Group Sequential and Adaptive Clinical Trial Designs

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

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

IMS Workshop on the Design and Analysis of Clinical Trials

National University of Singapore

October 2011

Outline of talk

- Group sequential tests for Phase III clinical trials

  Distribution theory

  Computation

  Benefits of group sequential testing

- Error spending tests
- A survival data example
- Group sequential tests with a delayed response
- From group sequential to adaptive designs
- Adapting the target population: Enrichment designs
- Conclusions

1. Group sequential tests for Phase III clinical trials

The setting for this lecture is a Phase III clinical trial, comparing a new treatment against the current standard.

Two positive Phase III trials are usually required to support the case made to regulators for the approval of a new treatment.

Suppose the treatment effect θ represents the advantage of the new treatment over the control, so a positive value means the new treatment is effective.

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

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

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

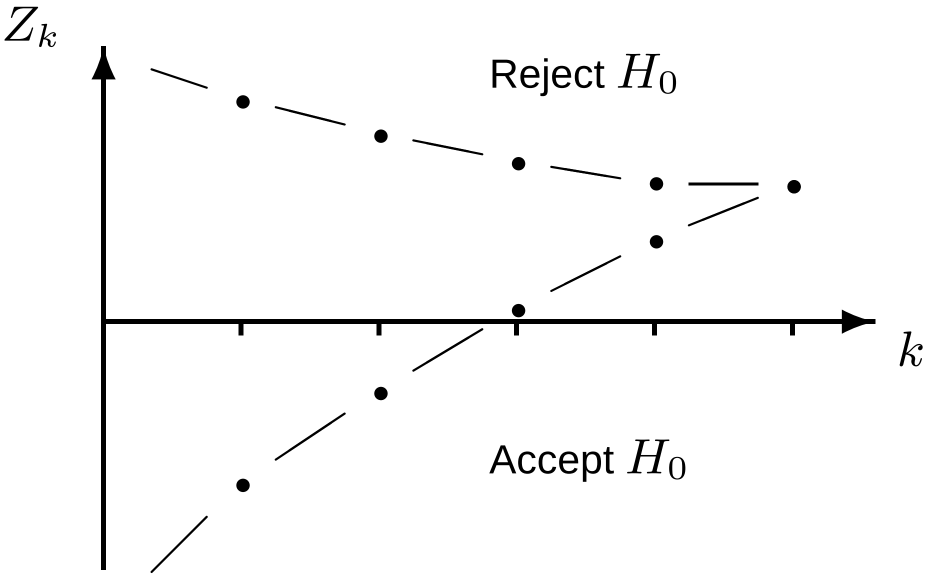
This could be done in a fixed sample size trial.

However, there are strong reasons (ethical, financial, and administrative) to monitor data as the study proceeds and possibly terminate the trial early.

Group sequential tests

In a Group Sequential clinical trial, standardized test statistics Z_1, Z_2, . . . , are computed at interim analyses and used to define a stopping rule for the trial.

A typical boundary for a one-sided test 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.

Joint distribution of parameter estimates

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

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

The information for θ at analysis k is

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

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

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

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

and

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

Sequential distribution theory

The joint distribution of $\hat{\theta}_1, \ldots, \hat{\theta}_K$ can be demonstrated 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

$$Z_k \;=\; \frac{\hat{\theta}_k}{\sqrt{\text{Var}(\hat{\theta}_k)}} \;=\; \hat{\theta}_k \sqrt{\mathcal{I}_k}.$$

For this,

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

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

$\text{Cov}(Z_{k_1}, Z_{k_2}) = \sqrt{\mathcal{I}_{k_1}/\mathcal{I}_{k_2}}$ 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,

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

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 (score statistics) for comparing two survival curves

— and to Z-statistics formed from these.

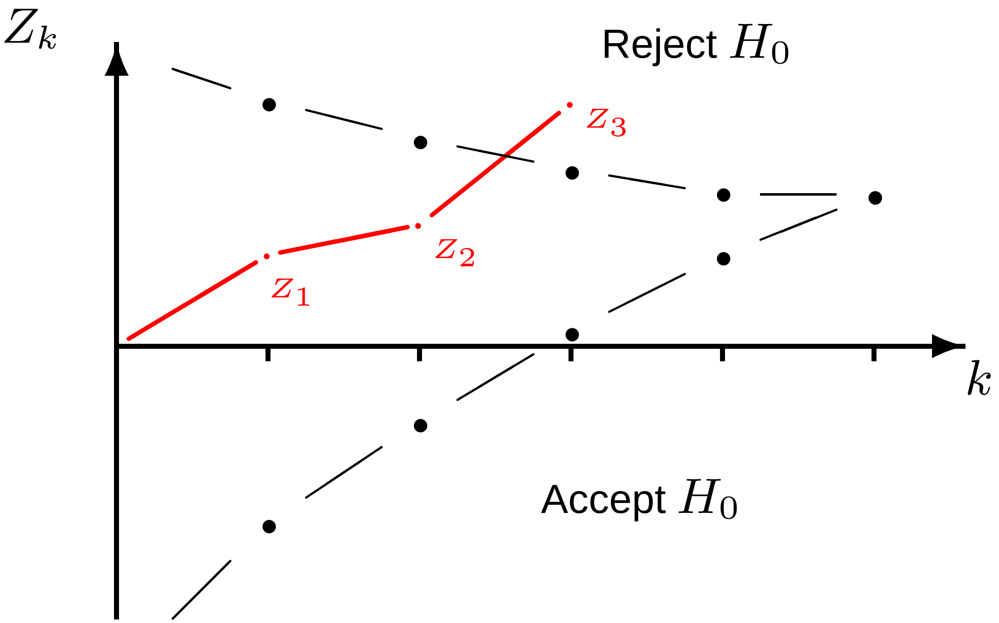
For survival data, observed information is roughly proportional to the number of failures.

Special types of group sequential test are needed to handle unpredictable and unevenly spaced information levels: see *error spending tests*.

Reference:

"Group-sequential analysis incorporating covariate information", Jennison & Turnbull (*J. American Statistical Association*, 1997).

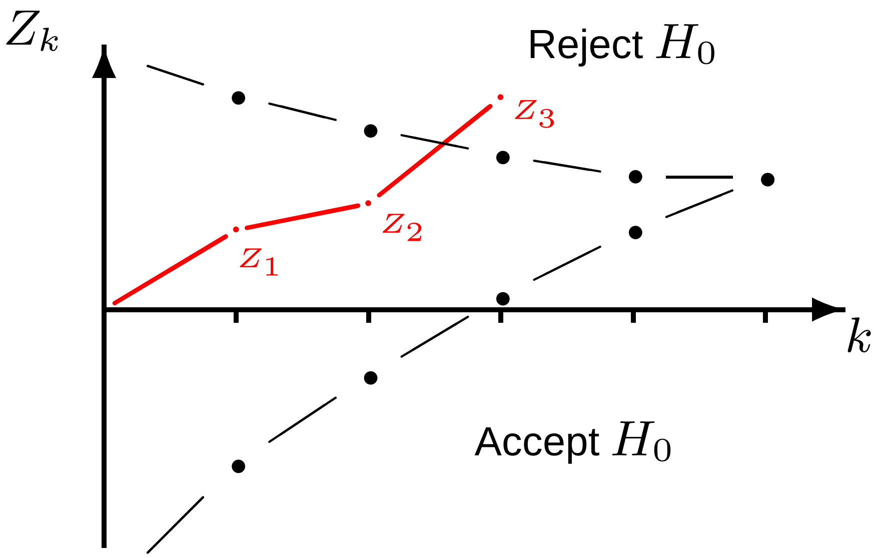
Computations for group sequential tests



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 (see JT, Ch. 19).

Combining such probabilities yields properties of a group sequential boundary.

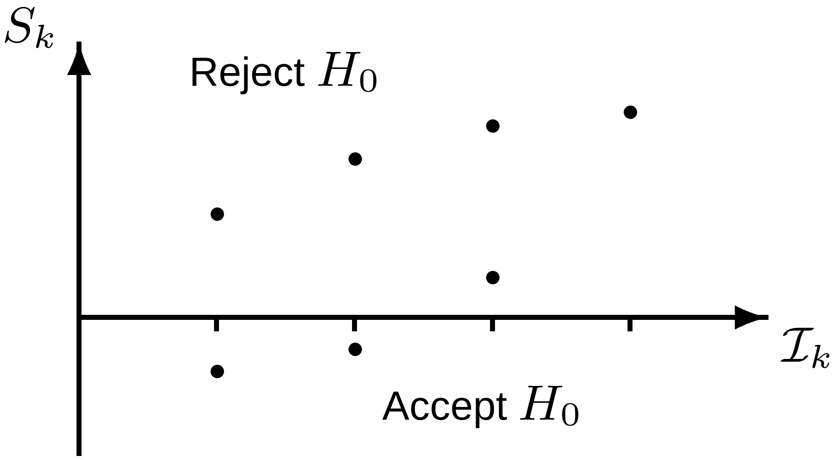
Constants and group sizes can be chosen to define a test with a specific type I error probability and power.

