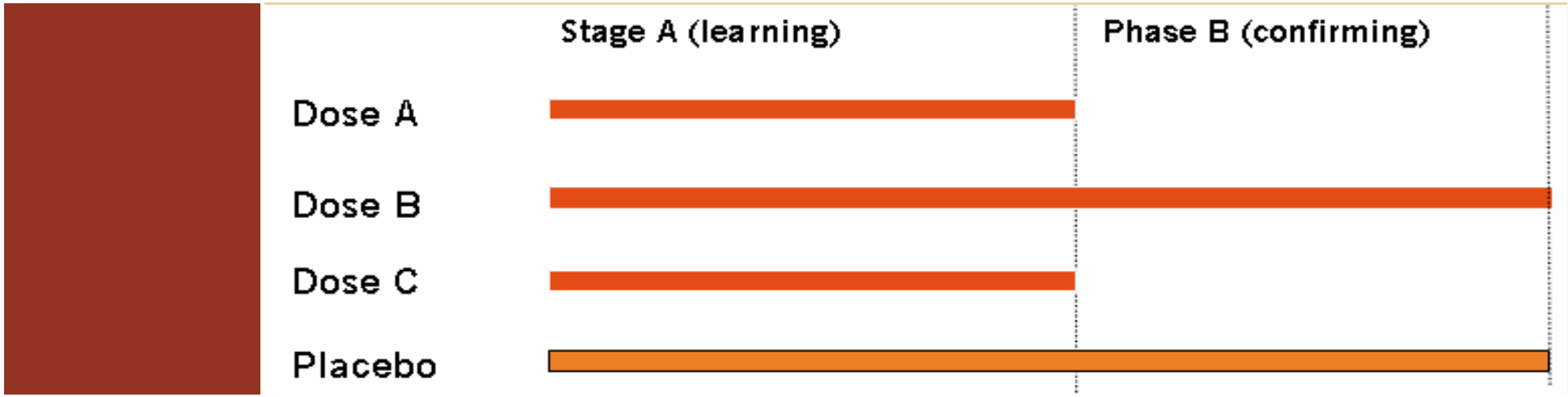


Lessons learned from designing an adaptive clinical trial with time-to-event as primary endpoint

