Adaptive Trials and Traditional Designs:

finding the best method for

your study objectives

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 2010

Overview of this talk

We shall discuss:

Choosing between fixed sample, group sequential and adaptive designs

Waiting to specify the power curve: an adaptation too far

Seamless Phase II / Phase III: simplicity versus efficiency

Multiple treatments, multiple populations: when only adaptation works

The bigger picture: looking beyond a single stage of the drug development process.

1. Fixed sample, group sequential or adaptive design?

A fixed sample size study cannot always provide the power desired for a trial, and even then it may do so in an inefficient manner.

Example

Consider a two-treatment comparison with normally distributed responses on treatments A and B

$$X_{Ai} \sim N(\mu_A,\, \sigma^2) \quad \text{and} \quad X_{Bi} \sim N(\mu_B,\, \sigma^2).$$

Objective

It is desired to test H_0: $\theta = \mu_A - \mu_B \leq 0$ against $\theta > 0$ with type I error rate α.

The study should achieve power $1 - \beta$ at $\theta = \delta$.

In the case of known variance, the required sample size per treatment is

$$n \,=\, \{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2 \; 2\,\sigma^2/\delta^2. \qquad (1)$$

But what if σ^2 is unknown?

(a) Sample size re-estimation for a response variance

If the variance σ^2 is unknown, one can follow an adaptive procedure, using an estimate of σ^2 from early data to modify the initial choice of sample size.

Wittes & Brittain (*Statistics in Medicine*, 1990) suggest a general approach for incorporating an “internal pilot study” in a clinical trial design.

Let ϕ denote a nuisance parameter, e.g., the response variance.

> Suppose the sample size required under a particular value of ϕ is given by the function $n(\phi)$.
>
> From a pre-study estimate, $\hat{\phi}_0$, calculate an initial sample size, $n(\hat{\phi}_0)$.
>
> At an interim stage, find a new estimate $\hat{\phi}_1$ from the data observed so far and aim for the new target of $n(\hat{\phi}_1)$ observations.

Sample size re-estimation: Wittes & Brittain

The Wittes & Brittain approach can be used to achieve specified power in our example with normal responses of unknown variance. Here, $\phi = \sigma^2$.

At an interim analysis, σ^2 can be estimated from the usual sums of squares about the treatment means.

Alternatively, one may use a blinded variance estimate based on the pooled data without treatment labels being identified: see, e.g., Zucker et al. (*Statistics in Medicine*, 1999).

In general, variance estimates from the Wittes & Brittain procedure are biased downwards. However, Jennison & Turnbull (2000, Ch. 14) show the type I error rate is only slightly perturbed.

A Bauer & Köhne (*Biometrics*, 1994) combination test can attain a specified power while achieving type I error rate α exactly.

Sample size re-estimation: Bauer & Köhne

Initial design

We specify a two-stage adaptive design, using the inverse χ^2 combination rule to test H_0: $\theta \leq 0$ against $\theta > 0$.

The test will combine P-values P_1 and P_2 from the two stages, rejecting H_0 for low values of $P_1\,P_2$ — which is distributed as $e^{-\frac{1}{2}\chi^2_4}$ under H_0.

An initial estimate σ_0^2 is used in the formula (1) to obtain a sample size of n_0 per treatment. This would give the desired power if indeed $\sigma^2 = \sigma_0^2$.

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

Stage 1

Yields estimates $\hat{\theta}_1$ and $\hat{\sigma}_1^2$, plus the t-statistic t_1 for testing H_0 vs $\theta > 0$.

We convert t_1 to a one-sided P-value, $P_1 = Pr_{\theta=0}\{T_{2\,n_1-2} > t_1\}$.

Sample size re-estimation: Bauer & Köhne

Stage 1 . . .

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

This estimate can be substituted in the original calculation, in place of σ_0^2.

Or, say, the stage 2 sample size might be chosen to give conditional power $1-\beta$, given P_1, assuming $\theta = \hat{\theta}_1$ and $\sigma^2 = \hat{\sigma}_1^2$.