A parametric family of one-sided tests

Reference: Pampallona & Tsiatis (*J. Statistical Planning and Inference*, 1994).

Stopping boundaries can be defined with a particular shape.

The computational methods just described can be used to find the parameter values needed to satisfy type I error rate and power requirements.



Pampallona & Tsiatis (1994) propose a family of boundaries with varying degrees of early stopping.

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 test needs information

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

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

A group sequential test with K analyses will need to be able to continue to a maximum information level $\mathcal{I}_K$ which is greater than $\mathcal{I}_{fix}$.

The benefit is that, on average, the sequential test can stop earlier than this and expected information on termination, $E_\theta(\mathcal{I})$, will be considerably less than $\mathcal{I}_{fix}$, especially under extreme values of θ.

We term the ratio $R = \mathcal{I}_K/\mathcal{I}_{fix}$ the “inflation factor” for a group sequential design.

Benefits of group sequential testing

In specifying a group sequential test's boundary, one can aim to minimise the expected information $E_\theta(\mathcal{I})$ under effect sizes of θ of most interest, subject to a fixed number of analyses K and inflation factor R.

Eales & Jennison (*Biometrika*, 1992) and Barber & Jennison (*Biometrika*, 2002) report on designs optimised for criteria of the form $\sum_i w_i E_{\theta_i}(\mathcal{I})$ or

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

where f is a normal density.

These optimal group sequential designs can be used in their own right.

They also serve as benchmarks for other methods which may have additional useful features: see later comments on the efficiency of "error spending" designs.

Benefits of group sequential testing

One-sided tests, $\alpha = 0.025$, $1 - \beta = 0.9$, 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
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: $E(\mathcal{I}) \searrow$ as $K \nearrow$ but with diminishing returns,

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

2. Error spending tests

The sequence $\mathcal{I}_1, \mathcal{I}_2, \ldots$ is often unpredictable.

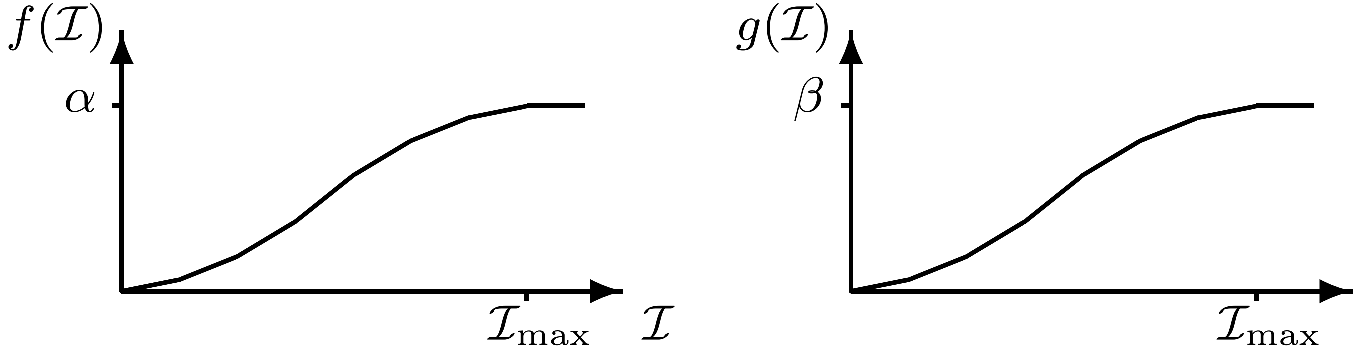
Lan & DeMets (*Biometrika*, 1983) presented two-sided tests of H_0: $\theta = 0$ against $\theta \neq 0$ which "spend" type I error probability as a function of observed information.

For a one-sided test of H_0: $\theta \leq 0$ against $\theta > 0$, we need two functions to spend

Type I error probability α under $\theta = 0$,

Type II error probability β under $\theta = \delta$.

A ***maximum information design*** works towards a target information level $\mathcal{I}_{\max}$.



Type I error probability α is spent according to the function $f(\mathcal{I})$, and type II error probability β according to $g(\mathcal{I})$.

One-sided error-spending tests

Analysis 1:

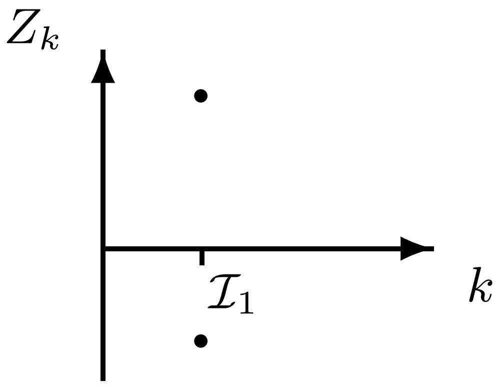
Observed information $\mathcal{I}_1$.

Reject H_0 if $Z_1 > b_1$, where

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

Accept H_0 if $Z_1 < a_1$, where

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



One-sided error-spending tests

Analysis 2:

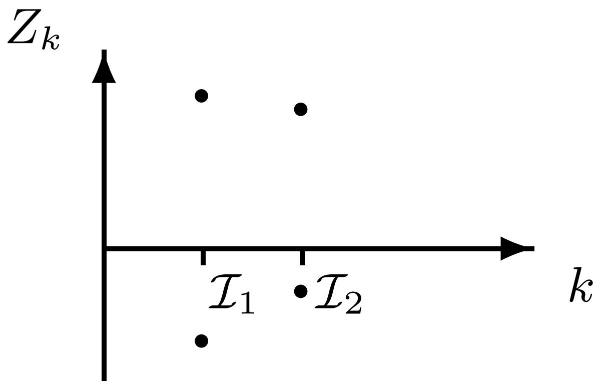
Observed information $\mathcal{I}_2$.

Reject H_0 if $Z_2 > b_2$, where

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

Accept H_0 if $Z_2 < a_2$, where

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



One-sided error-spending tests

Analysis k:

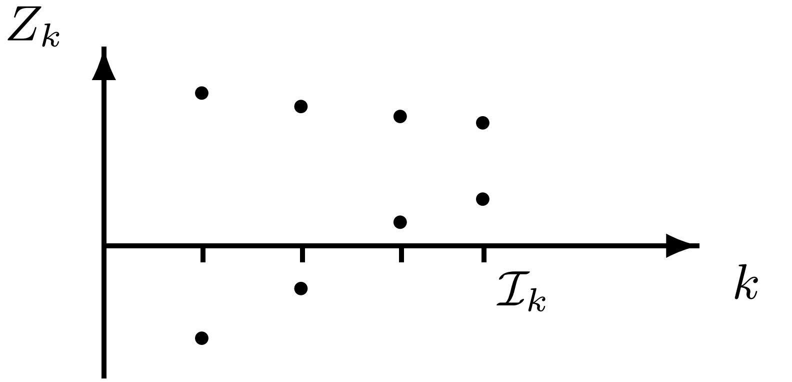
Find a_k and b_k to satisfy

$$P_{\theta=0}\{a_1 < Z_1 < b_1, \, \dots \, , a_{k-1} < Z_{k-1} < b_{k-1}, Z_k > b_k\}$$

$$= f(\mathcal{I}_k) - f(\mathcal{I}_{k-1}),$$

$$P_{\theta=\delta}\{a_1 < Z_1 < b_1, \, \dots \, , a_{k-1} < Z_{k-1} < b_{k-1}, Z_k < a_k\}$$

$$= g(\mathcal{I}_k) - g(\mathcal{I}_{k-1}).$$

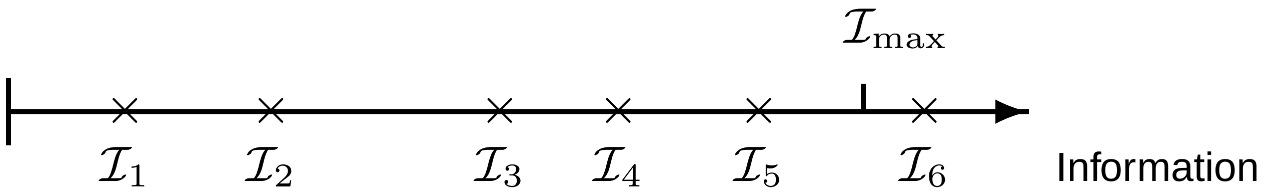


Remarks on error spending tests

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 $\mathcal{I}_{\max}$.



