Clinical trials: Past, present and future

Christopher Jennison

Dept of Mathematical Sciences,

University of Bath, UK

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

AstraZeneca, Advanced Analytics Centre

Symposium

Alderley Edge, January 2015

Outline of talk

1. The traditional approach to Phase III clinical trial design

   Keeping Phase III trials simple

   Early adaptive methods and group sequential designs

2. Recent developments in adaptive Phase III designs

   Combining data across stages

   Testing multiple hypotheses

3. Case studies

   A clinical trial with a survival endpoint and treatment selection

   An adaptive enrichment trial

4. Overarching design of Phase II and Phase III trials

1. The role of a Phase III clinical trial

Phase III trials are conducted at the end of the drug development process, or the development of a new medical treatment. Then,

The treatment has been refined and tested in earlier development and in Phase I and II trials,

A substantial body of work supports the investigators' belief that the new treatment is effective and safe.

The aim of the Phase III trial is to compare the new treatment with the current standard treatment or a placebo, when given to the target patient population.

The need for a clear, unambiguous comparison leads to the desire for a simple Phase III clinical trial.

All aspects of the Phase III trial design are pre-defined and written into the protocol and statistical analysis plan.

Traditional Phase III clinical trials (pre 2000 approx.)

A trial protocol specifies:

The experimental treatment and the control or placebo treatment,

The patient population (eligibility criteria, etc.),

Sample size for the trial,

Statistical analysis plan.

Interim analyses may be conducted to:

Monitor safety,

Stop early for futility if the new treatment is not effective,

Stop early if there is overwhelming evidence of efficacy.

Many of the decisions taken in creating such a design would benefit from further knowledge of the treatment, the patients, or patient responses.

Early examples of adaptive methods

There is a long history of “Adaptive” statistical methods.

Adaptive randomisation

In a trial comparing two treatments, adaptive randomisation can be used to increase the proportion of patients allocated to the better of two treatments.

However, once randomisation becomes unequal, ethical issues may arise as to whether it is permissible to randomise at all.

Adaptive randomisation highlights the role of “equipoise” in a randomised clinical trial and just what this term should mean.

Ethical and statistical concerns were clearly evident in two Harvard trials in the 1970s and 1980s which investigated ECMO treatment of critically ill, new-born babies (Ware, *Statistical Science*, 1989).

Similar issues, and design solutions, are emerging in the context of rare diseases.

Early examples of adaptive methods

Sample size re-estimation

The sample size needed to achieve a specific power under a given treatment effect is proportional to the response variance — which is typically unknown when planning a trial.

Wittes & Brittain (*Statistics in Medicine*, 1990) suggested choosing an initial sample size based on a plausible response variance, then updating the sample size as better estimates of response variance are obtained.

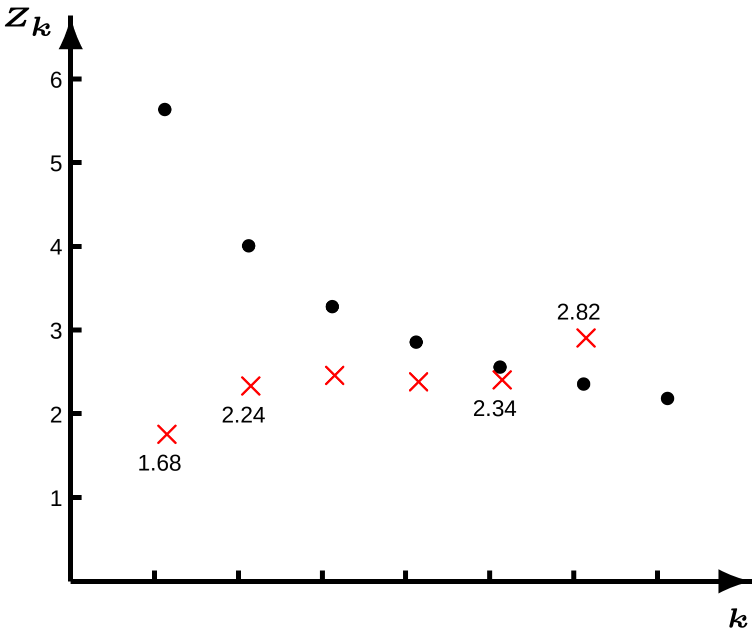
The same approach can be used to handle an unknown baseline hazard rate for survival data.

Sample size re-estimation in the light of estimates of “nuisance parameters” is still one of the most commonly used adaptive methodologies.

Sequential analysis of clinical trials

Group sequential methods, introduced in the late 1970s, allow early stopping for either a positive or negative final decision.

An early example, the the Beta-Blocker Heart Attack Trial, compared propanolol with placebo. (DeMets et al., *Controlled Clinical Trials*, 1984)



The trial stopped with a positive outcome after the 6th of 7 planned analyses.

Sequential analysis of clinical trials

In their book *Group Sequential Methods with Applications to Clinical Trials*, Jennison & Turnbull (2000) gave a unified treatment of group sequential methods, including:

General theory of group sequential analysis

Early stopping for futility or for a positive outcome

Survival data

Error spending designs that adapt to unpredictable information levels

Sample size re-estimation as nuisance parameters (but not the treatment effect) are estimated

Multiple endpoints or multiple treatments

Although some key papers started to appear in the mid 1990s, Jennison & Turnbull did not discuss modern adaptive methods.

2. Adaptive designs for Phase III clinical trials

We noted that many of the decisions taken in designing a clinical trial would benefit from further knowledge of the treatment, the patients, or patient responses.

What is the best dose for the new treatment?

What is the best method of delivery for the new treatment?

Does the treatment have greater benefit for a sub-population of patients?

For a normally distributed response, what is the variance?

Or, for time-to-event data, what is the baseline hazard rate?

How large a treatment effect is clinically significant?

How large a treatment effect is anticipated?

