ADAPTIVITY IN CLINICAL TRIAL DESIGNS:

OLD AND NEW

Christopher Jennison,

Dept of Mathematical Sciences, University of Bath, UK

and

Bruce Turnbull,

Cornell University, Ithaca, New York

London

2 December 2004

0. A Cautionary Tale

Class action lawsuit *vs* Fred Hutchinson Cancer Center, Seattle.

Protocol 126: Bone marrow transplant for leukemia, 1981–1993

Civil trial: Feb-April 2004.

Plaintiffs: Estates of 5 deceased subjects.

Attorney: Alan Milstein — well-known scourge of clinical trialists including IRB and DSMB members (successful cases vs U. Penn, U. Oklahoma, . . .)

Charges: breach of the right to dignity, fraud, assault and battery, product liability, violations of laws governing research and consumer rights.

Protocol 126

Of concern to us: Statements made by plaintiff's expert statistical witness

1. Phase 1: Sample size was 12, but no rationale given in protocol or SAP. Actual sample size was 22 — no documentation for change, nor any approval by IRB.

   "Unexplained change was substandard, because *once you state the sample size you should abide by it.*"

2. Phase 3: No formal plan for interim monitoring or stopping rules in protocol. Only "vague" statement that study would be stopped if there was

   "cumulative evidence of toxicity or lack of efficacy"

3. There was enough statistical evidence at a meeting on Feb 8, 1984 to warrant stopping the trial, which would have saved the enrollment of 50 additional subjects.

4. Relapse-free survival curves to be used to calculate monetary damages.

Protocol 126

Note: In 1983, President of Fred Hutchinson Cancer Center turned down a request from IRB to establish an independent DSMB on the grounds of cost and that it would “reveal secrets to competitors”.

Now we have:

- FDA (1998) Guidance for Industry: E9 Statistical Principles for Clinical Trials
- FDA (2001) Draft Guidance for Clinical Trial Sponsors: On the Establishment and Operation of Clinical Trial Data Monitoring Committees.

Adaptivity

- Adaptive choice of test statistic as information on assumptions emerge; e.g. adaptive scores in a linear rank test, logrank vs Gehan
- Adaptive allocation to achieve balance within strata
- Adaptive allocation to assign fewer patients to inferior treatment arm
- Adaptivity to accruing information on nuisance parameters
- Adaptivity to accruing information on safety/secondary endpoints
- Adaptivity to adjust power based on accruing information on primary endpoints
- Adaptivity to to drop arms in multi-arm study based on accruing information on primary endpoints
- Others ....

Outline of Presentation

1. Interim monitoring of clinical trials

   Adapting to observed data

2. Distribution theory, the role of “information”

3. Error-spending tests

   Adapting to unpredictable information

   Adapting to nuisance parameters

4. Most efficient group sequential tests

   Adapting optimally to observed data

5. More recent adaptive proposals

6. Example of inefficiency in an adaptive designs

7. Conclusions

1. Interim monitoring of clinical trials

It is standard practice to monitor progress of clinical trials for reasons of *ethics*, *administration* (accrual, compliance) and *economics*.

Special methods are needed since multiple looks at accumulating data can lead to over-interpretation of interim results

Methods developed in manufacturing production were first transposed to clinical trials in the 1950s.

Traditional sequential methods assumed continuous monitoring of data, whereas it is only practical to analyse a clinical trial on a small number of occasions.

The major step forward was the advent of *Group Sequential* methods in the 1970s.

Pocock's repeated significance test (1977)

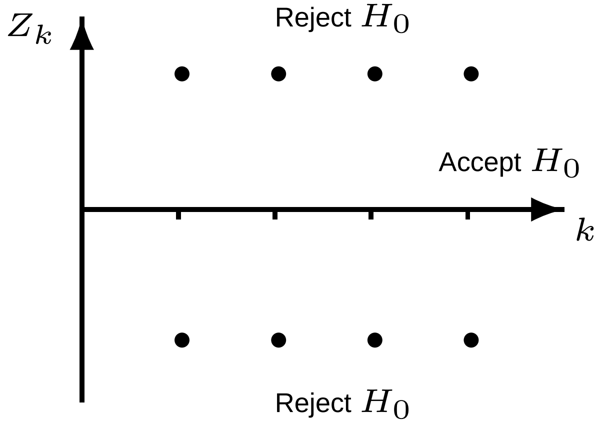
To test H_0: $\theta = 0$ vs $\theta \neq 0$, where θ represents the treatment difference.

Use standardised test statistics $Z_k,\ k = 1, \ldots, K$.

Stop to reject H_0 at analysis k if $|Z_k| > c$,

if H_0 has not been rejected by analysis K, stop and accept H_0.

Choose c to give overall type I error rate = α.



Types of hypothesis testing problems

Two-sided test:

testing H_0: $\theta = 0$ against $\theta \neq 0$.

One-sided test:

testing H_0: $\theta \leq 0$ against $\theta > 0$.

Equivalence tests:

one-sided — to show treatment A is as good as treatment B, within a margin δ (non-inferiority).

two-sided — to show two treatment formulations are equal within an accepted tolerance.

Types of early stopping

1. Stopping **to reject** H_0: *No treatment difference*

- Allows progress from a positive outcome
- Avoids exposing further patients to the inferior treatment
- Appropriate if no further checks are needed on treatment safety or long-term effects.

2. Stopping **to accept** H_0: *No treatment difference*

- Stopping “ for futility” or “abandoning a lost cause”
- Saves time and effort when a study is unlikely to lead to a positive conclusion.

One-sided tests

To look for superiority of a new treatment, test

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

If the new treatment if not effective, it is not appropriate to keep sampling to find out whether $\theta = 0$ or $\theta < 0$.

