f_1(z_1(i_1))\, w_2(i_2) f_2(z_2(i_2)|z_1(i_1))$$

$$\ldots\; w_k(i_k)\, f_k(z_k(i_k)|z_{k-1}(i_{k-1})).$$

Numerical integration

In the multiple summation

$$\sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \ldots \sum_{i_k=1}^{n_k} w_1(i_1)\, f_1(z_1(i_1))\, w_2(i_2) f_2(z_2(i_2)|z_1(i_1))$$

$$\ldots\; w_k(i_k)\; f_k(z_k(i_k)|z_{k-1}(i_{k-1})),$$

the structure of the k nested summations is such that the computation required is of the order of $k-1$ double summations.

Using Simpson's rule with 100 to 200 grid points per integral can give accuracy to 5 or 6 decimal places.

For details of efficient sets of grid points, see Ch. 19 of *Group Sequential Methods with Applications to Clinical Trials* by Jennison and Turnbull (2000).

Finding optimal group sequential tests

Recall, we want a group sequential test of H_0: $\theta \leq 0$ vs $\theta > 0$ with

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

$Pr_{\theta=\delta}\{\text{Accept } H_0\} = \beta$,

Analyses at $\mathcal{I}_k = (k/K)\,\mathcal{I}_{max}, \;\; k = 1, \ldots, K$,

Minimum possible value of $\{E_0(\mathcal{I}) + E_\delta(\mathcal{I})\}/2$.

We deal with constraints on error rates by introducing Lagrangian multipliers to create the *unconstrained problem* of minimising

$$\{E_0(\mathcal{I}) + E_\delta(\mathcal{I})\}/2 + \lambda_1 Pr_{\theta=0}\{\text{Reject } H_0\} + \lambda_2\, Pr_{\theta=\delta}\{\text{Accept } H_0\}.$$

We shall find a pair of multipliers $(\lambda_1,\, \lambda_2)$ such that the solution has type I and II error rates α and β, then this design will solve the *constrained problem* too.

Bayesian interpretation of the Lagrangian approach

Suppose we put a prior on θ with $Pr\{\theta = 0\} = Pr\{\theta = \delta\} = 0.5$ and specify costs of

1	per unit of information observed,
$2\,\lambda_1$	for rejecting H_0 when $\theta = 0$,
$2\,\lambda_2$	for accepting H_0 when $\theta = \delta$.

Then, the total Bayes risk is

$$\{E_0(\mathcal{I})+E_\delta(\mathcal{I})\}/2+\lambda_1\, Pr_{\theta=0}\{\text{Reject } H_0\}+\lambda_2\, Pr_{\theta=\delta}\{\text{Accept } H_0\},$$

just as in the Lagrangian problem.

An advantage of the Bayes interpretation is that it can give insight into solving the problem by using “Dynamic Programming” or “Backwards Induction”.

Solution by Dynamic Programming

Denote the posterior distribution of θ given $Z_k = z_k$ at analysis k by

$$p^{(k)}(\theta|z_k), \quad \theta = 0, \delta.$$

At the final analysis, K

There is no further sampling cost, so compare decisions

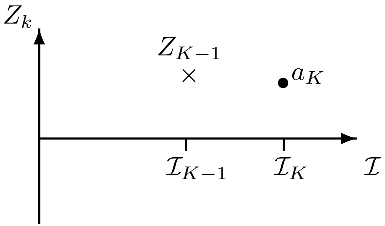
$$\begin{aligned} &\text{Reject } H_0: \qquad E(\text{Cost}) = 2\,\lambda_1\,p^{(K)}(0|z_K), \\ &\text{Accept } H_0: \qquad E(\text{Cost}) = 2\,\lambda_2\,p^{(K)}(\delta|z_K). \end{aligned}$$

The boundary point a_K is the value of z_K where these expected losses are equal.

The optimum decision rule is to reject H_0 for $Z_K > a_K$.

Dynamic Programming

At analysis $K-1$



If the trial stops at this analysis, there is no further cost of sampling and the expected additional cost is

Reject H_0: $\quad 2\,\lambda_1\,p^{(K-1)}(0|z_{K-1})$,

Accept H_0: $\quad 2\,\lambda_2\,p^{(K-1)}(\delta|z_{K-1})$.

