Designing an Adaptive Clinical Trial with Treatment Selection and a Survival Endpoint: Reflections on the GATSBY Study

Christopher Jennison

Dept of Mathematical Sciences, University of Bath, UK

http://cjennison1.github.io

Institut für Biometrie und Klinische Epidemiologie,

Charité – Universitätsmedizin Berlin

November 2026

Background information

Consider a Phase 3 confirmatory trial comparing a new treatment and a control.

Prior to around 2000, the general philosophy was to

- Specify everything:
  - Patient population,
  - Treatment,
  - Primary endpoint,
  - Sample size,
  - Method of analysis
- Conduct the trial as planned

Around 2000, adaptive trials started to become fashionable. With adaptation, one may re-visit initial plans in the light of trial data.

Background information

ICH E20: The International Council on Harmonisation has issued a draft Guideline on Adaptive Designs for Clinical Trials.

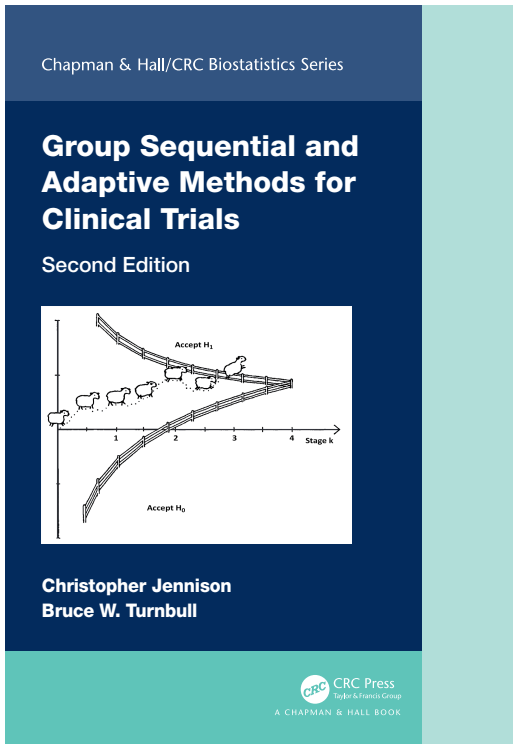
This sets out five key principles that adaptive trials should follow — but it does not stipulate particular methods to achieve this.

Many research papers and several books describe methods that can meet some of the ICH E20 guidelines.

Existing methods may have to be applied in particular ways

Some methods may need further development to satisfy E20 principles.

See my recently published book with Bruce Turnbull and references therein.



Background information

Roche's GATSBY trial (2012 to 2015) compared treatments for patients with HER2-positive advanced gastric cancer.

Initially, there were three treatment arms:

- Trastuzumab Emtansine (T-DM1), high dose every 3 weeks,
- Trastuzumab Emtansine (T-DM1), lower dose once a week,
- Control arm, Taxane

At an interim analysis, the independent Data Monitoring Committee (iDMC) selected one of the T-DM1 treatment arms.

Subsequent patients were randomised between the selected arm and control.

At a further interim analysis, the trial could stop futility.

Outline of talk

1. A study with a survival endpoint and treatment selection
2. Protecting the type I error rate in an adaptive design

   A closed testing procedure

   Combination tests
3. Properties of log-rank statistics
4. Applying a combination test to survival data
5. Avoiding error rate inflation in an adaptive trial

   (Jenkins, Stone & Jennison, *Pharmaceutical Statistics*, 2011)
6. Properties of the proposed adaptive design
7. Experiences as an iDMC member for adaptive trials
8. Conclusions