Specify type I error rate and power

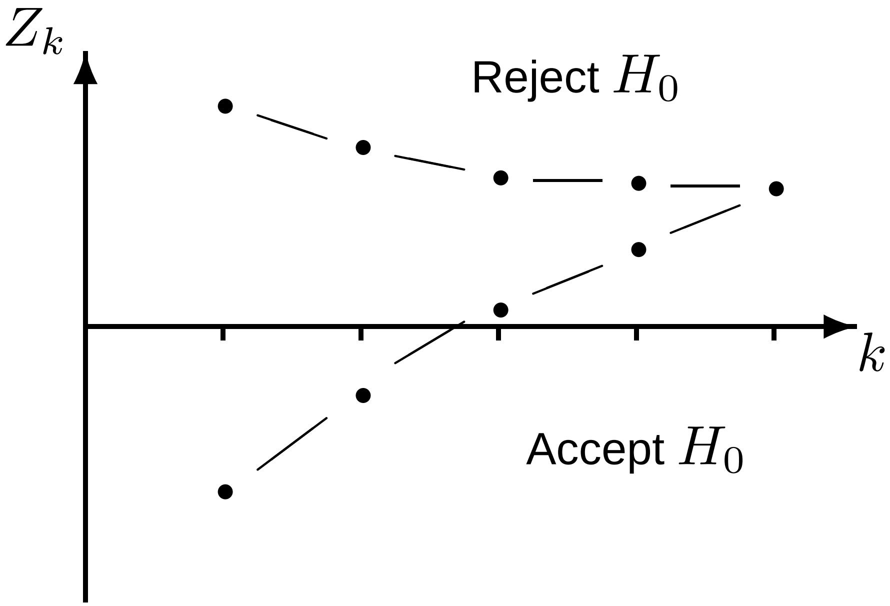
$$Pr\{\text{Reject } H_0 \,|\, \theta = 0\} = \alpha,$$

$$Pr\{\text{Reject } H_0 \,|\, \theta = \delta\} = 1 - \beta.$$

A sequential test can reduce expected sample size under $\theta = 0$, $\theta = \delta$, and at effect sizes in between.

One-sided tests

A typical boundary one-sided testing boundary:



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

— ***adapting to data***, stopping when a decision is possible.

2. Joint distribution of parameter estimates

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

The information for θ at analysis k is

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

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

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

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

Canonical joint distribution of z-statistics

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

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

For this,

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

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

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

Canonical joint distribution of score statistics

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

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

The score statistics possess the “independent increments” property,

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

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

Sequential distribution theory

The preceding results for the joint distribution of $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ can be demonstrated directly for:

θ a single normal mean,

$\theta = \mu_A - \mu_B,$ the effect size in a comparison of two normal means.

The results also apply when θ is a parameter in:

a general normal linear,

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

So, we have the theory to support general comparisons, including adjustment for covariates if required.

Survival data

The canonical joint distributions also arise for:

parameter estimates in Cox's proportional hazards regression model

a sequence of log-rank statistics (score statistics) for comparing two survival curves

— and to z-statistics formed from these.

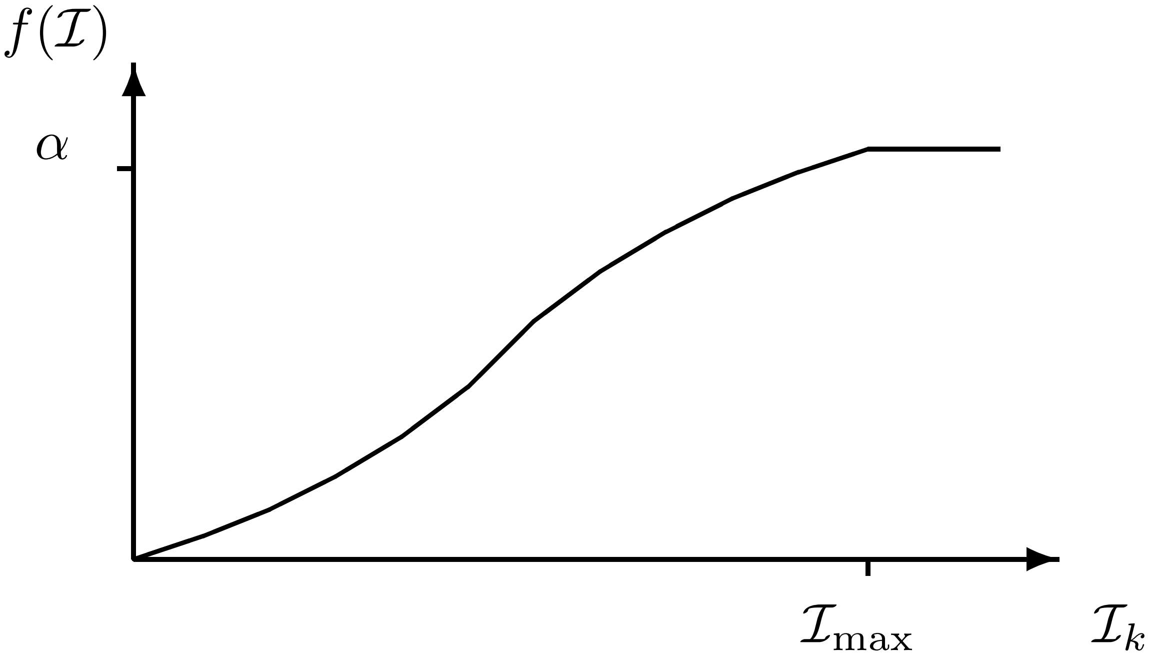
For survival data, observed information is roughly proportional to the number of failures seen.

Special types of group sequential test are needed to handle unpredictable and unevenly spaced information levels.

3. Error spending tests

Lan & DeMets (Biometrika, 1983) presented two-sided tests which “spend” type I error probability as a function of observed information.

