Adaptive Designs:

Why, How and When?

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

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

ISBS Conference

Shanghai, July 2008

Adaptive designs: Why?

Ordinarily, in a clinical trial one specifies at the outset:

Patient population,

Treatment,

Primary endpoint,

Hypothesis to be tested,

Power at a specific effect size.

Adaptive designs allow these elements to be reviewed during the trial.

Because . . . there may be limited information to guide these choices initially, but more knowledge will accrue as the study progresses.

Adaptive designs: How?

Two basic ingredients recur in many adaptive methods:

(i) Combination tests

The trial is conducted in two stages: the design of the second stage can depend on responses in the first stage.

Data from each stage data are summarised through a P-value or Z-statistic.

A *combination test* produces an overall decision from the two summary statistics with correct type I error probability.

(ii) Multiple testing procedures

Following design changes, multiple null hypotheses may need to be considered.

In order to avoid inflating error rates, testing procedures must control overall, "family-wise" error probabilities.

Adaptive designs: When?

We shall consider adaptive methods that allow the following modifications.

1. Changing ***sample size*** in response to estimates of a nuisance parameter
2. Changing ***sample size*** in response to interim estimates of the effect size
3. Switching to a ***patient sub-population***
4. Changing the ***null hypothesis*** (superiority/non-inferiority)
5. ***Treatment selection*** (combined Phase II/III trials)

For each, we shall outline an adaptive method and discuss when it may offer useful benefits.

How — Combination tests: The inverse χ^2 test

Reference: Bauer & Köhne, *Biometrics*, 1994.

Initial design

Stipulate that Bauer & Köhne's combination test will be used.

Define the null hypothesis H_0 (with a one-sided alternative).

Design Stage 1, fixing sample size and test statistic for this stage.

Stage 1

Observe the P-value, P_1.

Under H_0, $P_1 \sim U(0, 1)$.

Design Stage 2 in the light of Stage 1 data.

Stage 2

Observe the P-value, P_2.

Under H_0, $P_2 \sim U(0, 1)$ and P_2 is independent of P_1.

Bauer & Köhne's inverse χ^2 test

Overall test

If $P \sim U(0,\, 1)$, then

$$-\ln(P) \;\sim\; \mathsf{Exp}\,(1) = \frac{1}{2}\,\chi^2_2.$$

Thus, under H_0,

$$-\ln(P_1\,P_2) \;\sim\; \frac{1}{2}\,\chi^2_4$$

and we combine the two P-values, rejecting H_0 if

$$-\ln(P_1\,P_2) \;>\; \frac{1}{2}\,\chi^2_{4,\,1-\alpha}.$$

This χ^2 test was originally proposed for combining results of several studies by R. A. Fisher (1932) *Statistical Methods for Research Workers*.

How — Combination tests: The inverse normal test

Stipulate the ***inverse normal test*** with weights w_1 and w_2, where $w_1^2 + w_2^2 = 1$.

Stage 1

Compute $Z_1 = \Phi^{-1}(P_1)$.

Under H_0, $Z_1 \sim N(0, 1)$.

Design Stage 2 in the light of Stage 1 data.

Stage 2

Compute $Z_2 = \Phi^{-1}(P_2)$.

Under H_0, $Z_2 \sim N(0, 1)$ and Z_2 is independent of Z_1.

Overall test

Under H_0, $Z = w_1 Z_1 + w_2 Z_2 \sim N(0, 1)$.

Reject H_0 if $Z > \Phi^{-1}(1 - \alpha)$.

1. Sample size re-estimation for nuisance parameters

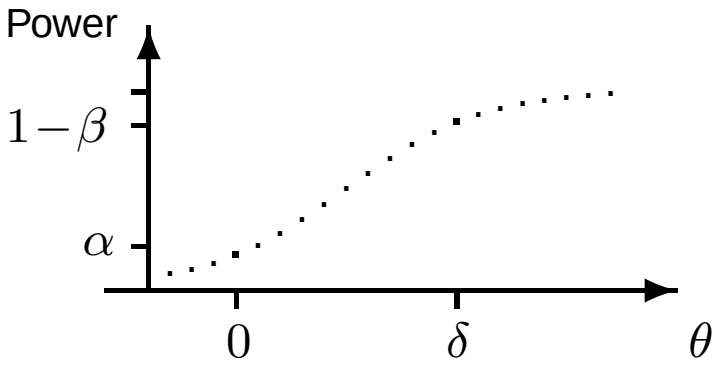
Type I error and power

Suppose θ represents the effect of a new treatment vs control.

A study is to test H_0: $\theta \leq 0$ against $\theta > 0$ with

one-sided type I error probability $\alpha = 0.025$, say.

The choice of sample size determines the power curve.



Power may also depend on a “nuisance parameter” such as a normal variance or overall failure rate for a survival response.

Sample size re-estimation for nuisance parameters

In a two-treatment comparison with normal response, power $1-\beta$ at effect size $\theta = \delta$ requires sample size per treatment of

$$n \;=\; (z_\alpha + z_\beta)^2\, 2\,\sigma^2/\delta^2, \qquad (1)$$

where z_p denotes $\Phi^{-1}(1-p)$.