Such questions are addressed throughout the development of a new treatment.

Adaptive designs allow final changes to be made as new information is gathered during a Phase III trial.

Adaptive designs for Phase III clinical trials

In the early 2000s, industry and regulators were aware of falling success rates in late stage trials — a “statistical” solution would be very welcome indeed!

The idea of shifting from rigidly defined Phase III clinical trials to a flexible, adaptive approach was both attractive and challenging.

Sceptics asked:

Can the results of an adaptive trial be statistically valid and credible?

Will regulators accept adaptive designs?

What features of a trial should be adapted?

What are the benefits of adaptation?

Some proposals seemed to violate fundamental statistical principles.

There was a need for critical appraisal of new methodologies.

The statistical building blocks of adaptive clinical trials

(i) Testing a null hypothesis by combining data across stages

A key piece of methodology for hypothesis testing in adaptive designs is the ***Combination test*** (Bauer & Köhne, *Biometrics*, 1994).

Initial design

Define the null hypothesis, H_0: $\theta \le 0$, and say a combination test will be used.

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

Stage 1

Observe P_1, the one-sided P-value for testing H_0 based on Stage 1 data.

Design Stage 2 in the light of Stage 1 data.

Stage 2

Observe P_2, the one-sided P-value for testing H_0 based on Stage 2 data.

Under $\theta = 0$: $P_1 \sim U(0,1)$, $P_2 \sim U(0,1)$, and P_1 and P_2 are independent.

Bauer & Köhne’s inverse $\boldsymbol{\chi}^{\mathbf{2}}$ combination test

The inverse $\boldsymbol{\chi}^{\mathbf{2}}$ test rejects H_0 for low values of $P_1\, P_2$.

If $P \sim U(0,1)$, then $-\ln(P) \sim$ Exp $(1) = \frac{1}{2}\,\chi^2_2$.

Thus, under $\theta = 0$,

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

Combining the two P-values in an overall test, we reject H_0 if

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

Despite the data-dependent adaptation, the overall type I error rate is still protected at level $\boldsymbol{\alpha}$ ***under*** $\boldsymbol{H_0}$***.***

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

Methods for combining data across stages of an adaptive trial

Other forms of combination test (Bauer & Köhne, 1994) are available, such as the “inverse normal” combination rule.

Or, methods can be based on preserving the conditional type I error probability:

Proschan & Hunsberger (*Biometrics*, 1995)

Denne (*Statistics in Medicine*, 2001)

Müller & Schäfer (*Biometrics*, 2001 and *Statistics in Medicine*, 2004)

L. D. Fisher (*Statistics in Medicine*, 1998) proposed a “variance spending” approach.

Adaptation can occur in a group sequential clinical trial:

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

Lehmacher & Wassmer (*Biometrics*, 1999)

Despite their varied descriptions and derivations, there is much in common between all of these methods.

Adaptive designs using a combination test

Let θ denote the treatment effect, e.g., the difference in mean response between patients on the new treatment and patients on control.

We test H_0: $\theta \leq 0$ vs $\theta > 0$, where $\theta > 0$ means the new treatment is superior.

A combination test may be used when sample size is re-estimated in response to a new estimate for a nuisance parameter, or an estimate of θ itself.

A combination test safeguards the overall type I error rate when sample size is re-estimated.

Jennison & Turnbull have noted that sample size re-estimation in response to estimates of θ has much in common with use of a group sequential stopping rule — the estimates, $\hat{\theta}_k$, at analyses $k = 1,\ 2,\ \ldots$, determine the final sample size.

Since efficient and well-understood group sequential designs are already available, there is no need to create adaptive designs to achieve the same goal.

The statistical building blocks of adaptive clinical trials

(ii) Testing multiple hypotheses

Adaptive designs have opened new horizons for trials in which **several null hypotheses** may be tested — and the design of the trial can be modified to align with the hypothesis of interest.

There can be a variety of reasons to change the null hypothesis or choose one (or more) from a set of possible null hypotheses:

Selecting one out of several versions of a treatment,

Restricting to a sub-group of the patient population,

Switching from a test of superiority to a test of non-inferiority.

When H_0 changes or is selected, attention focuses on the ***new*** null hypothesis. Care is needed to avoid ***selection bias*** as this hypothesis is data-generated.

Methods must protect the type I error rate when there are multiple hypotheses.

Testing multiple hypotheses

The familywise error rate

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

A procedure's ***familywise error rate*** under a set of values $(\theta_1, \ldots, \theta_h)$ is

$$P\{\text{Reject } H_i \text{ for some } i \text{ with } \theta_i \leq 0\} \;=\; P\{\text{Reject any true } H_i\}.$$

The familywise 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

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

Using such a procedure, the probability of choosing to focus on a parameter θ_{i^*} and then falsely claiming significance for null hypothesis H_{i^*} is at most α.

Testing multiple hypotheses: Closed testing procedures

Marcus et al. (*Biometrika*, 1976) define a ***closed testing procedure*** which combines level α tests of each H_i and of intersections of these hypotheses.

We have null hypotheses H_i, $i = 1, \ldots, h$.

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

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

Construct a level α test of each intersection hypothesis H_I, i.e., a test which 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 overall if, and only if, H_I is rejected for every set I containing index j.

It can be show (quite easily) that this procedure provides strong control, at level α, of the familywise error rate.

Putting the building blocks together

Selected references: Bauer & Köhne (*Biometrics*, 1994), Bretz, Schmidli et al. (*Biometrical Journal*, 2006), Schmidli, Bretz et al. (*Biometrical Journal*, 2006).

Closed testing procedures can be used to test multiple hypotheses in a single stage, non-adaptive design.

One may wish to test hypotheses about secondary endpoints or patient sub-groups after obtaining a positive result on the primary endpoint.

Positive results will be included in the labelling of the new treatment.

