Optimising Group Sequential Designs:

Where Frequentist meets Bayes

Christopher Jennison

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

Bruce Turnbull

Department of Statistical Science,
Cornell University, UK
https://people.orie.cornell.edu/bruce

Deming Conference, Atlantic City

December 2025

Plan for this talk

We shall consider:

(I) Frequentist and Bayesian designs for a group sequential trial

What we would like to do

What we can do

Convergence of approaches

(II) More complex adaptive trial designs

(I) Group sequential tests: Problem formulation

Consider a Phase III clinical trial comparing a new treatment against a control.

We denote the treatment effect by θ.

Examples: θ could be the difference in mean response or, in a time-to-event study, θ could be the log hazard ratio.

We wish to decide whether $\theta > 0$, in which case the new treatment is superior.

The trial will have K analyses.

The information for θ at analysis k will be $\mathcal{I}_k$.

Marginally,

$$\widehat{\theta}_k \sim N(\theta, \mathcal{I}_k^{-1})$$

and the score statistics, $S_k = \widehat{\theta}_k \, \mathcal{I}_k$, have independent increments.

1. Bayes decision theory: Berry & Ho (*Bmcs*, 1988)

Prior distribution $\pi(\theta)$.

Possible decisions

$$\begin{aligned} d_1 &: \text{ Do not pursue drug approval,} \\ d_2 &: \text{ Pursue drug approval.} \end{aligned}$$

Loss function for taking decision d_1 or d_2 when the true value of the treatment effect is θ

$$\begin{aligned} L(\theta, d_1) &= 0 \qquad\qquad \text{for all } \theta, \\ L(\theta, d_2) &= \begin{cases} -K\theta & \text{if } \theta > 0, \\ L & \text{if } \theta \leq 0. \end{cases} \end{aligned}$$

Sampling cost: 1 per subject in the trial, N in total, say.

Aim: minimise the expected loss

$$\int \pi(\theta) \int f(x \,|\, \theta) \, \{L(\theta,\, d(x)) + N(x)\} \, \mathsf{d}x \; \mathsf{d}\theta$$

Bayes decision theory formulation: Berry & Ho

Aim: Minimise the expected loss

$$\int \pi(\theta) \int f(x \,|\, \theta) \, \{L(\theta,\, d(x)) + N(x)\} \; \mathrm{d}x \; \mathrm{d}\theta$$

Berry & Ho found the optimal stopping rule and decision rule by dynamic programming.

They presented examples with priors

$$\theta \sim N(-1, 2), \quad \theta \sim N(0, 2), \quad \theta \sim N(1, 2),$$

$$K = 5,000 \quad \text{and} \quad L = 2,000,$$

and showed results for designs with 2 and 3 analyses.

Lorden (*Ann. Statistics*, 1976) had applied the numerical optimisation scheme described by Lai (*Ann. Statistics*, 1973) to solve similar problems in a frequentist setting.

2. Frequentist problem: Barber & Jennison (*Bmka*, 2002)

Test H_0: $\theta \le 0$ against $\theta > 0$ with type I and type II error rates

$$Pr_{\theta=0}\{\text{Reject } H_0\} \le \alpha, \quad Pr_{\theta=\delta}\{\text{Reject } H_0\} \ge 1-\beta.$$

So, a fixed sample size test requires information

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

Aim: In a group sequential test with K analyses, minimise

$$\int f(\theta)\, \mathbb{E}(\mathcal{I}_T)\, \mathrm{d}\theta$$

where $\mathcal{I}_T$ is the observed information on termination.

One may set $f(\theta) = \pi(\theta)w(\theta)$ where $\pi(\theta)$ is a prior for θ and $w(\theta)$ reflects the importance of early stopping for different values of θ.

Barber & Jennison found optimal stopping and decision rules by dynamic programming.

A frequentist problem: Barber & Jennison

Group sequential tests with K equally sized groups, $\mathcal{I}_{max} = R\,\mathcal{I}_{fix}$,

type I error probability $\alpha = 0.025$, power 0.9 at $\theta = \delta$.