Initial design

A Bauer & Köhne two-stage design is specified.

Sample size n_0 is determined using a preliminary estimate σ_0^2 in (1).

Stage 1 is planned with a sample size of $n_1 = n_0/2$.

Stage 1

Yields estimates $\hat{\theta}_1$ and $\hat{\sigma}_1^2$.

The t-statistic t_1 for testing H_0: $\theta \leq 0$ vs $\theta > 0$ is converted to a P-value, P_1.

Sample size re-estimation for nuisance parameters

Stage 1 . . .

Now use the variance estimate $\hat{\sigma}_1^2$ to re-calculate sample size.

One may simply substitute this value in (1).

Or, also take account of the interim estimate of treatment effect, $\hat{\theta}_1$.

This defines an additional sample size of n_2 in Stage 2.

Stage 2

Calculate the t-statistic t_2 for testing H_0 based on Stage 2 data alone and convert to a P-value, P_2.

The overall test — which has type I error rate exactly α — rejects H_0 if

$$-\ln(P_1\, P_2) \;>\; \frac{1}{2}\,\chi^2_{4,\,1-\alpha}.$$

Problem 1: Adaptive designs provide a successful solution.

Sample size re-estimation for nuisance parameters

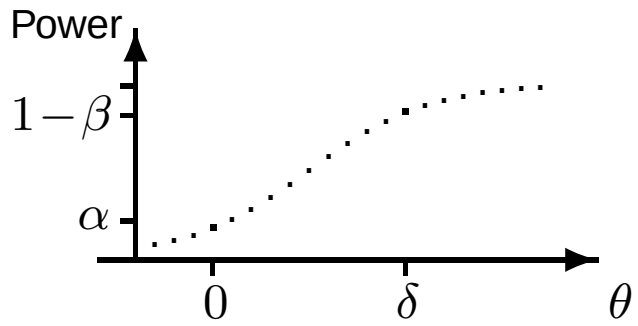
We have described a method of sample size re-estimation based on Bauer & Köhne's ***combination test***.

Other approaches are also available: designing in terms of "required information", then modifying sample size during the trial to reach the target information level.

Internal pilots: Wittes & Brittain (*Statistics in Medicine*, 1990)

Information monitoring: Mehta & Tsiatis (*Drug Information J.*, 2001)

These modifications are intended to achieve the power curve originally specified.



What if requirements for the power curve change during a trial?

2. Sample size modification to increase power

Investigators may start out optimistically and design a trial with power to detect a large treatment effect. Interim data may then suggest a smaller effect size — still clinically important but difficult to demonstrate with the chosen sample size.

- An adaptive design can allow sample size to be increased during the trial, ***rescuing*** an under-powered study.
- Some would advocate this approach as a way to "let the data say" what power and sample size should be chosen.
- Or, a ***group sequential design*** can achieve a desired power curve and save sample size through early stopping when the effect size is large.

Questions:

Is there a down-side to the "wait and see" approach?

How are the adaptive and group sequential approaches related?

Sample size modification to increase power

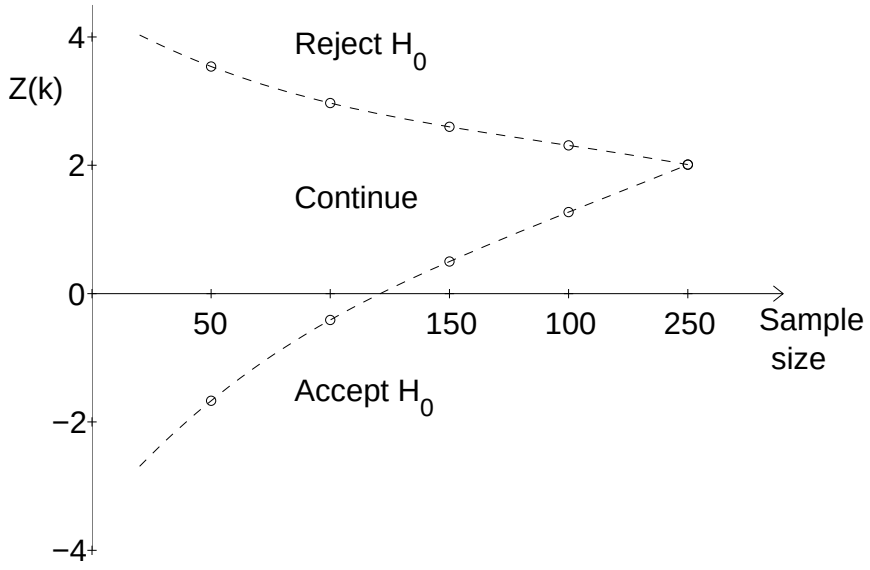
Example *(Jennison & Turnbull, Biometrika, 2006, Ex. 2)*

We start with a group sequential design with 5 analyses,

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

one-sided type I error probability $\alpha = 0.025$ and

Initial design: power $1 - \beta = 0.9$ at $\theta = \delta$.