This defines an additional sample size of n_2 per treatment arm in stage 2.

Stage 2

We calculate the t-statistic t_2 for testing H_0 vs $\theta > 0$ based on stage 2 data alone.

Convert t_2 to a P-value, $P_2 = Pr_{\theta=0}\{T_{2\,n_2-2} > t_2\}$.

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

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

(b) Efficiency benefits of group sequential tests

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

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

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

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

Standardized test statistics $Z_{(1)}$, $Z_{(2)}$, . . . , are computed at interim analyses and used to define a stopping rule for the trial.

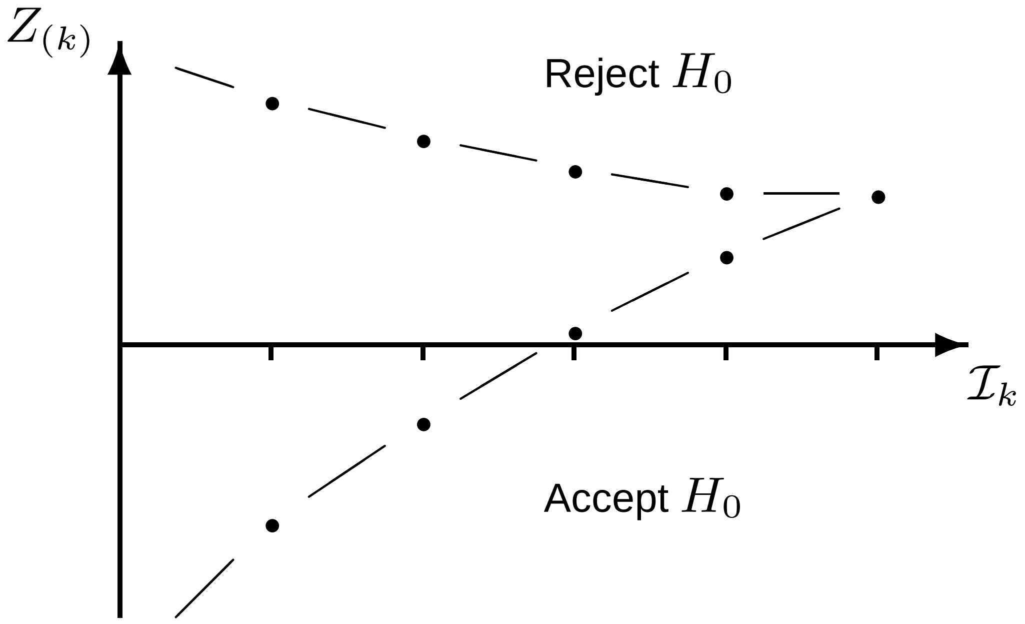
The information for θ at analysis k is denoted by

$$\mathcal{I}_k \; = \; \{\text{Var}(\hat{\theta}_k)\}^{-1}.$$

Note the notation: $Z_{(k)}$ denotes the Z-statistic based on *all* data up to analysis k, whereas Z_k represents the Z-statistic from data in group k alone.

Benefits of group sequential tests

A typical boundary for a one-sided test has the form:



Crossing the upper boundary leads to early stopping for a positive outcome, rejecting H_0 in favour of $\theta > 0$.

Crossing the lower boundary implies stopping for “futility” with acceptance of H_0.

Benefits of group sequential tests

In order to test H_0: $\theta \leq 0$ against $\theta > 0$ with type I error probability α and power $1 - \beta$ at $\theta = \delta$, a fixed sample size test needs information

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

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

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

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

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

Benefits of group sequential tests

One-sided tests, $\alpha = 0.025$, $1 - \beta = 0.9$, K analyses, $\mathcal{I}_{max} = R\,\mathcal{I}_{fix}$,

equal group sizes, minimising $\{E_0(\mathcal{I}) + E_\delta(\mathcal{I})\}/2$.

