On the Feasibility of Phase II/III Studies in Oncology

Werner Brannath, MUW, Vienna

Uli Burger, Hoffmann-La Roche, Basel
Maximo Carreras, Hoffmann-La Roche, Basel

BBS Early Spring Conference 2010, Basel

Scope of the presentation

- Dose finding in oncology
- Surrogates in oncolcogy and their problems
- Proposals and limitations for seamless phase II/III designs
- Comparison of different methods
- Summary

Dose finding in oncology

- Classical paradigm:
  - Dose defined in Phase I based on safety endpoints only (e.g. 3 plus 3 designs or CRM designs)
  - Chose highest dose with acceptable safety profile (MTD - maximal tolerated dose)
  - Explore in phase II if MTD is active
  - Proof in phase III that it is more active than standard
- Classical paradigm today often outdated
  - Not all new treatments are cytotoxics, i.e., highest dose not always solely most effective dose.
  - In this case MTD usually not well defined
    $\Rightarrow$ Breakdown of classical paradigm
    $\Rightarrow$ Other dose finding steps are needed
  - Efficacy plays a major role already for dose finding

Dose finding in oncology based on efficacy

- Potential endpoints are
  - response rate (RR; e.g. 50% reduction in tumor size)
  - progression-free survival (PFS; e.g. time till $\geq$ 25% increase in tumor size) or
  - overall survival (OS).
- RR and PFS usually viewed as surrogate for OS.
- OS only clinically relevant endpoint
- Level of surrogacy of RR and PFS for OS usually unknown.

Dose finding in oncology - Problems

- All three endpoints statistically insensitive, leading to large and time consuming phase II dose finding trials.
- One-sided $\alpha = \beta = 0.2$
  - RR: Test 10% difference in RR $\Rightarrow$ about 280 patients
  - PFS: Test 25% reduction in HR in PFS $\Rightarrow$ 137 events
- Solutions:
  - Using of more sensitive biomarkers
  - Use of adaptive design allowing larger phase II dose finding parts while still maintaining overall size of the program

  $\Rightarrow$ How should such designs look like?

Surrogates in oncology – Problems

- When using a surrogate for phase II we face:
  - How to describe the relationship between RR and OS?
  - What difference in RR do we need to observe to believe in a difference in OS?
- Example:
  - Usual thinking is that a difference in PFS or OS is induced by a difference in response.
  - When this would be true, most oncology treatments would not be sufficiently effective. Necessary response rate difference is above 30% under standard assumptions to yield meaningful difference in PFS solely caused by difference in RR.
  - Effect on PFS or OS is usually composite, increases response rate and effects survival within each response category.

Surrogates in oncology – Problems

- Surrogate model for RR for an experimental treatment E versus a standard treatment S can be described as

  - Cytotoxic effect: Increase in RR by $\Delta = \theta_E - \theta_S$

  - Cytostatic effect: Prolongation in OS within responder and non-responder subgroups

- If the effect of E on OS is working only through the response category, then RR is a perfect surrogate.
- But usually: Effect of E on OS is partially an effect working through change in response category and change in OS within category.

Surrogates in oncology - Model

$S_g(t)$ survival function,
$\lambda_g(t)$ hazard function in groups $g = E$ or S,

With $S_{g,R}(t)$, $\lambda_{g,R}(t)$ and $S_{g,N}(t)$, $\lambda_{g,N}(t)$ for responder (R) and non-responder (N):

$$\lambda_g(t) = \theta_g \, \lambda_{g,R}(t) + (1 - \theta_g) \, \lambda_{g,N}(t)$$

Assuming proportional hazards: $\lambda_{S,N}(t) = \lambda(t)$ and $\lambda_{S,R}(t) = r \, \lambda(t)$, $\lambda_{E,N}(t) = c\lambda(t)$, $\lambda_{E,R}(t) = cr\lambda(t)$,

$$\Longrightarrow \quad \text{HR}_{E/S} = \frac{\lambda_E(t)}{\lambda_S(t)} = c \cdot \left\{ 1 - \frac{(1-r)\,(\theta_E - \theta_S)}{1 - (1-r)\,\theta_S} \right\}$$

Surrogates in oncology – Examples

- $\theta_S = 20\%$ (RR in S); $r = 0.50$ (HR R/N).

  Cytotoxic effect, but no cytostatic effect:
  $\theta_E = \theta_S + 10\% = 30\%$ and $c = \lambda_{E,R}(t)/\lambda_{S,R}(t) \sim 1$,

  $$\Rightarrow \quad \text{HR}_{E/S} = 0.95$$

  (Capecitabine 1st line CRC monotherapy, true rate ~ 0.93?)