Sample size modification to increase power

Suppose, at analysis 2, a low interim estimate $\hat{\theta}_2$ prompts investigators to consider the trial's power at effect sizes below δ, where power 0.9 was originally set:

Lower effect sizes start to appear plausible,

Conditional power under these effect sizes, using the current design, is low.

Applying the method of Cui, Hung and Wang (*Biometrics*, 1999)

Sample sizes for groups 3 to 5 are multiplied by a factor γ.

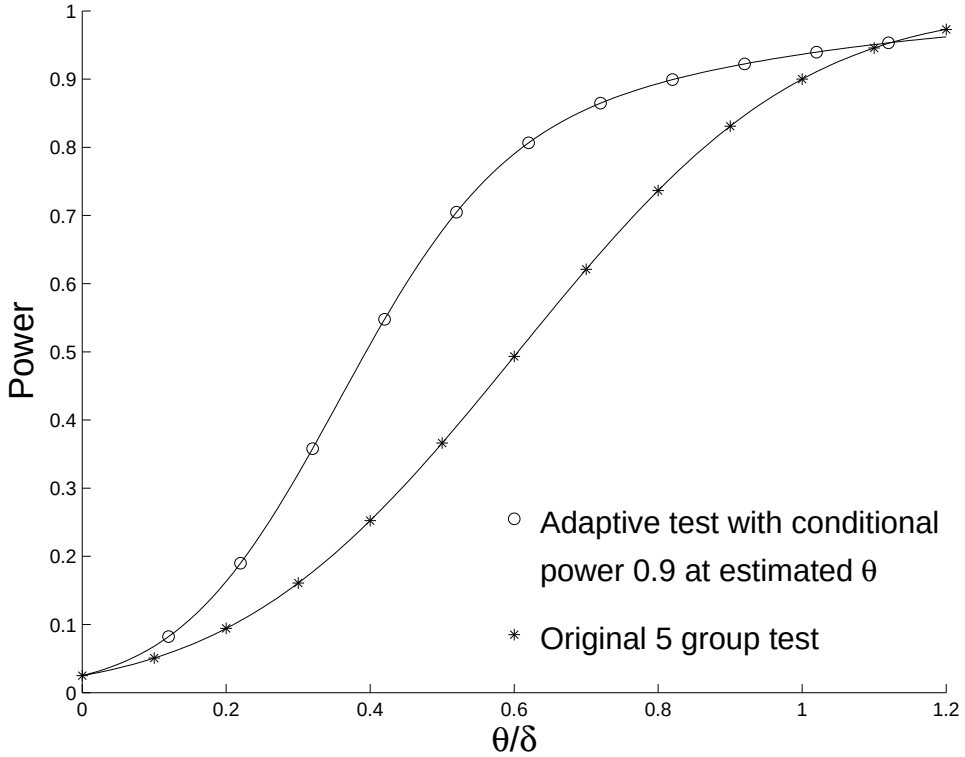
Sample sums from these groups are down-weighted by $\gamma^{-1/2}$. This gives the same variance as before but the mean is multiplied by $\gamma^{1/2}$.

Using the new weighted sample sum in place of the original sample sum maintains the type I error rate and increases power.

We choose the factor γ to give conditional power 0.9 **if $\boldsymbol{\theta}$ is equal to $\hat{\boldsymbol{\theta}}_{\mathbf{2}}$**, with the constraint $\gamma \leq 6$ so sample size can be at most 4 times the original maximum .

Sample size modification to increase power

Re-design has raised the power curve at all effect sizes.



Overall power at $\theta = \delta/2$ has increased from 0.37 to 0.68.

Sample size modification to increase power

Reasons for re-design arose purely from observing $\hat{\theta}_2$. A group sequential design responds to such interim estimates — in the decision to stop the trial or to continue.

Investigators could have considered at the design stage how they would respond to low interim estimates of effect size.

If they had thought this through and chosen the above adaptive procedure, they could also have examined its overall power curve.

Assuming this power curve were acceptable, how else might it have been achieved?