1. A survival study with treatment selection

Consider a Phase 3 trial of cancer treatments comparing

Experimental Treatment 1: Intensive dosing

Experimental Treatment 2: More frequent lower doses

Control treatment

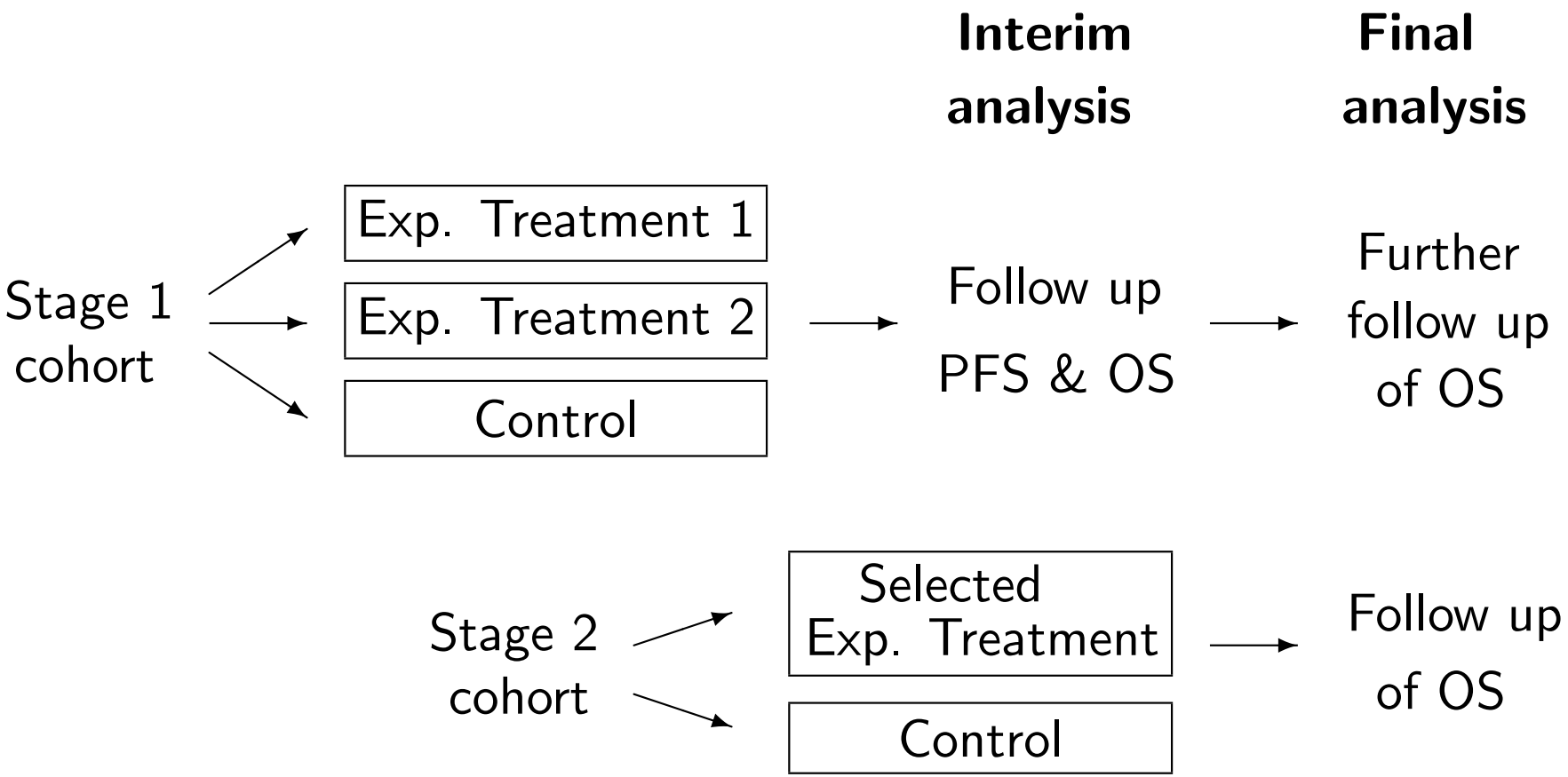
The primary endpoint is Overall Survival (OS).

At an interim analysis, information on OS, Progression Free Survival (PFS), PK measurements and safety will be used 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



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

2. Protecting the type I error rate

We shall assume a proportional hazards model for OS 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

$H_{0,1}$: $\theta_1 \leq 0$ vs $\theta_1 > 0$ *(Exp Treatment 1 superior to control)*,

$H_{0,2}$: $\theta_2 \leq 0$ vs $\theta_2 > 0$ *(Exp Treatment 2 superior to control)*.

In order to control the familywise error rate (FWER), we require

$$P_{(\theta_1,\theta_2)}\{\text{Reject any true null hypothesis}\} \leq \alpha$$

for all (θ_1, θ_2).

A closed testing procedure

Define level α tests of

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

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

and a level α test of the intersection hypothesis

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

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.

A closed testing procedure

Definition of a closed testing procedure

In a closed testing procedure of hypotheses H_i, $i = 1, \ldots, p$, with FWER at most α, we define level α tests of

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

for each subset I of $\{1, \ldots, p\}$.

In the closed testing procedure, the simple hypothesis H_j is rejected if H_I is rejected for every set I containing index j.