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

			R			*Minimum*
K	1.01	1.05	1.1	1.2	1.3	*over R*
2	80.8	74.7	73.2	73.7	75.8	73.0 at R=1.13
3	76.2	69.3	66.6	65.1	65.2	65.0 at R=1.23
5	72.2	65.2	62.2	59.8	59.0	58.8 at R=1.38
10	69.2	62.2	59.0	56.3	55.1	54.2 at R=1.6
20	67.8	60.6	57.5	54.6	53.3	51.7 at R=1.8

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

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

Flexible group sequential tests

Since the sequence $\mathcal{I}_1, \mathcal{I}_2, \ldots$ is often unpredictable, it is good to have a group sequential design that can adapt to observed information levels.

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

One-sided error spending tests can be defined similarly (see JT, Ch. 7).

Spending error as a function of observed information leads to “information monitoring” designs.

Mehta & Tsiatis (*Drug Information Journal*, 2001) use the information monitoring approach to accommodate nuisance parameters. In particular, they propose a group sequential analogue of the Wittes & Brittain method for normal responses with unknown variance. This method combines adaptation to an estimated variance and the efficiency gains of early stopping in a group sequential test.

Roles of adaptive and group sequential designs

So far, we have used “adaptation” to modify sample size in order to attain the information needed for a power criterion.

The role of group sequential designs has been to stop early when data allow in order to achieve type I error rate and power with as small a sample size as possible.

Bauer & Köhne’s combination test framework allows sample size modification for a variety of reasons:

> *One may increase the original sample size in order to enhance power, and so rescue an under-powered study.*
>
> *One can try to mimic features of a group sequential design to stop early when the data are favourable, and hence reduce average sample size.*

There have been proposals to delay power considerations until data have been observed, since adaptation allows this approach.

2. Waiting to specify the power curve: an adaptation too far

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?

Waiting to specify the power curve

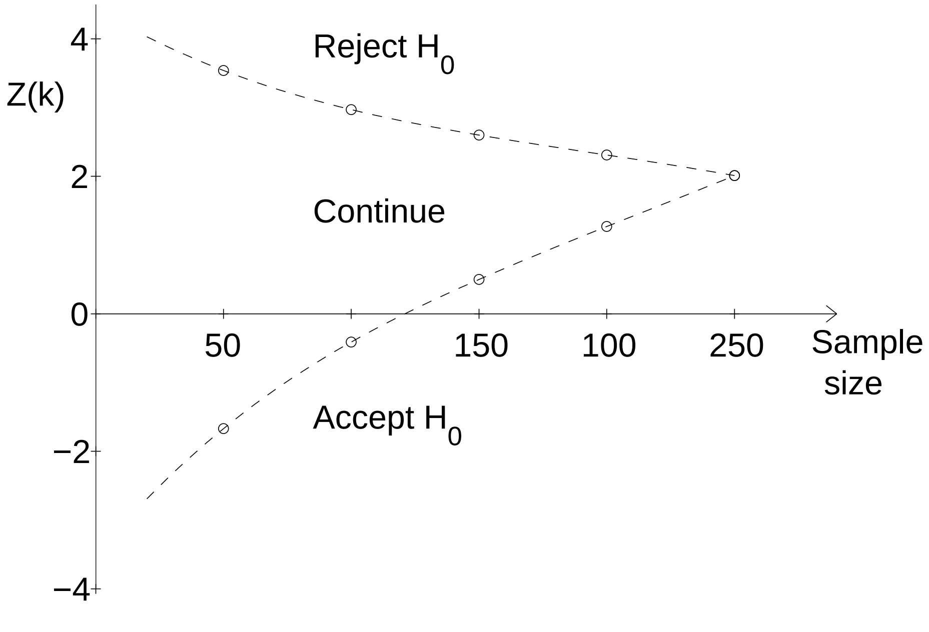
Example *(Jennison & Turnbull, Biometrika, 2006, Example 2)*

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 $\theta = \delta$.



Waiting to specify the power curve

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.