Minimum possible values of $\int f(\theta) E_\theta(\mathcal{I}_T)\,\mathrm{d}\theta$, where $f(\theta)$ is

the density of a $N(\delta, \delta^2/4)$ distribution,

Minimum values of $\int f(\theta) E_\theta(\mathcal{I}_T)\,\mathrm{d}\theta$, as a percentage of $\mathcal{I}_{fix}$

			R			*Minimum*
K	1.01	1.05	1.1	1.2	1.3	*over* R
2	79.3	74.7	**73.8**	74.8	77.1	73.8 at R=1.11
3	74.8	69.0	67.0	**66.1**	66.6	66.1 at R=1.20
5	71.1	65.1	62.7	60.9	**60.5**	60.5 at R=1.32
10	68.2	62.1	59.5	57.5	**56.7**	56.4 at R=1.46
20	66.8	60.6	58.0	55.8	**54.8**	54.2 at R=1.59

Recommend: $K = 5$, $R = 1.05$ or 1.1.

A frequentist problem: Barber & Jennison

For practical application:

Error spending tests with type I and type II error spending functions of the form

$$f_1(\mathcal{I}) \,=\, \alpha\,(\mathcal{I}/\mathcal{I}_{\max})^{\rho}, \quad f_2(\mathcal{I}) \,=\, \beta\,(\mathcal{I}/\mathcal{I}_{\max})^{\rho}$$

are almost optimal.

Error spending designs adapt to observed information levels, controlling the type I error rate and maintaining efficiency.

Rho-family designs with $\rho = 2$ have a sample size "inflation factor" around $R = 1.1$.

Designs with $\rho = 3$ have an "inflation factor" around $R = 1.05$.

Derivation of optimal frequentist tests

For a general function $f(\theta)$, the problem is to minimise

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

subject to

$$Pr_{\theta=0}\{\text{Reject } H_0\} \le \alpha, \quad Pr_{\theta=\delta}\{\text{Reject } H_0\} \ge 1-\beta. \qquad (1)$$

Following the Lagrangian approach, we minimise

$$\int f(\theta)\, E_\theta(\mathcal{I}_T)\, \mathrm{d}\theta + \lambda_1\, Pr_{\theta=0}\{\text{Reject } H_0\} + \lambda_2\, Pr_{\theta=\delta}\{\text{Accept } H_0\}.$$

Dynamic programming (backwards induction) does this efficiently.

Then we search for values λ_1 and λ_2 so that (1) is satisfied — and we have the solution to the problem with error rate constraints.

Derivation of optimal frequentist tests

Suppose $f(\theta)$ the density of a $N(\delta, \delta^2/4)$ distribution.

Then, in minimising

$$\int f(\theta)\, E_\theta(\mathcal{I}_T)\, \mathrm{d}\theta + \lambda_1\, Pr_{\theta=0}\{\text{Reject } H_0\} + \lambda_2\, Pr_{\theta=\delta}\{\text{Accept } H_0\},$$

we have solved a Bayes decision problem with prior

$\theta = 0$	with probability 1/3
$\theta = \delta$	with probability 1/3
$\theta \sim N(\delta/2, \delta^2/4)$	with probability 1/3

and a cost of sampling $\mathcal{I}_T$ when $\theta \sim N(\delta/2, \delta^2/4)$.

This is a Bayes optimal procedure — but for a rather odd looking problem.

3. Back to Bayesian proposals

Spiegelhalter, Freedman & Parmar (1994) proposed the following sequential procedure. Given a prior $\pi(\theta)$,

At each analysis $k = 1, \ldots, K-1$

if $Pr\{\theta > 0 \mid \textit{data}\} > 1 - \epsilon$	stop, declare $\theta > 0$
if $Pr\{\theta < 0 \mid \textit{data}\} > 1 - \epsilon$	stop, declare $\theta \le 0$
otherwise	continue to group $k+1$,

after group K

if $Pr\{\theta > 0 \mid \textit{data}\} > 1 - \epsilon$	stop, declare $\theta > 0$
otherwise	stop, do not declare $\theta > 0$.