Ekkehard Glimm and Lilla Di Scala, Novartis Pharma

BBS Seminar, 12 March 2010

NOVARTIS

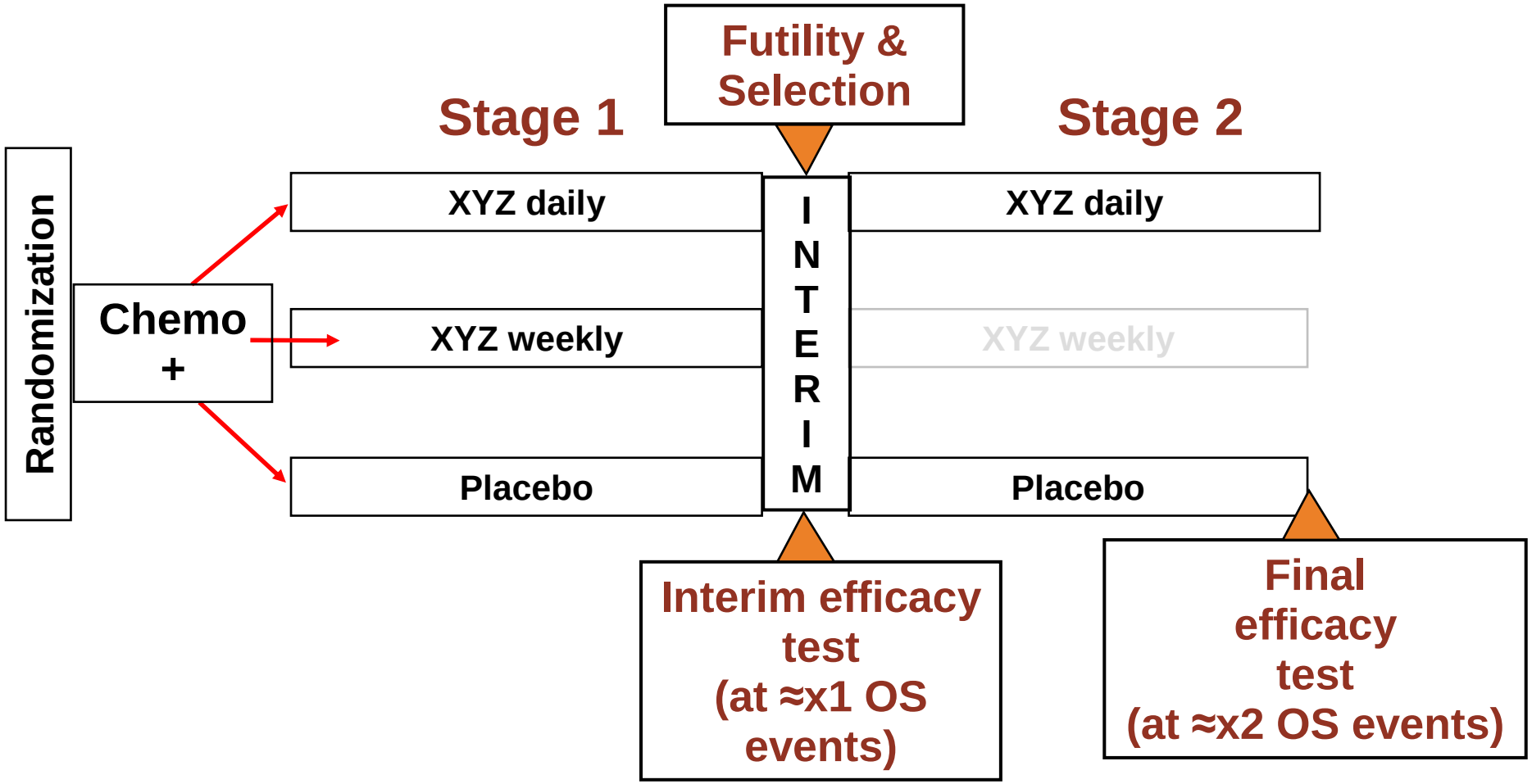
Introduction

Development setting

- Indication: Lung Cancer
- Experimental compound: Oral agent XYZ001A
- Experimental treatment: XYZ+ Chemotherapy
- Regimens: XYZ daily & XYZ weekly
- Trial Objectives:
  - Superiority of at least one regimen over chemotherapy
  - Treatment regimen selection
- Primary endpoint: Overall Survival (OS)
- Proposed design:

  Adaptive Phase 2/3 with interim selection of one XYZ arm

Design at a glance



NOVARTIS

Design elements

- **PFS** (progression-free survival) as surrogate endpoint (correlated with OS), associated with tumor growth
- Desire: treatment selection not only based on OS, but also on PFS
- Interim and final foresee decisions
  - interim: treatment selection & stop efficacy/futility
  - proof of efficacy must be on OS and must comply with confirmatory requirements (type I error control)
- Setting the interim decision criteria: „internal“ risk-taking, to be handled with care (regulatory and operationally -wise)

$\Rightarrow$ Strategy: Hybrid Bayesian-Frequentist Adaptive Design

Frequentist aspect: confirming efficacy

Notation:

$$l_{ij} = \frac{\sum_{k=1}^{d_i}(\delta_{kj} - p_{kj})}{\sqrt{\sum_{k=1}^{d_i} p_{kj}(1-p_{kj})}}$$

is the log-rank test statistic of **all** events d_i in treatment j vs control after stage i=1,2.

δ_{kj} = 1 if the kth event is in j, 0 otherwise. t_k time of kth event.

p_{kj} = (#patients at risk at t_k in j)/(#patients at risk at t_k in j or control)

$$i_{ij} = \sqrt{\sum_{k=1}^{d_i} p_{kj}(1-p_{kj})}$$ information on j at stage i.

Frequentist aspect: confirming efficacy

- Statistical test for OS benefit: log-rank test on OS only
- Stop for efficacy at interim is possible.
- Type I error control: O'Brien-Fleming type boundaries with Lan-deMets α-spending approach
- Dunnett's test applied at Interim to test for OS benefit of any of the two XYZ arms.
- If $\max(l_{11}, l_{12}) \geq c_1$, stop for efficacy.
  - c_1 critical value corresponding to the α-spending approach.
- If not, select a treatment arm *S* and claim efficacy if $l_{2S} \geq c_2$.
  - Different ways of calculating $c_2 \rightarrow$ discussed below

Bayesian aspect: futility and treatment selection

- Futility analysis at interim: look at the predictive probability of final OS benefit above a fixed futility threshold for any one of the treatment arms (e.g. 35%)
- Treatment selection at interim based on predictive power of claiming success in terms of OS benefit
- Treatment selection rules investigated are based on
  - PFS only (=surrogate endpoint only)
  - PFS+OS (utility function approach)
  - OS only (benchmark design)

It involves the joint modelling of vector [log HR_{PFS},log HR_{OS}] as normal multivariate vector and prior setting

Frequentist aspect: confirming efficacy

logrank test with two treatments and control

Independent increments:

$$\tilde{l}_{2j} = \frac{i_{2j} l_{2j} - i_{1j} l_{1j}}{\sqrt{i_{2j}^2 - i_{1j}^2}}$$

Approximate independent increments distribution:

$$\begin{pmatrix} l_{11} \\ l_{12} \\ \tilde{l}_{21} \\ \tilde{l}_{22} \end{pmatrix} \sim N\left(\begin{pmatrix} i_{11}\theta_1 \\ i_{12}\theta_2 \\ \sqrt{i_{21}^2 - i_{11}^2}\theta_1 \\ \sqrt{i_{22}^2 - i_{12}^2}\theta_2 \end{pmatrix}, \begin{pmatrix} 1 & v_{11,12} & 0 & 0 \\ v_{11,12} & 1 & 0 & 0 \\ 0 & 0 & 1 & v^*_{21,22} \\ 0 & 0 & v^*_{21,22} & 1 \end{pmatrix} \right)$$