Proof of strong control of familywise error rate

Let $\tilde{I}$ be the set of indices of all true hypotheses H_i.

Since $H_{\tilde{I}}$ is true, $P\{\text{Reject } H_{\tilde{I}}\} = \alpha$.

For a familywise error to be committed, $H_{\tilde{I}}$ must be rejected.

Hence, the probability of a familywise error is no greater than α.

Combining data across stages

Consider testing a generic null hypothesis H_0: $\theta \leq 0$ against $\theta > 0$.

Suppose Stage 1 data produce Z_1 where

$$Z_1 \sim N(0,1) \quad \text{if } \theta = 0.$$

After adaptations, Stage 2 gives Z_2 with *conditional* distribution

$$Z_2 \sim N(0,1) \quad \text{if } \theta = 0$$

and this is also the *unconditional* distribution of Z_2 when $\theta = 0$.

Weighted inverse normal combination test

With pre-specified weights w_1 and w_2 satisfying $w_1^2 + w_2^2 = 1$,

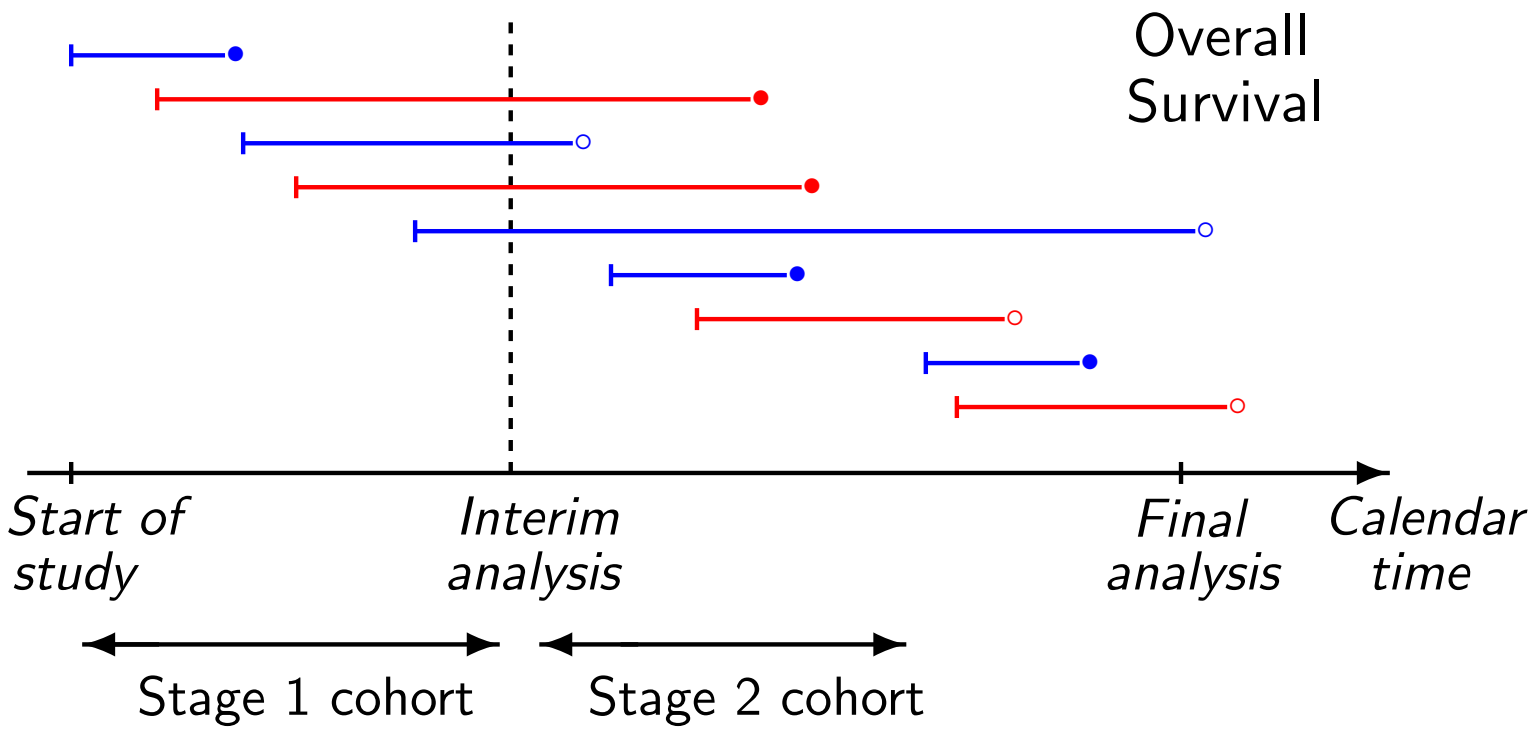
$$Z \;=\; w_1 Z_1 + w_2 Z_2 \;\sim\; N(0,1) \quad \text{if } \theta = 0,$$

and Z is stochastically smaller than $N(0,1)$ if $\theta < 0$.

