Group Sequential and Adaptive Designs:

Where Are We Now?

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

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

Merck Serono

Geneva, Switzerland

June 2010

Where have we come from?

Around 10 years ago:

Group sequential methods were well understood.

Adaptive designs were becoming a hot topic with influential papers by

Bauer & Köhne (*Biometrics*, 1994)

Proschan & Hunsberger (*Biometrics*, 1995)

Fisher (*Statistics in Medicine*, 1998)

Cui, Hung & Wang (*Biometrics*, 1999)

Lehmacher and Wassmer (*Biometrics*, 1999)

Denne (*Statistics in Medicine*, 2001)

Müller and Schäfer (*Biometrics*, 2001).

The role of Group Sequential methods

Group sequential designs allow termination of a trial as soon as there is sufficient evidence to answer the primary research questions.

Early termination may be for a positive result (success) or a negative outcome (stopping for futility).

“Error spending” designs can deal with unequal and unpredictable group sizes.

“Information monitoring” designs allow modification of sample size in response to information on nuisance parameters, to attain the desired power.

An efficient design will ***reduce average sample size*** or time to a conclusion while protecting the type I error rate and maintaining power.

There has been recent progress in group sequential designs for a delayed response, when there are treated subjects who have not yet responded at each interim analysis: see Lisa Hampson’s University of Bath PhD thesis.

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.

The excitement about adaptive methods

Claims of a new generation of statistical methods led to:

Hope that these new methods could reverse the trend of failures late in the development process,

Enthusiasm from companies for innovative statistical methods,

Even optimism from some regulators.

Statisticians were challenged to deliver on the expectations that had been raised.

This is in marked contrast to the practice of statisticians being reticent about their achievements — then disappointed their new methods go unused.

In this talk, I aim to assess how well we have met this challenge.

Adaptive applications

I shall discuss progress in the following areas:

1. Sample size re-estimation for a nuisance parameter,
2. Sample size modification to increase power,
3. Testing for both superiority and non-inferiority,
4. Switching to a patient sub-population,
5. Multiple objectives with hierarchical structure,
6. Adaptive dose finding designs,
7. Reducing “white space” between Phases II and III,
8. Combining data from Phases II and III,
9. Joint planning of Phases II and III.

A key tool: The combination test (Bauer & Köhne, 1994)

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 for testing H_0.

Design Stage 2 in the light of Stage 1 data.

Stage 2

Observe the P-value P_2 for testing H_0.

Clearly, P_1 has the usual $U(0,\,1)$ distribution under H_0.

Under H_0, $P_2 \sim U(0,\,1)$ conditionally — and hence unconditionally — on the Stage 2 design, and P_2 is independent of P_1.

The combination test

Since P_1 and P_2 are independent $U(0,\,1)$ variates under H_0, it follows that

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

Hence, the P-values can be combined in an overall test, rejecting H_0 if

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

Bauer & Köhne refer to this as the ***inverse χ^2 test*** (the test dates back to R. A. Fisher, 1932).

A similar approach combining two Z-statistics gives the ***inverse normal test***.

Lehmacher & Wassmer (1999) introduced an adaptive group sequential design combining Z-statistics from several stages — with the opportunity of re-design at each stage.

The inverse normal combination test

Stipulate this test will be used 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

Reject H_0 if $Z = w_1 Z_1 + w_2 Z_2 > z_\alpha$.

Modifications before Stage 2 may be “flexible” or follow a pre-specified rule.

Problem 1: Sample size re-estimation for a nuisance parameter

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

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

Initial design

Specify a Bauer & Köhne two-stage design using the inverse χ^2 combination test.

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 a nuisance parameter

Stage 1 . . .

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

One may substitute this value in (1) to get a new total sample size.

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

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 rejects H_0 if

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

This adaptive design modifies sample size to meet the power requirement

and maintains type I error rate exactly equal to α.

Sample size re-estimation for a nuisance parameter

Other methods have been proposed to solve this problem.

Wittes & Brittain (*Statistics in Medicine*, 1990) introduced “internal pilot” procedures.

Mehta & Tsiatis (*Drug Information Journal*, 2001) presented an “information monitoring” approach to use with error spending group sequential tests.

Since these methods may slightly inflate the type I error rate, the exact control offered by combination tests is welcome.

Achieving adequate power is a crucial requirement for any study:

tackling this issue, with whatever method, is a key step forward.