When several null hypotheses arise in a group sequential or adaptive trial

In constructing a closed testing procedure, we need to define a combination test for each simple hypothesis and each intersection of simple hypotheses.

Each of these tests will combine data across stages of the trial.

The key requirement is that, for each hypothesis test, we stipulate how the P value will be computed from data in the next stage *before that stage is carried out*.

3. Case study 1: A clinical trial with a survival endpoint and treatment selection

Consider a trial of cancer treatments comparing

Experimental Treatment 1: Intensive dosing

Experimental Treatment 2: Slower dosing

Control treatment

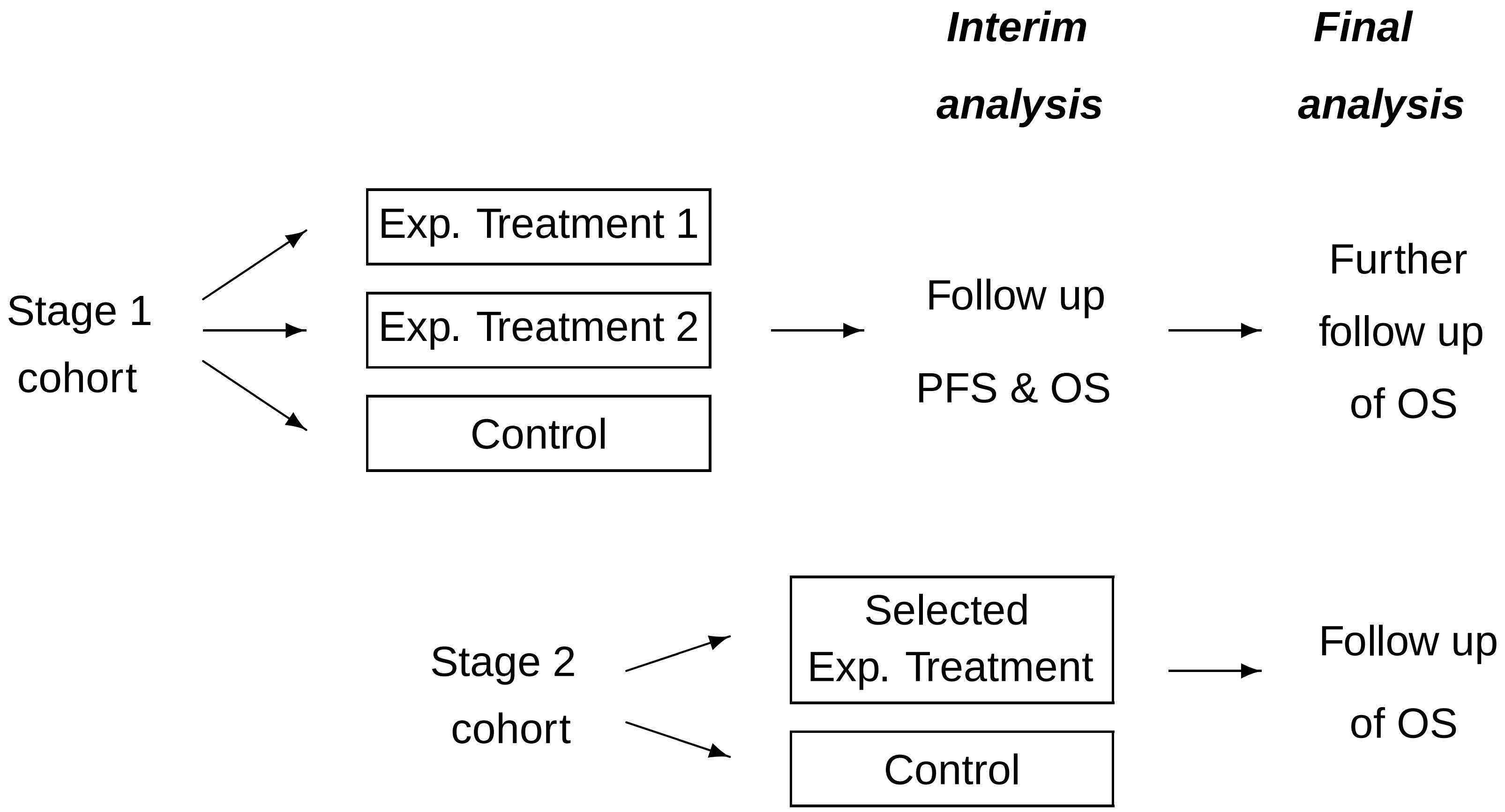
The primary endpoint is Overall Survival (OS).

Information on OS, Progression Free Survival (PFS) and safety will be used at an interim analysis to choose between the two experimental treatments.

Note that PFS is useful here as it is more rapidly observed.

After the interim analysis, patients will only be recruited to the selected treatment and the control.

Overall plan of the trial



Interim analysis

Final analysis

Stage 1 cohort

Exp. Treatment 1

Exp. Treatment 2

Control

Follow up

PFS & OS

Further follow up of OS

Stage 2 cohort

Selected Exp. Treatment

Control

Follow up of OS

At the final analysis, we test the null hypothesis that OS on the selected treatment is no better than OS on the control.