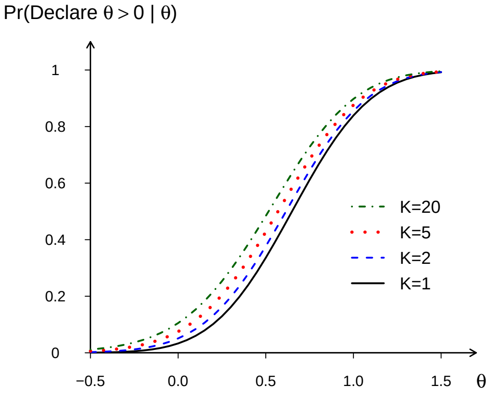
The trial stops early if the $1 - 2\epsilon$ **credible interval** for θ does not contain zero.

The credible interval is not affected by the fact that earlier analyses have been conducted.

Bayesian proposals: Spiegelhalter et al. (1994)

However, the frequentist properties of the Bayes procedure **are** affected by the number of analyses conducted.

Probability of declaring $\theta > 0$ for procedures with prior $\theta \sim N(0.5, 0.5)$, $\epsilon = 0.025$, and 1, 2, 5 and 20 analyses:



Bayesian proposals: Calibrated procedures

Spiegelhalter et al. (1994) proposed use of a "handicap prior", chosen so that the procedure has a particular type I error rate.

If the number of analyses is K and maximum information is $\mathcal{I}_K$, the handicap prior is

$$\theta \sim N(0, (h\mathcal{I}_K)^{-1}).$$

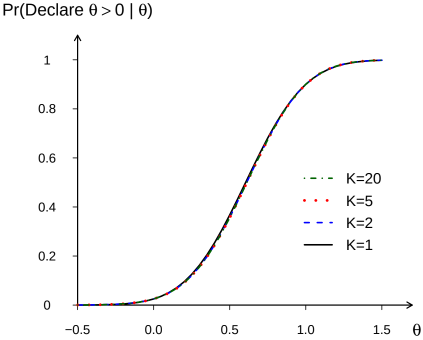
The "handicap" h depends on the number of analyses. Here are values of h for $\alpha = 0.025$.

Number of analyses K	*Handicap* h
1	0.000
2	0.163
5	0.271
20	0.382

Bayesian proposals: Calibrated procedures

We can choose $\mathcal{I}_K$ so that the procedure with a handicap prior has a specific power if the treatment effect is $\theta = \delta$.

Power functions of designs with $Pr\{\text{Reject } H_0 \,|\, \theta = 1\} = 0.9$



Note: When power curves are matched at $\theta = 0$ and $\theta = 1$, they are just about indistinguishable everywhere.

4. Bayesian proposals: Ventz & Trippa (2015)

Ventz & Trippa proposed optimising and calibrating at the same time. Their problem formulation has:

Possible decisions

$$\begin{array}{ll} d_1: & \text{Do not pursue drug approval,} \\ d_2: & \text{Pursue drug approval.} \end{array}$$

A "gain function" or "utility" comprising

$$\begin{aligned} G(\theta, d_1) &= 0 \qquad\qquad \text{for all } \theta, \\ G(\theta, d_2) &= \begin{cases} K(\theta) & \text{if } \theta > 0, \\ -L(\theta) & \text{if } \theta \leq 0, \end{cases} \end{aligned}$$

plus a term

$$B(\theta, d_i, T)$$

denoting the additional benefit from reaching decision d_i at analysis T minus the cost of treating patients in the trial.

Bayesian proposals: V&T Calibrated Bayes procedure

Assuming a prior distribution $\pi(\theta)$, Ventz & Trippa seek to minimise the expected gain

$$-\int_{-\infty}^{0} L(\theta)\, Pr(D = d_2 \,|\, \theta)\, \pi(\theta)\, d\theta + \int_{0}^{\infty} K(\theta)\, Pr(D = d_2 \,|\, \theta)\, \pi(\theta)\, d\theta$$

$$+ \int_{-\infty}^{\infty} E\{B(\theta, D, T)\}\, \pi(\theta)\, d\theta.$$

subject to error rate constraints

$Pr(D = d_2 \,|\, \theta = 0) = \alpha$ and $Pr(D = d_1 \,|\, \theta = \delta) = \beta$

for a specified value of δ ("pre-calibration").

