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, UK

ERCIM: 10th International Conference
London
December 2017
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
   The method of Jenkins, Stone & Jennison
   (Pharmaceutical Statistics, 2011)
6. Properties of the proposed adaptive design
7. Conclusions
A survival study with treatment selection

Consider a Phase 3 trial of cancer treatments comparing

Experimental Treatment 1: Intensive dosing
Experimental Treatment 2: Slower dosing
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

Stage 1 cohort
- Exp. Treatment 1
- Exp. Treatment 2
- Control

Stage 2 cohort
- Selected Exp. Treatment
- Control

Interim analysis
- Follow up
- PFS & OS

Final analysis
- Further 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 treatment.
Protecting the type I error rate

We shall assume a proportional hazards model for OS with

\[ \lambda_1 = \text{Hazard ratio, Control vs Exp Treatment 1} \]

\[ \lambda_2 = \text{Hazard ratio, Control vs Exp Treatment 2} \]

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

We test null hypotheses

\[ H_{0,1}: \theta_1 \leq 0 \ vs \ \theta_1 > 0 \ (\text{Exp Treatment 1 superior to control}), \]

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

In order to control the “familywise error rate”, we require

\[ P(\theta_1, \theta_2) \{\text{Reject any true null hypothesis}\} \leq \alpha \]

for all \((\theta_1, \theta_2)\).
A closed testing procedure

Define level $\alpha$ tests of

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

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

and a level $\alpha$ 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.
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

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

After adaptations, Stage 2 data produce $Z_2$ with conditional distribution

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

**Weighted inverse normal combination test**

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

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

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

So, for a level $\alpha$ test, we reject $H_0$ if $Z > \Phi^{-1}(1 - \alpha)$.
For now, consider Experimental Treatment 1 vs Control.

- Start of study
- Interim analysis
- Final analysis
- Calendar time

Overall Survival

Stage 1 cohort
Stage 2 cohort

Key:
- Red: Subjects randomised to Exp Treatment 1
- Blue: Subjects randomised to Control
- •: Death observed
- ○: Censored observation

Chris Jennison
Adaptive design with treatment selection and survival endpoint
Properties of log-rank tests

Comparing Experimental Treatment 1 vs Control, define

\[ S_1 = \text{Unstandardised log-rank statistic at interim analysis}, \]
\[ I_1 = \text{Information for } \theta_1 \text{ at interim analysis} \approx \frac{\text{Number of deaths}}{4} \]
\[ S_2 = \text{Unstandardised log-rank statistic at final analysis}, \]
\[ I_2 = \text{Information for } \theta_1 \text{ at final analysis} \approx \frac{\text{Number of deaths}}{4} \]

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

Then, approximately,

\[ S_1 \sim N(I_1 \theta_1, I_1), \]
\[ S_2 - S_1 \sim N(\{I_2 - I_1\} \theta_1, \{I_2 - I_1\}) \]

and \( S_1 \) and \( S_2 - S_1 \) are independent (independent increments).

A combination test for survival data

We create $Z$ statistics

Based on data at the interim analysis:

$$Z_1 = \frac{S_1}{\sqrt{I_1}},$$

Based on data accrued between the interim and final analyses:

$$Z_2 = \frac{S_2 - S_1}{\sqrt{I_2 - I_1}}.$$

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

\[ Z_2 = \frac{S_2 - S_1}{\sqrt{I_2 - I_1}} \sim N(0, 1) \quad \text{under } \theta_1 = 0, \]

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 prognoses are better for patients on Exp Treatment 1 than for those on Control, and the Stage 2 cohort size is reduced while 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.
In applying a Closed Testing Procedure, we require level $\alpha$ 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:

<table>
<thead>
<tr>
<th>Stage 1 data</th>
<th>Stage 2 data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{0,1}$</td>
<td>$Z_{1,1}$</td>
</tr>
<tr>
<td>$H_{0,2}$</td>
<td>$Z_{1,2}$</td>
</tr>
<tr>
<td>$H_{0,12}$</td>
<td>$Z_{1,12}$</td>
</tr>
</tbody>
</table>

The question is how we should 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: $Z_{1,12}$ is formed from $Z_{1,1}$ and $Z_{1,2}$, while $Z_{2,12} = Z_{2,j}$, where $j$ is the selected treatment.

However, treatment $j$ is selected because it has better PFS outcomes at the interim analyses, so it is likely that future OS for these patients will also be better.

This approach leads to a bias in the null distribution of $Z_{2,12}$. 
If we base a combination test on the two parts of the data accrued before and after the interim analysis, bias can result:

<table>
<thead>
<tr>
<th></th>
<th>$Z_1$</th>
<th>$Z_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stage 1 cohort</strong></td>
<td>Overall survival</td>
<td>Overall survival</td>
</tr>
<tr>
<td></td>
<td>(during Stage 1)</td>
<td>(during Stage 2)</td>
</tr>
<tr>
<td><strong>Stage 2 cohort</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>$Z_1$</th>
<th>$Z_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stage 1 cohort</strong></td>
<td>Overall survival</td>
<td>Overall survival</td>
</tr>
<tr>
<td></td>
<td>(during Stage 1)</td>
<td>(during Stage 2)</td>
</tr>
<tr>
<td><strong>Stage 2 cohort</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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, $Z_{2,12} = Z_{2,2}$. 
Discussion

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.