An alternative group sequential design

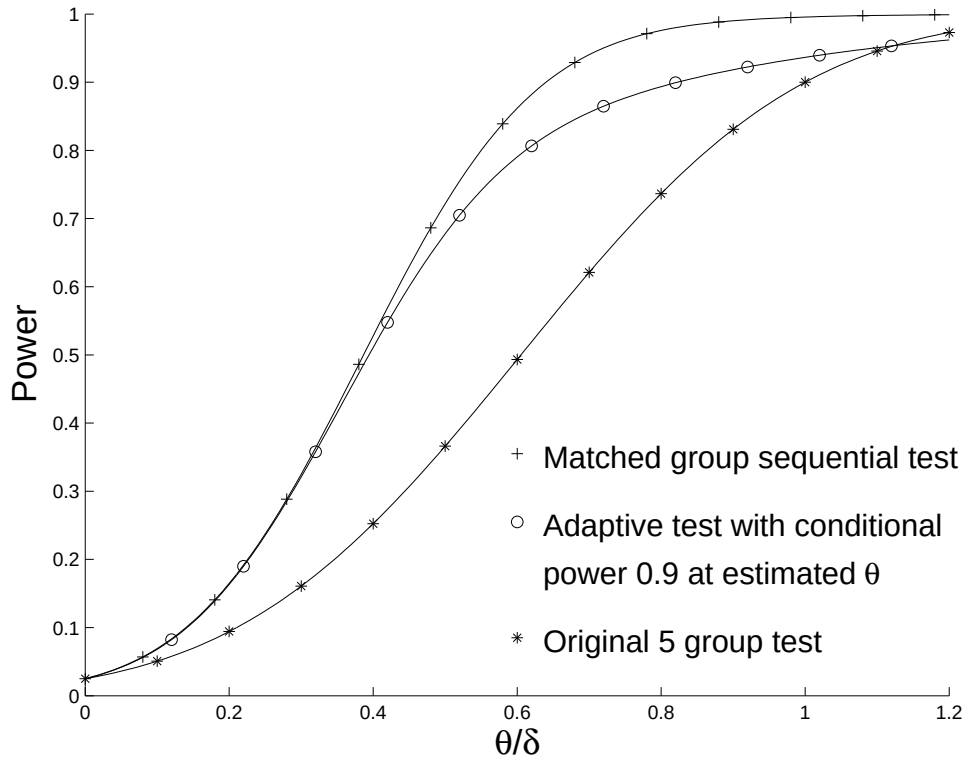
Five-group designs matching key features of the adaptive test can be found.

To be comparable, power curve should be as high as that of the adaptive design.

Can expected sample size be lower too?

Sample size modification to increase power

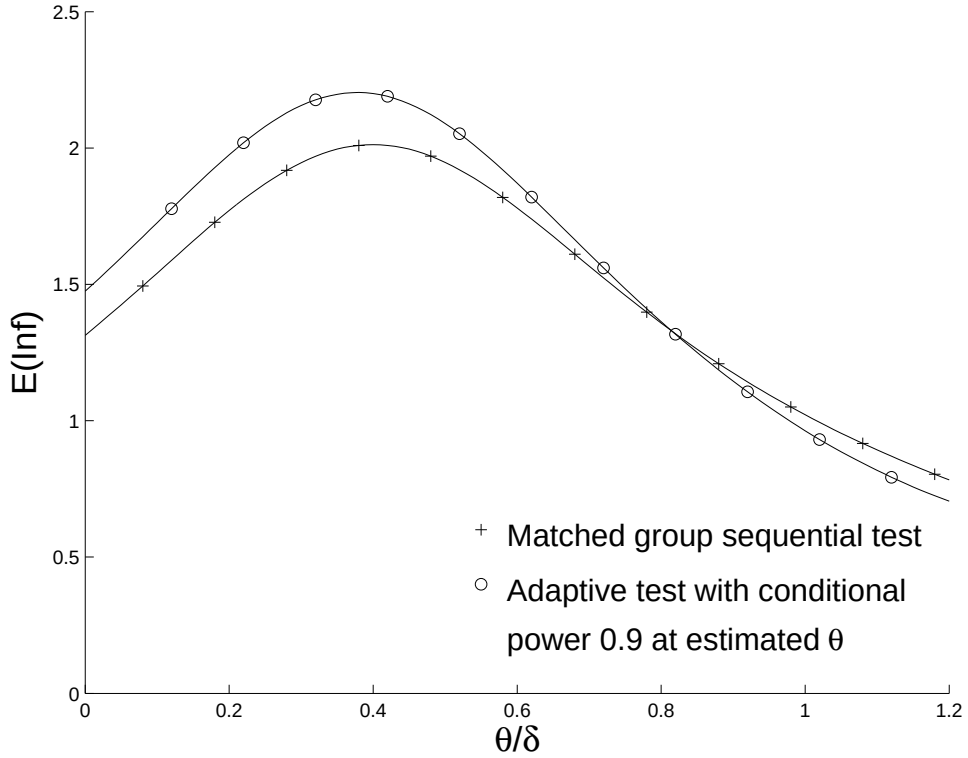
Power of our “matched” group sequential design is as high as that of the adaptive design at all effect sizes — and substantially higher at the largest θ values.



Sample size modification to increase power

The group sequential design has significantly lower expected information than the adaptive design over a range of effect sizes.

The group sequential design has slightly higher expected information for $\theta > 0.8\,\delta$, but this is where its power advantage is greatest.