An adaptive design

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 $\gamma^{1/2}$.

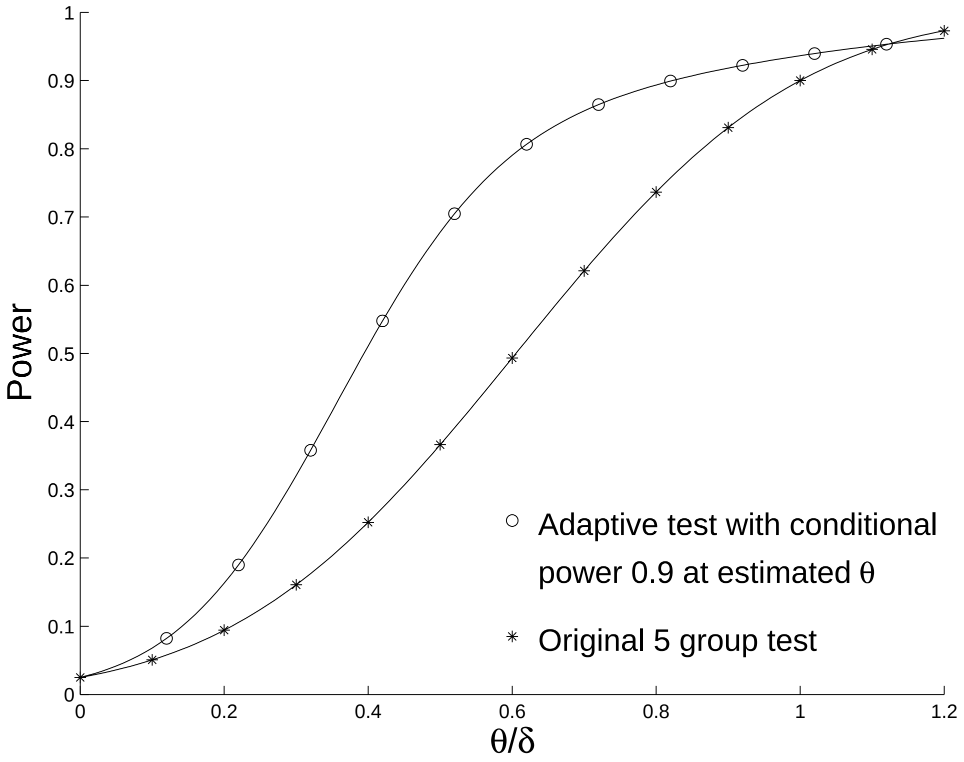
Using the new weighted sample sum in place of the original sample sum maintains the type I error rate and increases power.

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 .

An adaptive design

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.

The overall power curve

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

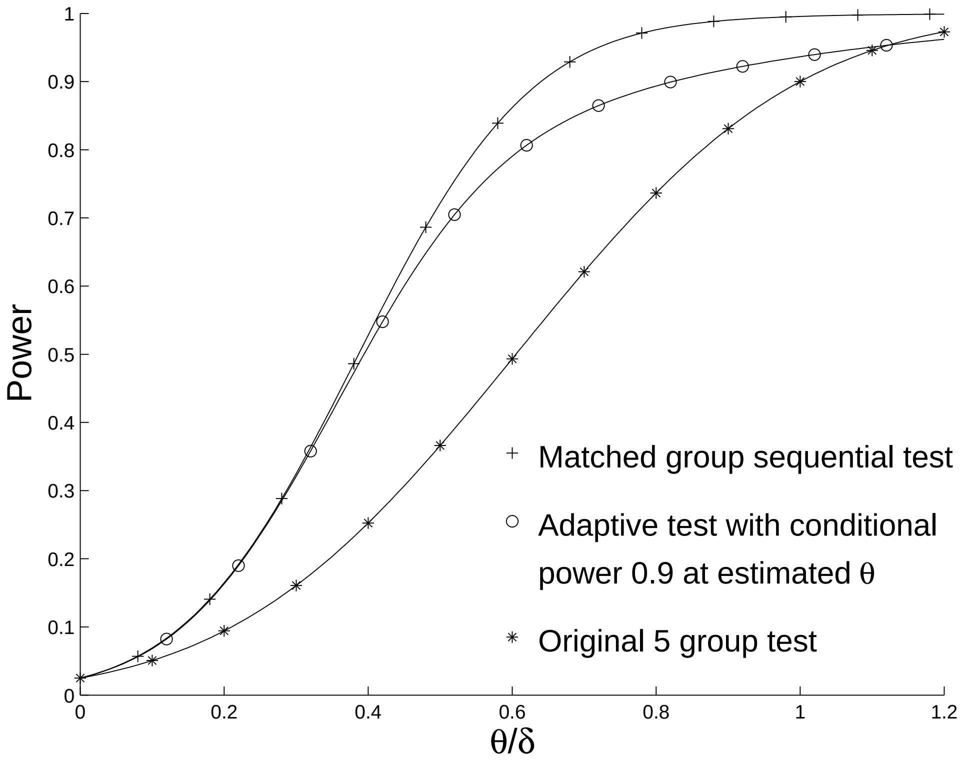
Five-group designs matching key features of the adaptive test can be found.

