Designing an Adaptive Trial with Treatment Selection and a Survival Endpoint

Christopher Jennison

Dept of Mathematical Sciences, University of Bath, UK

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

Martin Jenkins & Andrew Stone

AstraZeneca, Alderley Park, UK

SAMBa SLS

November 2016

Outline of talk

1. A clinical trial 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. Analysing an adaptive trial

   Method 1 — and why it may inflate the type I error rate

   Method 2 (Jenkins, Stone & Jennison, *Pharmaceutical Statistics*, 2011)
6. Properties of the proposed adaptive design
7. Related work
8. Conclusions

1. A clinical trial with treatment selection

Consider a trial of cancer treatments comparing

Experimental Treatment 1: Intensive dosing

Experimental Treatment 2: Slower dosing

Control treatment

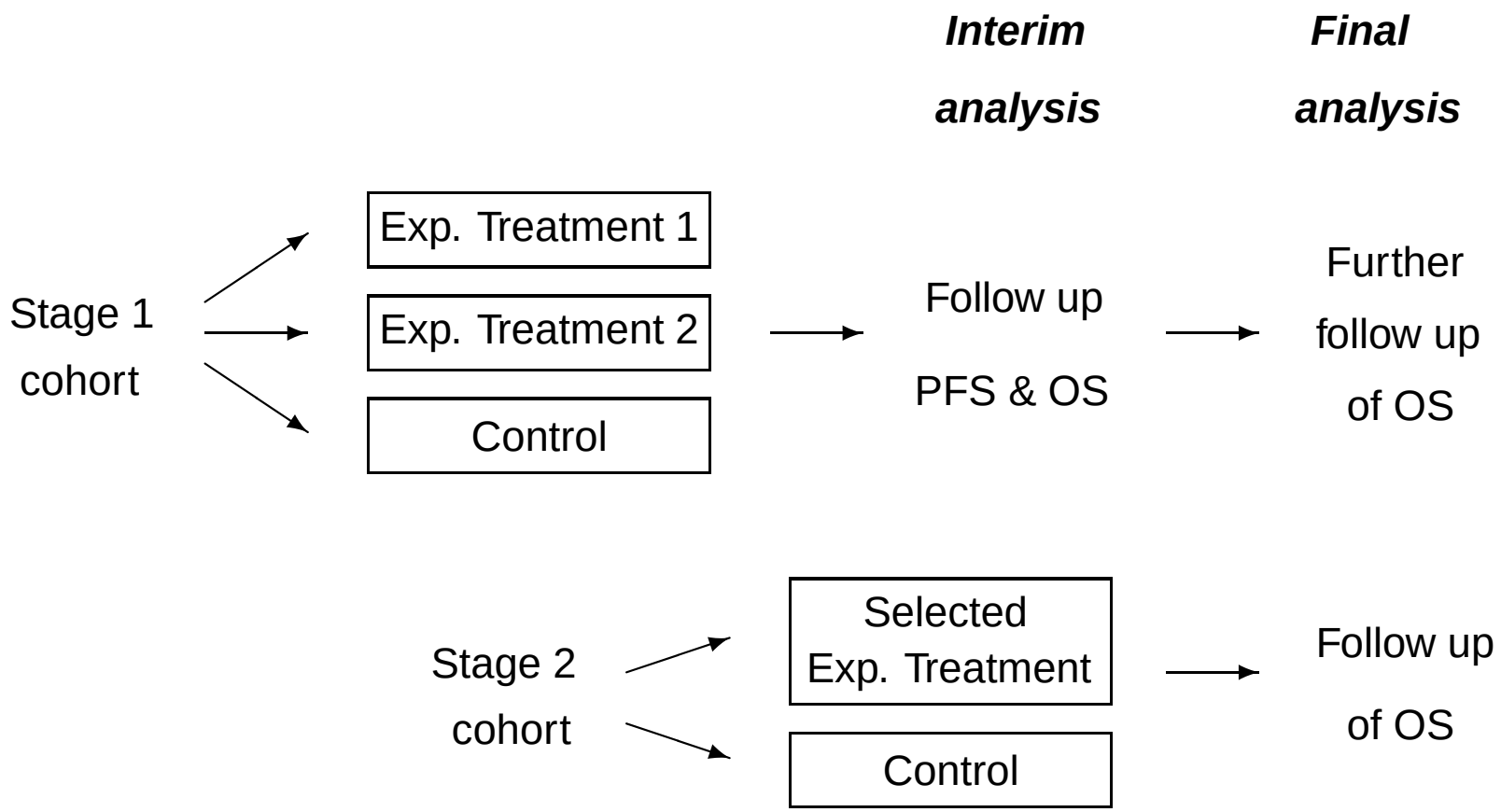
The primary endpoint is Overall Survival (OS).