At analysis $K-1$

If the trial continues to analysis K, the expected additional cost is

$$1 \times (\mathcal{I}_K - \mathcal{I}_{K-1})$$

$$+\, 2\,\lambda_1\, p^{(K-1)}(0|z_{K-1})\, Pr_{\theta=0}\{Z_K > a_K | Z_{K-1} = z_{K-1}\}$$

$$+\, 2\,\lambda_2\, p^{(K-1)}(\delta|z_{K-1})\, Pr_{\theta=\delta}\{Z_K < a_K | Z_{K-1} = z_{K-1}\}.$$

We can now define the optimal boundary points:

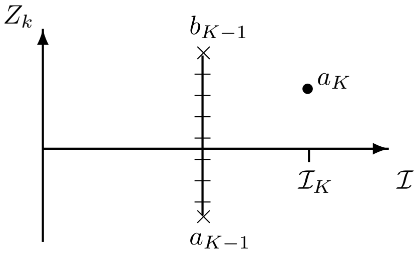
Set b_{K-1} to be the value of z_{K-1} where

E(Cost of continuing) $=$ E(Cost of stopping to reject H_0).

Set a_{K-1} to be the value of z_{K-1} where

E(Cost of continuing) $=$ E(Cost of stopping to accept H_0).

At analysis $K-1$



Before leaving analysis $K-1$, we set up a grid of points for use in numerical integration over the range a_{K-1} to b_{K-1}.

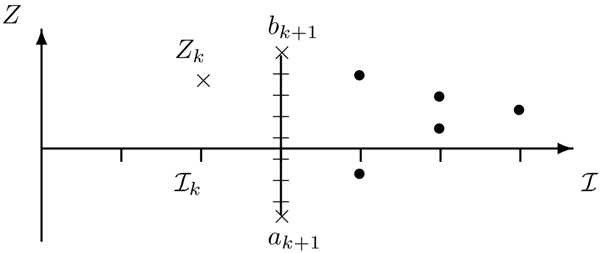
For each point, we sum over the posterior distribution of θ to calculate

$\beta^{(K-1)}(z_{K-1}) = E(\text{Additional cost when continuing} \,|\, Z_{K-1} = z_{K-1})$.

We are now ready to move back to analysis $K-2$.

Analyses 1 to $K - 2$

We work back through analyses $k = K - 2,\ K - 3,\ \ldots,\ 1$.



At each analysis, we find the optimal stopping boundary using knowledge of the optimal stopping rule at future analyses.

Then, for a grid of values of z_k, compute

$$\beta^{(k)}(z_k) \,=\, E(\text{Additional cost when continuing} \,|\, Z_k = z_k)$$

to use in evaluating the option of continuing at analysis $k - 1$.

Solving the original problem

For any given (λ_1, λ_2) we can find the Bayes optimal design and compute its type I and II error rates.

We now search for a pair (λ_1, λ_2) for which type I and type II error rates of the optimal design equal α and β, respectively.

The resulting design will be the optimal group sequential test, with the specified frequentist error rates, for our original problem.

Notes

1. The method of solving the overall problem demonstrates explicitly that good frequentist procedures should be similar to Bayes procedures.

2. The prior and costs in the final Bayes problem are a means to an end, rather than "true" costs of type I and type II errors, or costs of treating patients in the trial.

Properties of optimal designs

Tests with $\alpha = 0.025$, $1 - \beta = 0.9$, K analyses, $\mathcal{I}_{max} = R\,\mathcal{I}_{fix}$, equal group sizes, minimising $\{E_0(\mathcal{I}) + E_\delta(\mathcal{I})\}/2$.

Minimum values of $\{E_0(\mathcal{I}) + E_\delta(\mathcal{I})\}/2$, 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
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

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

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

Generalisations

Solutions can be obtained for a variety of related problems:

- Other optimality criteria such as a weighted sum

$$\sum_i w_i E_{\theta_i}(\mathcal{I})$$

  or an integral

$$\int f(\theta)\, E_\theta(\mathcal{I})\, d\theta$$

- Optimising a set of fixed group sizes in a group sequential test
- Data dependent group sizes in a group sequential test
- Group sequential tests for a delayed response
- Testing for either superiority or non-inferiority

Comments on optimal group sequential tests

We created an artificial Bayes problem in order to find the optimal frequentist design.

The prior probabilities for $\theta = 0$ and $\theta = \delta$ do not necessarily reflect beliefs about the likelihood of these values of θ, nor do λ_1 and λ_2 represent actual costs of type I and type II errors.