Problem 2. Sample size modification to increase power

Optimistic investigators may design a trial with power at a large treatment effect.

What if interim data suggest a smaller effect size — still clinically important but hard to demonstrate with the chosen sample size?

A Bauer & Köhne two-stage design allows sample size, and hence power, to be increased after the first stage of the trial.

- Advantage

  Adaptive methodology allows ***rescue*** of an under-powered study.

- Disadvantages

  It is more transparent to design for the power curve that is really desired.

  Then, a ***group sequential design*** can be used to provide early stopping, and save sample size, when the effect size is large.

Sample size modification to increase power

Many papers have appeared on adapting sample size to increase power.

These methods can be thought of as an extension of group sequential tests:

> *At each analysis, a decision is made to stop to reject or accept H_0, or to continue sampling,*
>
> ***In addition, the next group size can be chosen based on current data.***

There is intuitive appeal to adaptation of group size, e.g., taking smaller group sizes when close to a stopping boundary.

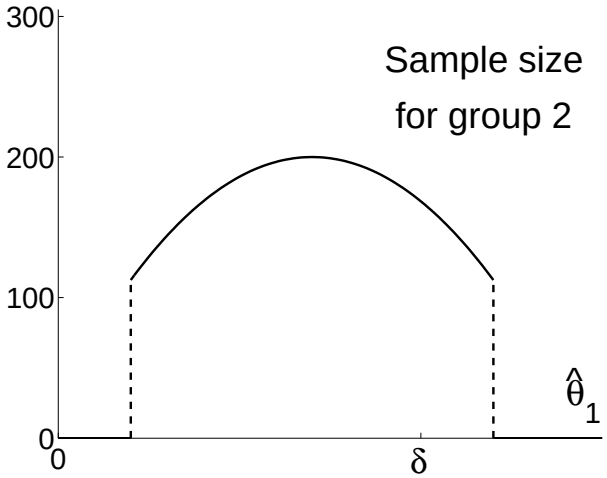
Such designs were proposed by Schmitz (1993) as “Sequentially Planned Sequential Tests”.

Jennison & Turnbull (*Biometrika*, 2006) derived optimal versions of these designs and found only small gains (around 1 or 2 per cent) over efficient non-adaptive group sequential tests with the same number of analyses.

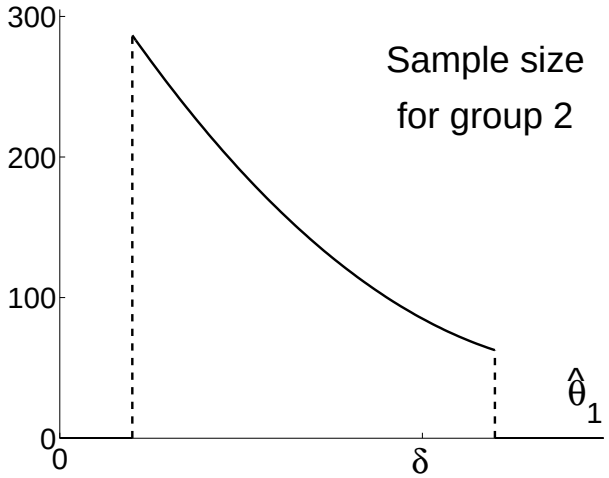
Sample size modification to increase power

In many proposals for adaptive designs, sample size is modified to give a specific conditional power, often at an effect size θ equal to the current estimate $\hat{\theta}$.

This is quite different from the optimal rules we have found.



Optimal sample size rule



Conditional power rule

Consequently, “conditional power” designs are less efficient than standard group sequential designs — by as much as 20 or 30 per cent.

Sample size modification to increase power

The adaptive debate has drawn attention to the benefits of group sequential designs — and these are starting to be labelled as “adaptive” designs.

We recommend group sequential designs with maximum sample size high enough to achieve power at small, but clinically significant, treatment effects.

If the treatment effect is larger, a group sequential design will stop early after only a fraction of this planned sample size.

Adaptive rules for modifying group sizes pose logistical difficulties: if the small percentage gain in efficiency is important, one can simply use a group sequential design with more frequent analyses.

There is no need to adapt for power if the correct power requirement is specified at the outset.

Problem 3: Testing for both superiority and non-inferiority

Other adaptive designs can involve a change in the hypothesis being tested.