The error spending function, $f(\mathcal{I})$, gives the type I error probability to be spent up to the current analysis



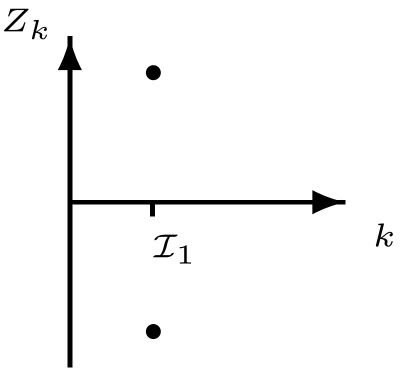
Maximum information design

- Specify the error spending function $f(\mathcal{I})$
- For each $k = 1$, 2, . . . , set the boundary at analysis k to give cumulative type I error probability $f(\mathcal{I}_k)$.
- Accept H_0 if $\mathcal{I}_{\max}$ is reached without rejecting H_0.

Precise rules are available to protect the type I error rate if the information sequence over-runs the target $\mathcal{I}_{\max}$, or if the study ends without reaching reaching this target. See slides 23 and 24 or Chapter 7 of Jennison & Turnbull (2000).

Implementing error spending tests

Analysis 1:

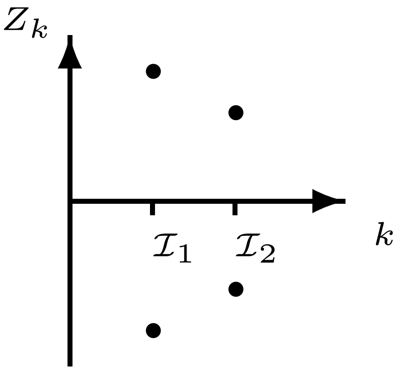


Observed information $\mathcal{I}_1$.

Reject H_0 if $|Z_1| > c_1$ where

$$Pr_{\theta=0}\{|Z_1| > c_1\} = f(\mathcal{I}_1).$$

Analysis 2:



Cumulative information $\mathcal{I}_2$.

Reject H_0 if $|Z_2| > c_2$ where

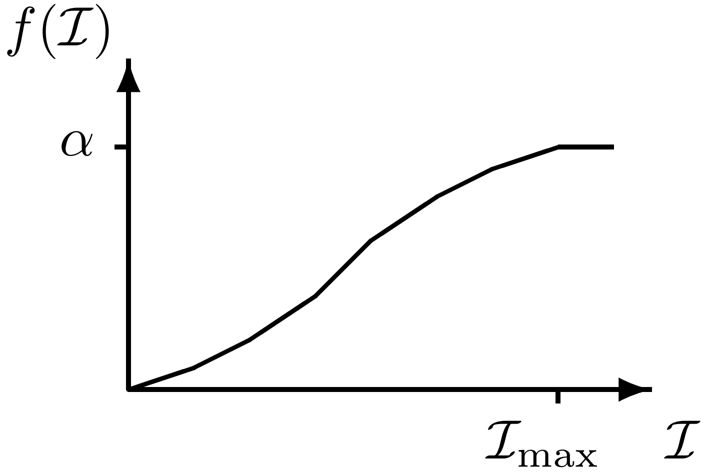
$$Pr_{\theta=0}\{|Z_1| < c_1, |Z_2| > c_2\} = f(\mathcal{I}_2) - f(\mathcal{I}_1).$$

etc.

Adapting to unpredictable information

One-sided error spending tests

Define $f(\mathcal{I})$ and $g(\mathcal{I})$ for spending type I and type II error probabilities.



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 $g(\mathcal{I}) \;\propto\; (\mathcal{I}/\mathcal{I}_{\max})^{\rho}$.

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

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

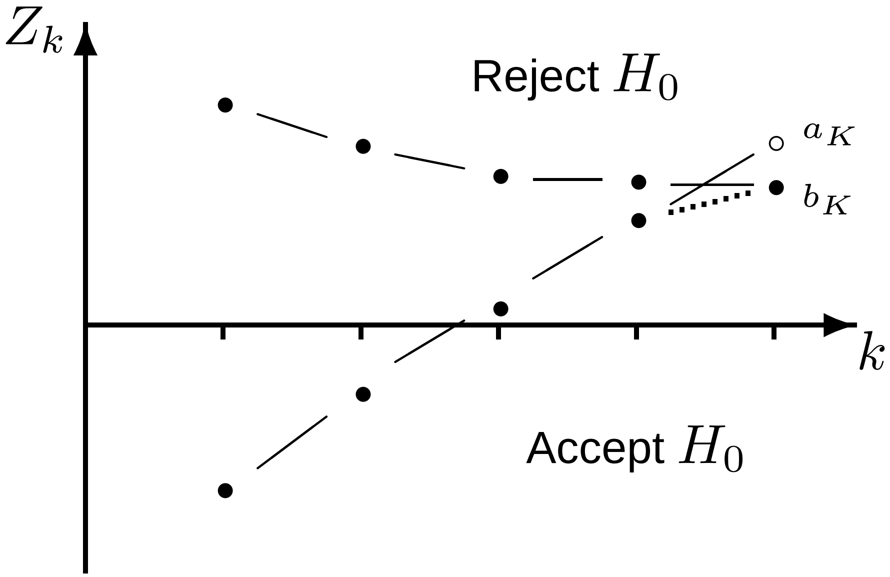
   For type I error rate α and power $1-\beta$ at $\theta = \delta$,

   $$\mathcal{I}_{\max} \;=\; R\,\frac{(z_\alpha + z_\beta)^2}{\delta^2},$$

   where R is the “inflation factor” for this design.