Protecting the type I error rate

We may assume a proportional hazards model with

$$\lambda_1 \;=\; \text{Hazard ratio, Control } \textit{vs} \text{ Exp. Treatment 1}$$

$$\lambda_2 \;=\; \text{Hazard ratio, Control } \textit{vs} \text{ Exp. Treatment 2}$$

$$\theta_1 = \log(\lambda_1), \quad \theta_2 = \log(\lambda_2).$$

We test null hypotheses

$$H_{0,1}\colon\ \theta_1 \le 0 \quad \text{vs} \quad \theta_1 > 0 \quad \textit{(Exp. Treatment 1 superior to control)},$$

$$H_{0,2}\colon\ \theta_2 \le 0 \quad \text{vs} \quad \theta_2 > 0 \quad \textit{(Exp. Treatment 2 superior to control)}.$$

We require

$$Pr\{\text{Reject any true null hypothesis}\} \;\le\; \alpha.$$

A closed testing procedure

Define level α tests of

$$H_{0,1}\colon\ \theta_1 \leq 0,$$

$$H_{0,2}\colon\ \theta_2 \leq 0$$

and of the intersection hypothesis

$$H_{0,12} \;=\; H_{0,1} \cap H_{0,2}\colon\ \theta_1 \leq 0 \text{ and } \theta_2 \leq 0.$$

Then:

Reject $H_{0,1}$ ***overall*** *if the above tests reject $H_{0,1}$ and $H_{0,12}$,*

Reject $H_{0,2}$ ***overall*** *if the above tests reject $H_{0,2}$ and $H_{0,12}$.*

The requirement to reject $H_{0,12}$ compensates for testing multiple hypotheses and the “selection bias” in choosing the treatment to focus on in Stage 2.

Combination tests

In the closed testing procedure, each null hypothesis is tested using a combination test to combine P-values from the two stages.

Overall survival data within each stage are analysed using a logrank test.

In testing the intersection hypothesis $H_{0,12}$: $\theta_1 \leq 0$ and $\theta_2 \leq 0$, P-values from logrank tests of Exp. Treatment 1 vs Control and Exp. Treatment 2 vs Control can be combined using, say, Simes' method or Dunnett's test.

There is an elegant theory for the behaviour of logrank statistics based on the increasing follow-up of a group of subjects (Tsiatis, *Biometrika*, 1981).

However, this theory may not be applicable in an adaptive trial (Bauer & Posch, *Statistics in Medicine*, 2004).

The problem can be solved by changing the definitions of "Stage 1" and "Stage 2" data (Jenkins, Stone & Jennison, *Pharmaceutical Statistics*, 2011).

Jenkins, Stone & Jennison (2011)

In constructing a combination test, it is natural to separate data into the parts accrued before and after the interim analysis:

	P_1	P_2
Stage 1 cohort	Overall survival (during Stage 1)	Overall survival (during Stage 2)
Stage 2 cohort		Overall survival (during Stage 2)

To avoid bias in a combination test, divide the data into parts from the two cohorts:

Stage 1 cohort	Overall survival (during Stage 1)	Overall survival (during Stage 2)	P_1
Stage 2 cohort		Overall survival (during Stage 2)	P_2

Assessing the adaptive design: Model assumptions

Overall Survival

	Log hazard ratio
Exp. Treatment 1 vs control	θ_1
Exp. Treatment 2 vs control	θ_2

Logrank statistics are correlated because of the common control arm.

Progression Free Survival

	Log hazard ratio
Exp. Treatment 1 vs control	ψ_1
Exp. Treatment 2 vs control	ψ_2

We suppose correlation between logrank statistics for OS and PFS $= \rho$.

Proportional hazards models for both endpoints are not essential (or reasonable?) — the implications for the joint distribution of logrank statistics are what matter.

Model assumptions

Log hazard ratios for OS: $\theta_1,\ \theta_2$.

Log hazard ratios for PFS: $\psi_1,\ \psi_2$.

We suppose

$\psi_1 = \gamma \times \theta_1$ and $\psi_2 = \gamma \times \theta_2$

Number of OS events for Stage 1 cohort $= 300$ (over 3 treatment arms)

Number of OS events for Stage 2 cohort $= 300$ (over 2 or 3* treatment arms)

Number of PFS events at interim analysis $= \lambda \times 300$.

*2 in the adaptive design, 3 in a non-adaptive design

From large sample theory, the standardised logrank statistic based on d observed events is, approximately,

$$N(\theta\sqrt{d/4}, 1)$$

when the log hazard ratio is θ.

Testing the intersection hypothesis

We have null hypotheses $H_{0,1}$: $\theta_1 \leq 0$ and $H_{0,2}$: $\theta_2 \leq 0$.

In the closed testing procedure we must also test the intersection hypothesis

$$H_{0,12} \ = \ H_{0,1} \cap H_{0,2}\text{:} \ \ \theta_1 \leq 0 \ \text{ and } \ \theta_2 \leq 0.$$

We shall use a ***Dunnett*** test to test the intersection hypothesis $H_{0,2}$.

Suppose P_1 and P_2 are the P-values for logrank tests of Exp. Treatment 1 vs control and Exp. Treatment 2 vs Control, so the corresponding normal deviates are

$$Z_1 = \Phi^{-1}(1 - P_1) \ \text{ and } \ Z_2 = \Phi^{-1}(1 - P_2).$$

If z_1 and z_2 are the observed values of Z_1 and Z_2, the ***Dunnett test*** of $H_{0,12}$ yields the P-value

$$P(\max(Z_1, Z_2) \geq \max(z_1, z_2))$$

where (Z_1, Z_2) are bivariate, standard normal with Corr$(Z_1, Z_2) = 0.5$.

Comparing adaptive and non-adaptive trial designs

Setting $\psi_1 = \theta_1,\ \psi_2 = \theta_2$ (i.e., $\gamma = 1$) with $\lambda = 1$ and $\rho = 0.6$, we simulated logrank statistics from their large sample distributions under the adaptive design.

For the adaptive design, we noted

$$\begin{aligned} P(1) &= P(\text{Select Treatment 1 and Reject } H_{0,1} \text{ overall}) \\ P(2) &= P(\text{Select Treatment 2 and Reject } H_{0,2} \text{ overall}) \\ E(\text{Gain}) &= \theta_1 \times P(1) + \theta_2 \times P(2). \end{aligned}$$

Here “Gain” represents a possible utility, in which the value of a positive outcome is proportional to the effect size of the recommended treatment.

