Critical Appraisal of Adaptive Methods

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

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

SMi, Adaptive Designs in Clinical Drug Development

London, February 2009

The role of adaptive methods

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.

Critical appraisal of adaptive designs

It is important to assess the benefits a new approach will be able to deliver.

How can the benefits be assessed?

Are these benefits real?

What "standard" methods should be used for comparison?

We shall use simulation to appraise adaptive methods for three applications:

1: Sample size modification,

2: Switching to a patient sub-population,

3: Treatment selection and testing in a combined Phase II/III design.

1. Sample size modification to increase power

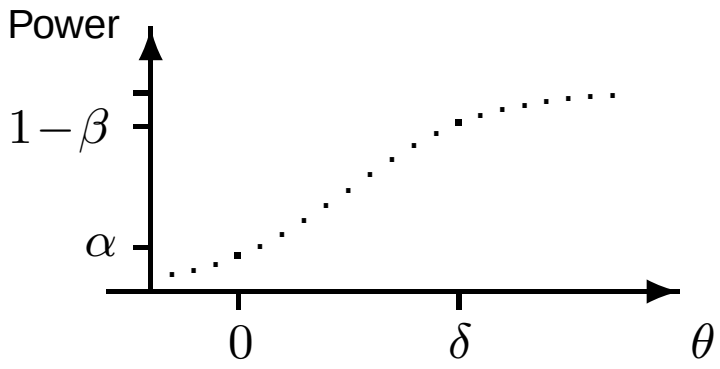
Type I error and power

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.

Sample size can be chosen to give a specific power curve.



(Assume there is no unknown "nuisance parameter" such as a normal variance or overall failure rate for survival data which appears in this relationship.)

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 *wait and see* 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

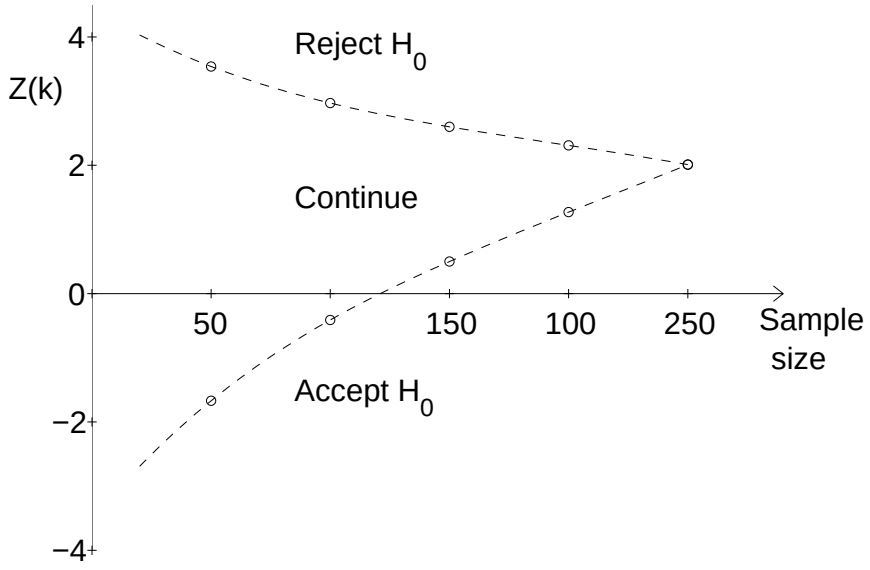
Example *(Jennison & Turnbull, Biometrika, 2006, Ex. 2)*

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.

Cui, Hung and Wang (*Biometrics*, 1999) cite instances of studies reporting to the FDA where such problems arose.

Special methods are needed in order to protect the type I error rate while making data-dependent modifications to sample size.

Cui, Hung and Wang developed a method which allows remaining group sizes to be increased in a group sequential design.

A variety of other methods for sample size modification is now available.

Sample size modification to increase power

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

Following a decision at analysis 2 to increase sample size:

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 preserves the variance of this term but the mean is multiplied by $\gamma^{1/2}$.