Over-running

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



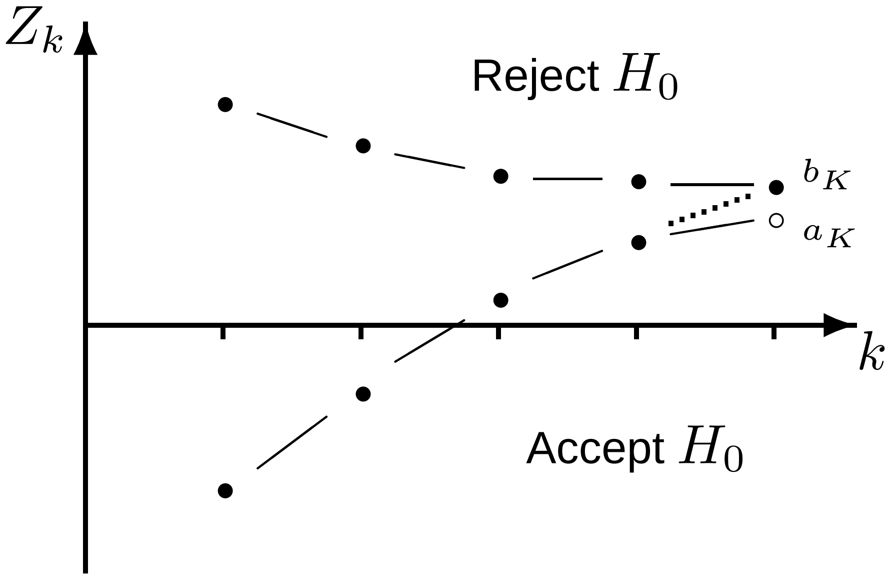
Keeping b_K as calculated guarantees type I error probability of exactly α.

So, reduce a_K to b_K — and gain extra power.

Over-running may also occur if $\mathcal{I}_K = \mathcal{I}_{\max}$ but the information levels deviate from the equally spaced values (say) used in choosing $\mathcal{I}_{\max}$.

Under-running

If a final information level $\mathcal{I}_K < \mathcal{I}_{\max}$ is imposed, 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 $1 - \beta$.

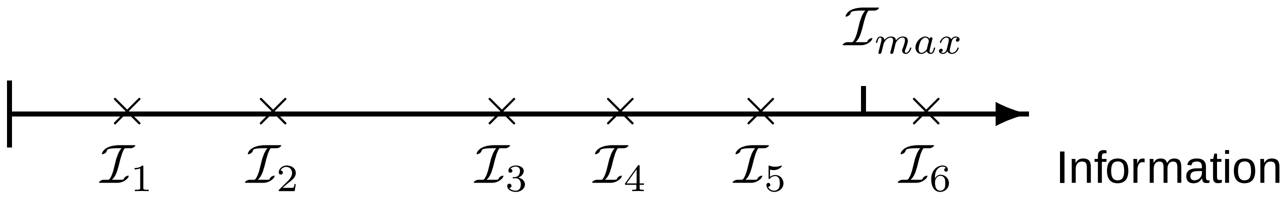
Error-spending designs and nuisance parameters

(1) Survival data, log-rank statistics

Information depends on the number of observed failures,

$$\mathcal{I}_k \approx \frac{1}{4}\{\text{Number of failures by analysis } k\}.$$

With fixed dates for analyses, continue until information reaches $\mathcal{I}_{\max}$.



If the overall failure rate is low or censoring is high, one may decide to extend the patient accrual period.

Changes affecting $\{\mathcal{I}_1, \mathcal{I}_2, \ldots\}$ can be based on observed information; they should **not** be influenced by the estimated treatment effect.

Error-spending designs and nuisance parameters

(2) Normal responses with unknown variance

In a two treatment comparison, a fixed sample test with type I error rate α and power $1-\beta$ at $\theta=\delta$ requires information

$$\mathcal{I}_f \;=\; \frac{(z_\alpha + z_\beta)^2}{\delta^2}.$$

A group sequential design with inflation factor R needs maximum information $\mathcal{I}_{\max} \;=\; R\mathcal{I}_f$.

The maximum required information is fixed — but the sample size needed to provide this level of information depends on the unknown variance σ^2.

Adapting to nuisance parameters

Adjusting sample size as variance is estimated

The information from n_A observations on treatment A and n_B on B is

$$\mathcal{I} \;=\; \left\{ \left(\frac{1}{n_A} + \frac{1}{n_B} \right) \sigma^2 \right\}^{-1}.$$

Initially: Set maximum sample sizes to give information $\mathcal{I}_{\max}$ if σ^2 is equal to an initial estimate, σ_0^2.

As updated estimates of σ^2 are obtained: Adjust future group sizes so the final analysis has

$$\left\{ \left(\frac{1}{n_A} + \frac{1}{n_B} \right) \hat{\sigma}^2 \right\}^{-1} \;=\; \mathcal{I}_{\max}.$$

NB, state $\mathcal{I}_{max}$ in the protocol, not initial targets for n_A and n_B.

At interim analyses, apply the error spending boundary based on observed (estimated) information.

4. Optimal group sequential tests

“Optimal” designs may be used directly — or they can serve as a benchmark for judging efficiency of designs with other desirable features.

Optimising a group sequential test:

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 θ s.

Derivation of optimal group sequential tests

Create a Bayes decision problem with a prior on θ, sampling costs and costs for a wrong decision. Write a program to solve this Bayes problem by backwards induction (dynamic programming).

Search for a set of costs such that the Bayes test has the desired frequentist properties: type I error rate α and power $1 - \beta$ at $\theta = \delta$.

This is essentially a Lagrangian method for solving a constrained optimisation problem — the key is that the unconstrained Bayes problem can be solved accurately and quickly.

Example of properties of optimal tests

One-sided tests, $\alpha = \beta = 0.05$, 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	80.9	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
20	67.6	60.5	57.4	54.6	53.3	52.0 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.

Assessing families of group sequential tests

One-sided tests:

Pampallona & Tsiatis

Parametric family indexed by Δ, boundaries for S_k involve $\mathcal{I}_k^{\Delta}$,
each Δ implies an “inflation factor” R such that $\mathcal{I}_{max} = R\mathcal{I}_{fix}$.