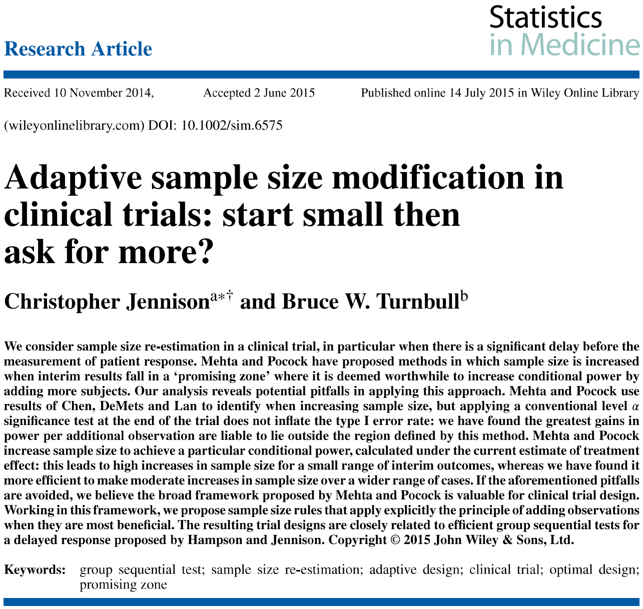
One can ask whether the Bayes decision problem that has been solved is realistic — but this requires costs associated with the trial sample size and final decisions to be put on a common scale.

Eales & Jennison (*Biometrika*, 1992) create such costs by considering overall benefit to patients inside and outside the trial.

A more basic observation is that the class of efficient frequentist designs is equal to the class of Bayes designs.

If a Bayes design is "calibrated" to have an acceptable type I error rate, it should be the same as an optimised frequentist design.

2. Sample size re-estimation



Research Article

Statistics in Medicine

Received 10 November 2014, Accepted 2 June 2015 Published online 14 July 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6575

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

Sample size re-estimation

In a group sequential trial, the final sample size depends on the observed data.

"Sample size re-estimation" gives another route to a similar end.

Example 1 of Mehta & Pocock (*Statistics in Medicine*, 2011) 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

$$Y_{Bi} \sim N(\mu_B, 7.5^2), \;\; i = 1, 2, \ldots, \;\; \text{on the new treatment,}$$

$$Y_{Ai} \sim N(\mu_A, 7.5^2), \;\; i = 1, 2, \ldots, \;\; \text{on the control arm,}$$

and the treatment effect is $\theta \; = \; \mu_B - \mu_A$.

Sample size re-estimation

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

In testing H_0: $\theta \le 0$ 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.

The sponsors are willing to increase sample size if interim results are "promising".

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

Delayed response: At this time a further 208 subjects will have been admitted to the trial, but treated for less than 26 weeks.

Sample size re-estimation

We consider the following variation on Mehta & Pocock's "Promising zone" design.

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

$$\widehat{\theta}_1 \; = \; \overline{Y}_B(1 : n_1/2) - \overline{Y}_A(1 : n_1/2)$$

and

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

In the remainder of the trial a further $n_2^* - n_1$ observations provide

$$\widehat{\theta}_2 \; = \; \overline{Y}_B(n_1/2+1 : n_2^*/2) - \overline{Y}_A(n_1/2+1 : n_2^*/2)$$

and

$$Z_2 = \frac{\widehat{\theta}_2}{\sqrt{\{4\sigma^2/(n_2^* - n_1)\}}}.$$

Sample size re-estimation

At the end of the trial, we test H_0: $\theta \leq 0$ with a combination test, rejecting H_0 if

$$\frac{1}{\sqrt{2}}\, Z_1 + \frac{1}{\sqrt{2}}\, Z_2 \; > \; 1.96.$$

In this framework, we are free to vary n_2^* and the final test will still have one-sided type I error rate $\alpha = 0.025$.

Given the 208 subjects "in the pipeline", we must take $n_2^* \geq 416$, but we can increase n_2^* beyond the planned value of 442 in order to increase power.

Questions:

What is an efficient way to choose n_2^* based on the observed $\widehat{\theta}_1$?

How should we formulate the problem to pose this question in a precise way?

Problem formulation (Jennison & Turnbull, *SiM*, 2015)

We specify γ, a "rate of exchange" between sample size and power.

Focusing on properties under $\theta = \tilde{\theta} = 1.6$, we aim to maximise

$$P_{\theta=\tilde{\theta}}\,(\text{Reject } H_0) \,-\, \gamma E_{\tilde{\theta}}(N). \tag{1}$$

Denote the conditional power under $\theta = \tilde{\theta}$ of the combination test, given $Z_1 = z_1$ and a total sample size of n_2^*, by

$$CP_{\tilde{\theta}}(z_1, n_2^*) \;=\; P_{\tilde{\theta}}\,\{\frac{1}{\sqrt{2}}\,Z_1 + \frac{1}{\sqrt{2}}\,Z_2 \;>\; 1.96 \mid Z_1 = z_1,\, n_2^*\}.$$

We aim to find the sample size function $n_2^*(z_1)$ that maximises (1). This objective can be written 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 $\theta = \tilde{\theta}$.

