Confirmatory Adaptive Group Sequential Trials: Treatment Arm and Subpopulation Selection Based on Surrogate Endpoints

Silke Jörgens
ICON Innovation Center
Generic Example

First part of trial

IA

Continue

Sample size recalculation?

FA

Stop for efficacy

Success

Stop for futility

Failure

1
Data Driven Changes

– The list of possible changes is in principle infinite

– Changes encountered in practice are mostly
  – Changes to the sample size (including changes to the allocation ratio)
  – Selection of treatment groups (from various doses or several treatment regimens)

– Selection of subpopulations also possible, but seen less frequently in practice:
  – Whenever the treatment is suspected to show benefit in a given subgroup only
    – Patients with a given biomarker (breast cancer example)
    – Severity of disease
    – Age of patients (especially interesting in studies in children)
Throughout this presentation, consider the inverse normal method for combining $p$-values over stages:

$$Z_k^* := \frac{w_1 \Phi^{-1}(1 - p_1) + \cdots + w_k \Phi^{-1}(1 - p_k)}{\sqrt{w_1^2 + \cdots + w_k^2}}$$

where

- $w_i, i = 1, \ldots, K$, denote prespecified weights for stages 1 to K
- $p_i, i = 1, \ldots, K$ denote the stagewise $p$-values for stages $i = 1, \ldots, K$

Applying this $p$-value combination method allows data driven changes to the design of the study.
Stagewise Data
The Situation in Time to Event Data

- Patient accrual
- Follow-up
- End of study
Time to Event Data

- Fundamental difference: Patients recruited into the first stage do not necessarily experience an event before the first interim analysis (true also for patients recruited into subsequent stages)
  → Each patient's data may contribute to multiple interim analyses

- Solution:
  - At each stage, the logrank test statistic is calculated based on overall data
  - Independent* increments from consecutive log rank test statistics can be calculated and used in inverse normal combination method (Wassmer, 2006)

- BUT: Data driven changes may only be informed by the current log rank test statistic (Bauer and Posch, 2004)

* Independency holds under $H_0$ only
Wishlist for Multiple Treatment Arms

- Consider many-to-one comparisons, e.g., G treatment arms and one control
- Consider one-sided testing
- In an interim stage treatment arms should be selected based on data observed so far
- Not only selection procedures, but also other adaptive strategies (e.g., sample size reassessment) should be allowed
Closed Testing Procedure

\[ H_0^1 \cap H_0^2 \cap H_0^3 \]

\[ H_0^1 \cap H_0^2 \]

\[ H_0^1 \]

\[ H_0^2 \]

\[ H_0^3 \]

\[ H_0^2 \cap H_0^3 \]

\[ H_0^3 \]
Closed Testing Procedure in Multi Stage Designs

\[ H_0^1 \cap H_0^2 \cap H_0^3 \]

\[ H_0^1 \]

\[ H_0^2 \]

\[ H_0^3 \]

... use inverse normal combination test for each pair of hypotheses (Bauer and Kieser, 1999; Posch et al., 2005)
Stage I

$\mu_0 = \mu_1 = \mu_2 = \mu_3$

$\mu_0 = \mu_1$  $\mu_0 = \mu_2$

$H_0^3$ can be rejected if all combination tests exceed the critical value $u_2$

Stage II

$\mu_0 = \mu_2 = \mu_3$

$\mu_0 = \mu_3$

$\mu_0 = \mu_3$
Population Selection

- **Stage 1 objective**
  - Stop for futility/efficacy
  - To continue with HER2- (Full) population
  - To confirm greater benefit in TNBC Subpopulation (Sub)
  - To adjust the sample size

- **Stage 2 data and the relevant groups from Stage 1 data combined**
Population Selection

- Testing procedure works in much the same way (testing hypotheses in multiple subgroups, instead of comparing multiple treatments to one control)

- Selection procedure can be performed to determine which subgroups should be tested in subsequent stages

→ This does not necessarily mean a change in enrollment criteria and patients' follow up (testing F+S uses the same patients as testing F only)
It’s All Very Nice in Theory

– But then reality happens:
Usefulness of Late Adaptations

– Sample size recalculation:
  – From theory, possible also after end of recruitment
  – From practice:
    – Sites may have been closed out already
    – Restarting recruitment gives information on interim analysis results
    – Recruitment gap prone to provoke heterogeneity over stages

– Treatment arm and patient population selection:
  – Again, possible from theory
  – Practical problems as above
  – In addition, paradoxical increase in sample size possible if all patients have been randomized before selection
Way Out Of This Dilemma

