ADAPTIVITY IN GROUP SEQUENTIAL DESIGNS

Professor Christopher Jennison,

Dept of Mathematical Sciences, University of Bath, UK

Professor Bruce Turnbull,

Cornell University, Ithaca, New York

Phase III Clinical Trials in Oncology:

From Design to Approval

IIR Life Sciences, Amsterdam, February 2006

http://www.bath.ac.uk/∼mascj

Outline of presentation

1. Group sequential designs for clinical trials

   Adapting to observed data

2. Error-spending tests

   Adapting to unpredictable information

   Adapting to nuisance parameters

3. Most efficient group sequential tests

   Optimal stopping boundaries

   Adapting group sizes to observed data

4. Flexibility for unplanned design changes

   Adapting to new objectives

5. An example of inefficiency in an adaptive designs

   How not to adapt!

1. Group sequential monitoring of clinical trials

A one-sided testing problem

Let θ be the treatment effect of a new treatment *vs* a standard, e.g.,

$\theta =$ difference in mean response for normal data, or

$\theta =$ log hazard ratio for survival data.

To look for superiority of the new treatment, test

$$H_0\colon \theta \leq 0 \quad \text{against} \quad \theta > 0.$$

Specify type I error rate $= \alpha$ and power $1 - \beta$ at $\theta = \delta$.

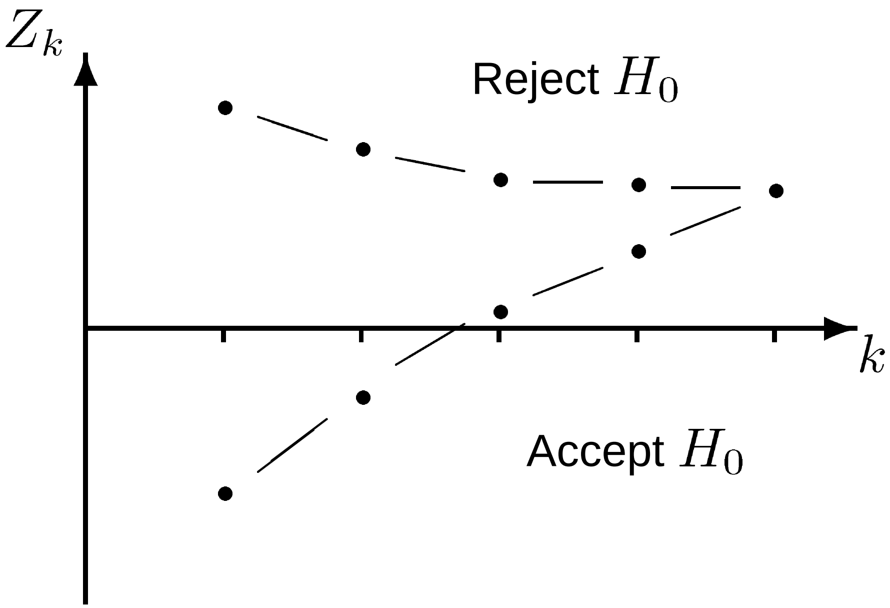
It is desirable to stop early

to accept H_0 — early stopping for futility,

to reject H_0 — early stopping for a positive outcome.

One-sided group sequential tests

A typical group sequential testing boundary has the form:



$E($Sample size$)$ can be around 50% to 70% of the fixed sample size,

see, e.g., Jennison & Turnbull (2000) “*Group Sequential Methods* . . . ”

Adapting to data, stopping when a decision is possible.

2. Error spending tests

Canonical joint distribution of parameter estimates

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

The *information* for θ at analysis k is

$$\mathcal{I}_k \;=\; \frac{1}{\text{Var}(\widehat{\theta}_k)}\,, \quad k = 1, \ldots, K.$$

In very 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

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

Spending error as a function of $\boldsymbol{\mathcal{I}_k}$

Observed information $\mathcal{I}_k$ depends on the number of subjects and other factors, e.g., for survival data, the overall failure rate.