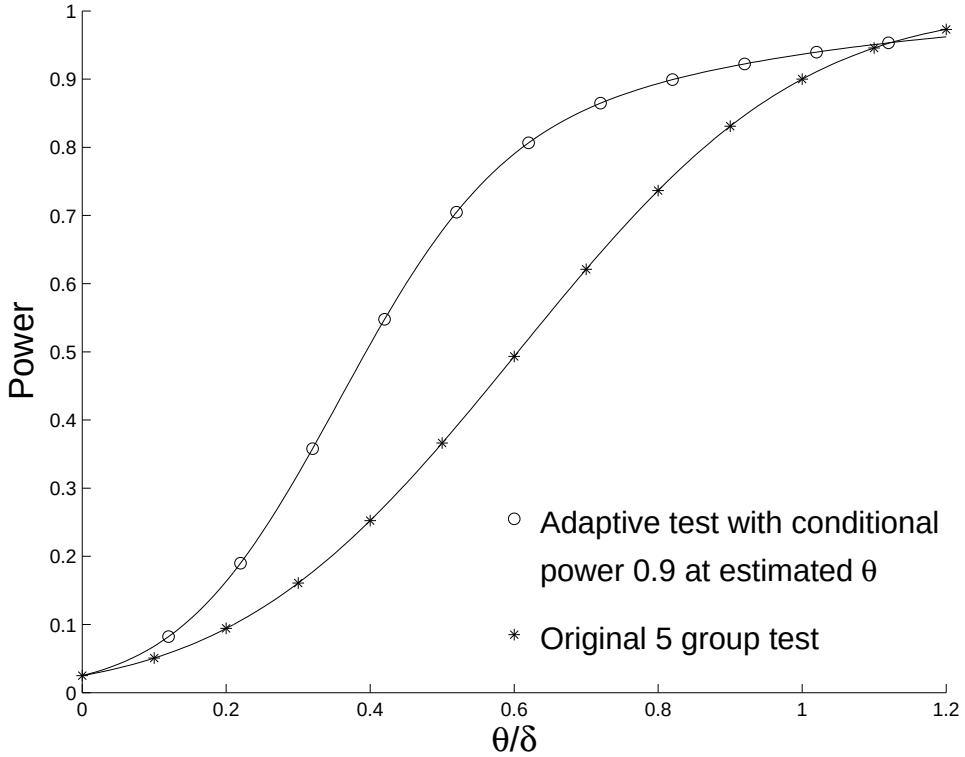
Using the new weighted sample sum in place of the original sample sum maintains the type I error rate and increases power.

In our example:

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

Simulations show that 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

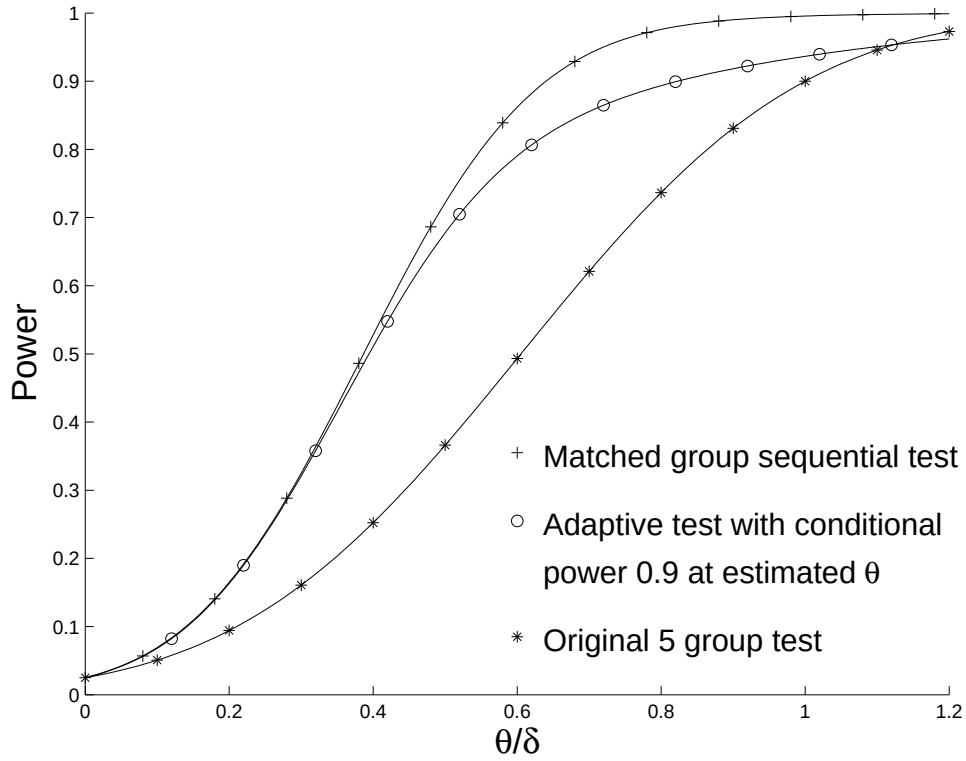
Five-group designs matching key features of the adaptive test can be found.