θ_j hazard ratio of treatment j vs control

i_{ij}^2 information accrued on treatment i (essentially, number of events)

$v_{11,12}$ correlation between l_{11} and l_{12} (approximately 0.5 under global H_0 and equal sample sizes)

$v^*_{21,22}$ correlation between $\tilde{l}_{21}$ and $\tilde{l}_{22}$ (approx. 0.5, but unobserved if one trt dropped)

Conservative Dunnett approach

Assume the study is continued into stage 2. One treatment arm S is selected.
How is treatment arm selection done? → See later.

- Calculate c_2 such that

$$Pr_{H_0}\left(max(l_{11}, l_{12}) < c_1, max(l_{21}, l_{22}) \geq c_2\right) =$$
$$\Phi_{v_{11,12}}(c_1, c_1) - \Phi_{\Sigma}(c_1, c_1, c_2, c_2) = \alpha - \alpha_1$$

  - Reject H_0 if $l_{2S} \geq c_2$

Remarks:

- The method is conservative: It compares l_{2S} to a critical value intended for max(l_{21}, l_{22}).
- The method is approximate (in addition to usual TTE-approximations): Need to approximate the unobserved correlation $v_{21,22}$ of two stage-2-test statistics.

Conditional Error Function Approach (König et al., SiM 2008)

- Calculate c_{12} (= c_2) and c_S such that

$$Pr_{H_0}\left(max(l_{11}, l_{12}) < c_1, max(l_{21}, l_{22}) \geq c_{12}\right) = \alpha - \alpha_1$$

and

$$Pr_{H_0}\left(max(l_{11}, l_{12}) < c_1, l_{21} \geq c_S\right) = \alpha - \alpha_1$$

- Calculate conditional rejection boundaries q_{12} and q_S

$$q_{12} = Pr_{H0}(max(l_{21}, l_{22}) \geq c_{12} | l_{11}, l_{12})$$
$$q_S = Pr_{H0}(l_{2S} \geq c_S | l_{1S})$$

- Reject H_0 if $p_{2S} < min(q_{12}, q_S)$

Remarks:

- Also needs approximations of correlations in joint asymptotic distribution of (l_{11},l_{12}, l_{21}, l_{22}).
- Not conservative.

Combination *p*-value approach (Lehmacher & Wassmer, Biometrics 1999)

- Fix weights $w_1>0$, $w_2>0$, $w_{12}^2+w_{22}^2=1$.
- Obtain p-values

$$p_{1,12} = 1 - \Phi_{0.5}\left(max(l_{11}, l_{12}), max(l_{11}, l_{12})\right)$$

$$p_{1S} = 1 - \Phi\left(l_{1S}\right)$$
$$p_{2S} = 1 - \Phi\left(\tilde{l}_{2S}\right)$$

- Reject H0 if

$$w_1 \cdot \Phi^{-1}(1 - max(p_{1,12}, p_{1S})) + w_2 \cdot \Phi^{-1}(1 - p_{2S}) \geq c_2$$

Remarks:

- $w_1^2 = \frac{d_1}{d_2}$, $w_2^2 = 1 - \frac{d_1}{d_2}$

- very similar to CEF-approach: This can be formulated as a CEF-approach (Posch & Bauer, 1999), but it is not the one from the previous slide.

What's more complicated than in non-TTE situations?

- Trial stages are not "automatically" stochastically independent, independent increments are only asymptotically independent.
- Many relevant quantities must be approximated:
  - information fractions
  - correlations
  - these in turn impact critical values.

  - For testing, approximations are required under the null hypothesis. But we are approximating under a global null which may not hold.

- How well do we keep the type I error in reality?
- Does the quick convergence of the usual log rank test to its asymptotic distribution still hold here?

Most critical of these: Independent increments.

Independent increments

- $\tilde{l}_{2j}$ is not simply the log-rank test of events after the interim.
- Independence is asymptotic.
- *Any* knowledge which is associated with the value of l_{2j} destroys the independent increments property if
  - it is available at stage 1 already,
  - it is not entirely captured by l_{1j}

An extreme example:

Two-arm trial with an interim analysis, death as event type.

- no selection, no stopping, no reassessment of total events accrued
- only the recruitment rate is changed based on auxiliary information (e.g.: slowed down if ratio of progression events in trt vs. control is "large", accelerated if it is "small")

$\Rightarrow$ No independent increments.

Type I error control

- Potential α-level violation is limited because only treatment arm selection is done
- Conservative Dunnett approach does keep α asymptotically.
- CEF and combination p-value approach do not, if PFS is used in decision making and is not stochastically independent of OS.
- However, inflation is very small ($\rightarrow$ simulations)
- Potential way out:
  - Split test statistic into stage-1- and stage-2-*recruits* rather than using independent increments.
  - Not done here because it would not allow stopping at interim.