Error spending, ρ-family

Error spent is proportional to $\mathcal{I}_k^{\rho}$, ρ determines the inflation factor R.

Error spending, γ-family (Hwang et al, 1994)

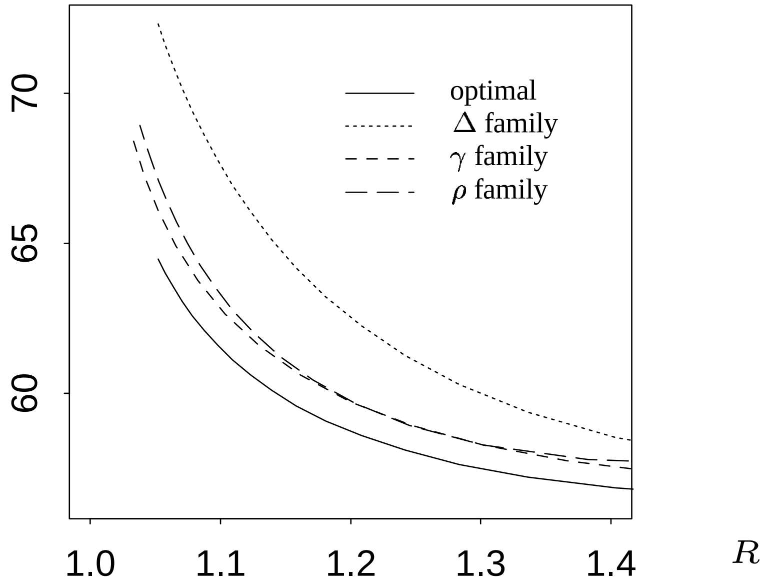
Error spent is proportional to

$$\frac{1 - e^{-\gamma \mathcal{I}_k / \mathcal{I}_{max}}}{1 - e^{-\gamma}}.$$

Families of tests

Tests with $K = 10$, $\alpha = 0.05$, $1 - \beta = 0.9$.

$\{E_0(\mathcal{I}) + E_\delta(\mathcal{I})\}/2$ as a percentage of $\mathcal{I}_{fix}$



Both error spending families are highly efficient but Pampallona & Tsiatis tests are sub-optimal.

Adapting optimally to observed data

Squeezing a little extra efficiency

Schmitz (1993) proposed group sequential tests in which group sizes are chosen adaptively:

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

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

and so forth.

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

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

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, optimal designs can be found by solving related Bayes decision problems.

Examples 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
8	56.6	58.0	58.3
10	55.9	57.2	57.5

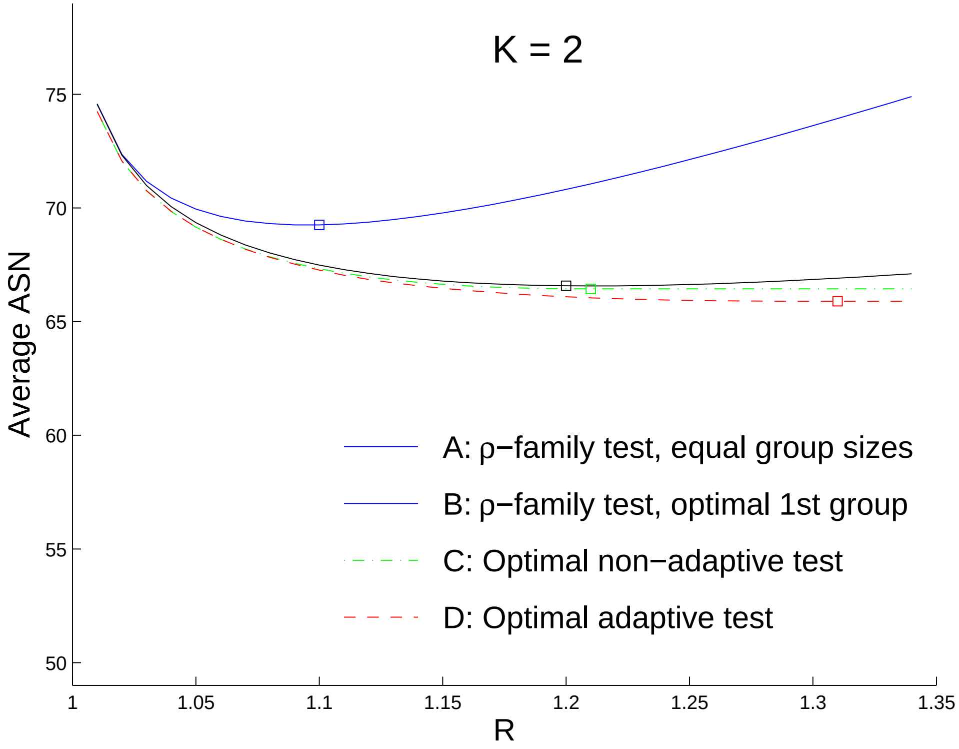
Varying group sizes *adaptively* makes for a complex procedure and the efficiency gains are slight.

Adapting super-optimally to observed data

Examples of “Schmitz” designs

Tests of H_0: $\theta = 0$ versus H_1: $\theta > 0$ with type I error rate $\alpha = 0.025$, power $1 - \beta = 0.8$ at $\theta = \delta$, and $K = 2$ analyses.