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

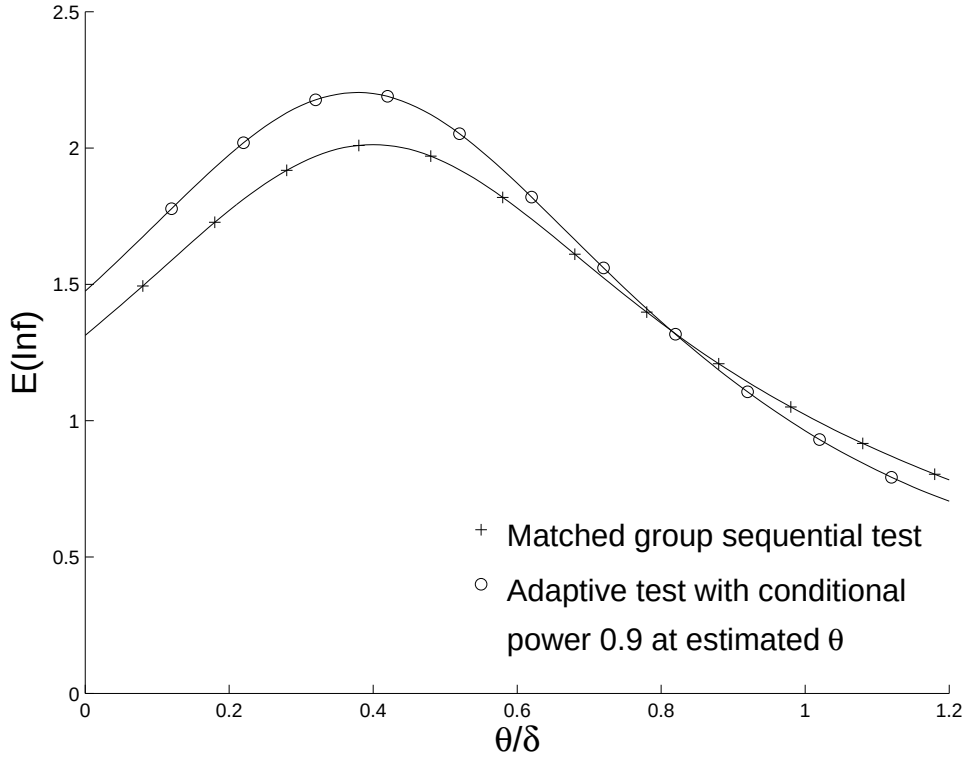
Power of our “matched” group sequential design is as high as that of the adaptive design at all effect sizes — and substantially higher at the largest θ values.



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.

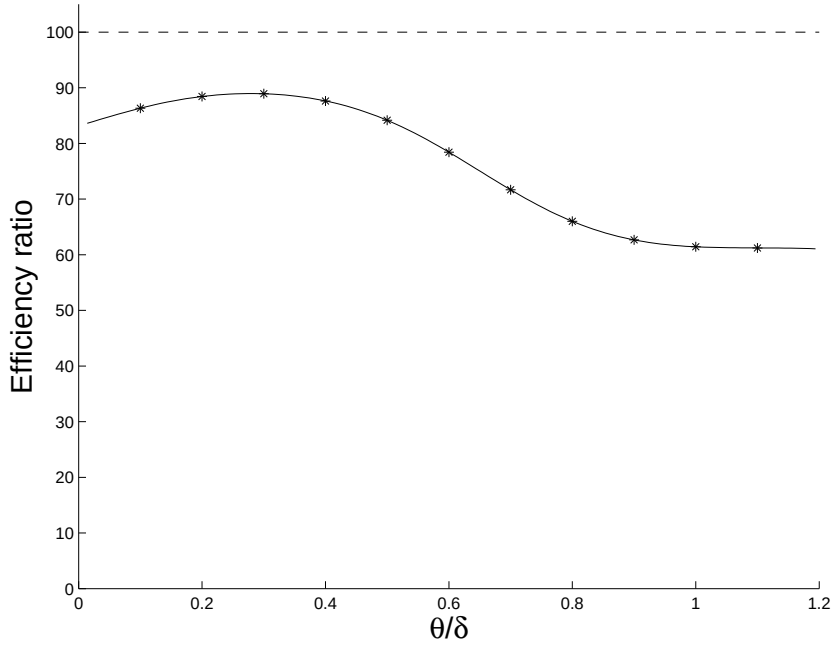


Sample size modification to increase power

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.

Application 1: Specify power properly at the outset: then, group sequential designs offer a simple and efficient option.

2. 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

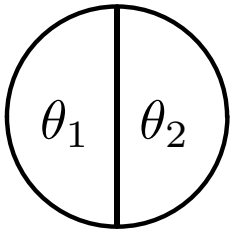
Restrict the remainder of the study to a sub-population.

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 $\theta_2 > 0$.