Suppose a study is designed to prove superiority of a new treatment against an active control, testing

$$\theta \leq 0 \quad \text{vs} \quad \theta > 0.$$

If this outcome starts to look unlikely, attention may shift to showing the new treatment is non-inferior, testing

$$\theta \leq -\delta \quad \text{vs} \quad \theta > -\delta.$$

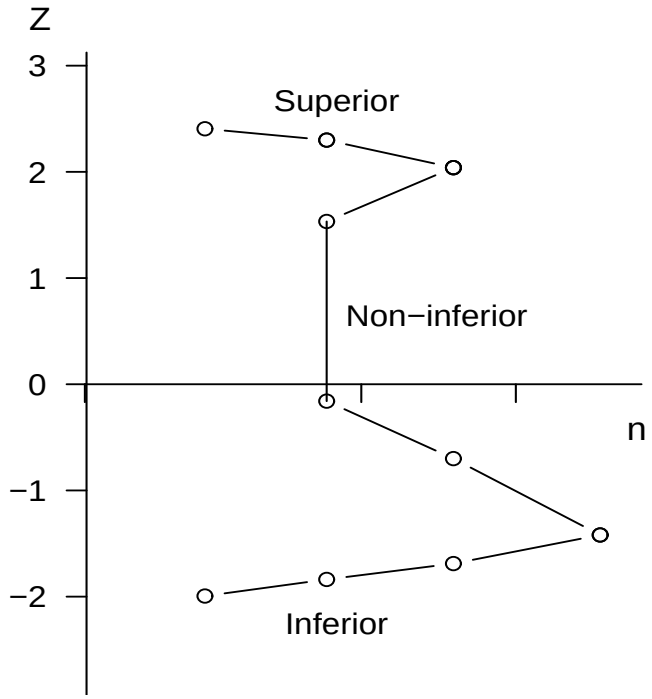
Power for the test of superiority is often set assuming a fairly large treatment effect, while the distance between hypotheses in the test of non-inferiority is smaller.

Since the test of non-inferiority requires a larger sample size, a change in objective is accompanied by a change in sample size.

See: Wang et al. (*Statist. in Medicine*, 2001), Shih et al. (*Statist. in Medicine*, 2004), Koyama et al. (*Statist. in Medicine*, 2005).

Testing for both superiority and non-inferiority

Another option is to employ a non-adaptive group sequential design with three possible decisions on termination.



Such designs are able to achieve a low expected sample size for each possible value of the treatment effect (Öhrn & Jennison, *Statist. in Medicine*, to appear).

A common task: Testing multiple hypotheses

Adaptation 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

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

so one can conclude at least one of the $H_{0,i}$ does not hold.

We shall focus on the second approach:

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.

Problem 4. 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

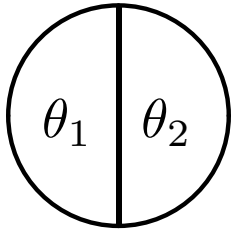
Restrict the remainder of the study to the defined sub-population.

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.,

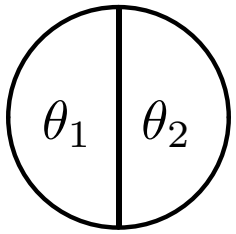
$$Pr\{\text{Reject any true } H_i\} \leq \alpha \quad \text{for all } (\theta_1, \ldots, \theta_k).$$

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.

The stage 1 test of H_{01} uses a combination of the Z-statistics for H_0 and H_1.

Simulation results: Power of non-adaptive and adaptive designs

				Non-adaptive	***Adaptive***		
	θ_1	θ_2	θ	*Full pop^n*	*Sub-pop 1 only*	*Full pop^n*	*Total*
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.

Problem 5. Even more treatments and endpoints

Some clinical trials have multiple objectives.

Investigators may hope to establish a drug is effective by

showing either superiority or non-inferiority,

for a range of endpoints,

at several doses.