So, for a level α test, we reject H_0 if $Z > \Phi^{-1}(1-\alpha)$.

3. Properties of log-rank tests

For now, consider Experimental Treatment 1 vs Control.



Key:
- Subjects randomised to Exp Treatment 1
- Subjects randomised to Control
- ● Death observed
- ○ Censored observation

Logrank statistic: Comparing Exp Treatment 1 vs Control

Suppose the observed number of deaths at the time of analysis is d.

Elapsed times between entry to the study and these deaths are

$$\tau_1 < \tau_2 < \ldots < \tau_d \quad \text{(assuming no ties)}.$$

Define variables at this analysis

r_{iT} and r_{iC}	Numbers at risk on Exp Trt 1 and Control at τ_i-
$r_i = r_{iT} + r_{iC}$	Total number at risk at τ_i-
O	Observed number of deaths on Control
$E = \sum_{i=1}^{d} r_{iC}/r_i$	"Expected" number of deaths on Control
$V = \sum_{1}^{d} r_{iT} r_{iC}/r_i^2$	"Variance" of O
$Z = (O - E)/\sqrt{V}$	Standardised logrank statistic

Properties of log-rank tests

Comparing Experimental Treatment 1 vs Control, define

S_1 = Unstandardised log-rank statistic at interim analysis,

$\mathcal{I}_1$ = Information for θ_1 at interim analysis $= V_1 \approx$ (No. of deaths)/4

S_2 = Unstandardised log-rank statistic at final analysis,

$\mathcal{I}_2$ = Information for θ_1 at final analysis $= V_2 \approx$ (No. of deaths)/4

Here, "Number of deaths" refers to the total number of deaths on Experimental Treatment 1 and Control arms only.

Then, approximately,

$$S_1 \sim N(\mathcal{I}_1 \theta_1, \mathcal{I}_1),$$

$$S_2 - S_1 \sim N(\{\mathcal{I}_2 - \mathcal{I}_1\} \theta_1, \{\mathcal{I}_2 - \mathcal{I}_1\})$$

and S_1 and $(S_2 - S_1)$ are **independent** (independent increments).

Reference: Tsiatis (*Biometrika*, 1981).

4. A combination test for survival data

We create Z statistics for comparing Exp treatment 1 and control

Based on data at the interim analysis:

$$Z_1 = \frac{S_1}{\sqrt{\mathcal{I}_1}},$$

Based on data accrued **between** the interim and final analyses:

$$Z_2 = \frac{S_2 - S_1}{\sqrt{\mathcal{I}_2 - \mathcal{I}_1}}.$$

If $\theta_1 = 0$, then $Z_1 \sim N(0,1)$ and $Z_2 \sim N(0,1)$ are independent.

If $\theta_1 < 0$, Z_1 and Z_2 are stochastically smaller than this.

So, we can use $Z = w_1 Z_1 + w_2 Z_2$ in an inverse normal combination test of $H_{0,1}$: $\theta_1 \leq 0$.

A combination test for survival data

The above distribution theory for logrank statistics of a single comparison requires

$$Z_2 = \frac{S_2 - S_1}{\sqrt{\mathcal{I}_2 - \mathcal{I}_1}} \sim N(0,1) \quad \text{under } \theta_1 = 0,$$

regardless of decisions taken at the interim analysis.

Bauer & Posch (*Statistics in Medicine*, 2004) note this implies that the conduct of the second part of the trial should not depend on the prognosis of Stage 1 patients at the interim analysis.

Suppose, after observing particularly good PFS data for patients on the selected treatment, the Stage 2 cohort size is reduced and follow up of Stage 1 patients is extended. Then, the distribution of Z_2 could be biased upwards.

Our example has another potential source of bias, depending on how the Stage 2 statistic for testing $H_{0,12}$ is defined — see later.

Analysing an adaptive survival trial

In applying a Closed Testing Procedure, we require level α tests of

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

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

$H_{0,12}$: $\theta_1 \leq 0$ and $\theta_2 \leq 0$.