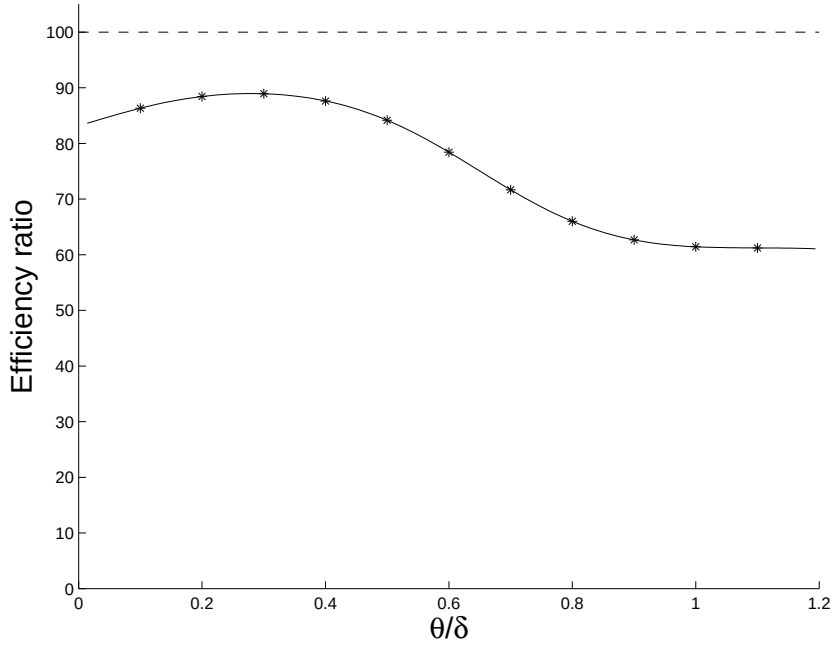


Sample size modification to increase power

Jennison & Turnbull (*Biometrika*, 2006) define an “Efficiency Ratio” to compare expected sample size, adjusting for differences in attained power.

By this measure, the adaptive design is up to 39% less efficient than the non-adaptive, group sequential alternative.

Efficiency ratio of adaptive design vs group sequential test



Sample size modification to increase power

We have found similar inefficiency relative to group sequential tests in a wide variety of proposed adaptive designs.

In general, adaptive designs have the advantage of extra freedom to choose group sizes in a response-dependent manner.

Jennison & Turnbull (*Biometrika*, 2006) show this adaptation can lead to gains in efficiency over non-adaptive group sequential tests — but the gains are very slight.

Sample size rules based on conditional power are far from optimal, hence the poor properties of adaptive designs using such rules.

Problem 2: Specify power properly at the outset: then, group sequential designs offer a simple and efficient option.

How — Multiple testing procedures

Adaptation to new treatments, new endpoints, etc., leads to consideration of multiple null hypotheses.

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

A procedure's ***family-wise 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 family-wise 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

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

With such strong control, 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)

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

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

We construct a level α test of each intersection hypothesis H_I: this test 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 family-wise error rate

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

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

3. Switching to a patient sub-population

A trial protocol defines a specific target population.

Suppose it is believed the treatment may be effective in a certain sub-population, even if it is ineffective in the rest of the population.

Enrichment: Restricting recruitment to a sub-population

At an interim analysis, the options are:

Continue as originally planned, or

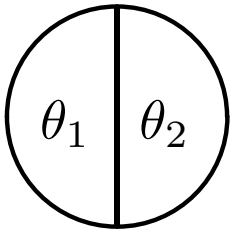
Restrict the remainder of the study to a sub-population.

This choice will affect the licence a positive outcome can support.

The possibility of testing several null hypotheses means a multiple testing procedure should be used.

Enrichment: Example

Sub-population 1 (proportion λ_1)



Sub-population 2 (proportion λ_2)

Overall treatment effect is $\theta = \lambda_1\theta_1 + \lambda_2\theta_2$.

We may wish to test:

- The null hypothesis for the full population, H_0: $\theta \leq 0$ vs $\theta > 0$,
- The null hypothesis for sub-population 1, H_1: $\theta_1 \leq 0$ vs $\theta_1 > 0$,
- The null hypothesis for sub-population 2, H_2: $\theta_2 \leq 0$ vs $\theta_2 > 0$.

Enrichment: Example

First, we look at a design that only considers the whole population effect, θ.

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

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

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

	Stage 1	*Stage 2*	*Overall*
H_0	$Z_{1,0}$	$Z_{2,0}$	$Z_0 = \frac{1}{\sqrt{2}} Z_{1,0} + \frac{1}{\sqrt{2}} Z_{2,0}$

Decision rules:

If $Z_{1,0} < 0$ — Stop at Stage 1, Accept H_0

If $Z_{1,0} \geq 0$ — Continue to Stage 2, then

- If $Z_0 < 1.95$ — Accept H_0
- If $Z_0 \geq 1.95$ — Reject H_0

Enrichment: Example

Assume equal sub-population proportions, so $\lambda_1 = \lambda_2 = 0.5$.

Properties of design for the whole population effect, θ:

θ_1	θ_2	θ	*Power for* H_0
20	20	20	0.90
10	10	10	0.37
20	0	10	0.37
0	20	10	0.37

Is it feasible to identify at Stage 1 that θ is low, but it would be worthwhile to switch resources to test a sub-population?