Note that type I and type II errors feature twice, in different guises — first in the gain function then in the constraints.

How does this relate to Barber & Jennison's problem formulation?

Bayesian proposals: V&T Calibrated Bayes procedure

On Slide 14, we saw an example of how, to a high degree of accuracy, power functions belong to a two parameter family.

Thus, the constraints

$$Pr(D = d_2 \,|\, \theta = 0) \,=\, \alpha \quad \text{and} \quad Pr(D = d_1 \,|\, \theta = \delta) \,=\, \beta$$

are essentially equivalent to

$$\int_{-\infty}^{0} L(\theta)\, Pr(D = d_2 \,|\, \theta)\, \pi(\theta)\, d\theta \,=\, \gamma_1$$

and

$$\int_{0}^{\infty} K(\theta)\, Pr(D = d_2 \,|\, \theta)\, \pi(\theta)\, d\theta \,=\, \gamma_2$$

for certain values of γ_1 and γ_2.

Bayesian proposals: V&T Calibrated Bayes procedure

Thus, when Ventz & Trippa minimise

$$-\int_{-\infty}^{0} L(\theta)\, Pr(D = d_2 \,|\, \theta)\, \pi(\theta)\, d\theta \;+\; \int_{0}^{\infty} K(\theta)\, Pr(D = d_2 \,|\, \theta)\, \pi(\theta)\, d\theta$$

$$+ \int_{-\infty}^{\infty} E\{B(\theta, D, T)\}\, \pi(\theta)\, d\theta$$

subject to

$$Pr(D = d_2 \,|\, \theta = 0) \;=\; \alpha \quad \text{and} \quad Pr(D = d_1 \,|\, \theta = \delta) \;=\; \beta.$$

they are effectively minimising

$$\int_{-\infty}^{\infty} E\{B(\theta, D, T)\}\, \pi(\theta)\, d\theta.$$

subject to

$$\int_{-\infty}^{0} L(\theta)\, Pr(D = d_2 \,|\, \theta)\, \pi(\theta)\, d\theta = \gamma_1 \quad \text{and} \quad \int_{0}^{\infty} K(\theta)\, Pr(D = d_2 \,|\, \theta)\, \pi(\theta)\, d\theta = \gamma_2.$$

Bayesian proposals: V&T Calibrated Bayes procedure

And, given the equivalence of the two types of constraint on the power function, they are also minimising

$$\int_{-\infty}^{\infty} E\{B(\theta, D, T)\}\,\pi(\theta)\,d\theta.$$

subject to

$$Pr(D = d_2 \,|\, \theta = 0) \,=\, \alpha \quad \text{and} \quad Pr(D = d_1 \,|\, \theta = \delta) \,=\, \beta.$$

This is exactly the type of problem that Barber & Jennison and others have tackled from a frequentist perspective.

A remark on the frequentist designs we saw earlier

I noted that the designs produced by Barber & Jennison were Bayes procedures for rather strange looking priors.

However, as noted above, we can replace the constraint

$$Pr(D = d_2 \,|\, \theta = 0) \,=\, \alpha \quad \text{and} \quad Pr(D = d_1 \,|\, \theta = \delta) \,=\, \beta$$

by an (almost) equivalent constraint of the form

$$\int_{-\infty}^{0} Pr(D = d_2 \,|\, \theta)\, f(\theta)\, d\theta = \gamma_1 \quad \text{and} \quad \int_{0}^{\infty} Pr(D = d_2 \,|\, \theta)\, f(\theta)\, d\theta = \gamma_2,$$

where $f(\theta)$ is a $N(\delta/2,\, (\delta/2)^2)$ density.

Then, we can obtain essentially the same optimal design by solving a more reasonable looking Bayesian problem.

CONVERGENCE!

Conclusions

Theoretical underpinnings

The convergence between frequentist and calibrated Bayes procedures demonstrates the “complete class theorems” of Brown, Cohen & Strawderman (*Annals of Statistics*, 1980) which state

{The set of admissible frequentist procedures}

= {The set of Bayes optimal procedures}.

