Designing an Adaptive Trial using a Combination Test for 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

Smart Trials 2014

London, April 2014

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. Related work

7. 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

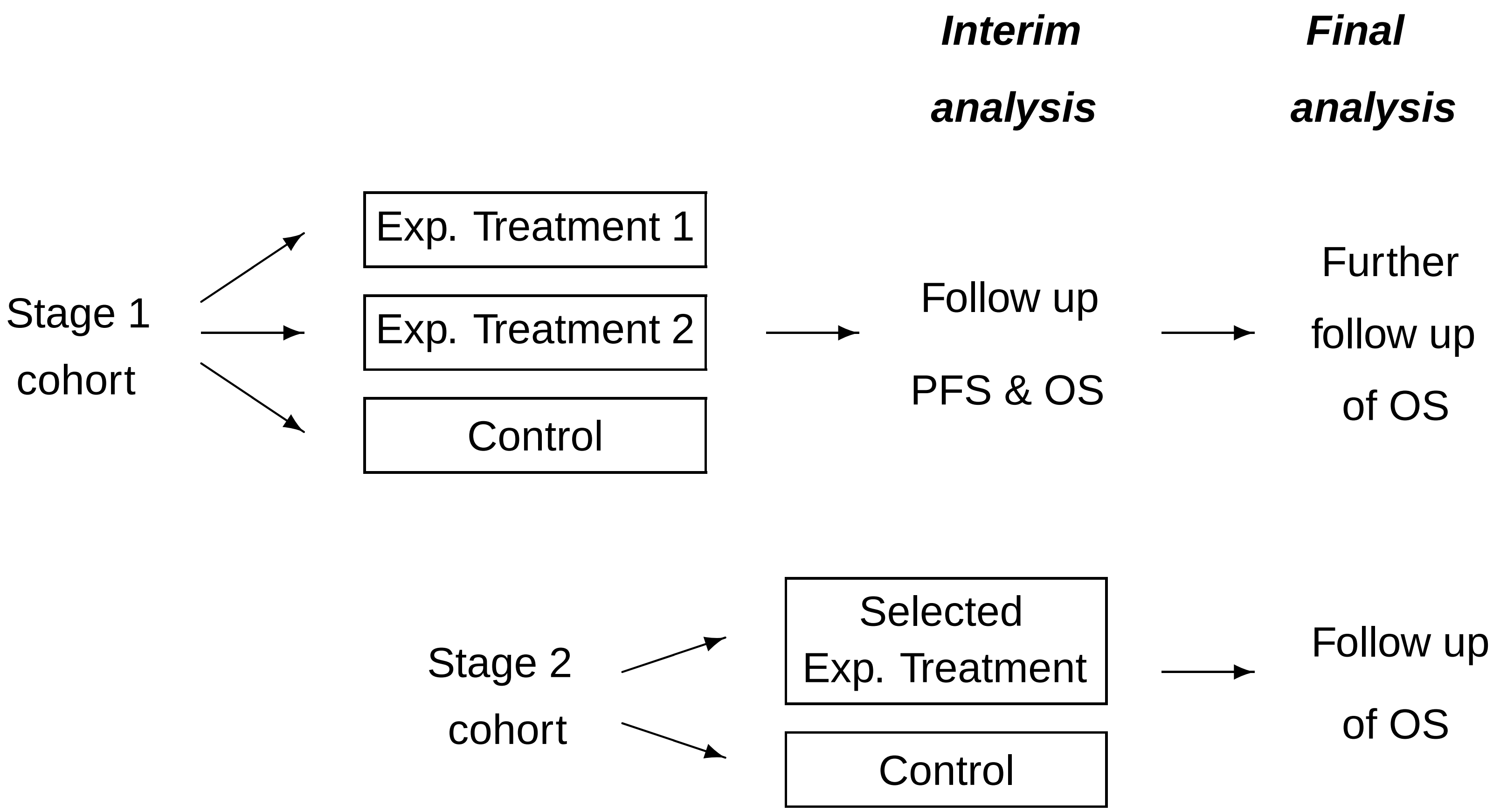
The primary endpoint is Overall Survival (OS).

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



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.

2. Protecting the type I error rate

We shall assume a proportional hazards model with

$$\begin{aligned} \lambda_1 &= \text{Hazard ratio, Control \textit{vs} Exp. Treatment 1} \\ \lambda_2 &= \text{Hazard ratio, Control \textit{vs} Exp. Treatment 2} \end{aligned}$$

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

We test null hypotheses

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

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

We require

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

A closed testing procedure

Define level α tests of

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

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

and of the intersection hypothesis

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

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

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

With data gathered in two stages, suppose Stage 1 data produce Z_1 where

$$Z_1 \sim N(0,1) \ \text{(or stochastically smaller)} \ \text{under } H_0.$$

After possible adaptations, Stage 2 data produce Z_2 with *conditional* distribution