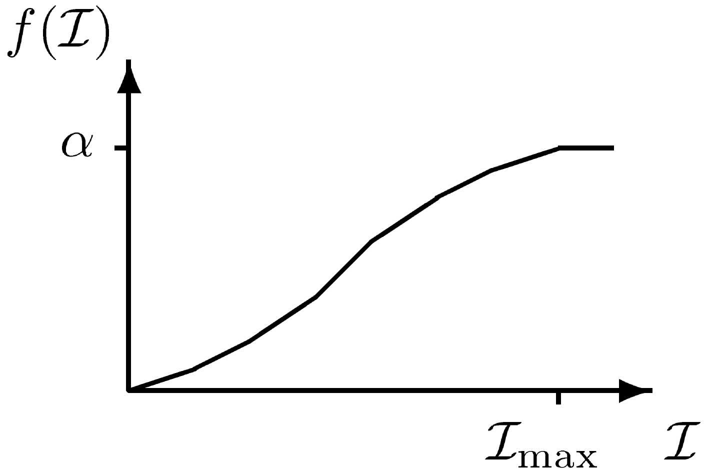
Thus, it may not be possible to predict the actual sequence of information levels, $\mathcal{I}_1, \mathcal{I}_2, \ \ldots$, in advance.

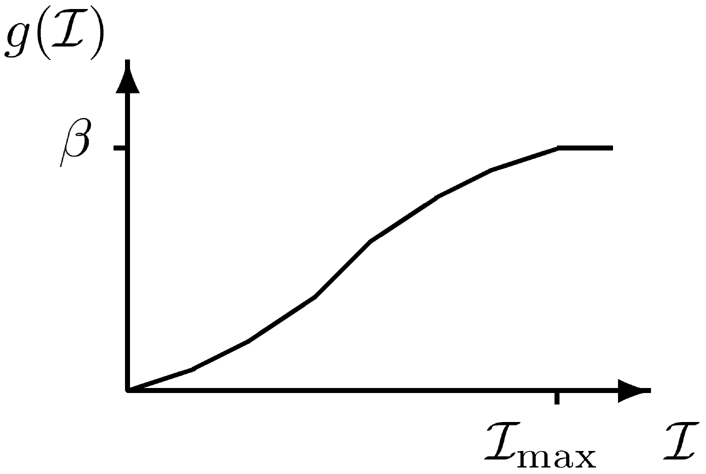
Lan & DeMets (1983) presented two-sided tests with the flexibility to "spend" type I error probability as a function $f(\mathcal{I})$ of the observed information:

> *at analysis k, the current boundary point is set so that*
> *the cumulative type I error probability is $f(\mathcal{I}_k)$.*

To extend to one-sided tests, define two functions, $f(\mathcal{I})$ and $g(\mathcal{I})$, for spending type I and type II error probabilities.

One-sided error spending tests





At analysis k, set boundary values $(a_k,\, b_k)$ so that

$$Pr_{\theta=0}\,\{\text{Reject } H_0 \text{ by analysis } k\} \;=\; f(\mathcal{I}_k),$$

$$Pr_{\theta=\delta}\,\{\text{Accept } H_0 \text{ by analysis } k\} \;=\; g(\mathcal{I}_k).$$

Power family of error spending tests: $f(\mathcal{I})$ and $g(\mathcal{I}) \;\propto\; (\mathcal{I}/\mathcal{I}_{\max})^{\rho}$.

Adapting to unpredictable information

Maximum information designs

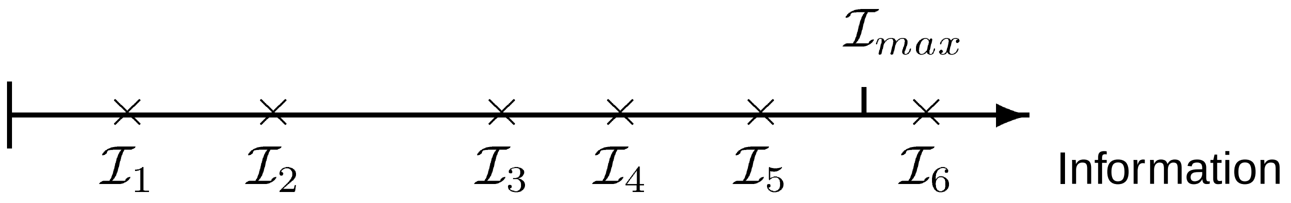
Design

Assume, say, K equally spaced information levels.

Find $\mathcal{I}_{max}$ such that boundaries meet up on reaching $\mathcal{I}_K = \mathcal{I}_{max}$.