Practical consequences

Both schools can learn from each other :

We are (or should be) solving the same problems.

The same optimisation techniques are good in both cases.

(II) More complex trial designs

A clinical trial may have

Multiple treatments,

Multiple populations,

Multiple endpoints.

With treatment effects $\theta_1, \ldots, \theta_p$, we may test

$$H_i\text{: } \theta_i \leq 0 \text{ vs } \theta_i > 0, \; i = 1, \ldots, p.$$

A common requirement is to control the family-wise error rate, so

$$P_{\boldsymbol{\theta}}\{\text{Reject any true } H_i\} \;\leq\; \alpha \quad \text{for all } \boldsymbol{\theta} = (\theta_1, \ldots, \theta_p).$$

This can be achieved by using a Closed Testing Procedure (CTP).

A CTP can be applied using Combination Tests to combine data summaries across stages of an adaptive trial.

More complex trial designs

Frequentist analysis

Applying a Closed Testing Procedure and Combination Tests, we can guarantee control of the family-wise error rate.

Bayesian design and analysis

If we specify a prior distribution for $(\theta_1, \ldots, \theta_p)$ and a gain function, it may be possible to find a Bayes optimal trial design that maximises the expected gain.

"Calibrating" this design to satisfy a condition on FWER is likely to be challenging:

The vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is high-dimensional,

There may be no obvious "least favourable configuration" where FWER is highest,

Simulations under many values of $\boldsymbol{\theta}$ will be needed (ICH E20).

More complex trial designs

A hybrid strategy

We can use frequentist tools (closed testing procedure and combination tests) to take care of the family-wise error rate.

Freedom will remain in defining decision rules, such as:

When to drop treatments in a multi-arm multi-stage design,

Whether to restrict attention to a patient subgroup in an enrichment design.

We can specify a prior distribution for unknown parameters and a gain function to be optimised.

Then we can define decision rules to give the Bayes optimal design within the specified class of (frequentist) designs.

We illustrate this process in the design of an enrichment trial.

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:

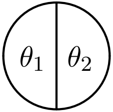
Continue recruiting from the full population, or

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

We want to control strongly the **family-wise error** rate.

Then, for all values of θ_1 and θ_3,

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

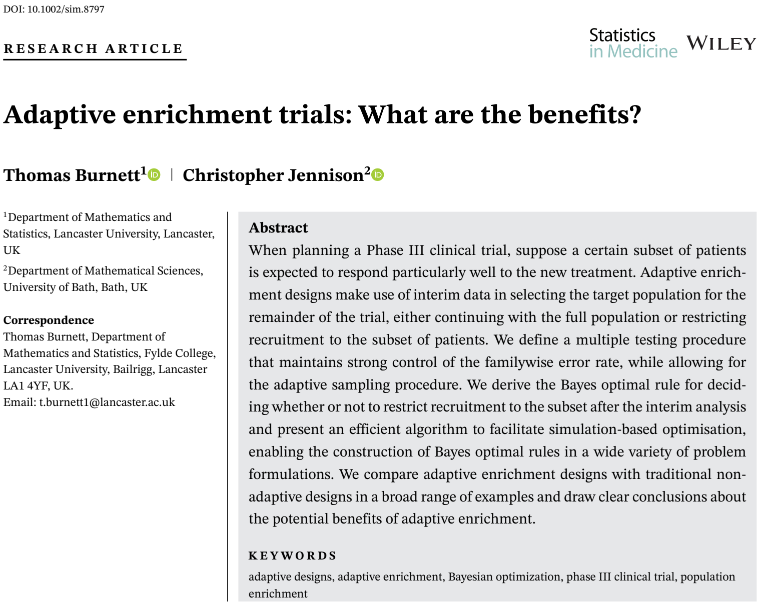
This can be achieved by a Closed Testing Procedure, involving level α tests of H_1, H_3 and the intersection hypothesis $H_1 \cap H_3$.

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



DOI: 10.1002/sim.8797

RESEARCH ARTICLE

Statistics in Medicine WILEY

Adaptive enrichment trials: What are the benefits?