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

Remarks on error spending tests

3. The value of $\mathcal{I}_{\max}$ can be chosen so that boundaries converge at the final analysis under a typical sequence of information levels, e.g.,

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

4. The ρ-family provides a convenient choice of error spending functions.

In the case of one-sided tests, type I error probability is spent as

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

and type II error probability as

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

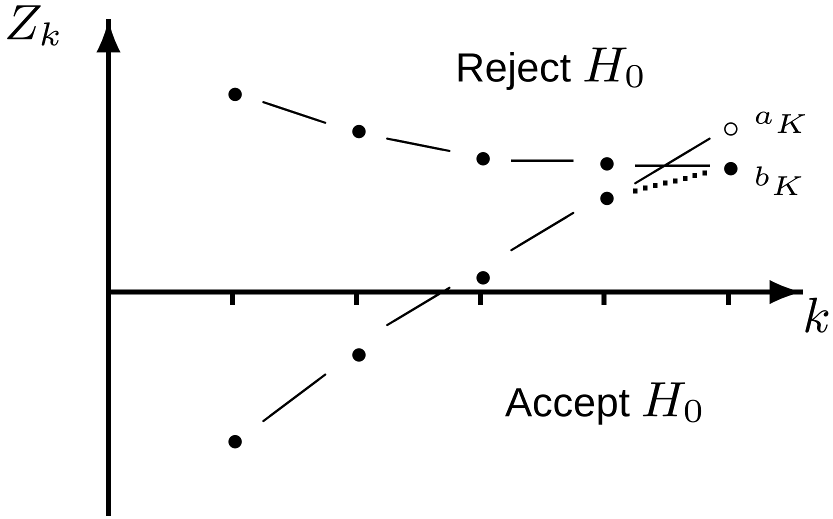
The value of ρ determines the inflation factor R.

Barber & Jennison (*Biometrika*, 2002) show ρ-family tests have excellent efficiency properties when compared with designs for the same number of analyses K and inflation factor R.

Error spending tests: over-running

Care is needed at the final analysis of a one-sided error spending test.

If one reaches $\mathcal{I}_K > \mathcal{I}_{\max}$, solving for a_K and b_K is liable to give $a_K > b_K$.



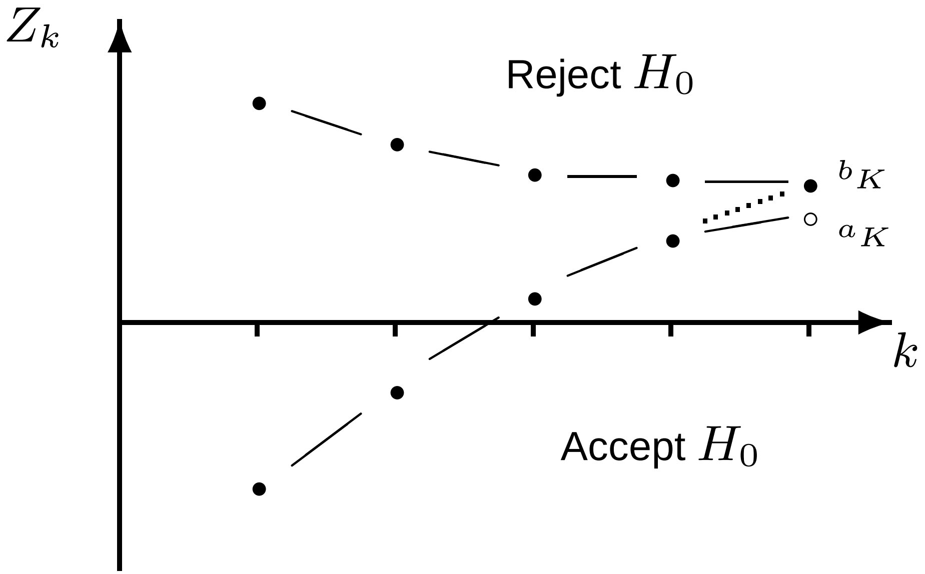
The calculated b_K guarantees type I error probability of α. So, reduce a_K to b_K — and gain extra power.

Even when $\mathcal{I}_K = \mathcal{I}_{\max}$, over-running may occur if information deviates from the equally spaced values (say) used in choosing $\mathcal{I}_{\max}$.

Error spending tests: under-running

A final information level $\mathcal{I}_K < \mathcal{I}_{\max}$ may be imposed when a final planned analysis is reached, e.g., at a maximum follow-up time in a survival study.

Then, solving for a_K and b_K is liable to give $a_K < b_K$.



Again, with b_K as calculated, the type I error probability is exactly α.

This time, increase a_K to b_K — attained power will be just below $1 - \beta$.

3. A survival data example

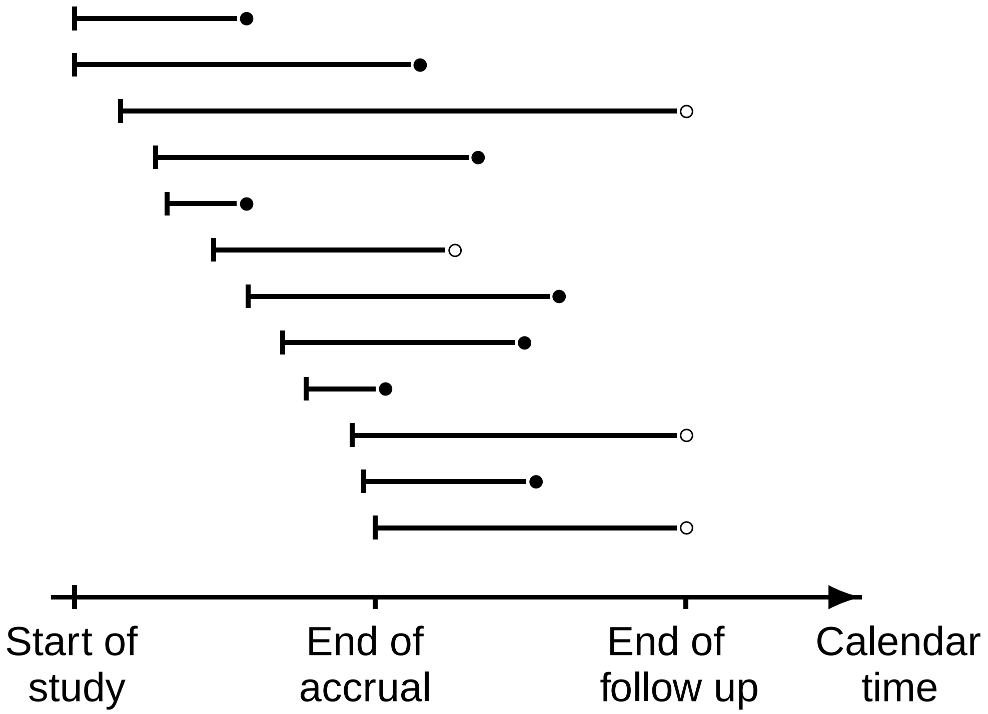
Example: Oropharynx Clinical Trial Data

Survival of patients on experimental Treatment A and standard Treatment B.

		Number entered		Number of deaths	
k	Date	Trt A	Trt B	Trt A	Trt B
1	12/69	38	45	13	14
2	12/70	56	70	30	28
3	12/71	81	93	44	47
4	12/72	95	100	63	66
5	12/73	95	100	69	73

From Kalbfleisch & Prentice (2002) *The Statistical Analysis of Failure Time Data, 2nd edition*, Appendix A, Data Set II. See also JT, Ch. 13.

Accrual and follow up in a survival study

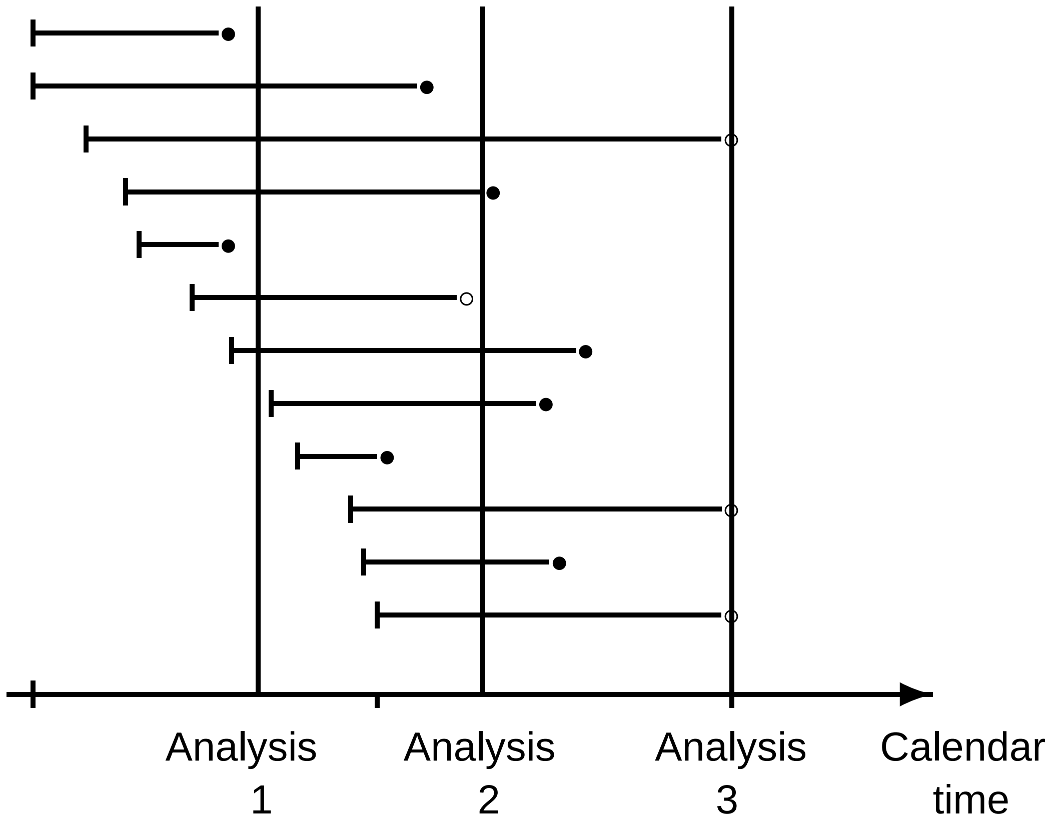


Subjects are randomised to a treatment group as they enter the study.