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



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

2. Protecting the type I error rate

We shall assume a proportional hazards model with

λ_1 = Hazard ratio, Control *vs* Exp. Treatment 1

λ_2 = Hazard ratio, Control *vs* 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", we require

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

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.

Combining data across stages

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

Data are gathered in two stages:

Stage 1 data produce the standardised statistic Z_1,

After possible adaptations, Stage 2 data produce Z_2.

In the particular case $\theta = 0$:

Then, Z_1 has the usual $N(0,1)$ distribution.

And $Z_2 \sim N(0,1)$ conditionally on Stage 1 data and the resulting Stage 2 design.

Since the conditional distribution of Z_2 is the same for all Stage 1 data, Z_2 is independent of the Stage 1 data (including Z_1).

Thus, when $\theta = 0$, Z_1 and Z_2 are independent $N(0,1)$ variables.

Combining data across stages

We have stage-wise statistics Z_1 and $Z_2 \sim N(0,1)$ independently under $\theta = 0$.

If $\theta < 0$, suppose we can show that Z_1 and Z_2 have distributions which are stochastically smaller than $N(0,1)$: then the Type I error probability under $\theta = 0$ will be higher than at any $\theta < 0$.

Weighted inverse normal combination test

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

$$Z \;=\; w_1 Z_1 + w_2 Z_2 \;\sim\; N(0,1) \text{ (or stochastically smaller) under } H_0.$$

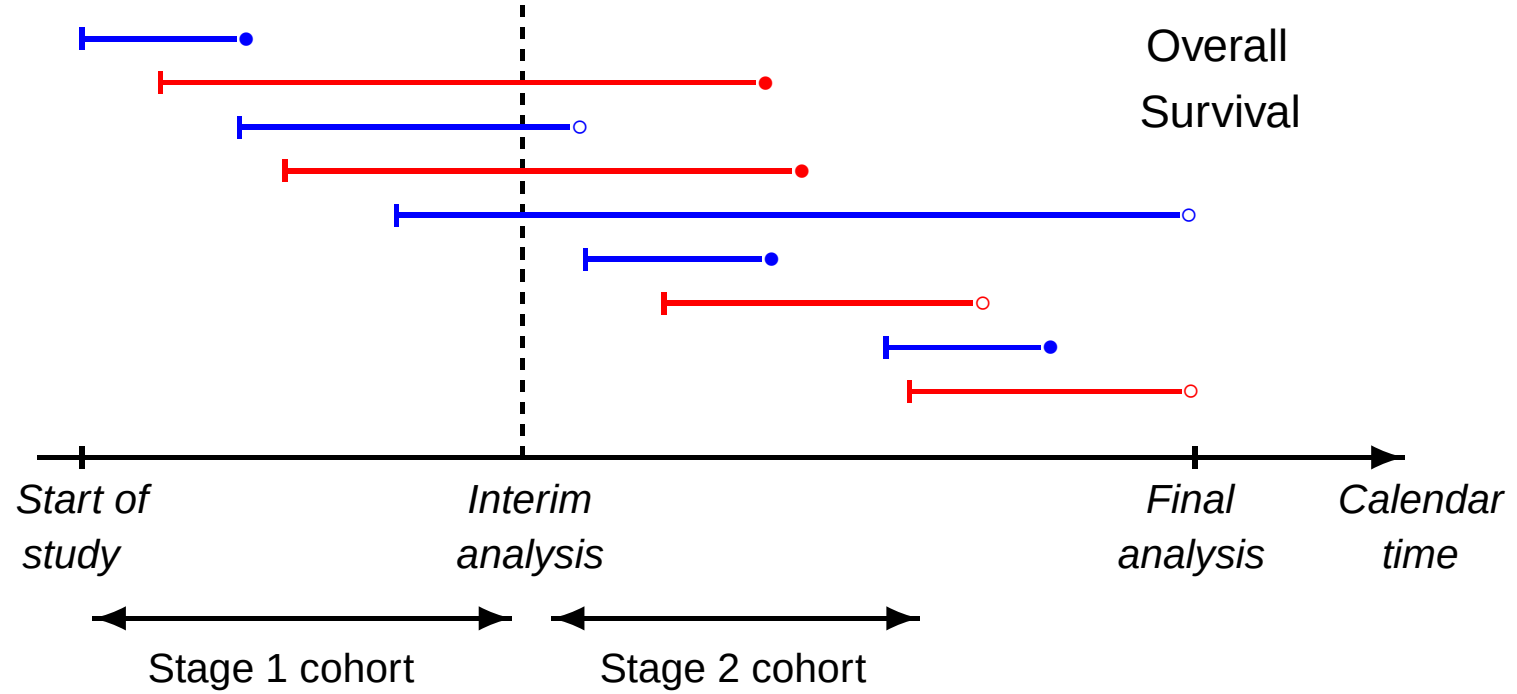
So, for a level α test, we reject H_0 if $Z > \Phi^{-1}(1-\alpha)$.