– Instead of using complete patients’ data, find an early outcome parameter to be used for selection
– Needs to be well correlated to primary outcome, but does not need to fulfill all requirements for a surrogate in the regulatory sense
– Examples are many:
  – Early readouts of primary parameter in case of multiple interim efficacy assessments in trial’s schedule of events
  – Oncology trials targetting overall survival
    – PFS
    – Early tumor response
  – Laboratory parameters predicting long-term outcomes
User’s Question

– Real life question in a project with early readout:

„Why do we need to adjust the analysis for the long-term responder rate for interim looks? The long-term responder rate was not looked at in the interim analysis, we analyzed the short-term responder rate only“

– The answer, as often, is selection bias
Selection Bias: Two Readouts

![Graph showing short-term and long-term data distributions](image-url)
Selection Bias: Only Patients With Positive Short-Term
... is an inflation of the type I error rate although we did not look at the long-term outcome in the interim.
Strategy For A Resolution Of This Dependency Issue

- We need to find a way to re-establish independency over stages

- Immediate solution: Only use short-term outcomes from patients participating in selection stage
  → This means that selection patients‘ long-term outcome is lost and in fact we are not better off than with separate trials

- Better solution: Allocate patients to the first interim analysis they entered, irrespective of when their long-term endpoint is observed
Strategy For Resolution Of Dependencies

\[ Z_K^* = w_1 \cdot \Phi^{-1}(1 - p_1) + w_2 \cdot \Phi^{-1}(1 - p_2) \]

Friede et al, 2011; Kunz et al, 2015
Drawbacks of This Strategy

– Selection is possible if short-term endpoint is thought to reasonably predict long-term endpoint
– Sample size recalculation is no longer an option in most cases:
  – Sample size recalculation based on available long-term outcomes not very reliable, as probably only few patients
  – Sample size recalculation for long-term endpoint based on short-term endpoint requires an exact knowledge of both variables’ link
  – Even if sample size recalculation is based on assumed effect (or other external information), the final value of the first stage’s test statistic is not known at the selection interim analysis and therefore conditional power cannot be calculated
The Situation in Time to Event Data

Patients

Patient accrual

Follow-up

Time $s$

End of study

24
How many events do we have to select on when the interim analysis is planned to take place?
Application to time to event data in principle straightforward

Instead of combining the independent increments of the logrank statistic (Wassmer, 2006), the p-values from the two populations, G1 and G2, are combined. For both populations, the p-values from accumulating data are used.

Formally, let $\tilde{p}_{k}^{G1}$ and $\tilde{p}_{k}^{G2}$ denote the p-values for testing the hypotheses in the closed system of hypotheses at stage k. For G1, this is a multiplicity adjusted p-value whereas for G2 it is multiplicity adjusted only if more than one subpopulation is considered for G2. At stage $k$ of the trial the combination test statistic

$$w\Phi^{-1}(1 - \tilde{p}_{k}^{G1}) + \sqrt{1 - w^2}\Phi^{-1}(1 - \tilde{p}_{k}^{G2})$$

is used, where $w$ is prefixed.
Application to Time To Event Data
Issues and Questions

- Three different ways to determine the G1 p-value
  1. Jenkins et al (2011)/Irle and Schäfer (2012): Prespecify end of follow-up, i.e., time point (calendar time) or number of events for calculation of $\tilde{p}_k^{G1}$: part of observed data ignored (mostly)
  3. Perform analysis at observed prespecified overall number of events (giving up the strict guarantee of Type I error rate control)

- The G1 population usually is small and hence yields large p-values. Therefore, for example, the use of the Bonferroni correction might yield adjusted p-value = 1, and the combination approach cannot reach rejection at a later stage irrespective of the outcome in Phase III. This defines a futility stop at interim.
Issues and Questions

- Asymptotic normality and the independent increments structure of the test statistic in both the G1 and the G2 population. Simulations show that the Type I error rate is controlled which is partly due to the conservatism of test procedures for the intersection hypotheses. → Can be resolved by not allowing early efficacy stopping

- A number of patients which have been randomized in a deselected subpopulation / to a deselected treatment arm are not used for further analysis.

- Furthermore, patients in G1 from deselected subpopulations / deselected treatment arms usually have discontinued follow-up (Friede et al., 2011). For these treatment populations, the test statistic is set equal $-\infty$ (or, equivalently, the p-value is set to 1).