Implementation

Use the error-spending construction with observed $\mathcal{I}_k$s. Continue up to $\mathcal{I}_{max}$ and make the boundaries converge, protecting type I error.



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

N.B. Changes affecting $\{\mathcal{I}_1, \mathcal{I}_2, \ldots\}$ should **not** be influenced by $\widehat{\theta}_k$s.

Error-spending designs and nuisance parameters

The target $\mathcal{I}_{max}$ is fixed but the sample size needed to achieve this can depend on parameters which are initially unknown.

(1) Normal responses with unknown variance

If $X_i \sim N(\mu_X,\, \sigma^2),\ Y_i \sim N(\mu_Y,\, \sigma^2)$ and $\theta = \mu_X - \mu_Y$,

$$\mathcal{I}_k \;=\; (\,\sigma^2/n_{X,k} + \sigma^2/n_{Y,k}\,)^{-1}.$$

(2) Survival data, log-rank statistics

Information depends on the number of observed failures,

$$\mathcal{I}_k \;\approx\; \{\text{Number of failures by analysis } k\}/4.$$

Error spending designs handle these issues automatically.

Adapting to nuisance parameters

3. Optimal group sequential tests

There is plenty of choice in defining a boundary to solve a particular testing problem. Thus, one can seek a boundary with an optimality property.

Formulate the testing problem:

fix type I error rate α and power $1-\beta$ at $\theta = \delta$,

fix number of analyses, K,

fix maximum sample size (information), if desired.

Find the design which minimises average sample size (information) at one particular θ or averaged over several θs.

This optimisation can be carried out by solving a related Bayes decision problem using backwards induction (dynamic programming).

Example of properties of optimal tests

One-sided tests, $\alpha = \beta = 0.05$, K analyses, $\mathcal{I}_{max} = R\,\mathcal{I}_{fixed}$, 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* $\mathcal{I}_{fixed}$

	R					Minimum
K	1.01	1.05	1.1	1.2	1.3	over R
2	80.9	74.5	72.8	73.2	75.3	72.7 at R=1.15
5	72.2	65.2	62.2	59.8	59.0	58.7 at R=1.4
10	69.1	62.1	59.0	56.3	55.2	54.3 at R=1.6

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

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

Adapting optimally to observed data

Squeezing a little extra efficiency

Schmitz (1993) proposed group sequential tests in which group sizes are chosen adaptively. We describe these on the score statistic scale:

Initially, fix $\mathcal{I}_1$, observe

$$S_1 \sim N(\theta\mathcal{I}_1, \mathcal{I}_1),$$

then choose $\mathcal{I}_2$ as a function of S_1, observe S_2 where

$$S_2 - S_1 \sim N(\theta(\mathcal{I}_2 - \mathcal{I}_1), (\mathcal{I}_2 - \mathcal{I}_1)),$$

etc, etc.

Specify sampling rule and stopping rule to achieve desired *overall* type I error rate and power.

Examples of “Schmitz” designs

To test H_0: $\theta = 0$ versus H_1: $\theta > 0$ with type I error rate $\alpha = 0.025$ and power $1 - \beta = 0.9$ at $\theta = \delta$.

Aim for low values of

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

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

Constraints:

Maximum sample information $= 1.2 \times$ fixed sample information.

Maximum number of analyses $= K$.

Again, find optimal designs by solving related Bayes decision problems.

Efficiency of “Schmitz” designs

Optimal average $E(\mathcal{I})$ as a percentage of the fixed sample information.

K	*Optimal adaptive design (Schmitz)*	*Optimal non-adaptive, optimised group sizes*	*Optimal non-adaptive, equal group sizes*
2	72.5	73.2	74.8
3	64.8	65.6	66.1
4	61.2	62.4	62.7
6	58.0	59.4	59.8
10	55.9	57.2	57.5

Varying group sizes *adaptively* makes for a complex procedure and the efficiency gains are slight.

Adapting group sizes optimally to observed data

4. Recent advances in flexible/adaptive methods

Mid study re-design to increase power

During the course of a study, reasons may arise to change the power. Suppose you design a study with power 0.9 at $\theta = \delta^*$. If a competing treatment is withdrawn, you may wish to increase sample size to attain power 0.9 at $\theta = \delta^{**} < \delta^*$.