Survival is measured from entry to the study.

Key:	●	death time observed,
	○	censored observation.

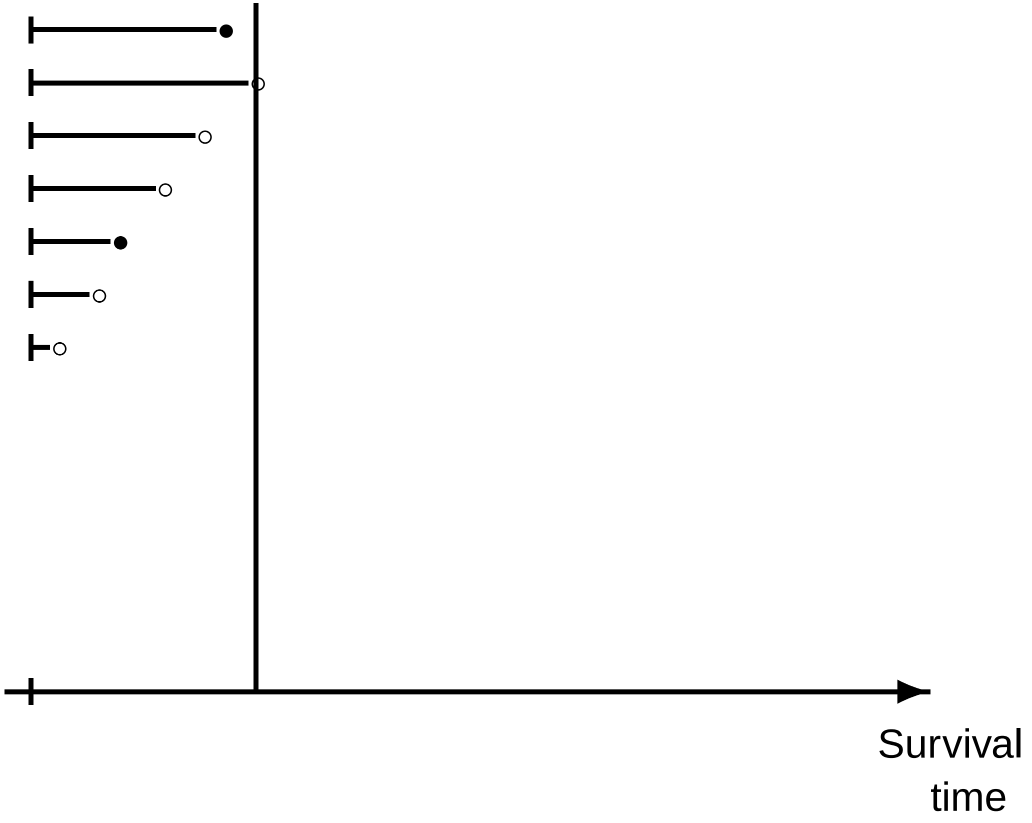
Interim analyses



At an interim analysis, subjects are censored if they are still alive at this point.

Information on such patients will continue to accrue at later analyses.

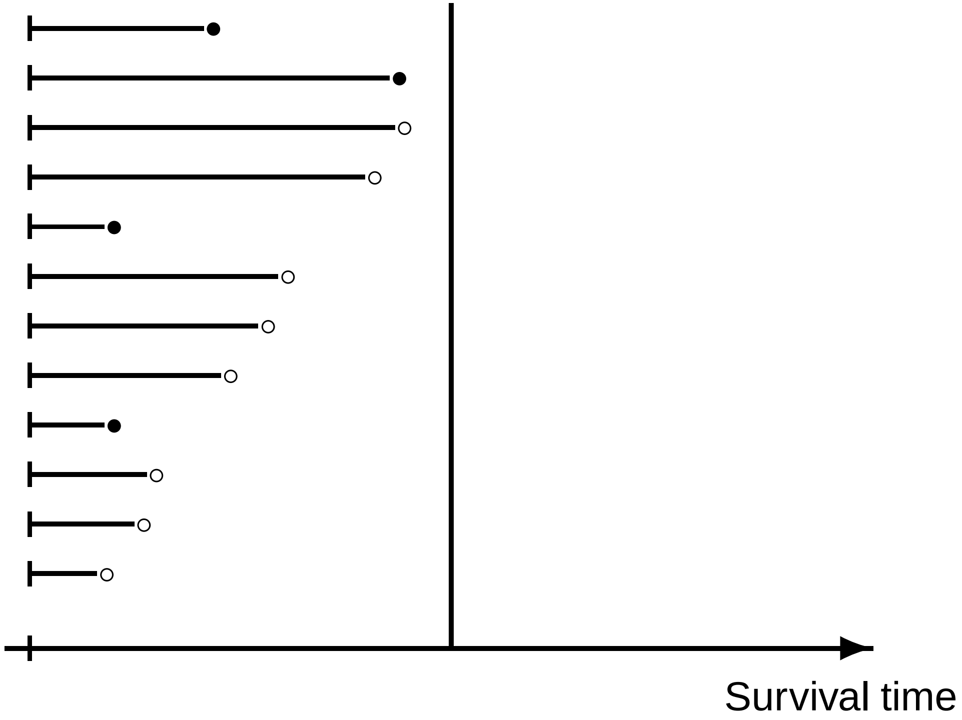
Interim analysis 1



At the first interim analysis, we analyse data on survival from randomisation time.

These times have a common starting point of zero and “analysis time” censoring occurs for subjects surviving past the first analysis.

Interim analysis 2



At interim analysis 2, we analyse data on survival from randomisation time.

These times have a common starting point of zero and “analysis time” censoring occurs for subjects surviving past the second analysis.

And so on, through further analyses . . .

The logrank statistic

At stage k, observed number of deaths is d_k.

Elapsed times between entry to the study and death for these cases are

$$\tau_{1,k} < \tau_{2,k} < \ldots < \tau_{d_k,k} \quad \text{(assuming no ties).}$$

Define

$r_{iA,k}$ and $r_{iB,k}$	Numbers at risk on Treatments A and B at $\tau_{i,k}-$
$r_{ik} = r_{iA,k} + r_{iB,k}$	Total number at risk at $\tau_{i,k}-$
O_k	Observed number of deaths on Trt B at stage k
$E_k = \sum_{i=1}^{d_k} r_{iB,k}/r_{ik}$	“Expected” number of deaths on Trt B at stage k
$V_k = \sum_1^{d_k} r_{iA,k} r_{iB,k}/r_{ik}^2$	“Variance” of O_k
$Z_k = (O_k - E_k)/\sqrt{V_k}$	Standardised logrank statistic at stage k

Proportional hazards model

Assume hazard rates h_A on Treatment A and h_B on Treatment B are related by

$$h_B(t) = \lambda\, h_A(t).$$

The log hazard ratio is $\theta = \ln(\lambda)$.

Then, with $\mathcal{I}_k = V_k$, we have approximately

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

$$\text{Cov}(Z_{k_1}, Z_{k_2}) = \surd(\mathcal{I}_{k_1}/\mathcal{I}_{k_2}) \quad \text{for } k_1 < k_2.$$

Here, V_k is the variance of the score statistic $Z_k\sqrt{\mathcal{I}_k}$.

Also, $\hat{\theta}_k = Z_k/\sqrt{\mathcal{I}_k} \sim N(\theta,\, \mathcal{I}_k^{-1})$ approximately.

For $\lambda \approx 1$, we have $\mathcal{I}_k = V_k \approx d_k/4$.

Design of the Oropharynx trial

To create: A one-sided test of H_0: $\theta \leq 0$ vs $\theta > 0$.

Note $\theta > 0 \;\Rightarrow\; \lambda > 1,$ i.e., Treatment A is better.