Bayesian aspect: futility and selection

Bayesian Decision-making at interim: futility and treatment selection

- **Futility analysis at interim**: look at the predictive probability of final OS benefit above a fixed futility threshold for any one of the treatment arms (e.g. 35%)
- **Treatment selection at interim** based on predictive power of claiming success in terms of OS benefit
- Treatment selection rules investigated are based on
  - PFS only (=surrogate endpoint only)
  - PFS+OS (utility function approach)
  - OS only (benchmark design)

Treatment Arm Selection and Futility: Predictive Power

Idea:

- Using Bayesian predictive power to decide on treatment arm or stop
- "Borrow strength" by including PFS events in the decision

Predictive power: Probability of rejecting after stage 2 given stage 1-data and a prior distribution for HRs (θ_1, θ_2).

With a vague prior $\binom{\theta_1}{\theta_2} \sim N(\theta_0, \mathbf{I}_0^{-1})$, $\mathbf{I}_0 \to 0$, this gives

$$prob_j = Pr\left(\hat{\theta}_{2j} > c \middle| \hat{\theta}_{1j}\right) = 1 - \Phi\left(\left(I_{1j}^{-1} + (I_{2j} - I_{1j})^{-1}\right)^{-1/2}\left(c - \hat{\theta}_{1j}\right)\right)$$

as predictive power for treatment j ($I_{ij}=i_{ij}^2$).

This is calculated for both PFS and OS.

Treatment Arm Selection and Futility: Predictive Power

Threshold selection:

$prob_{j,OS}$, $prob_{j,PFS}$ predictive probability for treatment j

- Threshold for futility: Stop if both

$$\max_j(prob_{j,OS}) < t_{OS} \text{ and } \max_j(prob_{j,PFS}) < t_{PFS}$$

(fixed threshold, e.g. $t_{OS} = t_{PFS} = 0.35$)

- Treatment arm selection:

$$util_j = w_j \cdot prob_{j,PFS} + (1 - w_j) \cdot prob_{j,OS}$$

Example for weights (we tried others as well, see simulation section):