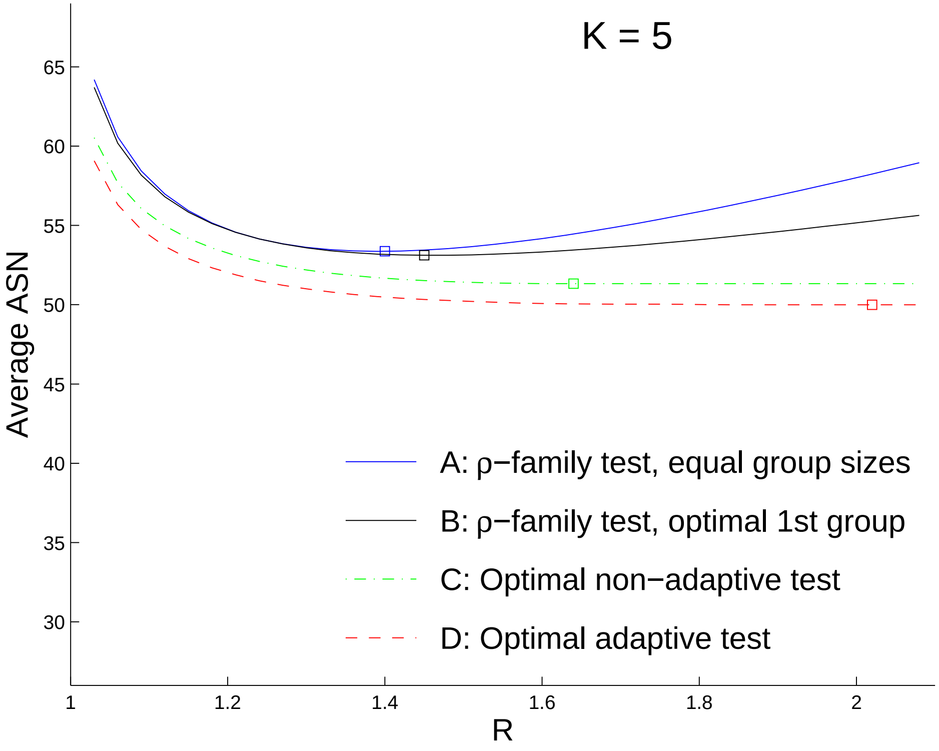
Designs minimise average ASN $\{E_{\theta=0}(\mathcal{I}) + E_{\theta=\delta}(\mathcal{I}) + E_{\theta=2\delta}(\mathcal{I})\}/3$.



Examples of “Schmitz” designs

Tests of H_0: $\theta = 0$ versus H_1: $\theta > 0$ with type I error rate $\alpha = 0.025$, power $1 - \beta = 0.8$ at $\theta = \delta$, and $K = 5$ analyses.

Designs minimise average ASN $\{E_{\theta=0}(\mathcal{I}) + E_{\theta=\delta}(\mathcal{I}) + E_{\theta=2\delta}(\mathcal{I})\}/3$.



5. Recent adaptive methods

“Adaptivity” $\neq$ “Flexibility”

- **Pre-planned extensions.** The way the design changes in response to interim data is pre-determined: Proschan and Hunsberger (1995), Li et al. (2002), Hartung and Knapp (2003) — very much like “Schmitz” designs.

- **Partially pre-planned.** The time of the first interim analysis is pre-specified, as is the method for combining results from different stages: Bauer (1989), Bauer & Köhne (1994).

- **Re-design may be unplanned.** The method of combining results from different stages is implicit in the original design and carried over into any re-design: Fisher (1998), Cui et al. (1999), Denne (2001) or Müller & Schäfer (2001).

Bauer (1989) and Bauer & Köhne (1994) . . .

. . . proposed mid-course design changes to one or more of

Treatment definition

Choice of primary response variable

Sample size:

— in order to maintain power under an estimated nuisance parameter

— to change power in response to external information

— to change power for internal reasons

a) secondary endpoint, e.g., safety

b) primary endpoint, i.e., $\hat{\theta}$.

Bauer & Köhne's two-stage scheme

Investigators decide *at the design stage* to split the trial into two parts.

Each part yields a one-sided P-value and these are combined.

- Run part 1 as planned. This gives

$$P_1 \sim U(0,\, 1) \quad \text{under } H_0.$$

- Make design changes.

- Run part 2 with these changes, giving

$$P_2 \sim U(0,\, 1) \quad \text{under } H_0,$$

  conditionally on P_1 and other part 1 information.

- Combine P_1 and P_2 by Fisher's combination test:

$$-\log(P_1\, P_2) \sim \frac{1}{2}\,\chi^2_4 \quad \text{under } H_0.$$

B & K: Major design changes before part 2

With major changes, the two parts are rather like separate studies in a drug development process, such as:

Phase IIb
Compare several doses and select the best.

Use a rapidly available endpoint (e.g., tumour response).

Phase III
Compare selected dose against control.

Use a long-term endpoint (e.g., survival).

Applying Fisher's combination test for P_1 and P_2 gives a meta-analysis of the two stages with a pre-specified rule.

Note: Each stage has its own null hypothesis and the overall H_0 is **the intersection** of these.

B & K: Minor design changes before stage 2

With only minor changes, the form of responses in stage 2 stays close to the original plan.

Bauer & Köhne’s method provides a way to handle this.

Or, an error spending test could be used:

Slight departures from the original design will perturb the observed information levels, which can be handled in an error spending design.

After a change of treatment definition, one can stratify with respect to patients admitted before and after the change. As long as the overall score statistic can be embedded in a Brownian motion, one can use an error spending test with a maximum information design.

B & K: Nuisance parameters

Example. Normal response with unknown variance, σ^2.

Aiming for type I error rate α and power $1-\beta$ at $\theta = \delta$, the necessary sample size depends on σ^2.

One can choose the second stage's sample size to meet this power requirement assuming variance is equal to s_1^2, the estimate from stage 1.

P_1 and P_2 from t-tests are independent $U(0,\ 1)$ under H_0 — exactly.

Other methods:

(a) Many "internal pilot" designs are available.

(b) Error spending designs can use estimated information (from s^2).

(c) The two-stage design of Stein (1945) attains both type I error and power precisely!