However, with only a little OS data available at the interim analysis, this is not a serious limitation.
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 $\theta_1$

Exp Treatment 2 vs control $\theta_2$

Logrank statistics are correlated due to the common control arm.

**Progression Free Survival**

Log hazard ratio

Exp Treatment 1 vs control $\psi_1$

Exp Treatment 2 vs control $\psi_2$

Denote correlation between logrank statistics for OS and PFS by $\rho$.

Proportional hazards models for both endpoints are not essential (or possible?) — the implications for the joint distribution of logrank statistics are what matter.
Log hazard ratios for OS: \( \theta_1, \theta_2 \).

Log hazard ratios for PFS: \( \psi_1, \psi_2 \).

We suppose \([ \ldots \) logrank statistics are distributed as if \( \ldots \)\]

\[ \psi_1 = \gamma \times \theta_1 \quad \text{and} \quad \psi_2 = \gamma \times \theta_2 \]

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 \( = \lambda \times 300 \).

When the log hazard ratio is \( \theta \), 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

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

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.

Our preliminary investigations showed the Dunnett test to give the most efficient overall testing versions of both adaptive and non-adaptive designs.
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 $Z_1 \sim N(0, 1)$, $Z_2 \sim N(0, 1)$ and $\text{Corr}(Z_1, Z_2) = 0.5$. 
Comparing adaptive and non-adaptive trial designs

With selected values of $\psi_1, \theta_1, \psi_2, \theta_2$ and $\rho$, 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

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

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

Hence, we define the overall expected “Gain” or utility measure

$$E(\text{Gain}) = \theta_1 \times P(1) + \theta_2 \times P(2).$$
Comparing adaptive and non-adaptive trial designs

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

$$

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

$$

<table>
<thead>
<tr>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>$P(1)$</th>
<th>$P(2)$</th>
<th>$E(\text{Gain})$</th>
<th>$P(1)$</th>
<th>$P(2)$</th>
<th>$E(\text{Gain})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.0</td>
<td>0.78</td>
<td>0.00</td>
<td>0.235</td>
<td>0.86</td>
<td>0.00</td>
<td>0.259</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1</td>
<td>0.78</td>
<td>0.01</td>
<td>0.234</td>
<td>0.82</td>
<td>0.02</td>
<td>0.247</td>
</tr>
<tr>
<td>0.3</td>
<td>0.2</td>
<td>0.70</td>
<td>0.11</td>
<td>0.234</td>
<td>0.69</td>
<td>0.16</td>
<td>0.238</td>
</tr>
<tr>
<td>0.3</td>
<td>0.25</td>
<td>0.60</td>
<td>0.26</td>
<td>0.244</td>
<td>0.58</td>
<td>0.30</td>
<td>0.249</td>
</tr>
<tr>
<td>0.3</td>
<td>0.295</td>
<td>0.47</td>
<td>0.43</td>
<td>0.267</td>
<td>0.47</td>
<td>0.44</td>
<td>0.274</td>
</tr>
</tbody>
</table>

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

The adaptive design has higher $P(1)$ when $\theta_1$ is well above $\theta_2$.

With $\theta_1$ and $\theta_2$ 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 reliable indicators of treatment effects on OS.
- There must be good information on PFS at the interim analysis.

We have investigated varying the parameters $\gamma$ and $\lambda$ where

$$\psi_1 = \gamma \times \theta_1, \quad \psi_2 = \gamma \times \theta_2, \quad \text{with } \theta_1 = 0.3 \text{ and } \theta_2 = 0.1$$

- 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 $\gamma$ 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 $= \lambda \times 300$.

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>$\lambda$</th>
<th>Non-adaptive</th>
<th>Adaptive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$P(1)$</td>
<td>$P(2)$</td>
</tr>
<tr>
<td>1.5</td>
<td>1.2</td>
<td>0.88</td>
<td>0.00</td>
</tr>
<tr>
<td>1.2</td>
<td>1.1</td>
<td>0.85</td>
<td>0.01</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>0.78</td>
<td>0.01</td>
</tr>
<tr>
<td>0.9</td>
<td>0.9</td>
<td>0.78</td>
<td>0.03</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>(PFS is not used)</td>
<td>0.74</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7</td>
<td>0.68</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Adaptation works well when there is enough PFS information for treatment selection at the interim analysis.
Conclusions about the benefits of an adaptive design

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

2. Overall, 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, both non-adaptive and adaptive trials compare the two forms of treatment: the conclusions from this comparison may be more broadly useful.