For the non-adaptive design

Patients are randomised to both treatments and control throughout, with the same total sample size. We used a closed testing procedure to protect the familywise error rate and regarded the treatment with higher estimated effect as “selected”.

Comparing adaptive and non-adaptive trial designs

We compare designs using a Dunnett test for the intersection hypothesis, with

$$\psi_1 = \theta_1, \quad \psi_2 = \theta_2, \quad \rho = 0.6, \quad \alpha = 0.025.$$

		Non-adaptive			Adaptive		
θ_1	θ_2	$P(1)$	$P(2)$	$E($Gain$)$	$P(1)$	$P(2)$	$E($Gain$)$
0.3	0.0	0.78	0.00	0.235	0.86	0.00	0.259
0.3	0.1	0.78	0.01	0.234	0.82	0.02	0.247
0.3	0.2	0.70	0.11	0.234	0.69	0.16	0.238
0.3	0.25	0.60	0.26	0.244	0.58	0.30	0.249
0.3	0.295	0.47	0.43	0.267	0.47	0.44	0.274

The adaptive design has higher $P(1)$ when θ_1 is substantially greater than θ_2.

When θ_1 and θ_2 are closer, the adaptive design still has the higher $E($Gain$)$.

Comparing adaptive and non-adaptive trial designs

The adaptive design can only be effective if there is appropriate information to select the correct treatment at the interim analysis.

This requires that

Treatment effects on PFS are reliable indicators of treatment effects on OS,

Sufficient information on PFS is available at the time of the interim analysis.

For the case $\theta_1 = 0.3$, $\theta_2 = 0.1$, we have investigated varying the parameters γ and λ where

$\psi_1 = \gamma \times \theta_1$ and $\psi_2 = \gamma \times \theta_2$

Number of OS events for Stage 1 cohort $= 300$ (over 3 treatment arms)

Number of OS events for Stage 2 cohort $= 300$ (over 2 or 3 treatment arms)

Number of PFS events at interim analysis $= \lambda \times 300$.

Comparing adaptive and non-adaptive trial designs

We compare designs with $\theta_1 = 0.3, \quad \theta_2 = 0.1, \quad \rho = 0.6, \quad \alpha = 0.025,$

PFS log hazard ratios: $\psi_1 = \gamma\,\theta_1, \quad \psi_2 = \gamma\,\theta_2,$

Number of PFS events at interim analysis $= \lambda \times 300$.

		Non-adaptive			Adaptive		
γ	λ	$P(1)$	$P(2)$	$E($Gain$)$	$P(1)$	$P(2)$	$E($Gain$)$
1.5	1.2				0.88	0.00	0.264
1.2	1.0				0.85	0.01	0.256
1.0	**1.0**	**0.78**	**0.01**	**0.234**	**0.82**	**0.02**	**0.247**
0.9	0.9	for all γ and λ			0.78	0.03	0.238
0.8	0.8	(PFS is not used)			0.74	0.04	0.225
0.7	0.7				0.68	0.05	0.208

Adaptation works well if there is enough PFS information for treatment selection.

Conclusions from Case Study 1

1. The adaptive design offers the chance to select the better treatment and focus on this treatment in the second stage of the trial.

2. Overall, the adaptation is beneficial as long as there is sufficient information to make a reliable treatment selection decision.

   The challenge is to know the likely level of information when deciding whether to implement an adaptive design.

3. Other evidence could be used in reaching this decision:

   Safety data

   Pharmacokinetic data

   Overall survival

4. In addition to reaching a final decision, the adaptive trial compares the two forms of treatment — and the conclusions may be useful in other settings.

Case study 2. An adaptive population enrichment trial:

Switching to a sub-population in response to interim data

Consider a new treatment developed to disrupt a disease's biological pathway.

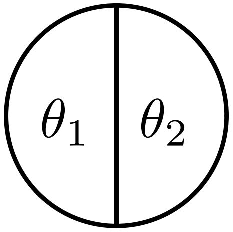
Patients with high levels of a biomarker for this pathway should gain particular benefit; the treatment's wider action may also help the broader patient population.

In a clinical trial with *enrichment* we

> Start by comparing the new treatment against control in the full population.
>
> Examine responses at an interim stage.
>
> If there is no evidence of treatment effect, stop for futility.
>
> If the new treatment appears effective in the full population, continue as before.
>
> If the new treatment appears to benefit just the subgroup, recruit only from the subgroup and increase the numbers in this subgroup.

Results may support a licence for the full population or just the sub-population.

Enrichment: Switching to a patient sub-population



Sub-population

(proportion λ_1)

Rest of the population

(proportion λ_2)

The treatment effect (difference in mean response between new treatment and standard) is θ_1 in the identified sub-population and θ_2 in the complement of this sub-population.

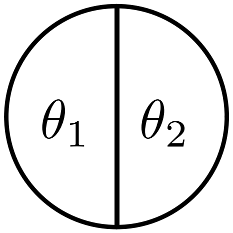
The overall treatment effect is $\theta_3 = \lambda_1\theta_1 + \lambda_2\theta_2$.

We may wish to test either or both of:

The null hypothesis for the full population, H_3: $\theta_3 \leq 0$ vs $\theta_3 > 0$,

The null hypothesis for the sub-population, H_1: $\theta_1 \leq 0$ vs $\theta_1 > 0$.

Enrichment: Example



Sub-population
(proportion λ_1)

Rest of the population
(proportion λ_2)

First, consider a design testing for a ***whole population effect***, $\theta_3 = \lambda_1\theta_1 + \lambda_2\theta_2$.

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

Sample size is set to achieve power 0.9 at $\theta_3 = 20$.

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

	Stage 1	*Stage 2*	*Overall*
H_3: $\theta_3 \leq 0$	$Z_{1,3}$	$Z_{2,3}$	$Z_3 = \frac{1}{\sqrt{2}} Z_{1,3} + \frac{1}{\sqrt{2}} Z_{2,3}$

Enrichment: Example

Two stage design testing for a whole population effect, θ_3.