Problem formulation

In order to maximise

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

for each z_1, we need to choose $n_2^*(z_1)$ to maximise

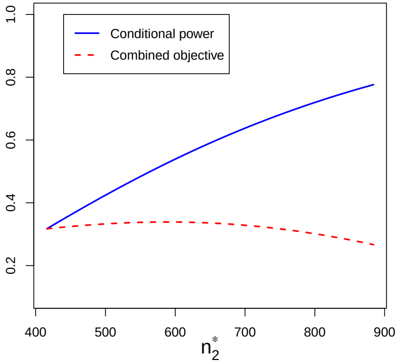
$$CP_{\tilde{\theta}}(z_1, n_2^*(z_1)) - \gamma n_2^*(z_1).$$

Here, the optimisation can be done by numerical calculation under a range of possible values for n_2^*.

Combining these results for different values of z_1 gives the optimised sample size rule $n_2^*(z_1)$.

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

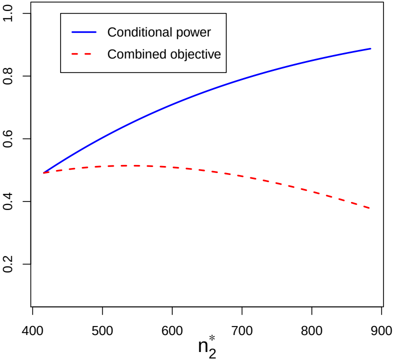
Plots for $\tilde{\theta} = 1.6, \;\; \gamma = 0.245/(4\,\sigma^2)$ **and** $\widehat{\theta}_1 = 0.75$



The function $CP_{\tilde{\theta}}(z_1, n_2^*) - \gamma(n_2^* - 442)$ attains its maximum at $n_2^* = 589$.

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

Plots for $\tilde{\theta} = 1.6$, $\gamma = 0.245/(4\,\sigma^2)$ and $\widehat{\theta}_1 = 1.25$



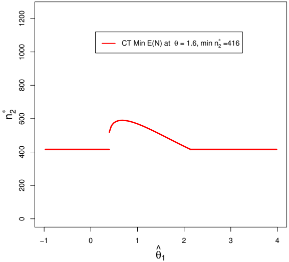
For $\widehat{\theta}_1 = 1.25$, conditional power rises less steeply as n_2^* increases.

Now, the function $CP_{\tilde{\theta}}(z_1, n_2^*) - \gamma(n_2^* - 442)$ has its maximum at $n_2^* = 570$.

Optimal sample size rule

Optimal sample size rule for combination test design

with n_2^* in $(416, 884)$, $\tilde{\theta} = 1.6$, $\gamma = 0.245/(4\,\sigma^2)$



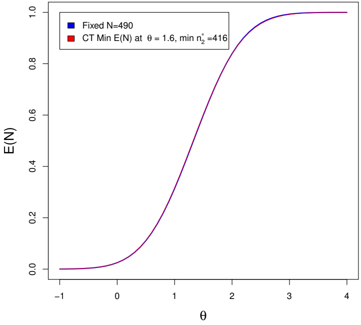
With $\gamma = 0.245/(4\,\sigma^2)$, overall power is 0.658 at $\theta = 1.6$ (the same as for the MP Promising Zone design).

By construction, the procedure has minimum $E_{\theta=1.6}(N)$ among all normal combination test designs with $n_1 = 208$ and $n_2^* \geq 416$ that achieve the same power.

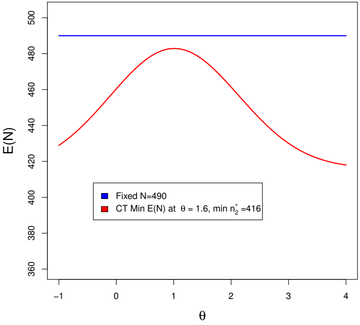
Properties of the optimal sample size rule

Combination test design with optimal sample size rule: $\boldsymbol{n_2^* \geq 416,\ \tilde{\theta} = 1.6,\ \gamma = 0.245/(4\,\sigma^2)}$

Power curve



$E_\theta(N)$ curves



The optimised design has the same power curve as a fixed sample size design with 490 patients — and lower $E_\theta(N)$ at all θ values.

Other options

1. We could reduce the value of γ, leading to an increase in sample size and higher power.

2. We could optimise other criteria, replacing $E_{\tilde{\theta}}(N)$ in $P_{\theta=\tilde{\theta}}\,(\text{Reject } H_0) \,-\, \gamma E_{\tilde{\theta}}(N)$ by a weighted sum or integral,

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

In the integral case, treating the power function in a similar way, we seek to maximise

$$\int w(\theta)\, P_\theta\,(\text{Reject } H_0)\, d\theta \;-\; \gamma \int w(\theta)\, E_\theta(N)\, d\theta.$$

If $w(\theta)$ is a prior distribution, representing investigators' beliefs about likely values of θ, we have a Bayes decision problem.