> *Can you do this during a fixed sample or group sequential study without biasing the type I error rate?*

Denne (2001) and Müller & Schäfer (2001) show this is possible as long as the re-design **preserves the conditional type I error probability**.

The methods of Bauer & Köhne (1994), Fisher (1998), Cui, Hung & Wang (1999) are described differently, but they also possess this property.

Re-design in response to an interim estimate, $\widehat{\boldsymbol{\theta}}$

Sample size may be modified in response to an estimate of effect size, $\widehat{\theta}$.

Often, designs are set up to attain a given conditional power under $\theta = \widehat{\theta}$.

Motivation can be:

- to rescue an under-powered study,
- a “wait and see” approach to choosing a study’s power requirement,
- trying to be efficient.

The conditional type I error rate approach safeguards overall type I error.

It is good to be able to rescue a poorly designed study.

But, group sequential tests already base the decision for early stopping on $\widehat{\theta}$ *— and optimal GSTs do this optimally!*

5. Example of inefficiency in an adaptive design

Scenario (of the type described by Cui, Hung & Wang, 1999)

A test is to have type I error probability $\alpha = 0.025$.

Investigators are optimistic the effect size, θ, will be as high as $\delta^* = 20$. But, effect sizes as low as $\theta = \delta^{**} = 15$ are clinically relevant and worth detecting.

First, consider a fixed sample study attaining power 0.9 at $\theta = \delta^* = 20$. Suppose this requires a sample size $n_f = 100$.

An adaptive design starts out as a fixed sample test with $n_f = 100$ observations, but data are examined after the first 50 responses to see if there is a need to “adapt”.

Cui et al. adaptive design

At an interim stage, after 50 observations, the estimated effect size is $\widehat{\theta}_1$.

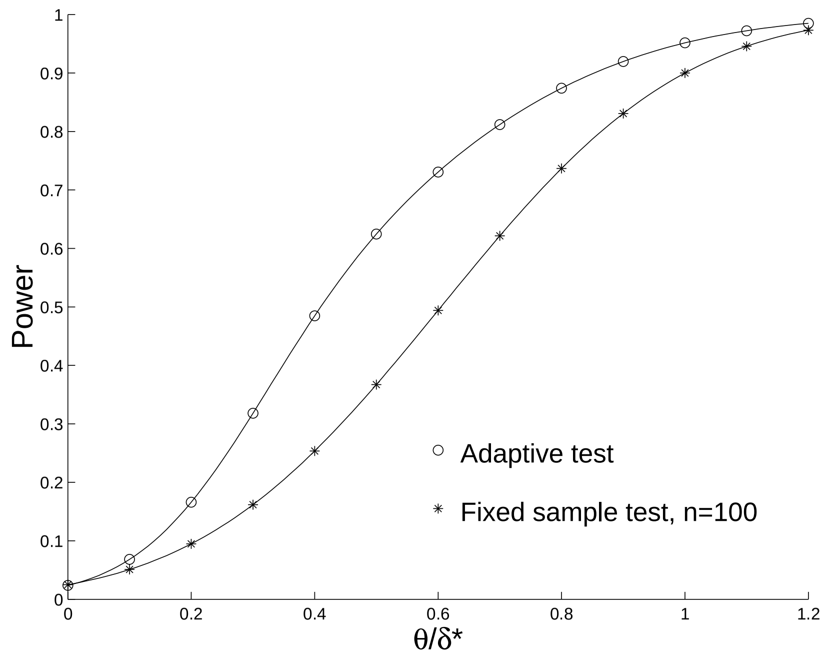
If $\widehat{\theta}_1 < 0.2\,\delta^* = 4,$ stop the trial for futility, accepting H_0.

Otherwise, re-design the remainder of the trial, preserving the conditional type I error rate given $\widehat{\theta}_1$:

choose the remaining sample size to give conditional power 0.9 if in fact $\theta = \widehat{\theta}_1$,

truncate this additional sample size to the interval (50, 500) — no decrease in sample size is allowed and the total sample size is at most 550.

Power of the Cui et al. adaptive test



The adaptive test improves on the power of the fixed sample test, achieving power 0.85 at $\theta = \delta^{**} = 15$ (i.e., $\theta/\delta^{*} = 0.75$).

If continuing past stage one, total sample size ranges from 100 to 550.