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

Can expected sample size be lower too?

A matched group sequential design

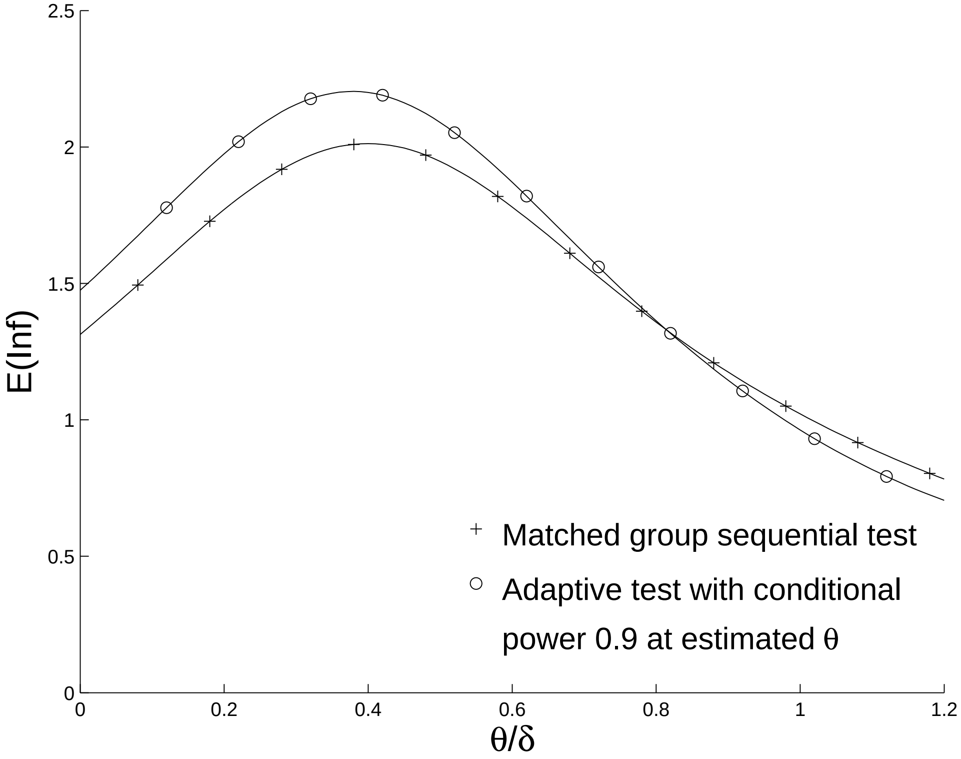
Power of our “matched” group sequential design is as high as that of the adaptive design at all effect sizes — and substantially higher at the largest θ values.



A matched group sequential design

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.