Other options

Suppose we wish to maximise

$$\int w(\theta)\, P_\theta\,(\text{Reject } H_0)\, d\theta \;-\; \gamma \int w(\theta)\, E_\theta(N)\, d\theta.$$

For each z_1, we must choose $n_2^*(z_1)$ to maximise

$$CP(z_1, n_2^*(z_1)) - \gamma n_2^*(z_1),$$

where $CP(z_1, n_2^*(z_1))$ is the conditional power *integrated over the posterior distribution of* θ given the stage 1 data summary z_1.

By working within the class of inverse normal combination tests, we automatically protect the type I error rate at level α.

We can choose γ to meet a specific power condition. Since the set of possible power curves is essentially a one-parameter family, precisely how power appears in the objective function is not crucial.

Other options

3. Additional constraints can be included in the design process:

   - One could set an upper limit on values for n_2^*.
   - Investors may only wish to consider an increase in sample size when results are deemed to be sufficiently "promising".

4. The sample size rule can be simplified by allowing just two possible values for n_2^*.

When optimised, this simpler design can achieve almost the same $E_\theta(N)$ function as designs with a more general form of $n_2^*(z_1)$.

Such a design is an example of a "Delayed Response Group Sequential Design"; see Hampson & Jennison, (*JRSS, B*, 2013).

Comments on “optimising” the Mehta-Pocock design

Controlling the frequentist type I error rate

In this case, type I error is controlled by using a combination test: we do not need a Lagrangian method and artificial Bayes problem.

The objective function

We set out to maximise the objective

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

where the constant γ represents the relative importance of sample size and the type II error rate.

This has the advantage that the problem has a simple solution.

As noted under case (2) of “Other options”, the approach can be extended to handle a more complete objective such as

$$\int w(\theta)\, P_\theta\,(\text{Reject } H_0)\, d\theta \,-\, \gamma \int w(\theta)\, E_\theta(N)\, d\theta.$$

3. Optimising a Phase 2/3 seamless design

During Phase 2 and Phase 3 of the drug development process,

The final decision is made on the treatment specification, including the dose level,

The selected treatment is tested against control.

A seamless Phase 2/3 trial design combines these two phases:

In stage 1 (Phase 2)

Compare K "treatments" against control

Select the best treatment and, if it has performed sufficiently well, proceed to stage 2.

In stage 2 (Phase 3)

Compare the selected treatment against the control.

Optimising a Phase 2/3 seamless design

After both stages are completed, we test the null hypothesis that the selected treatment is no better than the control.

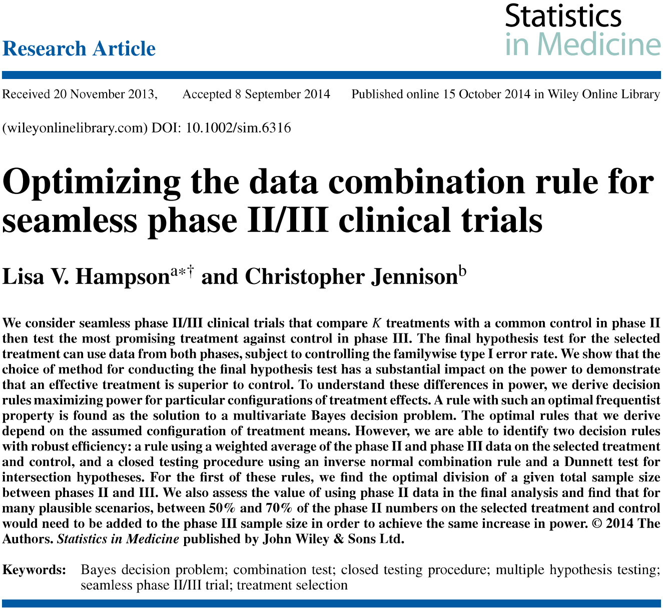
Since this treatment was selected based on data that will also be used in the final analysis, care must be taken to avoid inflating the overall type I error rate.

Design issues

We would like to optimise:

1. The way in which data on all treatments are combined in the final hypothesis test,
2. The way in which the total sample size is divided between the two stages.