$$Z_2 \sim N(0,1) \ \text{(or stochastically smaller)} \ \text{under } H_0.$$

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

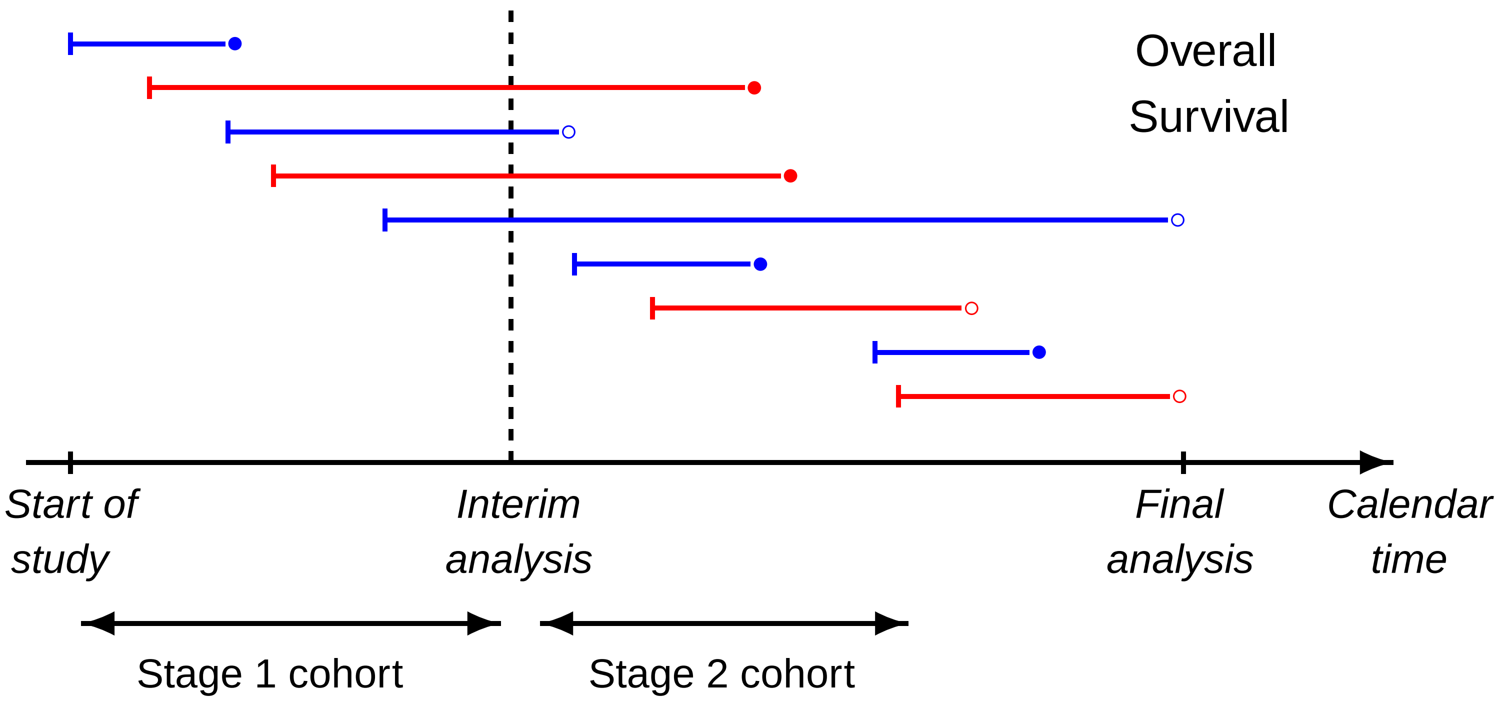
$$Z \;=\; w_1\, Z_1 + w_2\, Z_2 \;\sim\; N(0,1) \ \text{(or stochastically smaller)} \ \text{under } H_0.$$

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

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

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

$$\begin{aligned} S_1 &= \text{Unstandardised log-rank statistic an interim analysis,} \\ \mathcal{I}_1 &= \text{Information for } \theta_1 \text{ at interim analysis } \approx \text{ (Number of deaths)/4} \\ S_2 &= \text{Unstandardised log-rank statistic an final analysis,} \\ \mathcal{I}_2 &= \text{Information for } \theta_1 \text{ at final analysis } \approx \text{ (Number of deaths)/4} \end{aligned}$$

Here, “Number of deaths” refers to Experimental Treatment 1 and Control arms only.

Then, approximately,

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

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 the potential pitfalls.

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 increases the contribution of Stage 1 patients to Z_2, biasing it upwards.

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.

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

We shall take $Z_{1,1}$, $Z_{1,2}$, $Z_{2,1}$ and $Z_{2,2}$ to be standardised log-rank statistics.

For $Z_{1,12}$ we test the pooled Exp Tr 1 and Exp Tr 2 patients vs the Control group.

Method 1, continued

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 log-rank test of combined Exp Tr 1 and Exp Tr 2 vs Control.

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 will 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 log-rank test of combined Exp Tr 1 and Exp Tr 2 vs Control.

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 that is not a serious problem.

6. 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 designs 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.

7. Conclusions

1. Adaptive designs for trials with survival endpoints are desirable for 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 if design changes at an interim analysis are based on data that are also informative about the later survival of continuing patients.

   Significant bias can also arise from setting $Z_{2,12}$ to be one of $Z_{2,1}$ and $Z_{2,2}$ where the choice between these depends on the PFS outcomes for subjects whose overall survival will contribute to $Z_{2,1}$ and $Z_{2,2}$.

4. Defining the elements of a combination test in terms of the complete survival data for separate cohorts of patients leads to a valid testing procedure.