External factors or internal, secondary information

At an interim stage suppose, for reasons not concerning the primary endpoint, investigators wish to achieve power $1-\beta$ at $\theta = \tilde{\delta}$ rather than $\theta = \delta$ ($\tilde{\delta} < \delta$).

If this happens after part 1 of a B & K design, the part 2 sample size can be increased, e.g., to give conditional power $1-\beta$ at $\theta = \tilde{\delta}$.

Unplanned re-design

Recent work shows the same can be done within a fixed sample or group sequential design by

preserving conditional type I error rate under $\theta = 0$,

ensuring conditional power $1-\beta$ at $\theta = \tilde{\delta}$

— see Denne (2001) or Müller & Schäfer (2001).

Responding to $\widehat{\theta}$, an estimate of the primary endpoint

Motivation may be:

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

Many methods have been proposed to do this.

If re-design is unplanned, the conditional type I error rate approach is available.

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!

The variance spending method

L. Fisher (1998), Cui et al. (1999), Denne (2001), ...

As before, consider study with two parts.

B & K used R.A. Fisher's inverse χ^2 method to combine $P_1,\ P_2$.

Instead we use the weighted inverse normal method (Mosteller and Bush 1954):

Define $Z_1 = \Phi^{-1}(1 - P_1) \quad Z_2 = \Phi^{-1}(1 - P_2)$

Note $Z_2 \sim N(0, 1)$ under H_0 conditionally on P_1, Z_1 and other part 1 information.

Suppose $w_1,\ w_2$ are fixed weights with $w_1^2 + w_2^2 = 1$.

Then $Z = w_1 Z_1 + w_2 Z_2\ \sim\ N(0, 1)$ and test that rejects H_0 when $Z > z(\alpha)$ has level α.

Variance spending method — normal observations

Suppose we observe $X_1, X_2, \ldots$ i.i.d. $N(\theta, \sigma^2)$ in two consecutive stages:

Stage 1: $X_1, \ldots X_{m_1}$

Stage 2: $X_{m_1+1}, \ldots X_n, \quad n = m_1 + m_2$

If $H_0 : \theta = 0$, we have

$$Z_1 = \frac{X_1 + \ldots + X_{m_1}}{\sigma\sqrt{m_1}} \qquad Z_2 = \frac{X_{m_1+1}, + \ldots + X_n}{\sigma\sqrt{m_2}}$$

Choose weights $w_1 = \sqrt{m_1/n},\ w_2 = \sqrt{m_2/n}$

If no redesign

$$Z = w_1 Z_1 + w_2 Z_2 = \frac{\sum_{i=1}^n X_i}{\sigma\sqrt{n}}$$

Note: Usual efficient test statistic (unlike B&K).

Extending the study — downweighting

Suppose instead we decide to change m_2 to γm_2 at interim look.

This decision can be based on Stage 1 data — i.e. $\gamma = \gamma(Z_1)$

Now Z- statistic ($w_1 Z_1 + w_2 Z_2$) becomes

$$
\begin{aligned}
Z &= \sqrt{\frac{m_1}{n}}\,\frac{\sum_{i=1}^{m_1} X_i}{\sigma\sqrt{m_1}} + \sqrt{\frac{m_2}{n}}\,\frac{\sum_{i=m_1+1}^{m_1+\gamma m_2} X_i}{\sigma\sqrt{\gamma m_2}} \\
&= \frac{1}{\sigma\sqrt{n}}\left(\sum_{i=1}^{m_1} X_i + \gamma^{-\frac{1}{2}} \sum_{i=m_1+1}^{m_1+\gamma m_2} X_i\right)
\end{aligned}
$$

The test $Z > z(\alpha)$ retains type 1 error α.

But note that second stage observations are *down-weighted* (if $\gamma > 1$).

6. Example of inefficiency in an adaptive design

Example. A Cui, Hung & Wang (1999) style example.

Scenario.

We wish to design a test with type I error probability $\alpha = 0.025$.

Investigators are optimistic the effect, θ, could be as high as $\delta^* = 20$. However, effect sizes as low as about $\theta \geq \delta^{**} = 15$ are clinically relevant and worth detecting (cf the example cited by Cui et al).

First, consider a fixed sample study attaining power 0.9 at $\theta = \delta^* = 20$. We 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 the data are examined after the first 50 responses to see if there is a need to “adapt”.

Cui et al. adaptive design

Denote the estimated effect based on the first 50 observations by $\widehat{\theta}_1$.

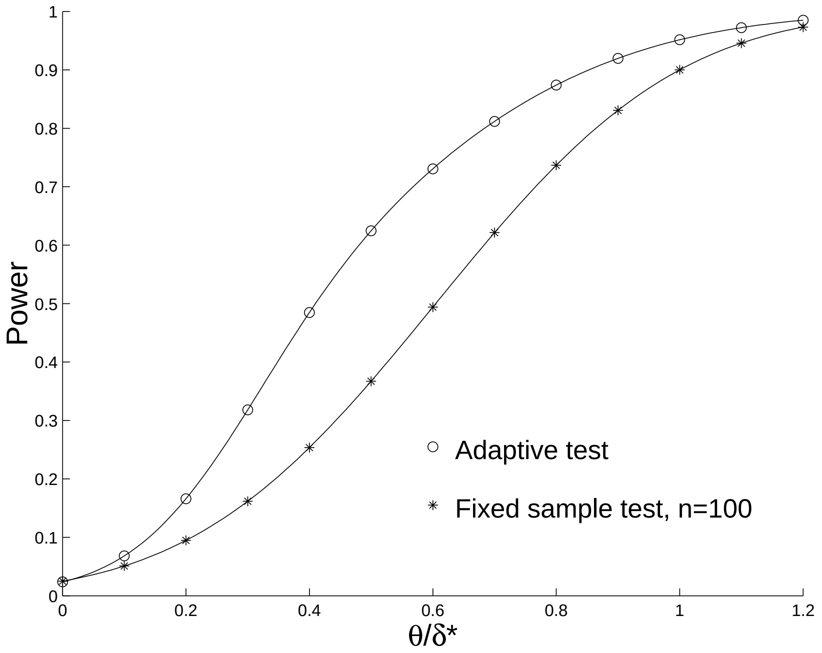
If $\widehat{\theta}_1 < 0.2\,\delta^* = 4$, stop the trial for futility, accepting H_0.