Multiple testing procedures

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

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

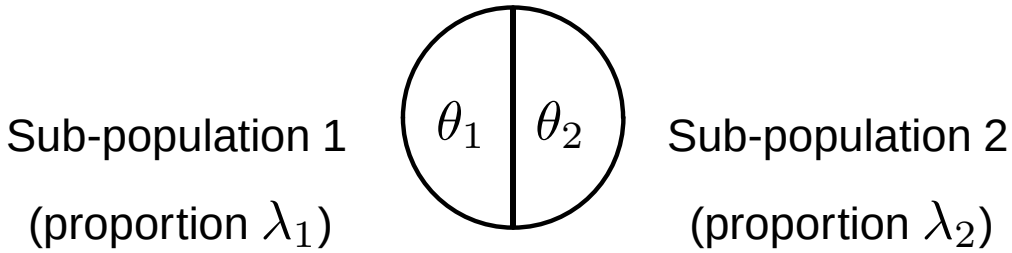
The simple hypothesis H_j: $\theta_j \leq 0$ is rejected if, and only if, H_I is rejected for every set I containing index j.

Proof of strong control of family-wise error rate

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

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.

Enrichment: Example



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

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

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

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

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

Enrichment: Example

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

	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.

Simulation results: 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.

Application 2: Restricting attention to a sub-population can be effective in improving power.

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

3. 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.

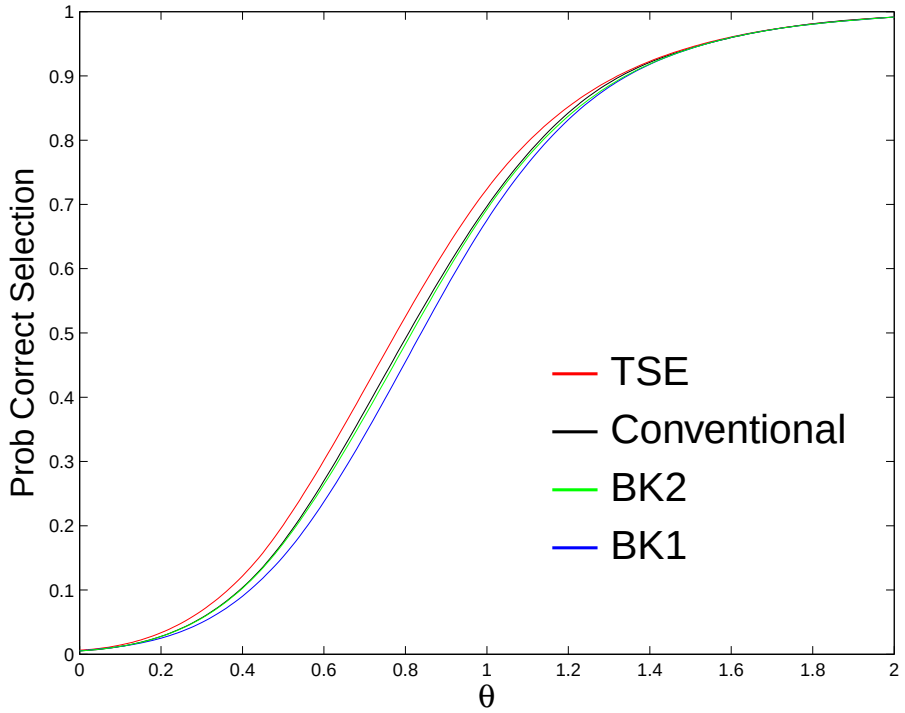
If the treatment performing best in Phase II meets a minimum threshold, it proceeds to further testing in Phase III, with 500 observations for this treatment and control.

They report simulations of 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.

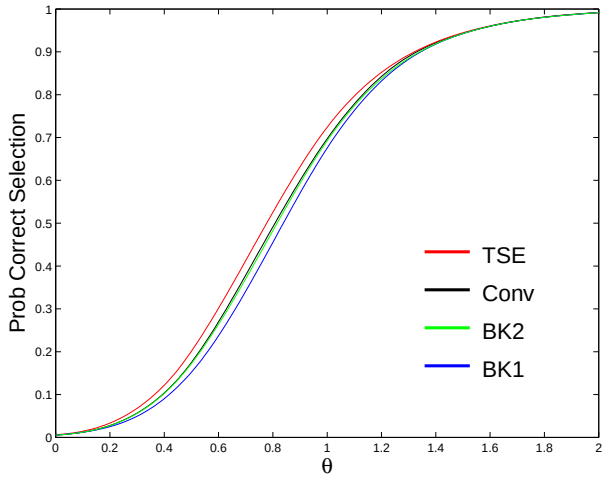
Combined Phase II/III trials: Example

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: only slightly worse than the Thall, Simon & Ellenberg design, and superior to inverse χ^2 and inverse normal combination tests.

Application 3: 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. However, they are an inferior option to *Group Sequential Tests* if one wishes to respond to estimates of the ***primary endpoint***.

Adaptive methods lead to moderate efficiency gains when ***restricting to a sub-population***.

Special methods can yield modest 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.