A conventional group sequential test

Similar overall power can be obtained by a non-adaptive GST with $K = 2$ analyses, designed to attain power 0.9 when $\theta = 14$.

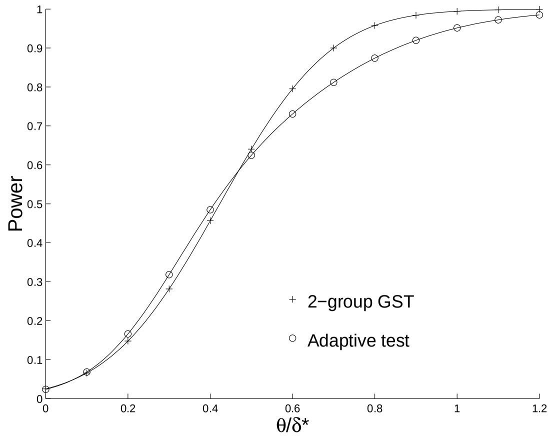
We have compared a power family, error spending test with $\rho = 1$:

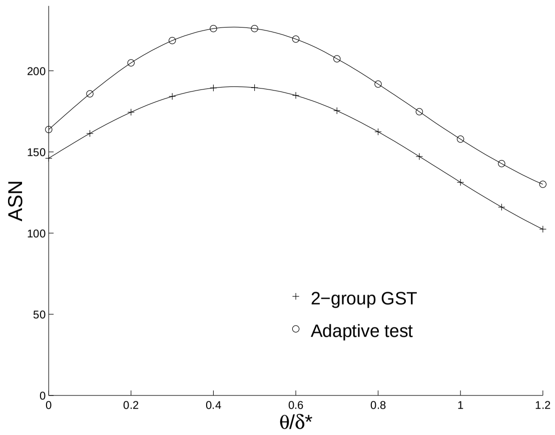
type I error rate is $\alpha = 0.025$,

taking the first analysis after 68 observations and the second analysis after 225 gives a test with power 0.9 at $\theta = 14$.

This test dominates the Cui et al. adaptive design with respect to both power and ASN. It also has a much lower maximum sample size — 225 compared to 550.

Cui et al. adaptive test *vs* non-adaptive GST





The conventional GST has:

higher power,

lower average sample size function,

much smaller maximum sample size.

Many other proposals for adaptive designs show similar inefficiencies.

6. Conclusions

Error Spending tests using Information Monitoring can adapt to

- unpredictable information levels,
- nuisance parameters,
- observed data, i.e., efficient stopping rules.

In addition, ***recent adaptive methods*** allow

- re-design in response to external developments,
- re-sizing to rescue an under-powered study,
- an on-going approach to study design.

But, these adaptive designs will not improve on the efficiency of “standard” Group Sequential Tests — **and they can be substantially inferior.**

References

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50**, 1029–1041. Correction *Biometrics* **52**, (1996), 380.

Cui, L., Hung, H.M.J. and Wang, S-J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55**, 853–857.

Denne, J.S. (2001). Sample size recalculation using conditional power. *Statistics in Medicine* **20**, 2645–2660.

Fisher, L.D. (1998). Self-designing clinical trials. *Statistics in Medicine* **17**, 1551–1562.

Jennison, C. and Turnbull, B.W. (2000). *Group Sequential Methods with Applications to Clinical Trials*, Chapman & Hall/CRC, Boca Raton.

Jennison, C. and Turnbull, B.W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* **23**, 971–993.

Jennison, C. and Turnbull, B.W. (2005). Meta-analyses and adaptive group sequential designs in the clinical development process. *J. Biopharmaceutical Statistics* **15**, 537–558.

Jennison, C. and Turnbull, B.W. (2006). Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine* **25**, 917–932.

Jennison, C. and Turnbull, B.W. (2006). Adaptive and non-adaptive group sequential tests. *Biometrika* **93**, xx–xx.

(Preprints available at `http://www.bath.ac.uk/~mascj/` or `http://www.orie.cornell.edu`)

Lan, K.K.G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659–663.

Müller, H-H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential procedures. *Biometrics* **57**, 886–891.

Schmitz, N. (1993). *Optimal Sequentially Planned Decision Procedures.* Lecture Notes in Statistics, 79, Springer-Verlag: New York.