Enrichment: A closed testing procedure

We wish to be able to consider three null hypotheses:

		On rejection, conclude:
H_0:	$\theta \leq 0$	Treatment is effective in the whole population
H_1:	$\theta_1 \leq 0$	Treatment is effective in sub-population 1
H_2:	$\theta_2 \leq 0$	Treatment is effective in sub-population 2

To apply a *closed testing procedure*, we need tests of intersection hypotheses:

H_{01}:	$\theta \leq 0$ and $\theta_1 \leq 0$
H_{02}:	$\theta \leq 0$ and $\theta_2 \leq 0$
$H_{12} = H_{012}$:	$\theta_1 \leq 0$ and $\theta_2 \leq 0$

(Since $\theta = \lambda_1\theta_1 + \lambda_2\theta_2$, it is clear that H_{012} is identical to H_{12}.)

Enrichment: An adaptive design

At Stage 1, if $\hat{\theta} < 0$, stop to accept H_0: $\theta \leq 0$.

If $\hat{\theta} > 0$ 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, H_{01}, H_{12}, H_{012}.
If $\hat{\theta}_1 < 0$ and $\hat{\theta}_2 > \hat{\theta}_1 + 8$,	Restrict to sub-population 2 and test H_2 only, needing to reject H_2, H_{02}, H_{12}, H_{012}.
Else,	Continue with full population and test H_0, needing to reject H_0, H_{01}, H_{02}, H_{012}.

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

Enrichment: 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

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

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

	Stage 1	*Stage 2*
H_0	$Z_{1,0}$	$Z_{2,0}$
H_{01}	$Z_{1,0}$	$Z_{2,0}$
H_{02}	$Z_{1,0}$	$Z_{2,0}$
H_{012}	$Z_{1,0}$	$Z_{2,0}$

where $Z_{i,0}$ is based on $\hat{\theta}$ from responses in Stage i.

So, there is no change from the original test of H_0.

Enrichment: An adaptive design

When switching to sub-population 1, we use:

	Stage 1	*Stage 2*
H_1	$Z_{1,1}$	$Z_{2,1}$
H_{01}	$Z_{1,0}$	$Z_{2,1}$
$H_{12} = H_{012}$	$Z_{1,0}$	$Z_{2,1}$

When switching to sub-population 2, we use:

	Stage 1	*Stage 2*
H_2	$Z_{1,2}$	$Z_{2,2}$
H_{02}	$Z_{1,0}$	$Z_{2,2}$
$H_{12} = H_{012}$	$Z_{1,0}$	$Z_{2,2}$

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

The need to reject intersection hypotheses adds to simple tests of H_1 or H_2.

Enrichment: Power of non-adaptive and adaptive designs

				Non-adaptive	***Adaptive***			
	θ_1	θ_2	θ	*Full pop^n*	*Sub-pop 1 only*	*Sub-pop 2 only*	*Full pop^n*	*Total*
1.	30	0	15	**0.68**	0.43	0.00	0.41	**0.85**
2.	20	20	20	**0.90**	0.03	0.03	0.84	**0.90**
3.	20	10	15	**0.68**	0.11	0.01	0.59	**0.71**

Case 1: Testing focuses (correctly) on H_1, but it is still possible to find an effect (wrongly) for the full population.

Case 2: Restricting to a sub-population reduces power for finding an effect in the full population.

Case 3: Adaptation improves power overall, but there is a small probability of restricting to the wrong sub-population.

Enrichment: Example

The rules for sticking or switching to a sub-population can be adjusted, but we cannot eliminate the probability of making an error in these decisions.

This is to be expected since the standard error of 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 results are found if only one sub-population is specified as a candidate for restricted sampling.

Problem 3: Switching to a sub-population can improve power.

However, higher overall sample size is needed for accurate sub-population inference.

4. Switching between tests for superiority and non-inferiority

When there is already an accepted treatment for a condition, it is not appropriate to test a new treatment against placebo.

A trial using the standard treatment as an active control has two possible positive outcomes:

Showing the new treatment is *superior* to the current standard,

Showing the new treatment is *non-inferior* to the standard.

Investigators may start a trial intending to show superiority, then decide to adapt to a new goal of non-inferiority if results are not as good as expected.

Having two hypotheses is not an issue as the two tests are nested:

Superiority — Null hypothesis: $\theta \leq 0$,

Non-inferiority — Null hypothesis: $\theta \leq -d$.

Testing for both superiority and non-inferiority

Differing sample size requirements

Wang, Hung, Tsong & Cui (*Statistics in Medicine*, 2001) note the non-inferiority margin d is often smaller than the effect size δ at which power for declaring superiority is specified.

Thus, a larger sample size is needed to test adequately for non-inferiority.

If Stage 1 data indicate that the key issue is to test for non-inferiority, one may wish to increase the Stage 2 sample size.

Adaptive re-design

Wang et al. propose a group sequential test with group size determined by the power for superiority.

They then use the method of Cui, Hung & Wang (1999) to increase group sizes if interest shifts to proving non-inferiority.

Testing for both superiority and non-inferiority

A non-adaptive group sequential approach

In principle, one may embed testing for both superiority and non-inferiority in a group sequential design.