Thomas Burnett[1] | **Christopher Jennison**[2]

[1] Department of Mathematics and Statistics, Lancaster University, Lancaster, UK

[2] Department of Mathematical Sciences, University of Bath, Bath, UK

Correspondence
Thomas Burnett, Department of Mathematics and Statistics, Fylde College, Lancaster University, Bailrigg, Lancaster LA1 4YF, UK.
Email: t.burnett1@lancaster.ac.uk

Abstract

When planning a Phase III clinical trial, suppose a certain subset of patients is expected to respond particularly well to the new treatment. Adaptive enrichment designs make use of interim data in selecting the target population for the remainder of the trial, either continuing with the full population or restricting recruitment to the subset of patients. We define a multiple testing procedure that maintains strong control of the familywise error rate, while allowing for the adaptive sampling procedure. We derive the Bayes optimal rule for deciding whether or not to restrict recruitment to the subset after the interim analysis and present an efficient algorithm to facilitate simulation-based optimisation, enabling the construction of Bayes optimal rules in a wide variety of problem formulations. We compare adaptive enrichment designs with traditional non-adaptive designs in a broad range of examples and draw clear conclusions about the potential benefits of adaptive enrichment.

KEYWORDS

adaptive designs, adaptive enrichment, Bayesian optimization, phase III clinical trial, population enrichment

Creating an efficient enrichment design

CJ has 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 enrich or not 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

$$E(\theta_1, \theta_2) \;=\; (12, 2)$$

and

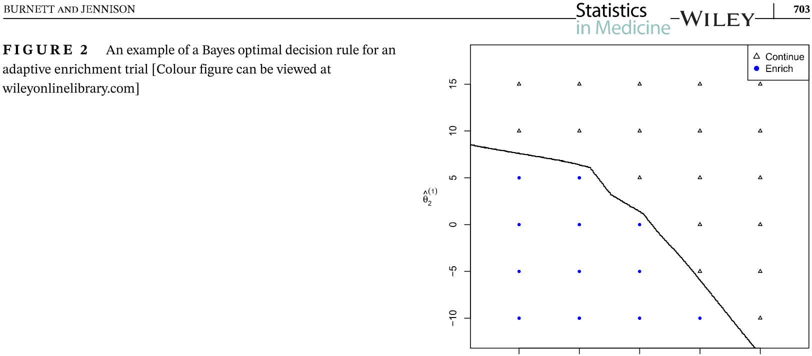
$$\text{Var}(\theta_1, \theta_2) \;=\; \begin{pmatrix} 25 & 18.75 \\ 18.75 & 25 \end{pmatrix}.$$

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:



BURNETT AND JENNISON

Statistics in Medicine — WILEY | 703

FIGURE 2 An example of a Bayes optimal decision rule for an adaptive enrichment trial [Colour figure can be viewed at wileyonlinelibrary.com]

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:

$$\begin{aligned} Pr(\text{Enrich}) &= 0.49 \\ Pr(\text{Reject } H_1 \text{ only}) &= 0.29 \\ Pr(\text{Reject } H_3) &= 0.44 \\ E(\text{Gain}) &= \mathbf{6.23} \end{aligned}$$

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

$$E(\text{Gain}) = \mathbf{6.09}$$

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

$$E(\text{Gain}) = \mathbf{5.57}$$

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 — or by regulators.

The adaptive approach allows enrichment when there is evidence to confirm that this is appropriate.

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 are able to find Bayes-optimal adaptive enrichment designs.

An outer layer of optimisation

Other design features that may be investigated 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.

Overall conclusions

(I) Designing a group sequential trial

Requirements on type I and type II error rates lead to calibration of Bayesian designs.

In this process, carefully chosen priors and gain functions are discarded.

When the objective for early stopping is clearly defined, optimised frequentist and Bayes designs should coincide.

(II) More complex adaptive trial designs

Frequentist tools produce designs that can be shown to protect the family-wise error rate without extensive simulations.

Bayesian thinking can then be applied to optimise within the class of such designs.

A hybrid design combines the strengths of both approaches.