Or, the test can be defined in terms of $P_1 = 1 - \Phi(Z_1)$ and $P_2 = 1 - \Phi(Z_2)$.

We shall apply such combination rules in testing our three null hypotheses, $H_{0,1}$, $H_{0,2}$ and $H_{0,12}$.

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.

Properties of log-rank tests

Comparing Experimental Treatment 1 vs Control, define

S_1 = Unstandardised log-rank statistic an interim analysis,

$\mathcal{I}_1$ = Information for θ_1 at interim analysis $\approx$ (Number of deaths)/4

S_2 = Unstandardised log-rank statistic an final analysis,

$\mathcal{I}_2$ = Information for θ_1 at final analysis $\approx$ (Number of deaths)/4

Here, "Number of deaths" refers to 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** — the "independent increments" property (Tsiatis, *Biometrika*, 1981). NB This result holds for staggered entry.

4. A combination test for survival data

We can create Z statistics

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 a combination test of $H_{0,1}$: $\theta_1 \leq 0$.

A combination test for survival data

In the above, it is crucial that

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

regardless of decisions taken at the interim analysis.

For this to be true, the conduct of the second part of the trial should not depend on the prognosis of Stage 1 patients at the interim analysis.

Bauer & Posch (*Statistics in Medicine*, 2004) note that this condition may fail to hold

Suppose, for example,

- PFS at the interim analysis is better for patients on Exp. Treatment 1 than Control, implying better prospects for OS on the Exp. Treatment 1 arm.
- Stage 2 cohort size is reduced and Stage 1 patients are followed up longer.

The change to Stage 2 cohort size increases the contribution of Stage 1 patients to Z_2, biasing it upwards — we shall need to avoid such potential biases.

5. Analysing an adaptive survival trial

Recall, we wish to apply a Closed Testing Procedure based on 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*
$H_{0,1}$	$Z_{1,1}$	$Z_{2,1}$
$H_{0,2}$	$Z_{1,2}$	$Z_{2,2}$
$H_{0,12}$	$Z_{1,12}$	$Z_{2,12}$

The question is how we should define $Z_{1,1}$, $Z_{2,1}$, etc.

Analysing an adaptive survival trial: Method 1

A natural choice is to:

Base $Z_{1,1}$, $Z_{1,2}$ and $Z_{1,12}$ on data available 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 shall take $Z_{1,1}$ and $Z_{1,2}$ to be standardised log-rank statistics, and $Z_{2,1}$ and $Z_{2,2}$ to be standardised increments in the log-rank statistics between analyses.

For $Z_{1,12}$ we could compare the pooled Exp Tr 1 and Exp Tr 2 patients with the Control group. However, for "consonance", it is probably better to combine $Z_{1,1}$ and $Z_{1,2}$ through, say, Simes' test or Dunnett's test (which we shall define later).

If we select Experimental Treatment 1 at the interim analysis, we no longer wish to test $H_{0,2}$ — we do not need $Z_{2,2}$ and we can set $Z_{2,12} = Z_{2,1}$.

Similarly, if we select Experimental Treatment 2, we no longer need $Z_{2,1}$ and we can set $Z_{2,12} = Z_{2,2}$.

Method 1, recap

Stage 1 statistics are calculated at the interim analysis:

$Z_{1,1}$ 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.

If Exp. Treatment 1 is selected at the interim analysis, Stage 2 statistics are

$Z_{2,1}$ from increment in log-rank statistic testing Exp Tr 1 vs Control, combining Stage 1 and Stage 2 cohorts

$Z_{2,12} = Z_{2,1}$.

If Exp. Treatment 2 is selected, Stage 2 statistics are

$Z_{2,2}$ from increment in log-rank statistic testing Exp Tr 2 vs Control, combining Stage 1 and Stage 2 cohorts

$Z_{2,12} = Z_{2,2}$.

Method 1: What can go wrong?

The first stage statistics are fine.

Suppose Experimental Treatment 1 is selected at the interim analysis.

Then, $Z_{2,1}$ is the increment in the log-rank statistic testing Exp Tr 1 vs Control, combining Stage 1 and Stage 2 cohorts.