Require:

Type I error probability $\alpha = 0.025$,

Power $1 - \beta = 0.8$ at $\theta = 0.5,$ i.e., at $\lambda = 1.65$.

Information needed for a fixed sample study is

$$\mathcal{I}_f \;=\; \frac{\{\Phi^{-1}(\alpha) + \Phi^{-1}(\beta)\}^2}{0.5^2} \;=\; 31.40.$$

Under the approximation $\mathcal{I} \approx d/4$, the total number of failures to be observed is $d_f = 4\mathcal{I}_f \approx 126$.

Since increments in information between analyses are unpredictable, an error spending design is a natural choice.

A one-sided, error spending design

Specification:

One-sided test of H_0: $\theta \leq 0$ vs $\theta > 0$,

Type I error probability $\alpha = 0.025$,

Power $1 - \beta = 0.8$ at $\theta = \ln(\lambda) = 0.5$.

At the design stage, assume $K = 5$ equally spaced information levels.

Use a power-family test with $\rho = 2$, i.e., spending error $\propto (\mathcal{I}/\mathcal{I}_{\max})^2$.

Information for a fixed sample test has to be inflated by $R = 1.098$.

So, we require $\mathcal{I}_{\max} = 1.098 \times 31.40 = 34.48$, which needs a total of $4 \times 34.48 \approx 138$ deaths.

Summary data and critical values for the Oropharynx trial

We construct error spending boundaries using the observed information levels.

This gives boundary values $(a_1,\, b_1), \ldots, (a_5,\, b_5)$ for the standardised logrank statistics $Z_1, \ldots, Z_5$.

k	Number entered	Number of deaths	$\mathcal{I}_k$	a_k	b_k	Z_k
1	83	27	5.43	-1.41	3.23	-1.04
2	126	58	12.58	-0.21	2.76	-1.00
3	174	91	21.11	0.78	2.43	-1.21
4	195	129	30.55	1.68	2.16	-0.73
5	195	142	33.28	2.14	2.14	-0.87

This design would have led to termination at analysis 2 with acceptance of H_0.

Covariate adjustment in the Oropharynx trial

Covariate information was recorded for subjects:

institution (6), gender, initial condition,

T-staging, N-staging, tumour site (3).

Initial condition, T-staging and N-staging are continuous variables.

Proportional hazards regression model

Include treatment effect β_1, strata $l = 1, \ldots, 6$ for the six participating institutions, and coefficients $\beta_2, \ldots, \beta_7$ to model other variables.

The hazard rate for patient i is modelled as

$$h_{il}(t) \;=\; h_{0l}(t)\; e^{\{\beta_1 I(\text{Patient } i \text{ on Trt B}) + \Sigma_{j=2}^{7} x_{ij}\beta_j\}}.$$

The objective is to test H_0: $\beta_1 = 0$ against the one-sided alternative $\beta_1 > 0$.

Covariate adjustment in the Oropharynx trial

Standard software for Cox regression will provide the maximum partial likelihood estimate of the parameter vector, β, and its estimated variance.

We are interested in the treatment effect represented by the first component of β.

At stage k we have

$$\widehat{\beta}_1^{(k)}$$

$$v_k = \widehat{\mathsf{Var}}(\widehat{\beta}_1^{(k)})$$

$$\mathcal{I}_k = v_k^{-1}$$

$$Z_k = \widehat{\beta}_1^{(k)} / \sqrt{v_k}.$$

Theory: The standardised statistics $Z_1, \ldots, Z_5$ have, approximately, the canonical joint distribution.

Covariate-adjusted analysis of the Oropharynx trial

Constructing the error spending test gives boundary values $(a_1,\, b_1), \ldots, (a_5,\, b_5)$ for $Z_1, \ldots, Z_5$.

k	$\mathcal{I}_k$	a_k	b_k	$\widehat{\beta}_1^{(k)}$	Z_k
1	4.11	$-$1.75	3.39	$-$0.79	$-$1.60
2	10.89	$-$0.44	2.85	$-$0.14	$-$0.45
3	19.23	0.59	2.50	$-$0.08	$-$0.33
4	28.10	1.45	2.24	0.04	0.20
5	30.96	2.23	2.23	0.01	0.04

Under this model and stopping rule, the study would have terminated — just — at analysis 2.

NB: β_1 is the log hazard ratio after covariate adjustment. For a positive treatment effect, we should expect $\beta_1 > \lambda$.

Information monitoring in error spending designs

In a maximum information error spending design, the intent is to continue until information level $\mathcal{I}_{\max}$ is reached (unless a stopping boundary is crossed first).

For survival data, one may

a) conduct interim analyses at fixed calendar times,

b) specify analyses after given numbers of events.

In either case, it may be difficult to achieve $\mathcal{I} \geq \mathcal{I}_{\max}$ if there is

- slow patient accrual,
- low failure rate,
- high loss of subjects to follow up.

One can specify a calendar time at which to terminate the trial and spend all remaining error probability.

If $\mathcal{I} < \mathcal{I}_{\max}$ at this point, “under-running” occurs and power is reduced.

Flexibility of information monitoring designs

Error spending tests protect the type I error rate *conditional on* the sequence $\{\mathcal{I}_k\}$.

It is legitimate to make design changes which affect the observed $\mathcal{I}_k$s — as long as these changes are not influenced by observed values of the Z_ks.

One might

- add more recruitment centres,
- extend the recruitment period,
- extend the duration of follow up.

To avoid suspicion of information levels being modified in response to observed values Z_k, the study protocol should state the strategy that will be followed.

Investigators may also wish to state a maximum calendar time at which the trial will terminate, whatever the attained information level.

4. Group sequential tests with a delayed response

Survival data

In a survival study, information continues to accrue as long as there are subjects alive and uncensored. Our analyses of the oropharynx clinical trial data show it is still possible to stop early and reduce the number of subjects recruited.

Even when a survival study continues beyond the accrual period, it can be advantageous to reach a decision sooner, especially when the outcome is positive.

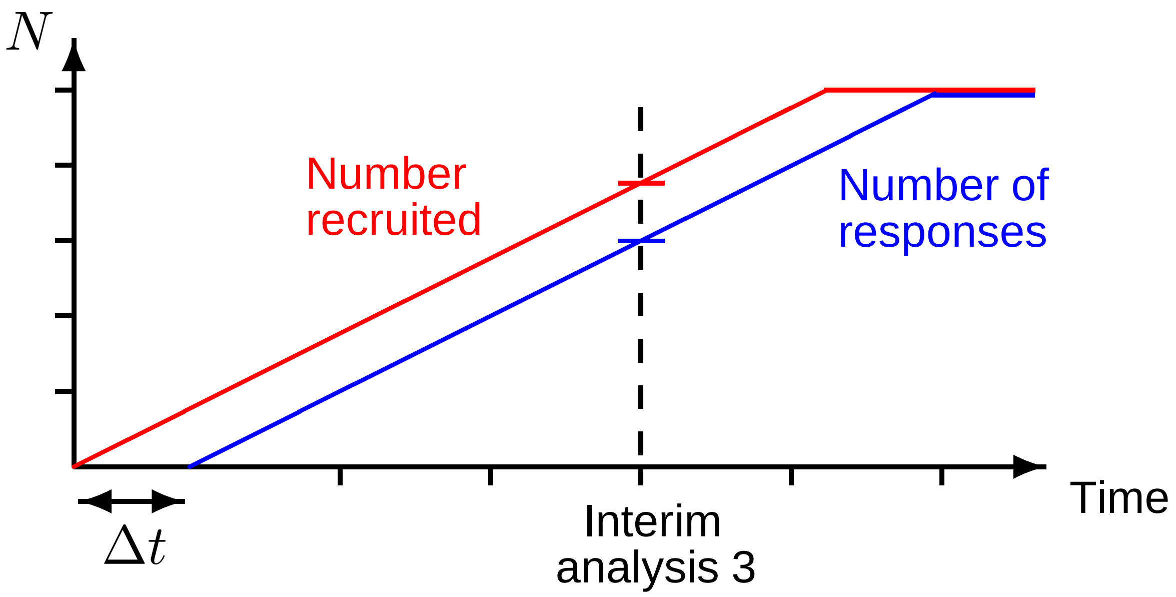
Other response types

Group sequential tests (GSTs) often assume a rapidly observed endpoint, so responses are available from all treated patients at each interim analysis.

However, this is not always the case. Consider, for example, a study comparing treatments for heart failure, where the primary endpoint is re-admission to hospital or death within 30 days: if 50 patients are recruited per month, there will be about 50 treated patients with unknown responses at each interim analysis.

Delayed response: General framework

Consider a trial with a delayed response, observed at time Δt after treatment.



Suppose a response is collected for each treated patient, even if there is a stopping decision at an interim analysis when a patient has been treated but not observed.