	Stage 1	*Stage 2*	*Overall*
H_3: $\theta_3 \leq 0$	$Z_{1,3}$	$Z_{2,3}$	$Z_3 = \frac{1}{\sqrt{2}} Z_{1,3} + \frac{1}{\sqrt{2}} Z_{2,3}$

Decision rules:

If $Z_{1,3} < 0$		Stop at Stage 1, Accept H_3
If $Z_{1,3} \geq 0$		Continue to Stage 2, then
	If $Z_3 < 1.95$	Accept H_3
	If $Z_3 \geq 1.95$	Reject H_3

Enrichment: Example

Assume the sub-population comprises half the total population, so $\lambda_1 = \lambda_2 = 0.5$.

Properties of design for the whole population effect, θ_3:

θ_1	θ_2	θ_3	*Power for* H_3: $\theta_3 \leq 0$
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_3 is low but θ_1 may be higher, so it would be advantageous to switch resources to test only the sub-population?

Enrichment: A closed testing procedure

We wish to be able to consider two null hypotheses:

H_3: $\theta_3 \leq 0$ Treatment is not effective in the whole population,

H_1: $\theta_1 \leq 0$ Treatment is not effective in the sub-population.

Since $\theta_3 = 0.5\,\theta_1 + 0.5\,\theta_2$, either of H_1 and H_3 may be true on its own.

To apply a ***closed testing procedure*** (Marcus et al, *Biometrika*, 1976) we also need a test of the intersection hypothesis:

$$H_{13}: \quad \theta_1 \leq 0 \text{ and } \theta_3 \leq 0.$$

Then to reject H_1 overall, while protecting the family-wise type I error rate, we need to reject both H_1 and H_{13} in individual tests at significance level α.

Similarly, we can reject H_3 overall if both H_3 and H_{13} are rejected in level α tests.

Enrichment: An adaptive design

At Stage 1, if $\hat{\theta}_3 < 0$, stop to accept H_3: $\theta_3 \leq 0$.

If $\hat{\theta}_3 > 0$ 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_{13}.
Else,	Continue with full population and test H_3, needing to reject H_3 and H_{13}.

The same *total* sample size for Stage 2 is retained in both cases, increasing the numbers for the 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

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

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

	Stage 1	*Stage 2*
H_3	$Z_{1,3}$	$Z_{2,3}$
H_{13}	$Z_{1,3}$	$Z_{2,3}$

where $Z_{i,3}$ is based on $\hat{\theta}_3$ from responses in Stage i.

So, there is no change from the original test of H_3.

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 the sub-population, we use:

	Stage 1	*Stage 2*
H_1	$Z_{1,1}$	$Z_{2,1}$
H_{13}	$Z_{1,3}$	$Z_{2,1}$

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

The need to reject the intersection hypothesis H_{13} 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_3	*Full pop^n*	*Sub-pop^n only*	*Full pop^n*	*Total*
1.	30	0	15	**0.68**	0.43	0.42	**0.85**
2.	20	0	10	**0.37**	0.24	0.26	**0.51**
3.	20	20	20	**0.90**	0.03	0.87	**0.90**
4.	20	10	15	**0.68**	0.11	0.60	**0.71**

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

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

Case 4: Adaptation improves overall power a little.

Using a “gain function” to compare trial designs

In assessing the possible benefits of an adaptive design, it is helpful to specify a “gain function” which represents the perceived benefit from each trial outcome.

The “gain function” can also be used in creating an efficient trial design.

A company might consider “gain” to be proportional to the number of patients likely to receive their new treatment after a successful Phase III trial.

When Sub-population 1 makes up a proportion λ_1 of the total population, set

$$G_1 = \begin{cases} k & \text{if } H_3 \text{ is rejected,} \\ \lambda_1 k & \text{if only } H_1 \text{ is rejected,} \\ 0 & \text{otherwise,} \end{cases}$$

where k reflects the size of the target population and the income generated per patient treated.

Expected gain for non-adaptive and adaptive designs

With gain function G_1 based on numbers of patients treated if the new treatment is successful:

			Non-adaptive		***Adaptive***		
θ_1	θ_2	θ_3	*Reject* H_3	$E(G_1)$	*Reject* H_1 *only*	*Reject* H_3	$E(G_1)$
30	0	15	0.68	**0.68** k	0.43	0.42	**0.64** k
20	0	10	0.37	**0.37** k	0.24	0.26	**0.38** k
20	20	20	0.90	**0.90** k	0.03	0.87	**0.89** k
20	10	15	0.68	**0.68** k	0.11	0.60	**0.66** k

The adaptive design does not look particularly good in terms of G_1.

But note this gain function penalises the adaptive design for only rejecting H_1 when $\theta_2 = 0$ and the new treatment only has benefit for Sub-population 1.

Specification of the “Gain function”

We could define “gain” to be proportional to the benefit received by the patient.

So, for example, if $\theta_2 = 0$, patients in Sub-population 2 receive no benefit and make no contribution to “gain” when H_3 is rejected.

This leads to a gain function of the form

$$G_2 = \begin{cases} (\lambda_1\theta_1 + \lambda_2\theta_2)\, c & \text{if } H_3 \text{ is rejected,} \\ \lambda_1\theta_1\, c & \text{if only } H_1 \text{ is rejected,} \\ 0 & \text{otherwise,} \end{cases}$$

where c reflects the size of the target population and the income generated per unit of improved response, per patient treated.

From a company perspective, this function reflects the fact that a new drug is more likely to be adopted if physicians see evidence of its efficacy when they prescribe it.

Also, being able to claim a new drug is proven effective for a particular patient group may help in a competitive market.