- Choice of weight $w$ → often chosen based on assumed population sizes
Simulation of Such Trials

- Simulation, as always, needs a lot of setting which aim at mirroring real life

- Such trials are not automatons. Decisions will in reality made by humans, and conduct of the trial is influenced by operational constraints

- Software packages need to provide options which can cover a variety of scenarios, which makes them somewhat generic
Simulation of Such Trials
ADDPLAN – Multi Armed Time To Event Trials
Simple Simulation Result (Seamless Phase II/III vs Separate)
Simple Simulation Result (Seamless Phase II/III vs Separate)
Expected Number of Events (Seamless Phase II/III vs Separate)
Accrual period = 60
Accrual period = 10
Introducing Surrogate For Selection

– Selection based on a surrogate needs the same assumptions plus a specification of the surrogate

– Two options: Either continuous or binary surrogate
  – In the continuous case:
    – Specify magnitude of effect and variability
    – Specify correlation with primary event probability at given time point
  – In the binary case:
    – Specify surrogate event probability
    – Specify prediction or sensitivity based correlation between event probabilities
Surrogate Options

Weight for combination test
Surrogate and correlation type
Surrogate effect size assumptions
Accrual period = 60
Saving Patients By Selection!

Accrual period = 10
Continuous Surrogate

- Spearman’s rank correlation used between continuous outcome
- We assume throughout that treatment arms (and populations) with high values of the surrogate are selected
- Direction of meaningful correlation depends on direction of power:
  - If power is directed towards $\omega > 1$: Correlation should be negative (because a low time to event is desirable)
  - If power is directed towards $\omega < 1$: Correlation should be positive (because a high time to event is desirable)
Impact Of The Correlation

![Graph showing the relationship between Power and Drift with different correlation values. The graph includes lines for various correlation coefficients, indicating how power changes with drift at different correlation levels.](image-url)
Is Not Always There!
And Sometimes, It’s Not The Correlation
Binary Surrogates

- Binary surrogates form one fourfold table with the primary endpoint (in this case, event probability within specified time frame) within each treatment/subgroup

- Two forms of correlation can be given:
  - Prediction: How many of the patients with a surrogate event experience a primary event within given time frame?
  - Sensitivity: How many of the patients with a primary event within given time frame have experienced the surrogate event?

- In both cases, margin probabilities and one of the four cells are fixed

- This means that there may be specifications without a solution
Consider a three-stage design with O’Brien & Fleming boundaries

- One subgroup S (with prevalence 40%), selecting either S or F in first interim analysis, based on better hazard ratio (i.e., assume a “positive” event)

- Assume that the predictive value for the surrogate is good (e.g., 80%)

- Planned number of patients:
  - Maximum of 400 in total
  - 50 for subpopulation selection
  - Weight for phase II is 0.125
Binary Surrogates - Example

- In the primary endpoint, we assume
  - a control event rate at 12 months of 53% for both $S$ and $S^c$
  - a treatment group event rate in $S$ of 70%
  - a treatment group event rate in $S^c$ of 59%

- In the surrogate, we assume
  - a control event rate at 12 months of 10% for both $S$ and $S^c$
  - a treatment group event rate in $S$ of 50%
  - a treatment group event rate in $S^c$ of 30%

- Need to construct three 2x2 tables (control, active in $S$, active in $S^c$)
## Binary Surrogates - Example

<table>
<thead>
<tr>
<th>Primary</th>
<th>Surrogate</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
<td>0.47</td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td>0.08</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.10</td>
<td></td>
</tr>
</tbody>
</table>
### Binary Surrogates - Example

<table>
<thead>
<tr>
<th>Primary</th>
<th>Surrogate</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>0.45</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>0.45</td>
<td><strong>0.08</strong></td>
</tr>
<tr>
<td></td>
<td><strong>0.90</strong></td>
<td>0.10</td>
<td></td>
</tr>
</tbody>
</table>
## Binary Surrogates - Example

<table>
<thead>
<tr>
<th>Primary</th>
<th>Surrogate</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td></td>
<td></td>
<td>0.30</td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td>0.40</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
<td>0.50</td>
</tr>
</tbody>
</table>
## Binary Surrogates - Example

<table>
<thead>
<tr>
<th>Active in S</th>
<th>Surrogate</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>0.20</td>
</tr>
<tr>
<td>Yes</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
</tr>
<tr>
<td>Primary</td>
<td>Surrogate</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td></td>
<td>No</td>
</tr>
<tr>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.70</td>
</tr>
</tbody>
</table>

Active in $S^c$
|| Surrogate | No | Yes |
|---|---|---|
| Primary | No | 0.35 | 0.06 |
| | Yes | 0.35 | 0.24 |
| | | 0.70 | 0.30 |
Overall Power

Power vs. Effect S1

- Correlation
  - red: corr = 0.6000
  - green: corr = 0.8000
  - blue: corr = 0.9000
Power for Subgroup

Effect S1

P_R reject S1 vs. Effect S1

Correlation:
- pred = 0.6000
- pred = 0.8000
- pred = 0.9000
Power for Subgroup