Optimising a Phase 2/3 seamless design



Research Article

Statistics
in Medicine

Received 20 November 2013, Accepted 8 September 2014 Published online 15 October 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6316

Optimizing the data combination rule for seamless phase II/III clinical trials

Lisa V. Hampson[a*†] and Christopher Jennison[b]

We consider seamless phase II/III clinical trials that compare *K* treatments with a common control in phase II then test the most promising treatment against control in phase III. The final hypothesis test for the selected treatment can use data from both phases, subject to controlling the familywise type I error rate. We show that the choice of method for conducting the final hypothesis test has a substantial impact on the power to demonstrate that an effective treatment is superior to control. To understand these differences in power, we derive decision rules maximizing power for particular configurations of treatment effects. A rule with such an optimal frequentist property is found as the solution to a multivariate Bayes decision problem. The optimal rules that we derive depend on the assumed configuration of treatment means. However, we are able to identify two decision rules with robust efficiency: a rule using a weighted average of the phase II and phase III data on the selected treatment and control, and a closed testing procedure using an inverse normal combination rule and a Dunnett test for intersection hypotheses. For the first of these rules, we find the optimal division of a given total sample size between phases II and III. We also assess the value of using phase II data in the final analysis and find that for many plausible scenarios, between 50% and 70% of the phase II numbers on the selected treatment and control would need to be added to the phase III sample size in order to achieve the same increase in power. © 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

Keywords: Bayes decision problem; combination test; closed testing procedure; multiple hypothesis testing; seamless phase II/III trial; treatment selection

Optimising a Phase 2/3 seamless design

Denote the K treatment effects vs control by $\theta_1, \ldots, \theta_K$.

Stage 1

Randomise m_1 subjects to each of the K treatments and the control and observe their responses.

Denote the estimated treatment effects by $\widehat{\theta}_{1,i},\ i = 1, \ldots, K$.

Treatment i^* with the highest $\widehat{\theta}_{1,i}$ is selected for stage 2.

Stage 2

Treatment i^* is compared against control, with m_2 observations on each. The estimated treatment effect is $\widehat{\theta}_{2,i^*}$.

Conclusion

A final decision is made, based on $\widehat{\theta}_{1,1}, \ldots, \widehat{\theta}_{1,K}$ and $\widehat{\theta}_{2,i^*}$.

Optimising a Phase 2/3 seamless design

There are K null hypotheses, H_i: $\theta_i \leq 0$, $i = 1, \ldots, K$.

If dose i^* is selected for Phase 3, we focus on testing H_{i^*}: $\theta_{i^*} \leq 0$.

Family-wise error

We want strong control of the **family-wise error** rate. Then, for all vectors $\theta = (\theta_1, \ldots, \theta_K)$,

$$Pr_\theta\{\text{Reject } \textit{any} \text{ true } H_i\} \leq \alpha.$$

Power

When some θ_i are greater than zero, we can define power as

$$Pr\{\text{Select treatment } j \text{ with maximum } \theta_i \text{ and reject } H_j\text{: } \theta_j \leq 0\}.$$

More generally, we can define a gain function or utility that is positive when H_{i^*} is rejected, whichever treatment is selected, but the gain increases with θ_{i^*}.

Optimising a Phase 2/3 seamless design

Family-wise error can be controlled by a Closed Testing Procedure:

Define level α tests for each null hypothesis H_i, and for all intersections of sets of null hypotheses.

Reject H_i overall if all intersection hypotheses that include H_i are rejected.

Theory implies the family-wise type I error rate is at most α.

Each hypothesis test can be formed as a Combination Test across the two stages of the trial (Bauer & Köhne, *Biometrics*, 1984).

How should we test the intersection hypotheses in stage 1?

What type of combination test is best?

The best choice may depend on the K-dimensional parameter θ.

Optimising a Phase 2/3 seamless design

Hampson & Jennison (*Statistics in Medicine*, 2013) found optimal final decision rules that maximise power when $\theta = \delta\, v$, for various choices of vector v.

Two procedures were close to 100% efficient across a wide range of scenarios.

1. In the framework we have described, use a Dunnett test for each intersection hypothesis in stage 1 and combine Z values across stages with a weighted normal combination test.

2. Use the procedure proposed by Thall, Simon and Ellenberg (*Biometrika*, 1988).

The very best design does depend on the high-dimensional, treatment effect vector θ.

However, since we have such robustly efficient procedures, we do not need to consider Bayesian averaging over θ.

Optimising a Phase 2/3 seamless design

Hampson & Jennison also considered how best to divide a total sample size between stage 1 (m_1 observations on K treatments and control) and stage 2 (m_2 on selected treatment and control).

The choice that maximises power depends on the vector of treatment effects, θ, in particular, the largest effect $\max_i(\theta_i)$.

If $\max_i(\theta_i)$ is fairly small, a high stage 2 sample size, m_2, is needed to give adequate power in that stage.