$Z_{2,1}$ The issues raised earlier should be considered — might Stage 2 be modified in the light of interim data in a way that biases $Z_{2,1}$?

Regulators are likely to worry about such possibilities !

$Z_{2,12}$ Setting $Z_{2,12} = Z_{2,1}$ will cause bias.

Exp Tr 1 is selected when subjects on this arm have good PFS, so the Exp Tr 1 patients who continue to be followed for OS in Stage 2 are liable to have good prognoses.

This method is almost certain to inflate the overall type I error rate !!

Method 2: Jenkins, Stone & Jennison (2011)

In constructing a combination test, Method 1 separates data into the parts accrued before and after the interim analysis:

	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 arising 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

Method 2

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

The "Stage 1" statistics are based on the final OS data for the Stage 1 cohort

$Z_{1,1}$ 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.

The "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, $Z_{2,12} = Z_{2,2}$.

Method 2

Notes

Jenkins, Stone & Jennison (2011) introduced "Method 2" in a design where a choice is made between testing for an effect in the full population or a sub-population.

If the length of follow up of the Stage 1 cohort for OS can be influenced by interim information about the likely survival of continuing patients, error rate inflation could result (as noted by Bauer & Posch, 2004).

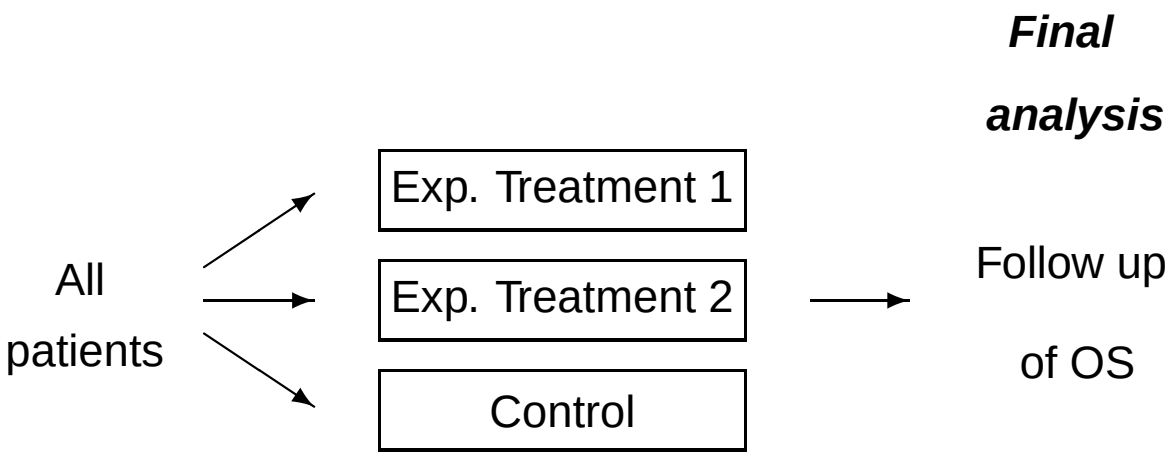
Hence, we stipulate the amount of follow up and require that this is not changed.

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

Our statistics $Z_{1,1}$, $Z_{1,2}$ and $Z_{1,12}$ are not known at the time of the interim analysis, so we cannot apply formal stopping rules defined in terms of these — but with a lack of OS data at the interim analysis, that is not a serious limitation.

6. Assessing the benefits of an adaptive design

We shall make comparisons with a non-adaptive trial design in which patients are randomised to both experimental treatments and control *throughout* recruitment.



Here, a closed testing procedure will be used to control the 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 because of the common control arm.

Progression Free Survival (for treatment selection)

	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.

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 suppose

$\psi_1 = \gamma \times \theta_1$ 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)

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

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

$$H_{0,1}\colon\ \theta_1 \leq 0 \quad \text{and} \quad H_{0,2}\colon\ \theta_2 \leq 0.$$

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

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

We can test $H_{0,12}$ by pooling the Exp. Treatment 1 and Exp. Treatment 2 patients and carrying out a logrank test vs the Control group.

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

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.

The corresponding normal deviates are

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

Testing the intersection hypothesis

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

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

Simes' test can be viewed as an extension of the Bonferroni test. It protects type I error conservatively when P_1 and P_2 are independent or positively associated.

Dunnett's test for comparisons with a common 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 $Z_1\sim N(0,1),\ Z_2\sim N(0,1)$ and Corr$(Z_1,Z_2)=0.5$.