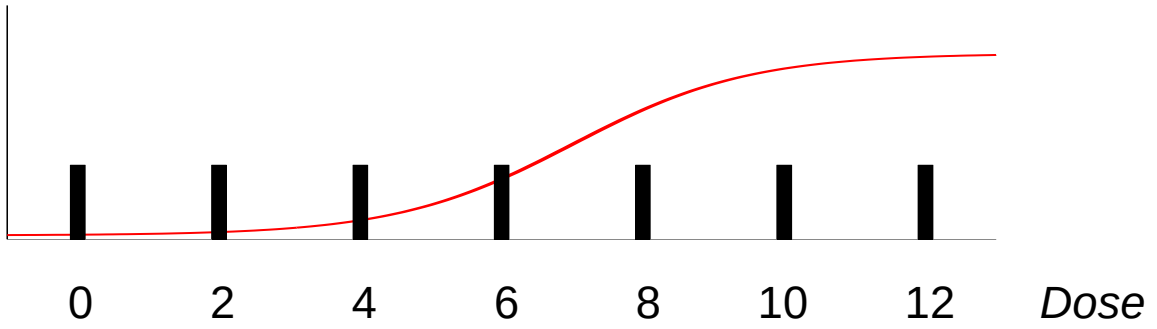
Trial design may include adaptation as answers for a hierarchy of aims arise during the course of a trial. As ever, it is essential to control the overall type I error rate.

Bretz et al. (*Statist. in Medicine*, 2009) and Burman et al. (*Statist. in Medicine*, 2009) describe a very usable graphical approach to “gatekeeping” and other multiple hypothesis testing procedures for hierarchical objectives.

Problem 6: Adaptive dose finding designs

Phase II clinical trials investigate dose-response, with the aims of establishing "proof of concept" and finding the best dose for a new drug.

A parallel group design allocates an equal number of subjects to each dose level:



Little information is gained from observations at:

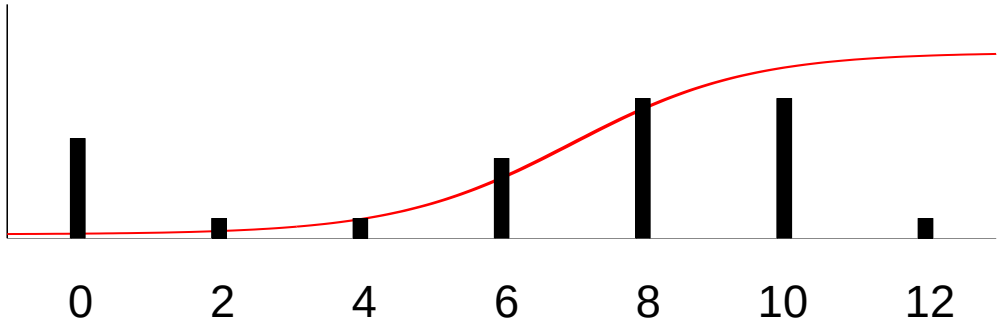
low doses with a very low response rate,

high doses at which the dose response curve has already levelled off.

Ideally, observations should be at dose levels in the "interesting part of the curve", with a proportion at dose zero for the control comparison.

Adaptive dose finding designs

An optimal allocation rule can be derived for a known dose response curve.



In reality, the dose response curve is unknown and an adaptive approach is needed.

The ASTIN trial (Krams et al., *Stroke*, 2003) used ***Bayesian adaptive allocation.***

Adaptive D-optimal designs can be used to learn about the entire dose response relationship, while ***c-optimal designs*** focus on a specific objective, e.g., finding the minimum dose that delivers 90% of the maximum effect (the ED90).

See, e.g., Dragalin et al., *Journal of Biopharmaceutical Statistics*, 2007.

Adaptive dose finding designs

For a review of work in this area, see the “PhRMA White Paper” by Bornkamp et al. (*Journal of Biopharmaceutical Statistics*, 2007) and accompanying discussion.

Research continues with the goal of finding adaptive designs that are effective yet simple to implement (both mathematically and logistically).

Designs should be robustly efficient across a variety of response curve shapes.

Efficiency can be defined relative to a chosen objective, e.g., identifying the ED90 with high accuracy.

Adaptive sampling is a natural choice in Phase II dose finding studies.

The role of this stage in the overall development process should motivate the choice of objective against which efficiency is sought.

Seamless Phase II/III designs

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.

Problem 7: 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 sample size 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.

Problem 8: 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.

Problem 9: Joint planning of Phases II and III

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.

So, where are we now?

There is plenty of progress to report in many areas.

Experience of implementing methods has informed research directions, generated new problems, and widened the scope of adaptive methodology.

I see a general trend towards planned adaptation according to pre-specified rules, rather than a free-for-all, flexible approach.

There is no doubt that statisticians have started to deliver on the promise of innovative designs to enhance the drug development process.

New forms of trial design are being used and experience gained. Further guidance from regulators will focus work on specific types of design.