Combination tests for these hypotheses are formed from:

	Stage 1 data	*Stage 2 data*
$H_{0,1}$	$Z_{1,1}$	$Z_{2,1}$
$H_{0,2}$	$Z_{1,2}$	$Z_{2,2}$
$H_{0,12}$	$Z_{1,12}$	$Z_{2,12}$

How should we define $Z_{1,1}$, $Z_{2,1}$, etc?

Analysing an adaptive survival trial

A natural choice is to:

Base $Z_{1,1}$, $Z_{1,2}$ and $Z_{1,12}$ on data at the interim analysis,

Base $Z_{2,1}$, $Z_{2,2}$ and $Z_{2,12}$ on the additional information accruing between interim and final analyses.

We could take $Z_{1,1}$ and $Z_{1,2}$ to be standardised log-rank statistics, and $Z_{2,1}$ and $Z_{2,2}$ standardised increments between analyses.

For intersection hypotheses:

Form $Z_{1,12}$ from $Z_{1,1}$ and $Z_{1,2}$,

Set $Z_{2,12} = Z_{2,j}$, where j is the selected treatment.

However, treatment j is selected because it has the better PFS outcomes at the interim analysis.

So, it is likely that future OS for these patients will be "better than average", leading to bias in the null distribution of $Z_{2,12}$.

5. The method of Jenkins, Stone & Jennison (2011)

If we base a combination test on the two parts of the data accrued before and after the interim analysis, bias can result:

	Z_1	Z_2
Stage 1 cohort	Overall survival (during Stage 1)	**Overall survival (during Stage 2)**
Stage 2 cohort		Overall survival (during Stage 2)

Instead, we divide the data into the parts from the two cohorts:

Stage 1 cohort	Overall survival (during Stage 1)	Overall survival (during Stage 2)	Z_1
Stage 2 cohort		Overall survival (during Stage 2)	Z_2

Partitioning data for a combination test

To avoid bias: All patients in the Stage 1 cohort are followed for overall survival up to a fixed time, shortly before the final analysis.

"Stage 1" statistics are based on Stage 1 cohort's **final** OS data

$Z_{1,1}$ from log-rank test of Exp Tr 1 vs Control

$Z_{1,2}$ from log-rank test of Exp Tr 2 vs Control

$Z_{1,12}$ from pooled log-rank test, or a Simes or Dunnett test.

"Stage 2" statistics are based on OS data for the Stage 2 cohort

If Exp Treatment 1 is selected:

$Z_{2,1}$ from log-rank test of Exp Tr 1 vs Control, $Z_{2,12} = Z_{2,1}$

If Exp Treatment 2 is selected:

$Z_{2,2}$ from log-rank test of Exp Tr 2 vs Control, $Z_{2,12} = Z_{2,2}$.

Partitioning data for a combination test

No early stopping for efficacy at the interim analysis

Jenkins, Stone & Jennison (2011) introduced the proposed method in a design where a choice is made between testing for an effect in the full population or a sub-population.

They stipulated that the amount of follow up for the Stage 1 cohort should be fixed at the outset to avoid any risk of inflating the type I error rate.

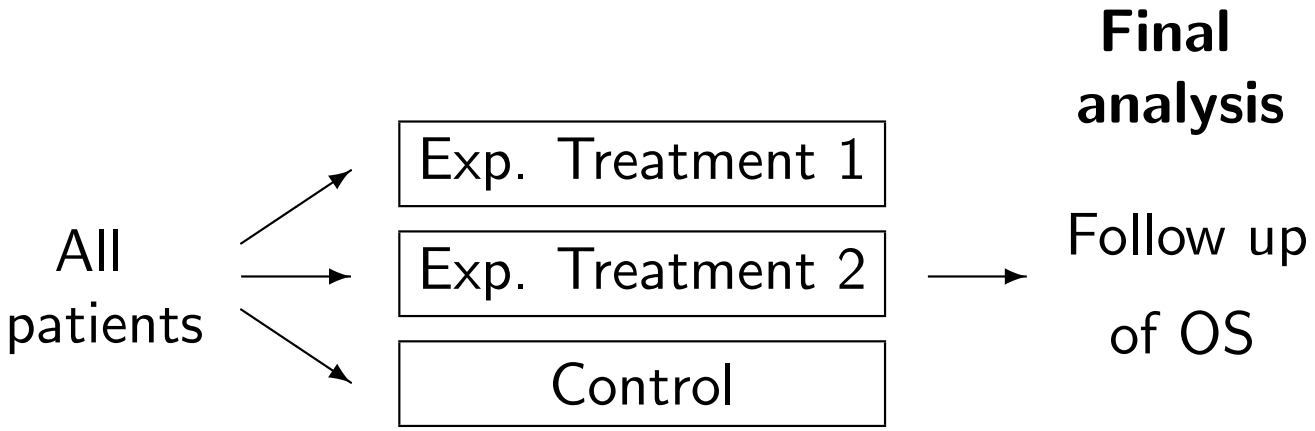
Some adaptive designs allow an early decision based on summaries of "Stage 1" data at an interim analysis.