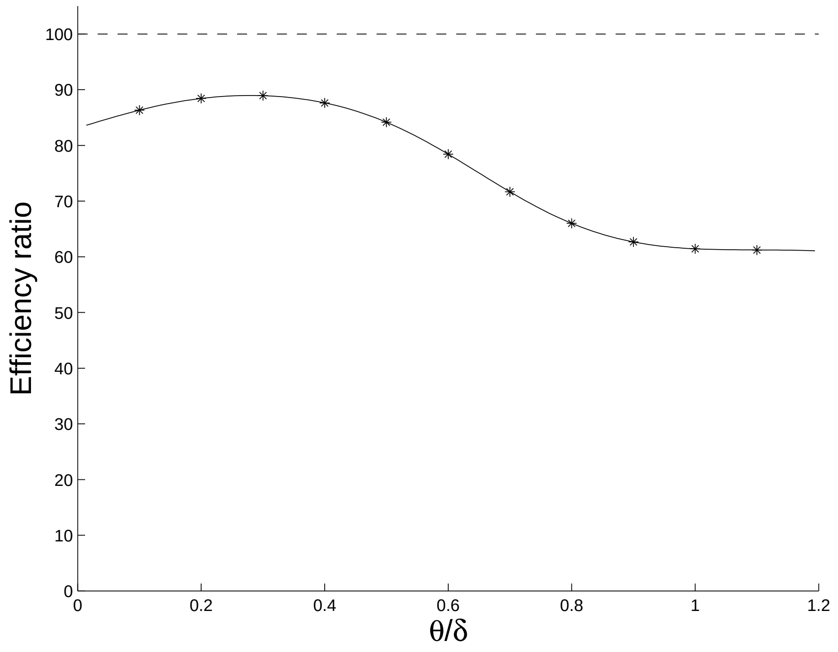


Efficiency comparison

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



Waiting to specify the power curve — conclusions

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.

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

3. Seamless Phase II / Phase III: simplicity vs efficiency

Phase IIb

The trial compares several dose levels of a treatment with a control in order to select a dose and provide evidence of improvement against the control.

A Phase III

The trial is run as a confirmatory study to demonstrate superiority against control of the treatment selected in Phase IIb.

The traditional approach has the following stages:

Write Phase IIb protocol, seek ethical and regulatory approval, (FDA, IRBs, . . .)

Run Phase IIb, analyse data, reach conclusions.

Write Phase III protocol, seek ethical and regulatory approval, (FDA, IRBs, . . .)

Run Phase III, analyse data, reach final conclusion.

Seamless Phase II/III designs

Planning the Phase III trial after Phase IIb allows investigators to make use of information gained in Phase IIb.

They may decide to modify:

Treatment definition,

Target population,

Primary endpoint,

Sample size.

Positive results in Phase IIb will help recruitment for participation in Phase III.

But, planning and gaining approval for the Phase III trial can be time-consuming.

If the final outcome is positive, the sooner this conclusion is reached, the better.

Reducing “white space” between Phases II and III

A ***seamless*** design has a single protocol combining the usual Phases II and III.

There are benefits of eliminating the gap between Phases:

Shorter time to reach a conclusion, so more patent lifetime remaining for a successful treatment,

No break in recruitment of subjects.

Difficulties are:

Regulators may well require sponsors to be blinded *for the whole study*,

The monitoring committee needs a complete set of rules to work by, specifying how aspects of the Phase III stage depend on results of the Phase II stage,

No mechanism for dealing with the unexpected.

Another option for Phases II and III

In the current paradigm, we

(a) Select a dose in Phase II,

(b) Run Phase III with this dose.

A key problem is that the dose going forward to Phase III may not be the best one for achieving high efficacy and good safety.

This problem can be exacerbated if Phase II has a short term endpoint and its sample size is too small to reveal rare occurrences of serious adverse events.

An alternative strategy:

Take forward 2 or 3 dose levels to Phase III, possibly eliminating doses during Phase III (particularly for poor safety).