$$w_j = \frac{d_{1j,PFS}}{d_{1j,PFS} + 2 \cdot d_{1j,OS}}$$

d_{1j} number of events, trt j

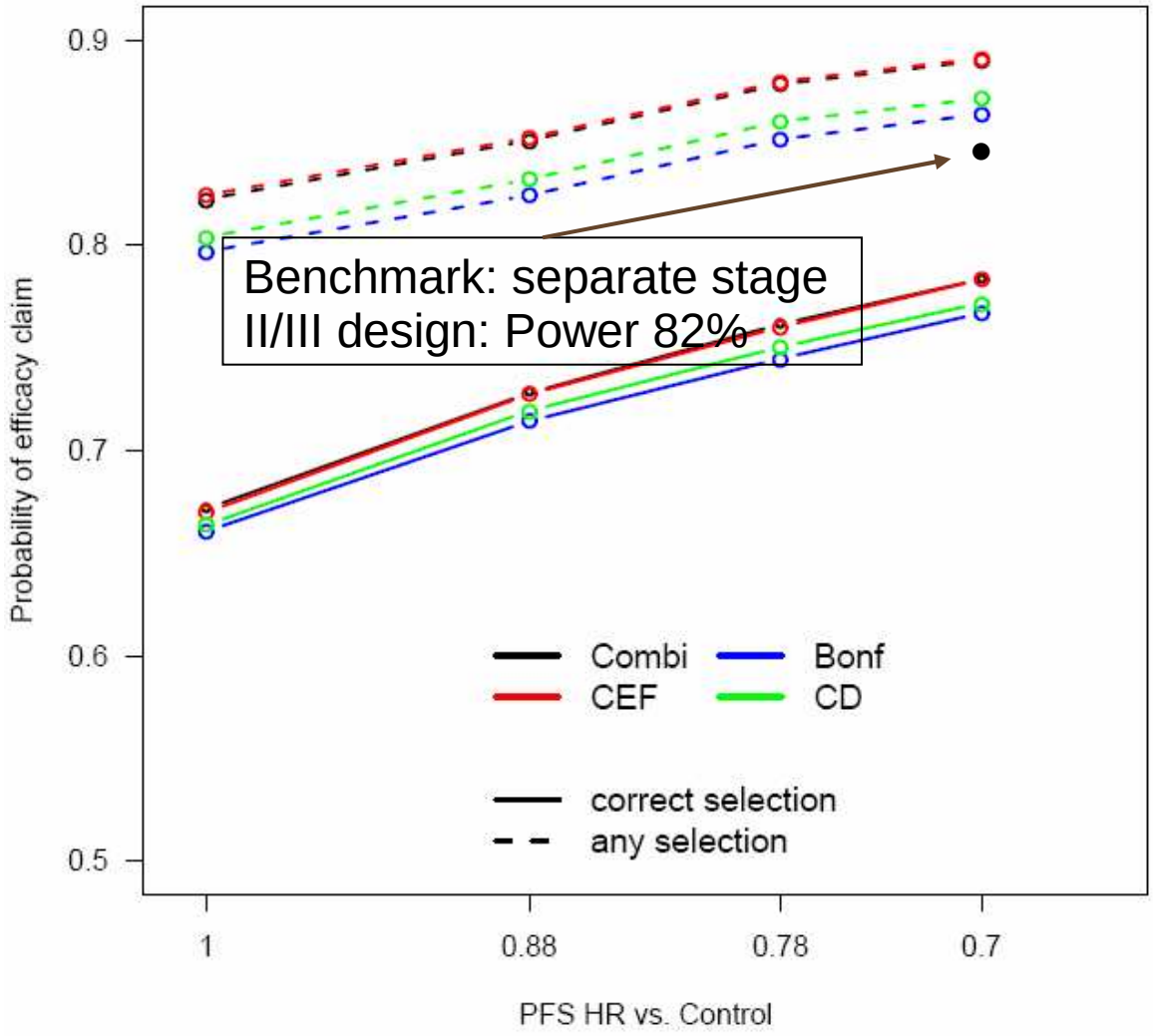
Simulations

Simulation set-up

- Total number of recruited patients: 1000; randomization ratio [1:1:1]
- event-driven setup: interim & final analysis at a prespecified number of OS events:
  - final at 600
  - interim at 20% or 30% of final
- Futility threshold for predictive power: 35%
- Treatment selection at interim: select „best“ experimental arm as judged by "utility" (combination of predictice powers of PFS and OS), varying combination rules.
- Number of trials simulated per scenario: 10'000
- Recruitment scheme: staggered, av. 0.5 patients x month x site, staggering based on site availability: 25m accrual; min. follow-up 6 m