In our three-treatment design, the statistics $Z_{1,1}$, $Z_{1,2}$ and $Z_{1,12}$ are not known at the time of the interim analysis, so we cannot define a formal stopping rule in terms of these.

However, with only a little OS data available at the interim analysis, this is not a serious limitation.

6. Assessing the benefits of an adaptive design

We compare with a non-adaptive trial in which randomisation is to both experimental treatments and control *throughout* the trial.



A closed testing procedure is used to control familywise error rate.

When the total numbers of patients and lengths of follow-up are the same in adaptive and non-adaptive designs,

Does the adaptive design provide higher power?

Are there other advantages?

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 due to the common control arm.

Progression Free Survival

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

Denote correlation between logrank statistics for OS and PFS by ρ.

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

Assessing the adaptive design: Model assumptions

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

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

We simulate logrank statistics distributed as if both OS and PFS follow proportional hazards models (!) and

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

Final 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 — for adaptive and non-adaptive cases)

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

When the log hazard ratio is θ, the standardised logrank statistic based on d observed events is, approximately, $N(\theta\sqrt{d/4}, 1)$.

Testing the intersection hypothesis $H_{0,12}$

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

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

Pooling: We could test $H_{0,12}$ by *pooling* the Exp Trt 1 and Exp Trt 2 patients and carrying out a logrank test vs the Control group.

Alternatively we could use a *Simes* test or a *Dunnett* test.

Simes' test:

Given observed values p_1 and p_2 of P_1 and P_2, Simes' test of $H_{0,12}$ yields the P-value

$$\min\left(\, 2\min(p_1, p_2),\ \max(p_1, p_2)\,\right).$$

Simes' test protects type I error conservatively when P_1 and P_2 are independent or positively associated.

Dunnett's test of an intersection hypothesis

Dunnett's test for comparisons with a common control

Suppose Z_1 and Z_2 are the Z-values for logrank tests of Exp Trt 1 vs control and Exp Trt 2 vs Control.

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) is bivariate normal with

$$Z_1 \sim N(0,1), \quad Z_2 \sim N(0,1), \quad \text{Corr}(Z_1, Z_2) = 0.5.$$

Our investigations of different tests of the intersection hypothesis showed the Dunnett test to give the most efficient overall testing versions of both adaptive and non-adaptive designs.

Comparing adaptive and non-adaptive trial designs

With selected values of ψ_1, θ_1, ψ_2, θ_2 and ρ, we simulate logrank statistics from their large sample distributions.

For the adaptive design, we define

$$P(1) = P(\text{Select Treatment 1 and Reject } H_{0,1} \text{ overall})$$

$$P(2) = P(\text{Select Treatment 2 and Reject } H_{0,2} \text{ overall})$$

For the non-adaptive design, we set

$$P(1) = P(\widehat{\theta}_1 > \widehat{\theta}_2 \text{ and } H_{0,1} \text{ is rejected overall})$$

$$P(2) = P(\widehat{\theta}_2 > \widehat{\theta}_1 \text{ and } H_{0,2} \text{ is rejected overall})$$

Hence, we define the overall expected "Gain" or utility measure

$$E(\text{Gain}) = \theta_1 \times P(1) + \theta_2 \times P(2).$$

Comparing tests of the intersection hypothesis

Intersection tests produce $Z_{1,12}$ in an adaptive trial design with

$$\psi_1 = \theta_1 \text{ and } \psi_2 = \theta_2 \ (\gamma = 1), \quad \lambda = 1, \quad \rho = 0.6, \quad \alpha = 0.025.$$

		$P(1)$			$E(\text{Gain})$		
θ_1	θ_2	Pooled	Simes	Dunnett	Pooled	Simes	Dunnett
0.3	0.0	0.77	0.85	0.86	0.232	0.254	0.259
0.3	0.1	0.78	0.81	0.82	0.238	0.245	0.247
0.3	0.2	0.68	0.68	0.69	0.238	0.237	0.238
0.3	0.25	0.58	0.58	0.58	0.250	0.249	0.249
0.3	0.295	0.48	0.47	0.47	0.275	0.274	0.274

All simulation results are based on 1,000,000 replicates.