Early stopping may be appropriate:

to reject $H_{0,S}$: $\theta \leq 0$ (establishing superiority),

to accept $H_{0,NI}$: $\theta \leq -d$ (failing even to show non-inferiority),

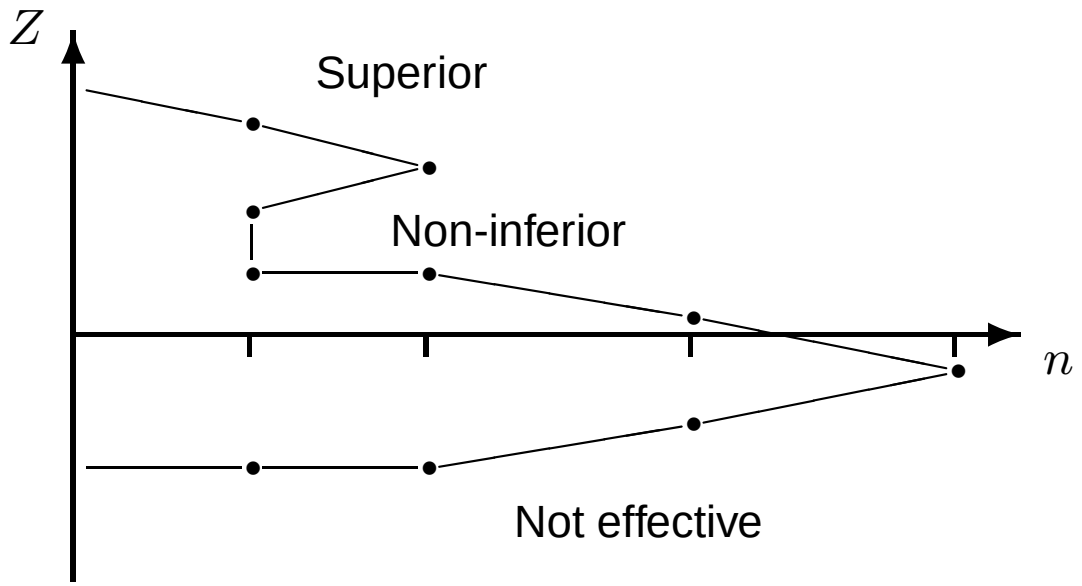
to declare non-inferiority only.

If power for declaring superiority is set at a higher effect size, δ, than the margin of non-inferiority, d, the stopping rule for declaring superiority will be more aggressive.

Question: Is there significant benefit to allowing adaptive specification of group sizes in such a group sequential design?

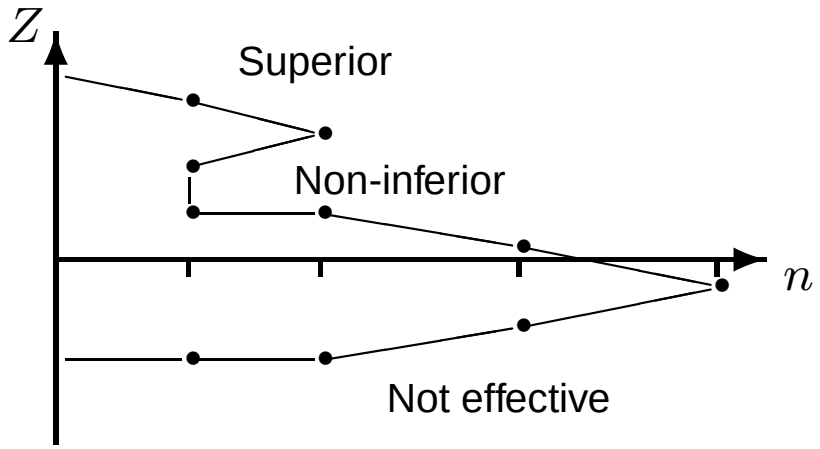
Testing for both superiority and non-inferiority

A group sequential design to test for either *Superiority* or *Non-inferiority* could have the general form:



In recent work with Fredrik Öhrn, we have derived optimal designs of this type and used these to assess the performance of other proposals.

Testing for both superiority and non-inferiority



The asymmetry of these designs is important when different fixed sample sizes are needed for superiority and non-inferiority goals.

We have found little further benefit from adaptively choosing group sizes.

Problem 4: One can test for superiority *and* non-inferiority.

We recommend a pre-planned group sequential design for this 3 decision problem.

5. Treatment selection: Combined Phase II / III trials

A ***seamless Phase II / III design*** facilitates progression from treatment selection to a large confirmatory trial without the usual “white space”.

Such a design may also combine Phase II and Phase III data in the final hypothesis test for efficacy of the selected treatment.

Since each treatment has its own null hypothesis, a multiple testing procedure is required to control the overall type I error rate.

Data may be merged through inverse χ^2 and inverse normal combination tests (e.g., Bretz, Schmidli, et al. and Schmidli, Bretz et al, *Biometrical Journal*, 2006).

Earlier proposals control type I error, and power, directly (e.g., Thall, Simon & Ellenberg, *Biometrika*, 1988).