Expected gain for non-adaptive and adaptive designs

With gain function G_2 based on numbers of patients treated and the improvement in their responses, if the new treatment is successful:

			Non-adaptive		***Adaptive***		
θ_1	θ_2	θ_3	*Reject* H_3	$E(G_2)$	*Reject* H_1 *only*	*Reject* H_3	$E(G_2)$
30	0	15	0.68	**10.2** c	0.43	0.42	**12.8** c
20	0	10	0.37	**3.7** c	0.24	0.26	**5.0** c
20	20	20	0.90	**18.0** c	0.03	0.87	**17.7** c
20	10	15	0.68	**10.2** c	0.11	0.60	**10.1** c

If the gain function G_2 is deemed an appropriate choice, the benefits of the adaptive design are now clear.

Optimising a design depends on a clear specification of the perceived benefits of the possible trial outcomes.

Conclusions from Case Study 2

1. The adaptive design offers the chance to modify recruitment in the second stage of the trial to pursue the most promising part of the patient population.

2. Overall, the adaptation can be beneficial — although whether or not this is the case will depend on how the sponsors view the different possible trial outcomes.

3. More generally, the process of designing an adaptive trial relies on a clear understanding of the value to sponsors and/or patients of different trial outcomes.

4. An overarching approach to designing sequential Phase II and Phase III trials

It hardly needs saying that phases of drug development occur sequentially.

While there is a great deal of work on “optimising” phases individually, much less attention has been devoted to optimising the overall development process.

I have been involved in work on this topic through a DIA (formerly PhRMA) Working Group, chaired by Carl-Fredrik Burman.

We have found that optimisation of the Phase II/Phase III part of the process is possible, given sufficient information.

I shall briefly summarise the input to such an optimisation problem and the conclusions we have reached.

The work I discuss here concerns dose-finding in a late Phase II or “Phase IIb” trial.

In his PhD thesis, Fredrik Öhrn investigated designs where a Phase II trial, based on a short -term endpoint, can determine whether to conduct the Phase III trial.

Joint design of Phase II and Phase III trials

Elements of the Phase IIb / Phase III process are

1. A dose response model,
2. A prior distribution for model parameters that reflects investigators' expectations
3. A model for the risk at each dose of losing the drug due to poor safety results,
4. Models for Phase IIb and Phase III response data,
5. The final decision rule that will determine whether investigators are able to reject the null hypothesis of no treatment effect at the selected dose,
6. The rule for:

   (i) Deciding whether to proceed to Phase III and, if so,

   (ii) Selecting the dose to test in Phase III and

   (iii) Choosing the Phase III sample size.

Joint design of Phase II and Phase III trials

The key challenge is to optimise the rule (6) for making decisions after observing Phase IIb data.

We have followed a Bayesian approach to account for uncertainty about model parameters and to integrate the average gain over a plausible set of scenarios.

In addition, we require information about:

The gain function after a successful final outcome,

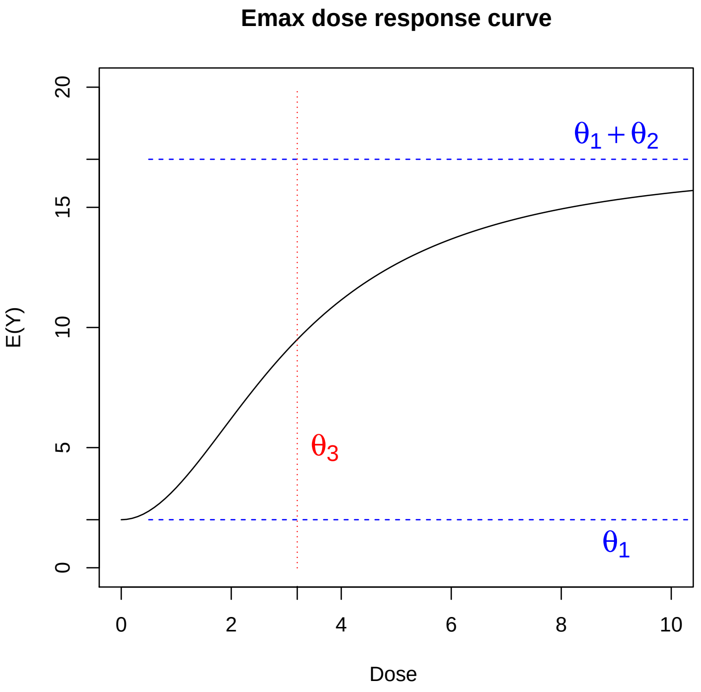
Sampling costs for Phase IIb and Phase III trials.

With all of the above in place, a substantial computational task remains.

The Emax dose response model

We assumed a 4 parameter Emax dose-response model, with mean response at dose d

$$\mu(d) \;=\; \theta_1 \,+\, \theta_2 \;\; \frac{d^{\,\theta_4}}{\theta_3^{\,\theta_4} + d^{\,\theta_4}}\;.$$



θ_1: Mean response at dose zero (placebo effect)

θ_2: Increase in mean response from dose zero to a very high dose

θ_3: ED50, the dose achieving half this maximum increase

θ_4: Governs the steepness of the dose response curve

We placed independent normal priors on the parameters θ_1, θ_2, θ_3 and θ_4.

Phase IIb and Phase III trial designs and response distributions

We assumed a normal response distribution for patient response in Phase IIb, and the same response distribution in Phase III.

In Phase IIb, patients are allocated equally to each of 7 active doses and at 3 times this rate to dose zero.

If it is decided to test dose j against control in Phase III, then two Phase III trials are conducted.

In each trial, we test the null hypothesis, H_{0j}, which states that the new treatment at dose j offers no improvement over the current treatment.

If H_{0j} is rejected at a significance level below $\alpha = 0.025$ in both trials, efficacy of the new treatment at dose j is established.

Gain function and sampling costs

We suppose a positive outcome in Phase III leads to approval of the new drug and a financial gain g.

Running the Phase IIb trial incurs a sampling cost of c_2 per subject.