- Bivariate exponential distribution with varying median times-to-event and correlations

Simulation results: *Power*

- **Optimistic scenario = fixed OS benefit of HR=0.75**
- **Power across different PFS assumptions (left-right=PFS good-bad surrogate)**
- **Selection criterion: PFS+OS-based on fixed weights**
- **Correlation ρ=0.4**



------ Combination *p*-value

------ CEF

------ Bonferroni

------ Dunnett

NOVARTIS

Simulation results: *Type I error*

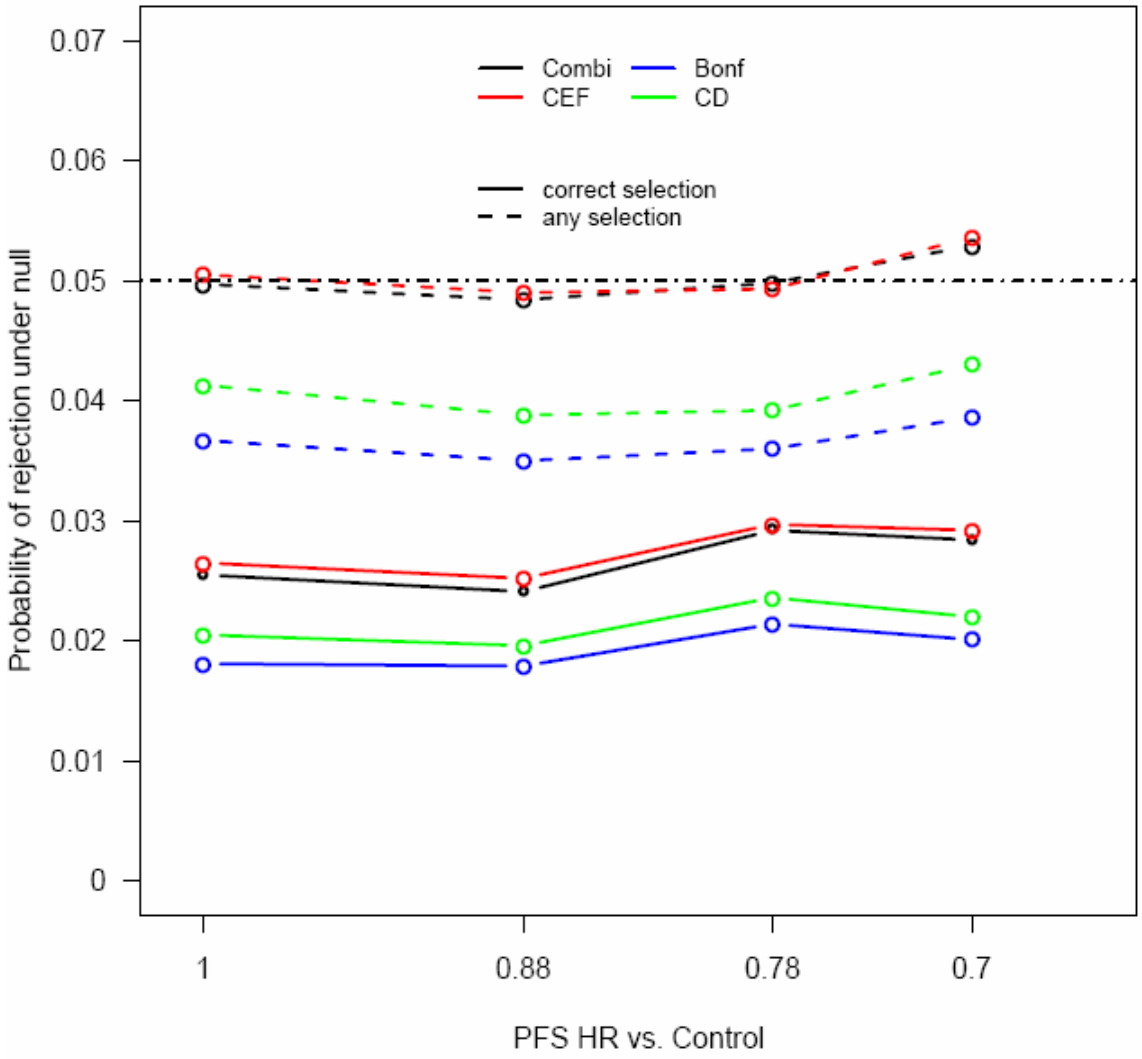
- **Null hypothesis=no OS benefit**
- **Rej prob across different PFS assumptions (left-right=PFS no-high PFS effect)**
- **Selection criterion: PFS+OS-based on fixed weights (same for each arm)**
- **Correlation ρ=0.4**
- **No stop for futility**