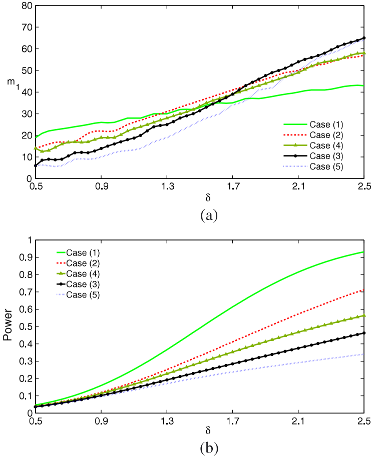
If $\max_i(\theta_i)$ is large, a lower m_2 may suffice and a higher m_1 increases the probability of selecting the best treatment in stage 1.

Bayesian averaging:

We do not know θ.

So, we express our expectations as a distribution for θ and choose a design with good properties averaged over this distribution.

Optimal Stage 1 group sizes in a seamless design



(a)

(b)

Figure 4. (a) Stage 1 group sizes maximizing the power of the TSE procedure when the total sample size is fixed at 448 and $\boldsymbol{\theta}$ is a random permutation (1) of $(0, 0, 0, 0, 1)\,\delta$, (2) of $(0.5, 0.5, 0.5, 0.5, 1)\,\delta$, (3) of $(0.75, 0.75, 0.75, 0.75, 1)\,\delta$, (4) of $(0.3, 0.475, 0.65, 0.825, 1)\,\delta$ and (5) of $(0.75, 0.8125, 0.875, 0.9375, 1)\,\delta$. (b) Power achieved by the optimized TSE procedures. Decision rules are listed in order of decreasing power. Designs are specified with $K = 5$, $\ell = 0$, $\sigma = 5.0$ and $\alpha = 0.025$. Results are based on 1 million simulations for each scenario.

Benefits of Phase 2/3 seamless designs

Regulators require a seamless Phase 2/3 trial to be conducted as a single trial, with a firewall between the data monitoring committee and the investigators.

Efficiency gains from using “Phase 2”data in the final hypothesis test must balance extra planning and organisational requirements.

With m_1 observations on each treatment and control in stage 1 and m_2 on the selected treatment and control in stage 2, what are the benefits of using the stage 1 data in the final analysis?

Hampson & Jennison show that:

> If only Phase 3 data are used in the final analysis, then in many plausible scenarios, m_2 needs to be increased by between $0.5m_1$ and $0.7m_1$, in order to achieve the same power as the seamless design.

Comments on Seamless Phase 2/3 Designs

Controlling the frequentist type I error rate

Use of a closed testing procedure (CTP) and combination test guarantees control of type I error.

Optimising within this class of designs

We can (very nearly) optimise the choice of CTP and combination test for all treatment effect vectors θ simultaneously.

However, the best choice of sample sizes in stage 1 and stage 2 does depend on the vector θ.

The Bayes solution is to specify a prior distribution for the unknown θ and optimise performance integrated over this distribution.

An outer layer

If the optimised value of m_1 leads to unacceptably low average power, consider a higher total sample size for the two stages.

4. Creating an efficient enrichment design

Consider a drug designed to disrupt a disease's biological pathway.

Patients with high levels of a biomarker for this pathway should gain particular benefit.

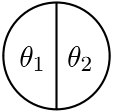
In a clinical trial with **enrichment** we

> Start by comparing the new treatment against control in the full population.
>
> At an interim analysis, we decide whether to:
>
> > Continue recruiting from the full population, or
> >
> > Recruit only from the subgroup — and increase their numbers.

Results may support a licence for the full population or just for the sub-population.

Creating an efficient enrichment design

Sub-population (proportion λ_1)



Rest of the population (proportion λ_2)

The treatment effect (difference in mean response between new treatment and control) is θ_1 in the sub-population and θ_2 in the complement of this sub-population.

The treatment effect over the full population is $\theta_3 = \lambda_1\theta_1 + \lambda_2\theta_2$.

We may wish to test either or both of:

The null hypothesis for the full population, H_3: $\theta_3 \leq 0$ vs $\theta_3 > 0$,

The null hypothesis for the sub-population, H_1: $\theta_1 \leq 0$ vs $\theta_1 > 0$.

Creating an efficient enrichment design

As in the adaptive seamless Phase 2/3 design, we want to control strongly the **family-wise error** rate.

Then, for all values of θ_1 and θ_3,

$$Pr_\theta\{\text{Reject } \textit{any} \text{ true } H_i\} \leq \alpha.$$

This can be achieved by a Closed Testing Procedure, involving level α tests of H_1, H_3 and the intersection hypothesis $H_1 \cap H_3$.

Each of these tests will be constructed as a Combination Test across the two stages of the trial.

Then, general theory implies that the family-wise type I error rate is controlled at level α.