Comparing tests of the intersection hypothesis

Setting $\psi_1 = \theta_1,\ \psi_2 = \theta_2$ (so PFS $\sim$ OS) and $\rho = 0.6$, we simulated logrank statistics from their large sample distributions under the adaptive design.

We noted

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

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

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

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

In the ***adaptive trial design***, we have compared use of

the test based on pooled data, Simes' test, and Dunnett's test

for determining the Stage 1 P-value $P_{1,12}$ and, hence, $Z_{1,12} = \Phi^{-1}(1 - P_{1,12})$.

Comparing tests of the intersection hypothesis

We compare intersection tests to produce $Z_{1,12}$ in an adaptive trial design with

$$\psi_1 = \theta_1, \quad \psi_2 = \theta_2, \quad \lambda = 1, \quad \rho = 0.6, \quad \alpha = 0.025.$$

		$P(1)$			$E(\text{Gain})$		
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 is most effective — it has good power and, unlike the pooled test, is well aligned (consonant) with individual tests of $H_{0,1}$ and $H_{0,2}$.

Comparing adaptive and non-adaptive trial designs

We shall compare designs using a Dunnett test for the intersection hypothesis.

For the non-adaptive design

The intersection hypothesis is tested at the final analysis as part of the closed testing procedure.

As for the adaptive design, we found the Dunnett test to give the best power.

If both $H_{0,1}$ and $H_{0,2}$ are rejected, we suppose the treatment with the higher observed effect size will be "chosen" for registration and marketing.

Accordingly, we define

$$\begin{aligned} P(1) &= P(\hat{\theta}_1 > \hat{\theta}_2 \text{ and } H_{0,1} \text{ is rejected overall}) \\ P(2) &= P(\hat{\theta}_2 > \hat{\theta}_1 \text{ and } H_{0,2} \text{ is rejected overall}) \\ E(\text{Gain}) &= \theta_1 \times P(1) + \theta_2 \times P(2). \end{aligned}$$

Comparing adaptive and non-adaptive trial designs

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

$$\psi_1 = \theta_1, \quad \psi_2 = \theta_2, \quad \lambda = 1 \text{ (\# PFS events} \sim \text{\# OS events)},$$

$$\rho = 0.6, \quad \alpha = 0.025.$$

		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

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 $E(\text{Gain})$.

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 $\psi_2 = \gamma \times \theta_2$

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)

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

NB It is quite plausible that γ should be greater than 1.

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	$P(1)$	$P(2)$	$E(\text{Gain})$	$P(1)$	$P(2)$	$E(\text{Gain})$
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 about the benefits of an adaptive design

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.

3. Other evidence may 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: the conclusions from this comparison may be useful in other settings.

7. Related work

1. Irle & Schäfer (*JASA*, 2012) propose similar adaptive designs for survival data.

Changes to the design and critical values for test statistics are made, preserving the conditional probability of rejecting a null hypothesis.

As the "Conditional Probability of Rejection" principle is related to combination tests, the method has much in common with that of Jenkins, Stone & Jennison (2011).

Irle & Schäfer's method imposes the same requirement of a fixed length of follow-up for "Cohort 1" patients.

Even with this condition in place, determining the conditional probability of a future event is problematic, since the final information level (in a log-rank statistic, say) is not known at the time this probability is calculated.

We recommend our combination test approach as simpler to explain and easier to implement.

Related work

2. Friede et al. (*Statistics in Medicine*, 2011) consider a seamless phase II/III trial design with treatment selection based on short-term and long-term responses.

In a study of treatments for multiple sclerosis, several experimental treatments are compared to a control. When the treatment selection decision is made, only a short-term response is available for some subjects but these will go on to provide a long-term response later.

Although the primary endpoint is not a time-to-event response, similar issues arise. When patients on the selected treatment are followed up, results are likely to be biased towards showing a positive treatment effect, given the short-term response data on which the treatment selection decision was based.

These authors follow a similar approach to Jenkins, Stone & Jennison (2011) and apply a combination test to the long-term response data from the *cohorts* of patients admitted before and after the interim decision point.

8. Conclusions

1. Adaptive designs for trials with survival endpoints can enable interim treatment selection or decisions about the population in which a treatment effect is to be sought.

2. A Closed Testing Procedure can be employed and Combination Tests used to carry out each level α hypothesis test with data from two (or more) stages.

3. The "independent increments" property of the log-rank statistic can fail, and other biases can arise, if design changes at an interim analysis are based on data that are also informative about the later survival of continuing patients.

4. The proposed design avoids this problem. Defining the elements of a combination test in terms of the complete survival data for separate cohorts of patients leads to a valid, and potentially efficient, testing procedure.