How useful is this data combination?

Combined Phase II/III trials: Example

Jennison & Turnbull (*J. Biopharmaceutical Statistics*, 2007) consider a Phase II comparison of 4 treatments, with 100 observations per treatment and control.

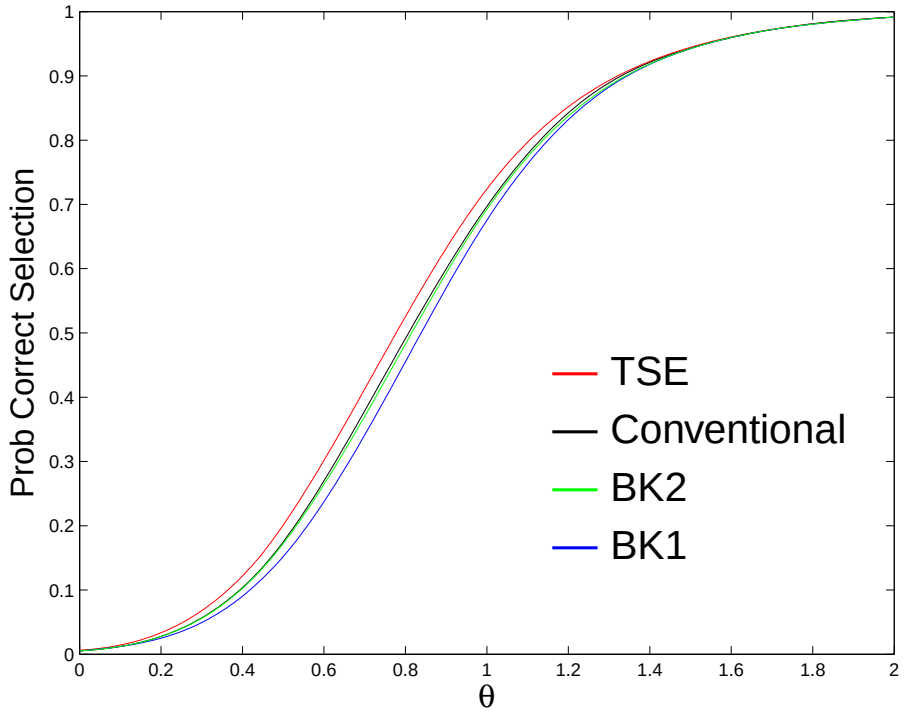
If the treatment performing best in Phase II meets a minimum threshold, it proceeds to further testing in Phase III, with 500 observations for this treatment and control.

They assess four different strategies for combining data:

1. **Conventional** Separate Phase II and Phase III trials: final decision is based on the Phase III data alone.
2. **BK1** Data from the two phases are combined by an inverse χ^2 test.
3. **BK2** Data from the two phases are combined by an inverse normal test.
4. **TSE** A Thall, Simon & Ellenberg design.

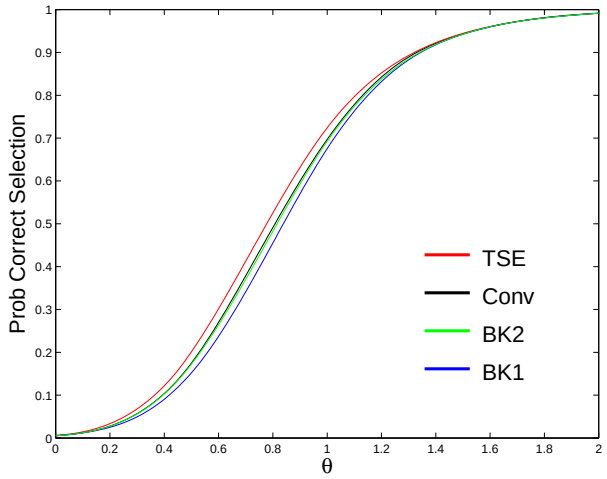
Combined Phase II/III trials: Example

Power of four 2-stage selection/testing procedures when three are ineffective and the other has effect size δ:



Differences in power curves correspond to differences in sample size needed for a given power of order 5% to 10%.

Combined Phase II/III trials: Example



The conventional procedure — with no data combination — is surprisingly efficient: slightly worse than the Thall, Simon & Ellenberg design, but superior to inverse χ^2 and inverse normal combination tests.

Problem 5: Joint planning of two phases of testing can be valuable — for a variety of reasons. Benefits of data combination may not be so great.

Conclusions

Adaptive methods provide a useful route to modifying sample size as a ***nuisance parameter*** is estimated.

Group sequential tests are ideal to respond to estimates of the ***primary endpoint***.

Adaptive methods lead to moderate efficiency gains when ***changing primary endpoint*** or ***restricting to a sub-population***.

Adaptive or group sequential methods are effective for switching between tests for ***superiority and non-inferiority***.

Special methods can yield benefits from ***combining data*** between phases.

But, remember that interim estimates will have high variance.

We recommend adaptation as part of a pre-planned and *pre-tested* trial design — "flexible adaptation" brings risks as well as opportunities.