Otherwise, re-design the remainder of the trial, preserving the conditional type I error rate given $\widehat{\theta}_1$ — thereby maintaining overall type I error rate α.

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

Then, truncate this additional sample size to the interval (50, 500), so no decrease in sample size is allowed and we keep the total sample size to 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 the first stage, total sample size ranges from 100 to 550.

A conventional group sequential test

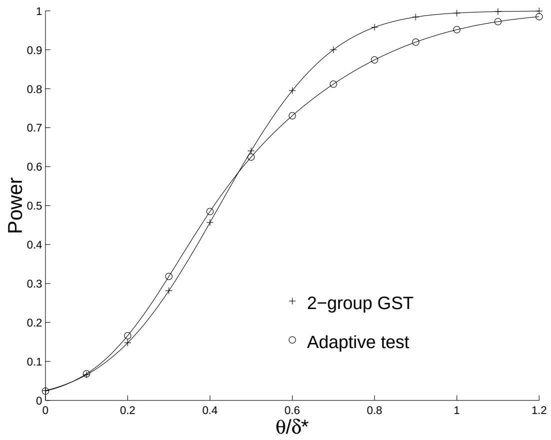
Similar overall power can be obtained by a non-adaptive GST designed to attain power 0.9 when $\theta = 14$.

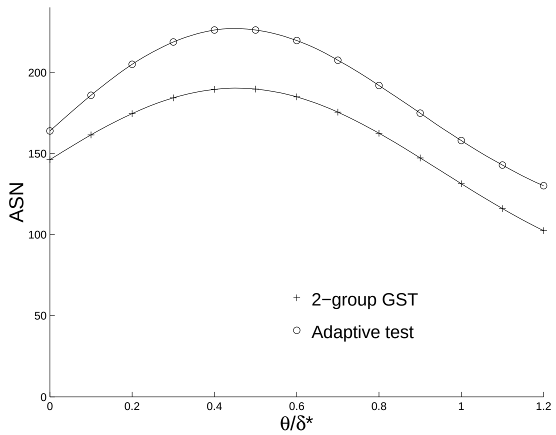
We have compared a power family, error spending test with $\rho = 1$:

> type I error rate is $\alpha = 0.025$,
>
> taking the first analysis after 68 observations and the second analysis after 225 gives a test meeting the requirement of 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 advantages of the conventional GST are clear. It has higher power, a lower average sample size function, and a much smaller maximum sample size.

We have found similar inefficiency in many more of the adaptive designs proposed in the literature.

Conditional power and overall power

It might be argued that only *conditional* power is important once a study is underway so overall power is irrelevant once data have been observed.

However:

Overall power integrates over conditional properties in just the right way.

It is overall power that is available at the design stage, when a stopping rule and sampling rule (even an adaptive one) are chosen.

As the example shows, “chasing conditional power” can be a trap leading to very large sample sizes when the estimated effect size is low — and, given the variability of this estimate, the true effect size could well be zero.

To a pharmaceutical company conducting many trials, long term performance is determined by overall properties, i.e., the power and average sample size of each study.

7. Conclusions

Error Spending tests using Information Monitoring can adapt to

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

Methods preserving conditional type I error allow re-design in response to external developments or internal evidence from secondary endpoints.

Recently proposed adaptive methods can

facilitate re-sizing for nuisance parameters,

support re-sizing to rescue an under-powered study,

allow an on-going approach to study design.

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

References

Bauer, P. (1989). Multistage testing with adaptive designs (with discussion). *Biometrie und Informatik in Medizin und Biologie* **20**, 130–148.

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

Chi, G.Y.H. and Liu, Q. (1999). The attractiveness of the concept of a prospectively designed two-stage clinical trial. *J. Biopharmaceutical Statistics* **9**, 537–547.

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.

ICH Topic E9. Note for Guidance on Statistical Principles for Clinical Trials. ICH Technical Coordination, EMEA: London, 1998.

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

Fisher, R.A. (1932). *Statistical Methods for Research Workers, 4th Ed.*, Oliver and Boyd, London.

Hartung, J. and Knapp, G. (2003). A new class of completely self-designing clinical trials. *Biometrical Journal* **45**, 3-19.

Hwang, I.K., Shih, W.J. and DeCani, J.S. (1990). Group sequential designs using a family of type I error probability spending functions. *Statist. Med.*, **9**, 1439–1445.

Jennison, C. and Turnbull, B.W. (1993). Group sequential tests for bivariate response: Interim analyses of clinical trials with both efficacy and safety endpoints. *Biometrics* **49**, 741–752.

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. (2004a,b,c). Preprints available at `http://www.bath.ac.uk/~mascj/` or at `http://www.orie.cornell.edu` (Click on “Technical Reports”, “Faculty Authors”, “Turnbull” and finally on “Submit”).

a. Efficient group sequential designs when there are several effect sizes under consideration.

b. Adaptive and non-adaptive group sequential tests.

c. Meta-analyses and adaptive group sequential designs in the clinical development process.

Li, G., Shih, W. J., Xie, T. and Lu. J. (2002). A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics* **3**, 277–287.

Mehta, C. R. and Tsiatis, A. A. (2001). Flexible sample size considerations using information-based interim monitoring. *Drug Information J.* **35**, 1095–1112.

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.

Pampallona, S. and Tsiatis, A.A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *J. Statist. Planning and Inference*, **42**, 19–35.

Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, **64**, 191–199.

Proschan, M.A. and Hunsberger, S.A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315–1324.

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