Running Phase III incurs a cost of c_3 per subject.

In our example, we took

$$c_2 = 1,$$

$$c_3 = 1,$$

$$g = 12{,}000.$$

The meaning of 1 cost or gain unit may be \$10,000 to \$50,000, depending on the condition being investigated — so g represents a multi-million dollar return.

Risk of failure for safety

Suppose the probability that dose d will eventually fail on safety grounds is $\gamma(d)$.

This could occur in Phase III or later on in post-marketing surveillance.

We assume $\gamma(d)$ is a known, increasing function of d.

The function $\gamma(d)$ is specified before Phase IIb and patient follow-up in Phase IIb is not long enough to learn more about the safety profile.

In our example, we took $\gamma(d)$ to be quadratic with $\gamma(7) = 0.2$. Thus, the risk for dose j is

$$\gamma_j \;=\; (j/7)^2 \times 0.2.$$

When Phase III has a positive outcome, we calculate the expected gain by discounting the gain function by a factor $1 - \gamma_j$.

Optimising the Phase IIb / Phase III design

Before Phase IIb

We choose the Phase IIb sample size, n_2.

At the end of Phase IIb

We decide whether to proceed to run Phase III and, if so, select

The dose to test in Phase III	d_j,
The Phase III sample size	n_3.

We wish to optimise:

The choice of n_2,

The rule for deciding whether to proceed to Phase III,

The rule for choosing d_j and n_3.

Optimisation algorithm

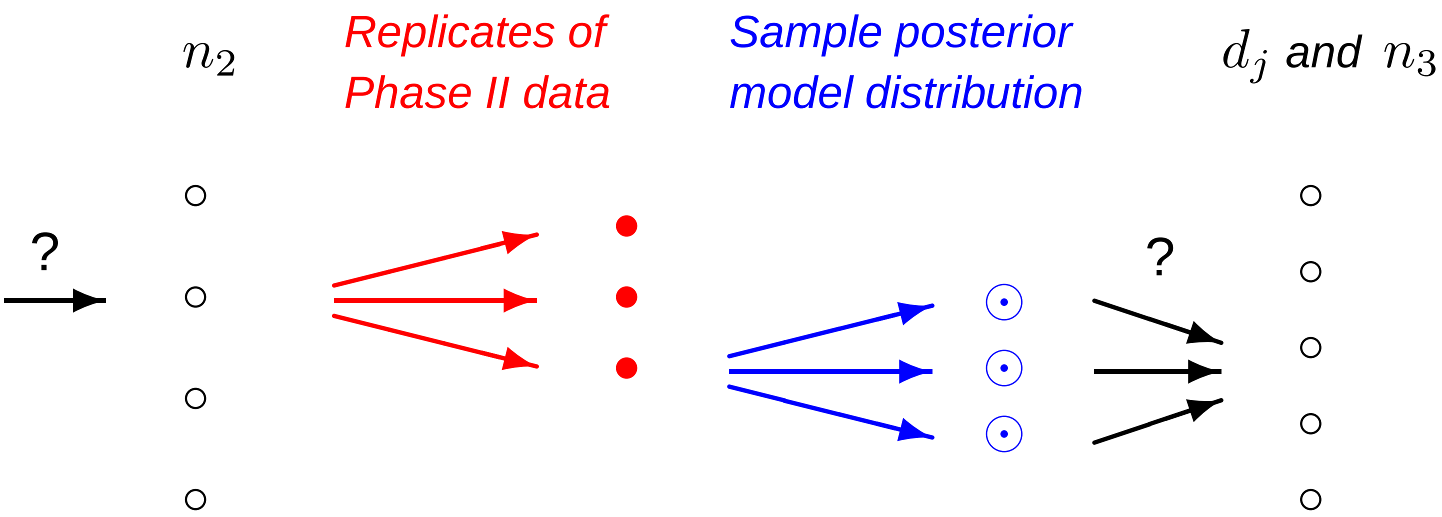
For a particular n_2:

We simulate $\boldsymbol{\theta}$, the vector of dose response curve parameters, from the prior.

Simulate Phase IIb data, given $\boldsymbol{\theta}$.

Evaluate Phase III options given the posterior for $\boldsymbol{\theta}$ and choose the best option.

Average over replicates to compute the expected net gain for this n_2.


n_2
Replicates of Phase II data
Sample posterior model distribution
d_j *and* n_3
?
?

Compare $E(\text{Net gain})$ over possible choices of n_2 and choose the best n_2.

Conclusions on the joint design of Phase II and Phase III trials

A full treatment of the Phase IIb/ Phase III design process is possible, with joint optimisation of both stages under a Bayesian model.

The result is guidance on Phase IIb sample size, and how to plan the Phase III trial given Phase II data.

The Bayesian approach allows propagation of uncertainty and provides a natural framework for decision making under uncertainty.

A clear conclusion is the ***benefit of applying group sequential Phase III designs:***

With group sequential Phase III trials, it is not so important to have an accurate estimate of the treatment to inform the choice of Phase III sample size.

Hence, the Phase IIb sample size only needs to be large enough to make a good choice of dose.

There are many directions in which to elaborate the problem we have studied.

However, a major challenge is eliciting the information needed to give a complete formulation of the optimisation problem.

Conclusions

1. Experience over 50 or 60 years led to reliable and well understood methodology for conducting Phase III trials.
2. Innovation had taken place but its potential impact was most probably lessened by cautious investigators and conservative regulators.
3. At the turn of the millennium, the announcement of adaptive methods caused great excitement and offered hope of a step change in the success rate of clinical trials.
4. We now have

   Clearer appraisal of the benefits of adaptive design,

   Practical experience of conducting adaptive trials.
5. Common requirements for further developments are:

   Statements of likely models with an assessment of uncertainty,

   A willingness of decision makers to quantify costs and likely benefits.