------ Combination *p*-value

------ CEF

------ Bonferroni

------ Dunnett



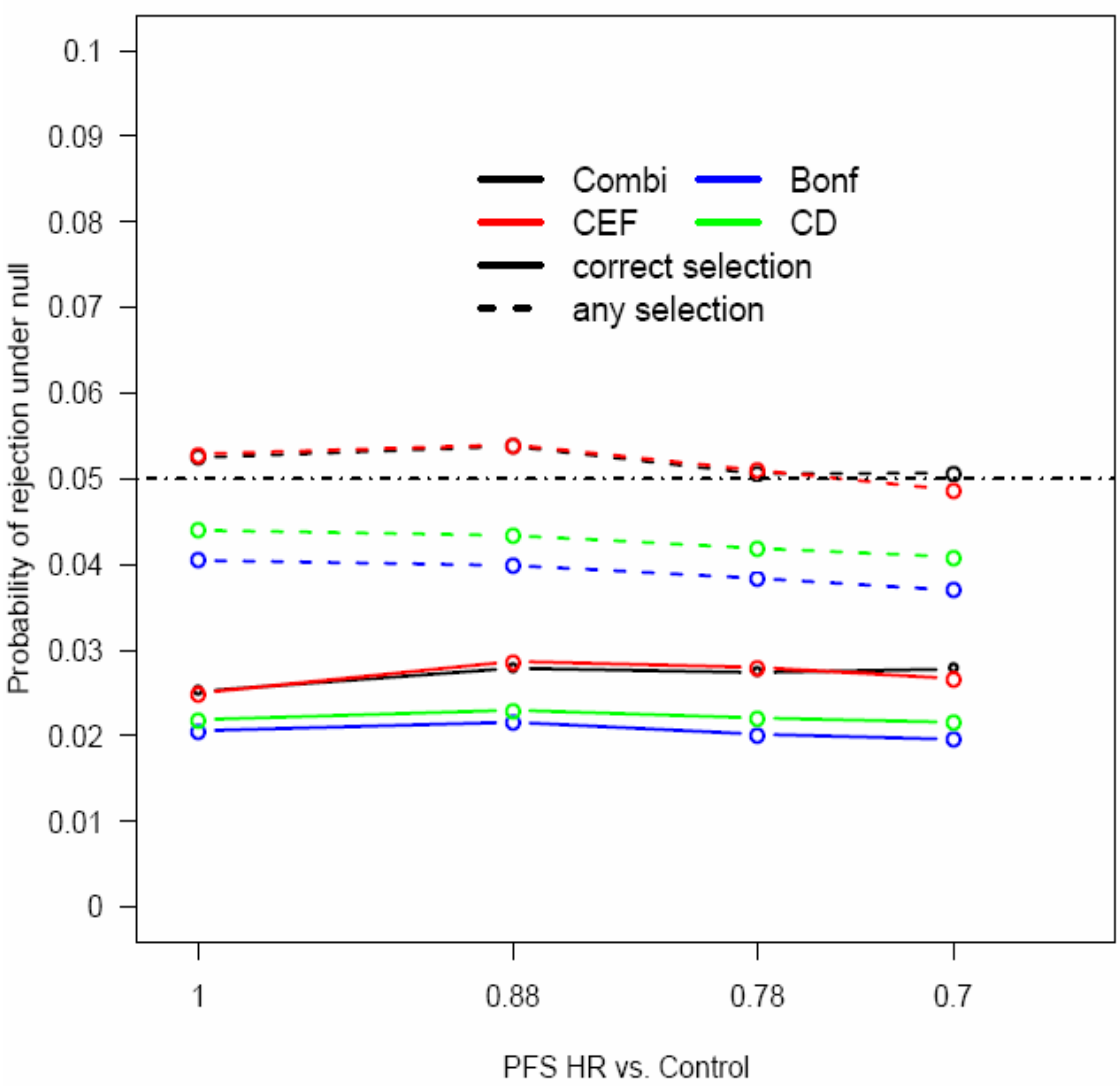
NOVARTIS

Simulation results: *Type I error*

- **Null hypothesis=no OS benefit**
- **Rej prob across different PFS assumptions (left-right=PFS no-high PFS effect)**
- **Selection criterion: PFS+OS-based on fixed weights (same for each arm)**

"worst case:"

- **Correlation ρ=0.9**
- **no stop for futility**

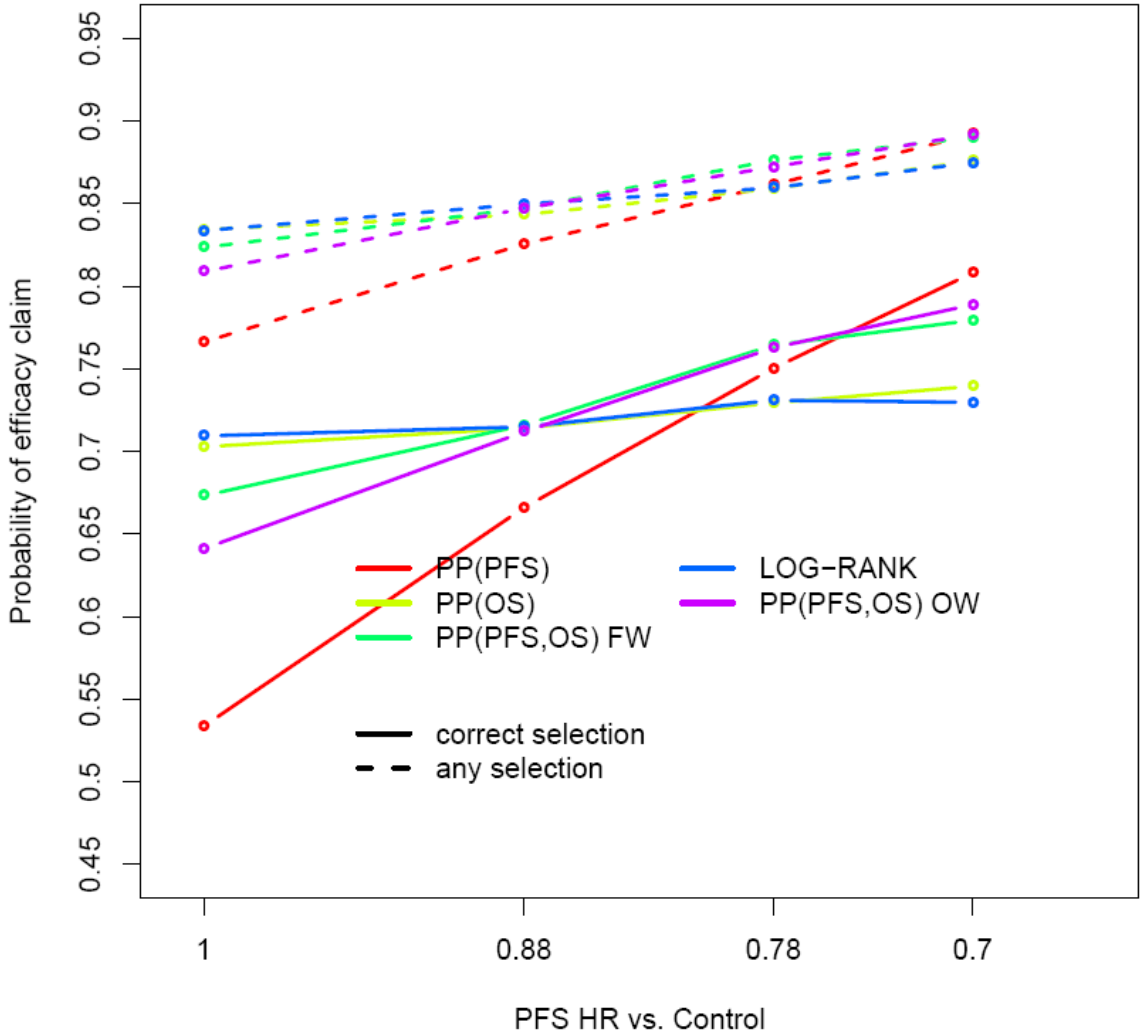


NOVARTIS

Simulation results: *comparison across decision criteria*

5 ways to decide at interim:

- PP(PFS): predictive power of PFS
- PP(OS): predictive power of OS
- FW= fixed weights (1/3 PFS, 2/3 OS)
- OW=observed weights
- ‘log-rank’=LR statistics-based (no Bayes) for OS



NOVARTIS

Discussion of simulation results

- Many other simulations results back up this summary
- Type I error is well controlled
  - irrespective of correlation between PFS and OS
  - for various ways of combining PFS and OS predictive power
- CEF and combination p-value approach give very similar results, are better than conservative Dunnett and Bonferroni
- Combined interim decision rules are better than decision on OS alone, gain in precision depends on correlation PFS to OS and on PFS hazard ratio.