This leaves freedom to define the rule for deciding whether or not to enrich at the interim analysis.

Creating an efficient enrichment design

I have worked on this problem with Thomas Burnett.

We chose to use Simes' test for the intersection hypothesis $H_1 \cap H_3$ and an inverse normal combination test.

We specified a utility or "gain function" to optimise:

$$\text{Gain} \;=\; \lambda_1\,\theta_1\,\mathcal{I}(\text{Reject } H_1 \text{ only}) \,+\, \theta_3\,\mathcal{I}(\text{Reject } H_3).$$

We placed a prior distribution on (θ_1, θ_2).

We then sought the adaptive decision rule that maximises the expected gain.

Given observed treatment effects, $\widehat{\theta}_1$ and $\widehat{\theta}_2$, at the interim analysis, the optimal decision (to continue in the full population or to enrich in the sub-population) is that which maximises the **conditional** expected gain.

Example: An optimal enrichment design

Consider a trial with total sample size that would provide power 0.9 to detect a treatment effect in the full population if $\theta_1 = \theta_2 = 10$.

Suppose $\lambda_1 = \lambda_2 = 0.5$.

Our prior distribution for (θ_1, θ_2) is bivariate normal with

$$E(\theta_1, \theta_2) \; = \; (15, 2)$$

and

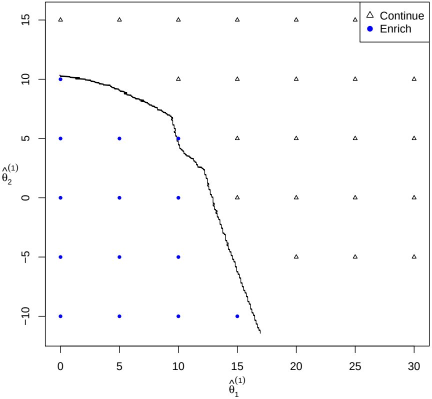
$$\mathsf{Var}(\theta_1, \theta_2) \; = \; \begin{pmatrix} 9 & 0 \\ 0 & 4 \end{pmatrix}.$$

We conduct an interim analysis after half the total number of subjects have been observed.

If the decision is to "enrich", all the remaining sample size is allocated to the sub-population.

Example: An optimal enrichment design

The optimal decision rule is:



The peculiar shape of the boundary reflects features of the Simes test applied to data at the interim analysis.

Example: An optimal enrichment design

Properties of the optimised enrichment design:

$$\begin{aligned} Pr(\text{Enrich}) &= 0.32 \\ Pr(\text{Reject } H_1 \text{ only}) &= 0.32 \\ Pr(\text{Reject } H_3) &= 0.60 \\ E(\text{Gain}) &= 7.74 \end{aligned}$$

The design with no enrichment that tests both H_1 and H_3 has

$$E(\text{Gain}) = 7.55$$

The design that recruits all subjects from the sub-population from the outset, and only tests H_1, has

$$E(\text{Gain}) = 7.40$$

Creating an efficient enrichment design

We have found examples of the gain function and prior for which the best adaptive design is superior to both simple, non-adaptive designs — but this is not always the case.

However, adaptive enrichment may have additional appeal:

If investigators differ in their prior beliefs, an optimal adaptive design for a "consensus" prior may be broadly acceptable.

An optimal design that recruits only from the sub-population may be deemed too restrictive by some investigators — and the adaptive approach allows a slower route towards this end.

When the optimal policy is to recruit from the full population (so no enrichment occurs and combination tests are not needed), the optimal adaptive design's $E(\text{Gain})$ is only slightly sub-optimal.

Comments on Enrichment Designs

Controlling the frequentist type I error rate

Use of a closed testing procedure and combination tests guarantees control of family-wise type I error.

Optimising within this class of designs

Given gain and cost functions, and a prior distribution for (θ_1, θ_2), we can compute Bayes-optimal adaptive enrichment designs.

An outer layer

Other design features that merit investigation include:

Details of the closed testing procedure and combination tests.

The timing of the interim analysis.

Preferential sampling of one population when the proportions λ_1 and λ_2 are away from 0.5.

5. Overall conclusions

Controlling the frequentist type I error rate

We can apply closed testing procedures and combination tests to protect family-wise error in complex, high-dimensional settings.

We can then work on optimising other aspects of a given design.

Optimising within a class of designs

Before trying to optimise, we need to understand which properties of a design are important to the investigators.

Typically, this is done through the elicitation of their gain function, cost function, and prior distribution for unknown parameters.

Then, we can optimise by analysis, calculation or simulation.

An outer layer

Once we can optimise the central component of a design, we may re-visit higher level aspects and question initial assumptions.