We shall describe Delayed Response Group Sequential Tests (DR GSTs), planned with these additional data in mind.

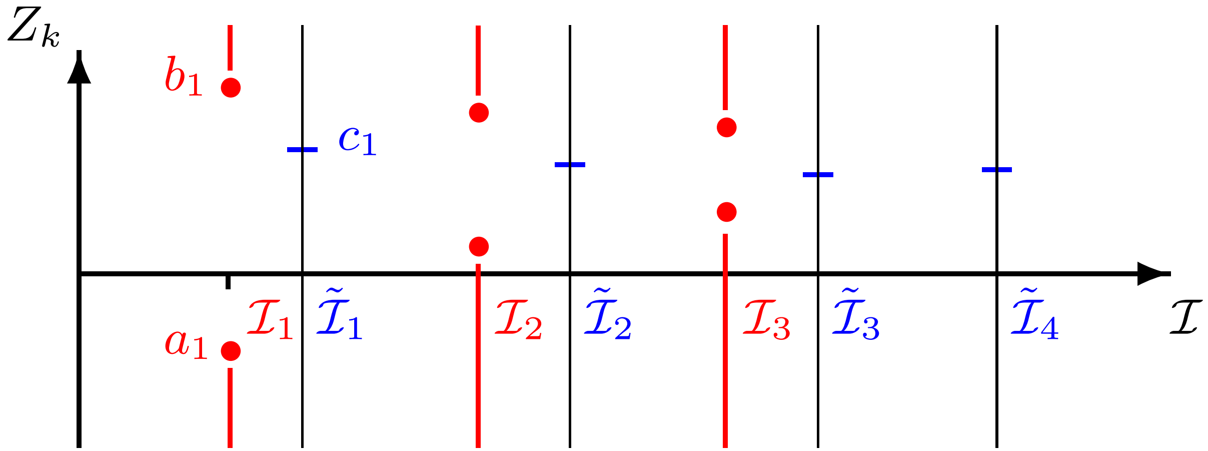
Formulating group sequential tests for a delayed response

Reference: Hampson & Jennison “Group sequential tests for delayed response”, *submitted for publication*.

At interim analysis k, with information $\mathcal{I}_k$, compare Z_k to values a_k and b_k.

If $Z_k < a_k$ or $Z_k > b_k$, cease recruitment of new patients and wait until responses have been obtained for all current patients.

At the final decision analysis, with information $\tilde{\mathcal{I}}_k$, reject H_0 if $\tilde{Z}_k > c_k$.



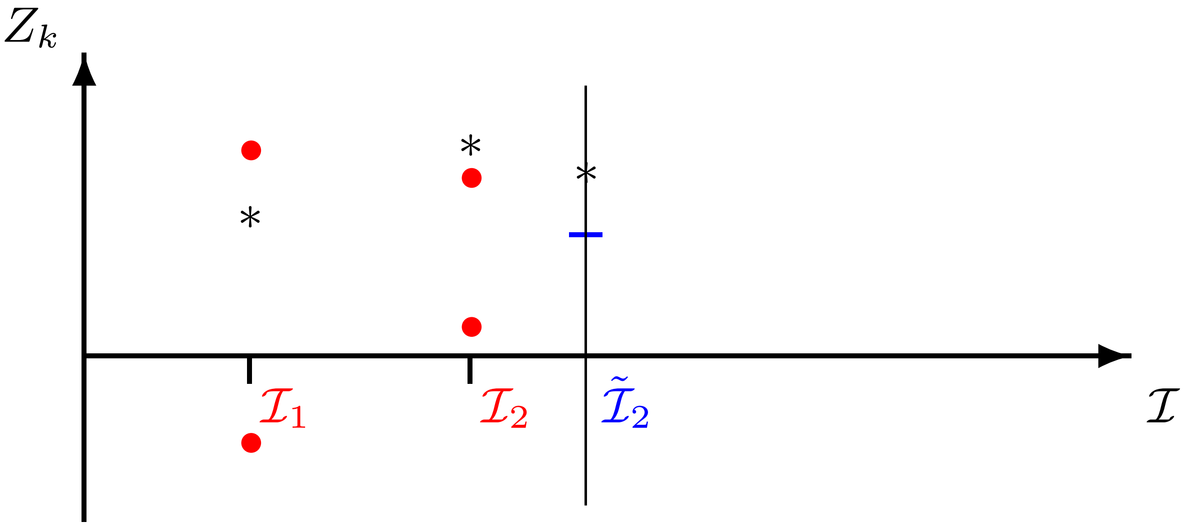
NB Whether $Z_k < a_k$ or $Z_k > b_k$ is only an indication of the likely final decision.

Delayed Response Group Sequential Tests (DR GSTs)

For a particular sequence of observed responses, we apply boundary points at a sequence of information levels of the form

$$\mathcal{I}_1,\ \ldots,\ \mathcal{I}_k,\ \tilde{\mathcal{I}}_k.$$

In the example below, recruitment ceases at the second analysis and the final decision is made with extra "pipeline" data bringing the information up to $\tilde{\mathcal{I}}_2$.



We can compute properties of a DR GST and optimise this type of design, using the same computational methods as for standard group sequential tests.

Example: Delayed Response Group Sequential Test

Hampson & Jennison (HJ) present an example of a trial comparing a new treatment for cholesterol reduction against a control.

The primary endpoint is reduction in serum cholesterol after 4 weeks of treatment.

Responses are assumed to be normally distributed with variance $\sigma^2 = 2$.

The treatment effect θ is the difference in mean response on treatment and control.

It is required to test H_0: $\theta \le 0$ against $\theta > 0$ with

Type I error rate $\alpha = 0.025$ at $\theta = 0$,

Power $1 - \beta = 0.9$ when $\theta = \delta = 1.0$.

A fixed sample test needs $n_{fix} = 86$ subjects divided between the two treatments.

HJ consider designs with a maximum sample size of 96, assuming a recruitment rate of 4 per week, giving $4 \times 4 = 16$ “pipeline” subjects at each interim analysis.

Example: Delayed Response Group Sequential Test

All 96 subjects will be recruited in 24 weeks and provide responses by 28 weeks.

Interim analyses are planned after $n_1 = 28$ and $n_2 = 54$ observed responses.

Stopping recruitment at interim analysis 1 will lead to a decision analysis with $\tilde{n}_1 = 44$ responses.

Stopping recruitment at interim analysis 2 leads to a decision analysis with $\tilde{n}_2 = 70$ responses.

No interim analysis is needed prior to the final decision analysis with 96 responses.

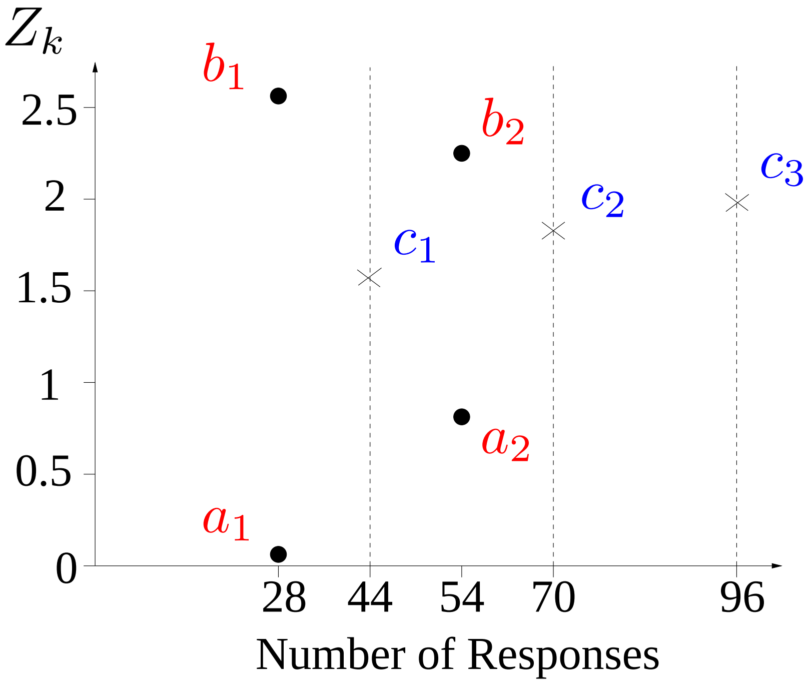
HJ derive a DR GST that minimises

$$F = \int E_\theta(N) f(\theta) d\theta,$$

where N is the total number of subjects treated and $f(\theta)$ is the density of a $N(0.5, 0.5^2)$ distribution. Optimisation is over all designs with the same interim and decision analysis times, achieving the specified type I error rate and power.

Example: Delayed Response Group Sequential Test

The critical values for statistics Z_k for the optimised DR GST are shown below.