This may not eliminate “white space”, but it tackles one of the major problems that can adversely affect the Phase III outcome.

Combining data from Phases II and III

We return to the traditional format with:

Phase II

K treatments and a control are compared, with m_1 observations on each.

Estimated treatment effects are $\hat{\theta}_{1,i},\ i = 1, \ldots, K$.

The treatment i^* with highest $\hat{\theta}_{1,i}$ is selected for Phase III.

Phase III

Treatment i^* is compared against control, with m_2 observations on each.

Estimated treatment effect is $\hat{\theta}_{2,i^*}$.

But now

A final decision is made, based on $\hat{\theta}_{2,i^*}$ ***and*** $\hat{\theta}_{1,1}, \ldots, \hat{\theta}_{1,K}$.

Combining data from Phases II and III

There are K null hypotheses, H_i: $\theta_i \leq 0$, $i = 1, \ldots, K$.

If dose i^* is selected for Phase III, we focus on testing H_{i^*}: $\theta_{i^*} \leq 0$.

Family-wise error

We wish to control **family-wise error**, so require $Pr\{$Reject *any* true $H_i\} \leq \alpha$ for all vectors $(\theta_1, \ldots, \theta_K)$.

Then, the probability of falsely claiming significance for the selected i^* is at most α.

Power

When some θ_i are greater than zero, we want a high probability of selecting an effective treatment and rejecting the associated null hypothesis.

Questions:

1. What is the best way to combine data from the two Phases?

2. What value do we get from the Phase II data?

Combining data from Phases II and III

Thall et al. (*Biometrika*, 1988) base the final decision just on $\hat{\theta}_{1,i^*}$ and $\hat{\theta}_{2,i^*}$.

Bretz et al. (*Biometrical Journal*, 2006) propose a closed testing procedure, using a Dunnett test to combine P-values in testing intersection hypotheses, with an inverse normal combination test across stages.

Lisa Hampson and I derived optimal decision rules for particular sets of treatment effects. We found both the above methods to have very high efficiency across a variety of treatment effect vectors.

Answer 1. For simplicity, we recommend Thall et al's method for combining data.

Answer 2. We have found the Phase II data on treatment i^* and the control to be worth around 50% of their face value.

For example, if Phase II has 100 observations per treatment and control, these improve power by the same amount as an extra 50 observations on treatment i^* and control in Phase III.

Combining data from Phases II and III

We have seen that the benefits of data combination are eroded by the multiplicity adjustment made to the Phase II data.

Moreover, the same organisational questions arise as before:

Regulators are liable to treat the combined study as a single trial and require blinding of the whole process.

Issues that might have addressed in the gap between Phases II and III must be anticipated and rules for how to proceed set up in the overall protocol.

The benefits of using Phase II data in the Phase III analysis come at the cost of a heightened administrative burden.

Data combination may still be desirable if observations are at a premium, e.g., in a rare illness with slow patient recruitment.

4. Multiple treatments, multiple populations: when only "adaptation" works

Re-defining the treatment or patient population can lead to several null hypotheses, $H_{0,1}, \ldots, H_{0,k}$, being tested in the course of a trial.

Bauer & Köhne (1994) present two approaches to testing multiple hypotheses:

1. If P-values P_i for hypotheses $H_{0,i}$ are combined through a combination test, a positive outcome is rejection of the intersection hypothesis

$$\cap_{i=1}^{k} H_{0,i}.$$

But, the conclusion that at least one of the $H_{0,i}$ does not hold is of limited value.

2. Use of a procedure that controls ***family-wise*** type I error rate allows a conclusion about a specific hypothesis, even though this choice is data generated.

We just saw this approach in selecting and testing a treatment in a Phase II/III trial.

Switching to a patient sub-population

A trial protocol defines a population of subjects who may benefit from the treatment.

Suppose it is believed the treatment could be particularly effective in a certain sub-population defined by a physiological or genetic biomarker.

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 the defined sub-population.