The Dunnett test has the highest power. Unlike the pooled test, it is well aligned (consonant) with individual tests of $H_{0,1}$ and $H_{0,2}$.

Comparing adaptive and non-adaptive trial designs

We compare designs using a Dunnett test for $H_{0,12}$ with

$$\psi_1 = \theta_1 \text{ and } \psi_2 = \theta_2 \ (\gamma = 1), \quad \lambda = 1, \quad \rho = 0.6, \quad \alpha = 0.025.$$

		Non-adaptive			Adaptive		
θ_1	θ_2	$P(1)$	$P(2)$	$E(\text{Gain})$	$P(1)$	$P(2)$	$E(\text{Gain})$
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

Here, $\lambda = 1$ implies there are 300 PFS events at the interim analysis.

The adaptive design has higher $P(1)$ when θ_1 is well above θ_2.

With θ_1 and θ_2 closer, the adaptive design still has higher $E(\text{Gain})$.

Comparing adaptive and non-adaptive trial designs

The adaptive design can only succeed if there is adequate information to select the correct treatment at the interim analysis:

Treatment effects on PFS should be be reliable indicators of treatment effects on OS,

There must be good information on PFS at the interim analysis.

We have investigated varying the parameters γ and λ where

$\psi_1 = \gamma \times \theta_1,\ \psi_2 = \gamma \times \theta_2,$ with $\theta_1 = 0.3$ and $\theta_2 = 0.1$

Final number of OS events for Stage 1 cohort $= 300$ (over 3 arms)

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

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

NB It is quite plausible that γ should be greater than 1, i.e., a larger treatment effect on PFS than on OS.

Comparing adaptive and non-adaptive trial designs

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

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

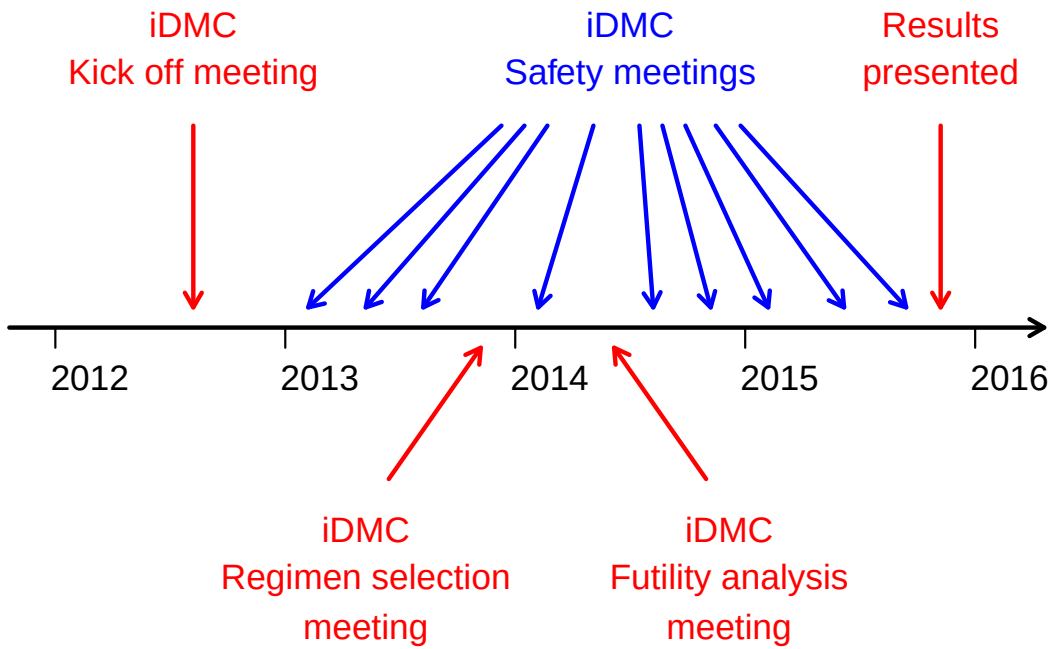
Number of PFS events at interim analysis $= \lambda \times 300.$

		Non-adaptive			Adaptive		
γ	λ	$P(1)$	$P(2)$	$E(\text{Gain})$	$P(1)$	$P(2)$	$E(\text{Gain})$
1.5	1.2				0.88	0.00	0.264
1.2	1.1				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 when there is enough PFS information for treatment selection at the interim analysis — but not otherwise.