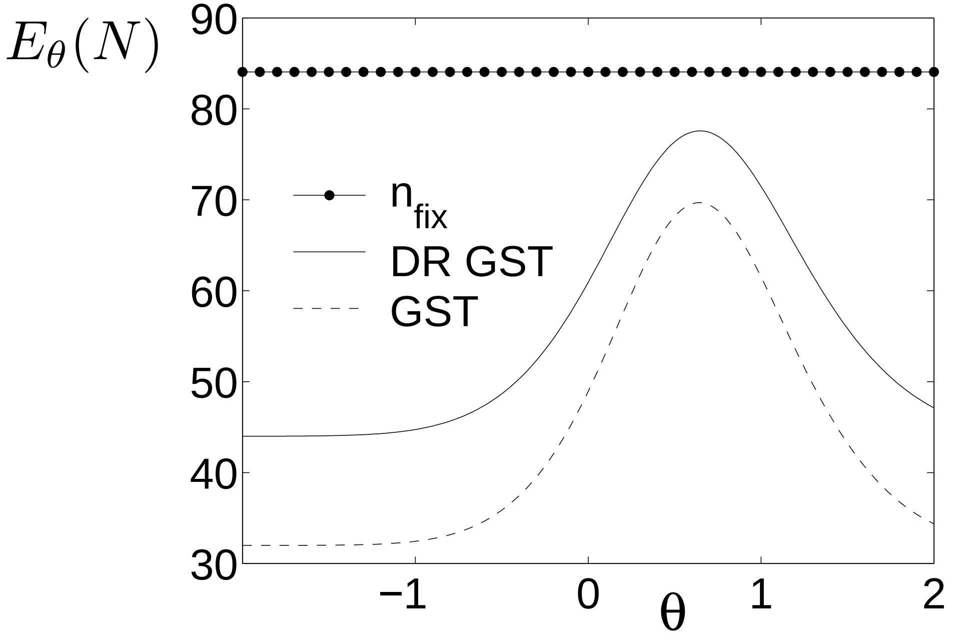


1. Critical values c_1 and c_2 at decision analyses are well below b_1 and b_2, so the probability of reversing the outcome expected when stopping recruitment is small.

2. Both c_1 and c_2 are less than 1.96. If desired, these values can be raised to 1.96 with little change to the design's power curve.

Example: Delayed Response Group Sequential Test

The figure shows expected sample size curves for the fixed sample design with $n_{fix} = 85$ patients, the optimised DR GST, and the GST for immediate response with analyses after 32, 64 and 96 responses, optimised for the same criteria.



The DR GST achieves savings in $E_\theta(N)$ below the fixed sample size, n_{fix} at all effect sizes θ. However, the delay in response means these savings are smaller than they would be in the case of an immediate response.

Group sequential tests for a delayed response

Hampson & Jennison (2011) assess how much of the reduction in expected sample size achieved by group sequential testing is lost as the volume of “pipeline” data increases.

Substantial savings are still present for a small number of pipeline subjects.

However, as this number increases to 25% of the total sample size, about half the benefits of group sequential testing are lost.

Strategies are available to restore some of this efficiency:

Recruiting subjects more slowly,

Incorporating data on short term responses which are correlated with the longer term, primary endpoint.

5. An alternative type of group sequential test

Reference: Lehmacher and Wassmer (*Biometrics*, 1999)

Let $Z_{(i)}$ denote the Z-statistic from data in group i alone, $i = 1, \ldots, K$.

Define the Z-statistic based on *all* the data up to analysis k to be

$$Z_k = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} Z_{(i)}. \qquad (1)$$

Under $\theta = 0$, each $Z_{(i)} \sim N(0, 1)$ and the sequence of statistics $\{Z_k\}$ has the joint distribution that arises when group sizes are equal and each Z_k is the usual statistic based on the cumulative data at analysis k.

Thus, we can use constants from a standard group sequential test to define a boundary $\{(a_k, b_k)\}$ for the $\{Z_k\}$ giving a test with specified type I error rate α.

The definition (1) can be used for statistics $Z_{(i)}$ with quite general definitions.

This provides a tool that enables flexible and adaptive sequential design.

Lehmacher and Wassmer's method

Group sequential t-tests

For normal data with unknown variance, we can compute a t-statistic from the group i data, convert this to a one-sided P-value P_i, and take the normal deviate

$$Z_{(i)} = \Phi^{-1}(1 - P_i).$$

The $Z_{(i)}$ are then independent and distributed as $N(0,1)$ under H_0.

Sample size adaptation

It is still the case that the $Z_{(i)}$ are independent $N(0,1)$ under H_0 if future group sizes are modified on the basis of estimates of the response variance.

This gives a method for sample size re-estimation to achieve a pre-specified power (but note that groups of different size are given equal weight in the overall Z_k).

A combination test

This way of combining the group summaries, $Z_{(i)}$, produces a K-stage version of the *combination tests* proposed by Bauer & Köhne (*Biometrics*, 1994).

6. Adapting the target population: Enrichment designs

Consider a new treatment developed to disrupt a disease's biological pathway.

Patients with high levels of a biomarker associated with this pathway should gain particular benefit, but the treatment's wider action may also help the general patient population.

As an example, it is recognised that only a portion of the patient population appears to respond to some current cancer treatments. However, we are only just learning how to identify such sub-populations through genetic characteristics.

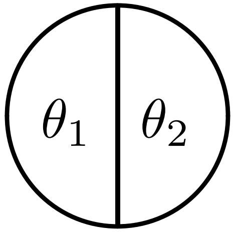
For new therapies, a target population may be specified — and also a smaller sub-population, in which the treatment is expected to be particularly effective.

The aim in an "enrichment design" is to learn whether there is a differential treatment effect in patient subgroups and, if appropriate, change the focus of the trial to those subgroups in which there is greatest potential benefit.

Enrichment designs

Sub-population

(proportion λ_1)



Rest of the population

(proportion λ_2)

In a clinical trial with enrichment we

Start by comparing the new treatment against control in the full population.

Examine responses at an interim stage.

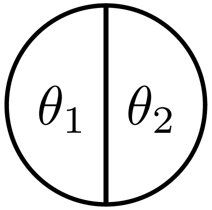
If there is no evidence of treatment effect, stop for futility.

If the new treatment appears effective in the full population, continue as before.

If the new treatment appears to benefit just the subgroup, recruit only from the subgroup and increase the numbers in this subgroup.

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

Enrichment designs



Sub-population
(proportion λ_1)

Rest of the population
(proportion λ_2)

Denote the treatment effect:

In the sub-population by θ_1,

In the complement of the sub-population by θ_2,

Aggregated over the whole population by $\theta_3 = \lambda_1\theta_1 + \lambda_2\theta_2$.

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

The null hypothesis for the full target population is H_3: $\theta_3 \leq 0$.

We may wish to test either of the two null hypotheses H_1 and H_3 against one-sided alternatives, $\theta_1 > 0$ and $\theta_3 > 0$.

Testing multiple hypotheses

Closed testing procedures

Suppose there are k null hypotheses, H_i: $\theta_i \leq 0$ for $i = 1, \ldots, k$.

A procedure's ***familywise error rate*** under a set of values $(\theta_1, \ldots, \theta_k)$ is

$$Pr\{\text{Reject } H_i \text{ for some } i \text{ with } \theta_i \leq 0\} \;=\; Pr\{\text{Reject any true } H_i\}.$$

The familywise error rate is controlled strongly at level α if this error rate is at most α for all possible combinations of θ_i values. Then

$$Pr\{\text{Reject any true } H_i\} \leq \alpha \quad \text{for all } (\theta_1, \ldots, \theta_k).$$

Using such a procedure, the probability of choosing to focus on the parameter θ_{i^*} and then falsely claiming significance for null hypothesis H_{i^*} is at most α.

Closed testing procedures (Marcus et al, *Biometrika*, 1976) provide strong control by combining level α tests of each H_i and of intersections of these hypotheses.

Closed testing procedures

For each subset I of $\{1, \ldots, k\}$, define the intersection hypothesis

$$H_I = \cap_{i \in I} H_i.$$

Construct a level α test of each intersection hypothesis H_I, i.e., a test which rejects H_I with probability at most α whenever all hypotheses specified in H_I are true.

Closed testing procedure

The simple hypothesis H_j: $\theta_j \leq 0$ is rejected if, and only if, H_I is rejected for every set I containing index j.

Proof of strong control of familywise error rate

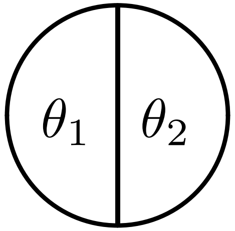
Let $\tilde{I}$ be the set of indices of all true hypotheses H_i. For a familywise error to be committed, $H_{\tilde{I}}$ must be rejected.

Since $H_{\tilde{I}}$ is true, $Pr\{$Reject $H_{\tilde{I}}\} = \alpha$ and, thus, the probability of a familywise error is no greater than α.

The enrichment design problem

A trial is to investigate whether a new treatment is beneficial to the full population or, failing that, in a sub-population.