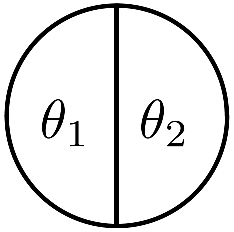
This choice will affect the licence a positive outcome can support.

The possibility of testing more than one null hypothesis means a multiple testing procedure must be used.

Enrichment: Example

Sub-population 1

(proportion λ_1)



Rest of the population

(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 $\theta_1 > 0$.

Closed testing procedures (Marcus et al, *Biometrika*, 1976)

With hypotheses H_i: $\theta_i \leq 0,\ i = 1, \ldots, k,$ define the intersection hypothesis $H_I = \cap_{i \in I} H_i$ for each subset I of $\{1, \ldots, k\}$.

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 hypothesis H_i: $\theta_i \leq 0$ is rejected overall if, and only if, H_I is rejected for every set I containing index i.

This procedure controls the family-wise error rate strongly at level α, i.e.,

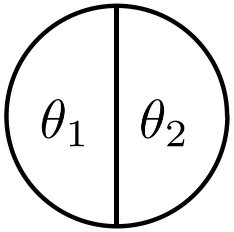
$$Pr\{\text{Reject any true } H_i\} \leq \alpha \quad \text{for all } (\theta_1, \ldots, \theta_k).$$

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.

Enrichment: Example

Sub-population 1

(proportion λ_1)



Rest of the population

(proportion λ_2)

First, consider a design testing for a whole population effect, $\theta = \lambda_1\theta_1 + \lambda_2\theta_2$, in the case $\lambda_1 = \lambda_2 = 0.5$.

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

If $\hat{\theta} < 0$ at the interim analysis, stop for futility with acceptance of H_0.

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

Enrichment: Example

Properties of design for the whole population effect.

θ_1	θ_2	θ	*Power for* H_0*:* $\theta \leq 0$
20	20	20	0.90
10	10	10	0.37
20	0	10	0.37

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

A closed testing procedure will require tests for 3 hypotheses:

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_{01}: $\theta \leq 0$ and $\theta_1 \leq 0$.

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 and H_{01}.

Else, Continue with full population and test H_0,
needing to reject H_0 and H_{01}.

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

All 3 hypotheses are tested in two-stage tests, stopping to accept the hypothesis at stage 1 if the Z-statistic is less than zero.

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

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

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

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

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

The stage 1 test of H_{01} uses a combination of the Z-statistics for H_0 and H_1, $Z_{1,01} = (Z_{1,0} + Z_{1,1})/\sqrt{(2+\sqrt{2})}$, which is $N(0,1)$ under H_{01}.

Enrichment: An adaptive design

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

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

When switching to sub-population 1, we use:

	Stage 1	*Stage 2*
H_1	$Z_{1,1}$	$Z_{2,1}$
H_{01}	$Z_{1,01}$	$Z_{2,1}$

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

The need to reject the intersection hypothesis H_{01} adds an extra requirement to the simple test of H_1.

Simulation results: Power of non-adaptive and adaptive designs

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

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

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

Case 4: Adaptation improves overall power a little.

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 differences between θ_1 and θ_2 that interest us.

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

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

5. The bigger picture: looking beyond a single stage of the drug development process

Surprisingly little attention has been given to this topic.

Questions:

How should resources be distributed between Phases II and III?

Which factors may affect this preferred distribution?

Remember:

Phase II may use a different endpoint from Phase III (often a more rapidly observed response).

Phase III will assess both efficacy and safety.

Joint planning of Phases II and III

There has been some work in this area:

Thall et al. (1988) optimised over Phase II and Phase III sample sizes in their design.

Patel & Ankolekar (*Statistics in Medicine*, 2007) consider optimal sample size of a Phase III trial from an economic perspective.

A PhRMA Working Group was recently set up to explore this issue.

There is heightened interest in improving the effectiveness of the drug development framework.

This requires analysis of the over-arching process as well as improvement of individual stages.