Conclusions

Conclusions

- General methods of adjusting for treatment arm selection (CEF, combination p-values) are applicable in the TTE context as well
- Some care and some (more*) approximations are needed
- In this context, methods behave well in terms of type I error control
- These methods are flexible: no need to carry forward the "best" treatment in sense of efficacy only (*if safety concerns*).
- Borrowing strength is achieved from using a Bayesian predictive power for interim decision about futility and treatment selection, using PFS and OS events combined into a utility index.

*than in the ordinary non-adaptive TTE context

Issues up for discussion

- Which testing approach better suited for Final testing? Pros&Cons... (tradeoff between power gain and simplicity/communicability)
- Selection criteria for for interim decision making? Which is more efficient but also...easy to communicate?
  - relative importance of an OS event *vs.* a PFS event?
  - fixed or observed weighing scheme? same for each arm?
- Joint modeling of PFS+OS: how to be more informative?
- Regulatory issues:
  - In particular: What about type I error assessment by simulation?

THANK YOU all for the attention !!!!

Backup

Type I error control

- Potential α-level violation is limited because only treatment arm selection is done
- Conservative Dunnett approach does keep α asymptotically.
- CEF and combination p-value approach do not, if PFS is used in decision making and is not stochastically independent of OS.
- However, inflation is very small ($\rightarrow$ simulations)
- Potential way out:
  - Split test statistic into stage-1- and stage-2-*recruits* rather than using independent increments.
  - Then the two test statistics are truly (not only asymptotically) independent.
  - "Free" decision making at interim on stage-1-recruits.
  - Disadvantage: "interim" test can only be done at final analysis, no early efficacy stop possible.

- If only treatment arm selection:

  Dunnett actually does keep alpha, but CEF and p-val combi not.

  Refer to later sims for assessment

  Q to regulators
- Way out: Split into stage-1 and stage-2 *recruits* rather than using independent increments. Discuss pros and cons (or at least hint at this: Why didn't we do that?)
- Then Bayesian aspect.
- Then sim setup and sims.
- Need to talk about change in interpretation of stage-1 and stage-2 efficacy?

Some complications

- α level control:
  - independent increment property easily violated
  - asymptotics of approximation of logrank test via multivariate normal distribution
- interpretation of successful outcome (early stop for efficacy vs. final significance of selected arm)
- operational issues:
  - overrun in the deselected treatment arm
  - recruitment

Independent increments

$$l_{ij} = \frac{\sum_{k=1}^{d_i}(\delta_{kj} - p_{kj})}{\sqrt{\sum_{k=1}^{d_i} p_{kj}(1-p_{kj})}}$$

is the log-rank test statistic of **all** events d_i in treatment j vs control after stage i=1,2.

δ_{kj} = 1if the kth event is in j, 0 otherwise. t_k time of kth events.

p_{kj} = (#patients at risk at t_k in j)/(#patients at risk at t_k in j or control)

Independent increments:

$\tilde{l}_{2j} = \frac{i_{2j}l_{2j} - i_{1j}l_{1j}}{\sqrt{i_{2j}^2 - i_{1j}^2}}$ and l_{ij} are stochastically independent.

$i_{ij} = \sqrt{\sum_{k=1}^{d_i} p_{kj}(1-p_{kj})}$ information on j at stage i.

logrank test with treatment arm selection

Approximate independent increments distribution:

$$\begin{pmatrix} l_{11} \\ l_{12} \\ \tilde{l}_{21} \\ \tilde{l}_{22} \end{pmatrix} \sim N\left(\begin{pmatrix} i_{11}\theta_1 \\ i_{12}\theta_2 \\ \sqrt{i_{21}^2 - i_{11}^2}\theta_1 \\ \sqrt{i_{22}^2 - i_{12}^2}\theta_2 \end{pmatrix}, \begin{pmatrix} 1 & v_{11,12} & 0 & 0 \\ v_{11,12} & 1 & 0 & 0 \\ 0 & 0 & 1 & v^*_{21,22} \\ 0 & 0 & v^*_{21,22} & 1 \end{pmatrix} \right)$$

θ_j hazard ratio of treatment j vs control

i_{ij}^2 information accrued on treatment j (essentially, number of events)

$v_{11,12}$ correlation between l_{11} and l_{12} (approximately 0.5 under global H_0 and equal sample sizes)

$v^*_{21,22}$ correlation between $\tilde{l}_{21}$ and $\tilde{l}_{22}$ (approx. 0.5, but unobserved)