Sub-population (proportion λ_1)



Rest of the population (proportion λ_2)

The treatment effect is θ_1 in the sub-population, θ_2 in its complement, and the average effect in the full population is $\theta_3 = \lambda_1\theta_1 + \lambda_2\theta_2$.

We wish to test:

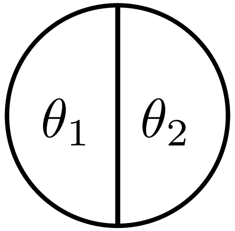
The null hypothesis for the full population, H_3: $\theta_3 \leq 0$ vs $\theta_3 > 0$,

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

The benefits of enrichment

Sub-population

(proportion λ_1)



Rest of the population

(proportion λ_2)

First, consider a design testing for a whole population effect, $\theta_3 \;=\; \lambda_1\theta_1 + \lambda_2\theta_2$.

The design has two analyses and one-sided type I error probability 0.025.

Sample size is set to achieve power 0.9 at $\theta_3 = 20$.

Data in each stage are summarised by a Z-value:

	Stage 1	*Stage 2*	*Overall*
H_3: $\theta_3 \leq 0$	$Z_{1,3}$	$Z_{2,3}$	$Z_3 \;=\; \frac{1}{\sqrt{2}}Z_{1,3} + \frac{1}{\sqrt{2}}Z_{2,3}$

The benefits of enrichment

Two stage design testing for a whole population effect, θ_3.

	Stage 1	*Stage 2*	*Overall*
H_3: $\theta_3 \leq 0$	$Z_{1,3}$	$Z_{2,3}$	$Z_3 \; = \; \frac{1}{\sqrt{2}} Z_{1,3} + \frac{1}{\sqrt{2}} Z_{2,3}$

Decision rules:

If $Z_{1,3} < 0$		Stop at Stage 1, Accept H_3
If $Z_{1,3} \geq 0$		Continue to Stage 2, then
	If $Z_3 < 1.95$	Accept H_3
	If $Z_3 \geq 1.95$	Reject H_3

The benefits of enrichment

Assume the sub-population comprises half the total population, so $\lambda_1 = \lambda_2 = 0.5$.

Properties of design for the whole population effect, θ_3:

θ_1	θ_2	θ_3	*Power for* H_3: $\theta_3 \leq 0$
20	20	20	0.90
10	10	10	0.37
20	0	10	0.37

Is it feasible to identify at Stage 1 that θ_3 is low but θ_1 may be higher, so one might switch resources to test a sub-population?

The benefits of enrichment

We wish to be able to consider two null hypotheses:

H_3: $\theta_3 \leq 0$ Treatment is not effective in the whole population,

H_1: $\theta_1 \leq 0$ Treatment is not effective in the sub-population.

Since $\theta_3 = 0.5\,\theta_1 + 0.5\,\theta_2$, either of H_1 and H_3 may be true on its own.

In applying a ***closed testing procedure***, we also test the intersection hypothesis

$$H_{13}\text{:}\quad \theta_1 \leq 0 \text{ and } \theta_3 \leq 0.$$

Then to reject H_1 overall, while protecting the family-wise type I error rate, we need to reject both H_1 and H_{13} in individual tests at significance level α.

Similarly, we can reject H_3 overall if both H_3 and H_{13} are rejected in level α tests.

An adaptive design

At Stage 1, if $\hat{\theta}_3 < 0$, stop to accept H_3: $\theta_3 \leq 0$.

If $\hat{\theta}_3 > 0$ and the trial continues:

If $\hat{\theta}_2 < 0$ and $\hat{\theta}_1 > \hat{\theta}_2 + 8$	Restrict to sub-population 1 and test H_1 only, needing to reject H_1 and H_{13}.
Else,	Continue with full population and test H_3, needing to reject H_3 and H_{13}.

The same *total* sample size for Stage 2 is retained in both cases, increasing the numbers for the sub-population when enrichment occurs.

An adaptive design

Each null hypothesis, H_i say, is tested in a 2-stage group sequential test.

With Z-statistics Z_1 and Z_2 from Stages 1 and 2, H_i is rejected if

$$Z_1 \geq 0 \quad \text{and} \quad \frac{1}{\sqrt{2}} Z_1 + \frac{1}{\sqrt{2}} Z_2 \geq 1.95.$$

When continuing with the full population, we use Z-statistics:

	Stage 1	*Stage 2*
H_3	$Z_{1,3}$	$Z_{2,3}$
H_{13}	$Z_{1,3}$	$Z_{2,3}$

where $Z_{i,3}$ is based on $\hat{\theta}_3$ from responses in Stage i.

With these definitions, there is no change from the original test of H_3. This should help maintain power to reject H_3 and identify an effect in the full population.

An adaptive design

With Z-statistics Z_1 and Z_2 from Stages 1 and 2, H_i is rejected if

$$Z_1 \geq 0 \quad \text{and} \quad \frac{1}{\sqrt{2}} Z_1 + \frac{1}{\sqrt{2}} Z_2 \geq 1.95.$$

When switching to the sub-population, we use:

	Stage 1	*Stage 2*
H_1	$Z_{1,1}$	$Z_{2,1}$
H_{13}	$Z_{1,3}$	$Z_{2,1}$

where $Z_{i,j}$ is based on $\hat{\theta}_j$ from responses in Stage i.

The need to reject the intersection hypothesis H_{13} adds an extra requirement to the simple test of H_1.

Simulation results: Power of non-adaptive and adaptive designs

				Non-adaptive	***Adaptive***		
	θ_1	θ_2	θ_3	*Full pop^n*	*Sub-pop^n only*	*Full pop^n*	*Total*
1.	30	0	15	**0.68**	0.43	0.42	**0.85**
2.	20	0	10	**0.37**	0.24	0.26	**0.51**
3.	20	20	20	**0.90**	0.03	0.87	**0.90**
4.	20	10	15	**0.68**	0.11	0.60	**0.71**

Cases 1 & 2: Testing focuses (correctly) on H_1, but it is still possible to find an effect (wrongly) for the full population. Overall power is increased.

Case 3: Restricting to the sub-population reduces power for finding an effect in the full population.

Case 4: Adaptation improves overall power a little.

Increasing power for finding a sub-population effect

In order to achieve greater power for finding an effect in the sub-population, we could use $Z_{1,1}$ rather than $Z_{1,3}$ as the Stage 1 statistic in the test of H_{13}.

However, this choice is detrimental to power when there is a good treatment effect across the whole population, as in the previous table's

Case 3: $\theta_1 = 20, \;\; \theta_2 = 20$,

Case 4: $\theta_1 = 20, \;\; \theta_2 = 10$.

A compromise between these two options is provided by

$$\tilde{Z}_{1,13} \; = \; (Z_{1,3} + Z_{1,1})/\sqrt{(2+\sqrt{2})},$$

which has a $N(0,1)$ distribution under H_{13}.

Increasing power for finding a sub-population effect

Taking the Stage 1 statistic for the test of H_{13} to be

$$\tilde{Z}_{1,13} = (Z_{1,3} + Z_{1,1})/\sqrt{(2+\sqrt{2})},$$

leads to the following results:

				Non-adaptive	*Adaptive*		
	θ_1	θ_2	θ_3	*Full pop^n*	*Sub-pop^n only*	*Full pop^n*	*Total*
1.	30	0	15	**0.68**	0.47	0.41	**0.88**
2.	20	0	10	**0.37**	0.33	0.25	**0.58**
3.	20	20	20	**0.90**	0.04	0.83	**0.87**
4.	20	10	15	**0.68**	0.15	0.57	**0.72**

Use of $\tilde{Z}_{1,13}$ has increased power to find a treatment effect in the sub-population in Cases 1 & 2 at the cost of a small drop in power for Case 3.

The benefits of enrichment

In defining an enrichment design, the rules for staying with the full population or switching to the sub-population can be adjusted to favor specific goals.

However, we cannot eliminate the probability of making an error in these decisions.

This is to be expected. The standard error of the interim estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ is 12.3 — much higher than the differences between θ_1 and θ_2 that interest us.

Similar problems are liable to arise in any adaptive procedure which uses noisy interim data as the basis of mid-study modifications.

So, although restricting attention to a sub-population can be effective in improving power, higher overall sample size is needed for accurate sub-population inference.

7. Conclusions

- Group sequential tests are valuable in monitoring clinical trials with a view to early stopping for efficacy or futility.
- The general framework of group sequential designs accommodates a wide variety of response distributions and types of stopping rule.
- Error spending designs can handle unpredictable increments in information about the primary endpoint while maintaining statistical efficiency.
- Extensions of the standard form of group sequential test have been developed to give efficient designs when there is a delay in observing patients' responses.
- Adaptive designs offer an alternative approach to updating sample size in response to estimates of nuisance parameters, such as the response variance.
- Combination tests used in conjunction with closed testing procedures provide a methodology for testing an adaptively selected hypothesis.