7. The independent Data Monitoring Committee

The IDMC for the GATSBY trial had three main tasks:

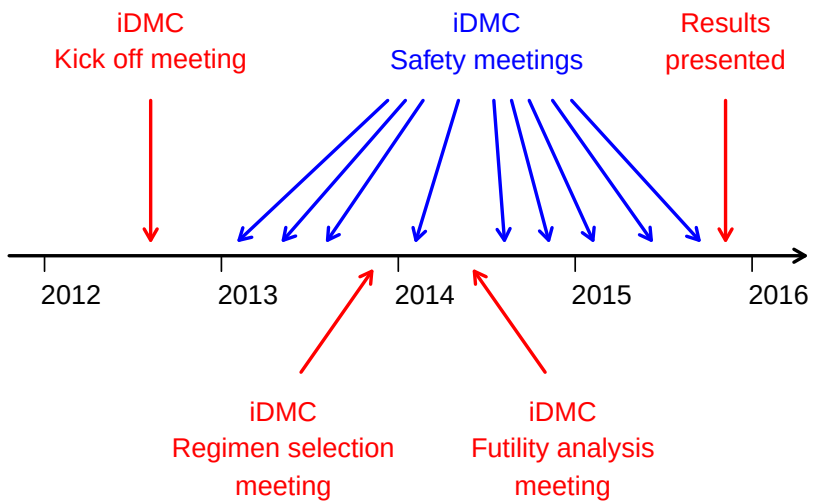


To monitor safety data,

To choose one of the two forms of the experimental treatment at the Regimen selection meeting,

To consider early stopping at the Futility analysis meeting.

Independent Data Monitoring Committee



The Kick off meeting was the final opportunity for the iDMC to discuss plans for the trial with the sponsor.

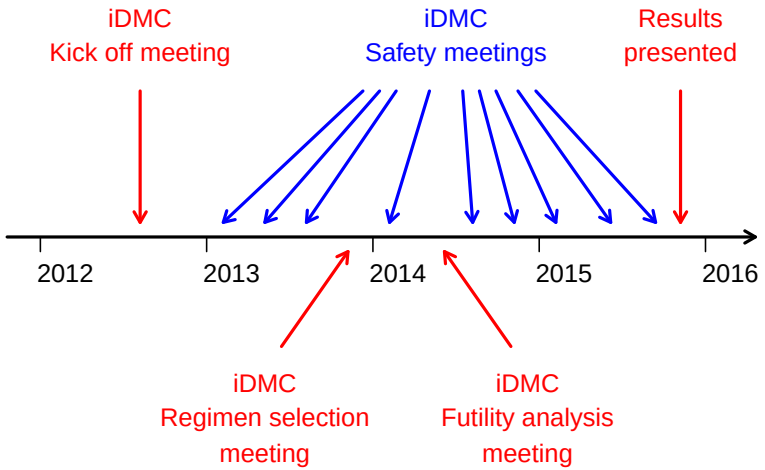
There was discussion about how to combine evidence about

Progression free survival, Overall survival, PK data

when choosing a treatment at the Regimen selection meeting.

After this meeting, a firewall was set in place and there could be no further discussion between the iDMC and the sponsor.

Independent Data Monitoring Committee



The final analysis showed no evidence of a treatment difference.

At the futility analysis, the data supported continuing the study.

At later meetings, the iDMC saw information about overall survival since Death was recorded as an event in the safety data.

There was no further provision to stop for futility and, with the firewall in place, the iDMC could not ask the sponsor for advice.

Since treatments were performing equally well, there was no reason to stop to ensure patient safety.

Conclusions

The GATSBY trial design offered the chance to compare two versions of T-DM1 treatment and focus on the superior form in the second stage of the trial.

In such an adaptive design, significant effort is needed to ensure the familywise type I error rate is controlled.

It is just as important to assess the properties of an adaptive design and compare with simpler non-adaptive options.

Adaptive treatment selection can be beneficial if there is enough information to make a reliable treatment selection decision.

In an adaptive trial, the kick off meeting is the time to discuss

- How an adaptation decision is to be made
- When, how and why early termination may occur.

Remember: Communication between sponsor and iDMC is restricted once the trial is under way.