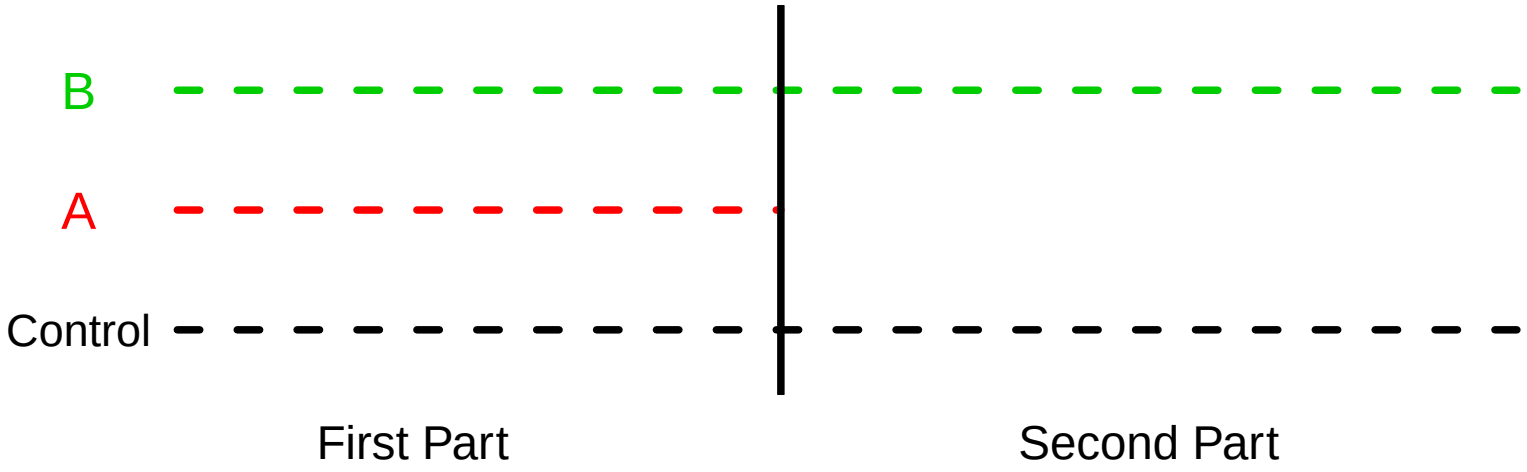
- $\theta_S = 57\%$ (RR in S), $r = 0.50$ (HR R/N)

  Cytotoxic effect and cytostatic effect:
  $\theta_E = \theta_S + 24\% = 81\%$ and $c \sim 0.6$,

  $$\Rightarrow \quad \text{HR}_{E/S} = 0.50$$

  (Rituximab-Chemo in NHL, observed HR =0.41)

Adaptive Designs with Treatment Selection



- *First part:* start with several treatment arms.
- *Interim analysis (IA):* select treatments, e.g., based on RR.
- *Second part:* continue recruitment only in selected arms.
- *Final analysis:* test treatment efficacy of the selected treatments from the data of both parts.

Testing efficacy of selected treatments

(BAUER & KIESER 1999, HOMMEL 2001, . . .)

Testing strategy which combines two approaches:

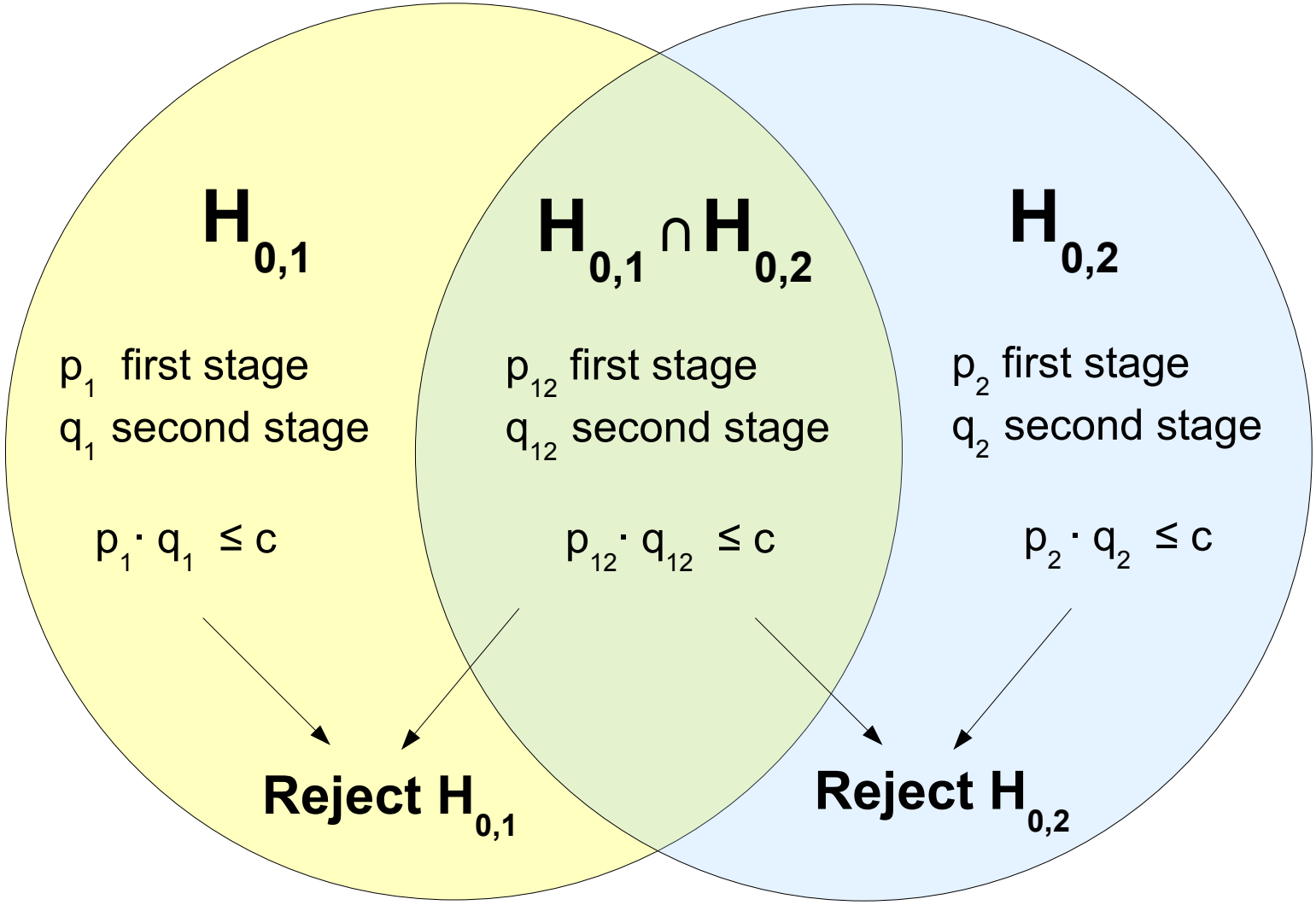
1) Closed Testing Principle
   to control the multiple type I error rate;

2) Combination Test (Conditional Error Function Approach)
   to cope with the adaptations

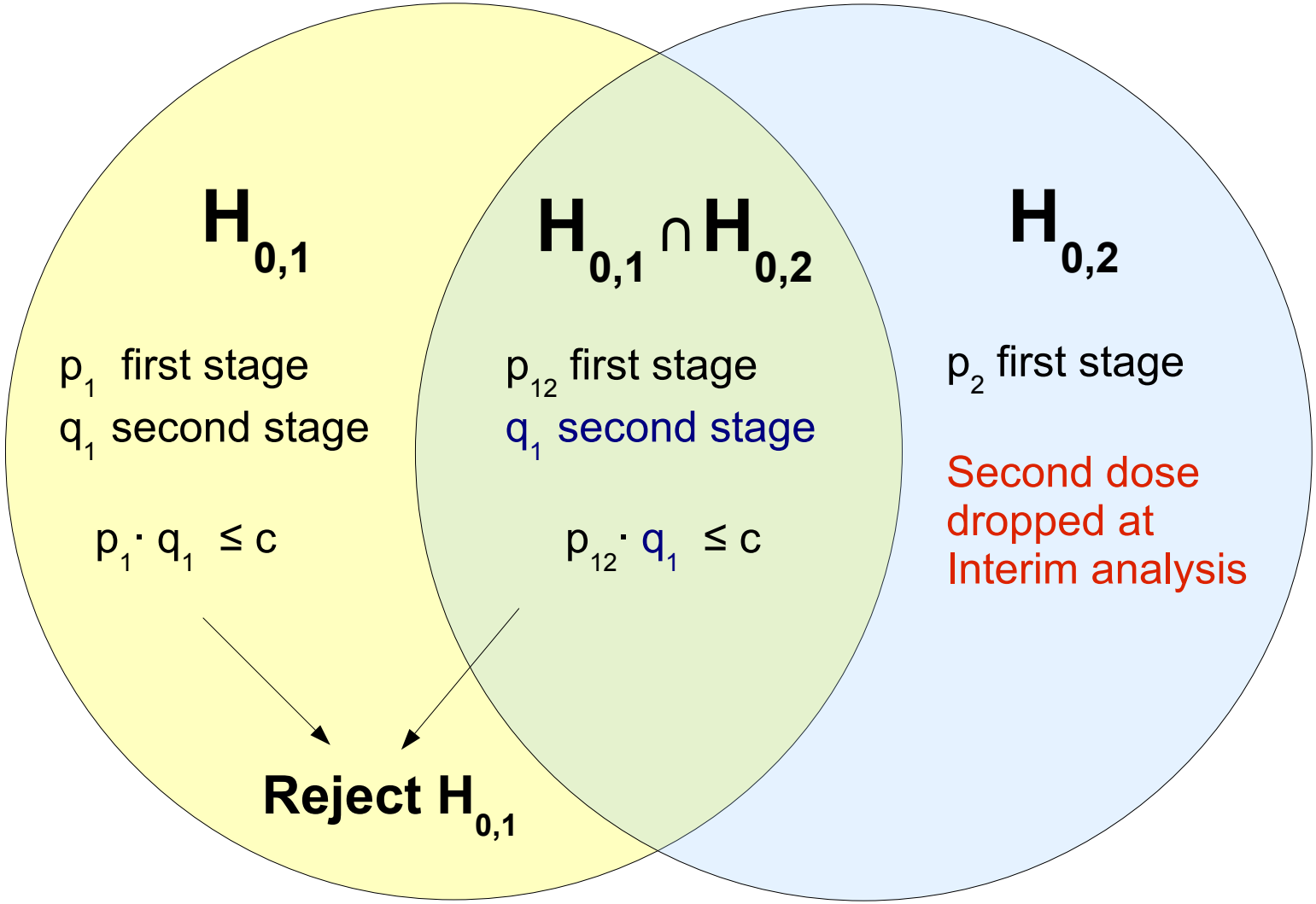
Methodology:

Use a combination test for each intersection hypothesis and apply the closed testing principle to these combination tests.

When selecting both treatments

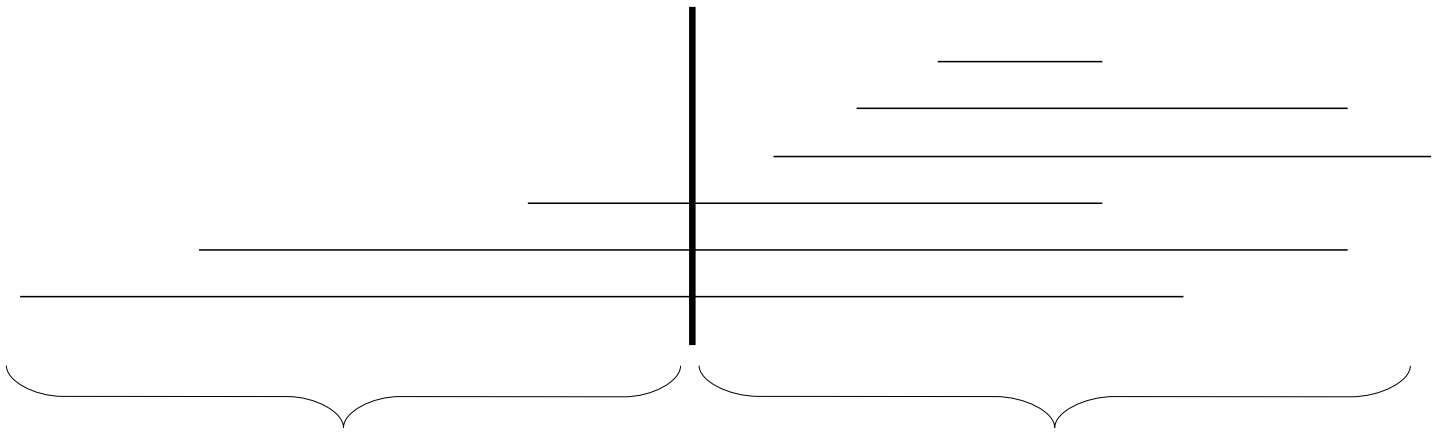


When selecting treatment 1 only



Combination test with follow-up wise stages

(MÜLLER & SCHÄFER '01, BAUER & POSCH '01, WASSMER '06, ..)



Stage 1: p_1, p_2, p_{12} from — log-rank tests, follow-up **till** time of IA, i.e., **right** censoring at time of IA

Stage 2: q_1, q_2, $q_{1,2}$ from — log-rank tests, follow-up **from** time of IA, i.e., **left** censoring at IA (or increments).

Combination test with follow-up wise stages

- Same type of stage-wise data splitting as in group sequential designs (Jennison and Turnbull, 2001).
- Gains power from both, *cytotoxic* and *cytostatic* effect.

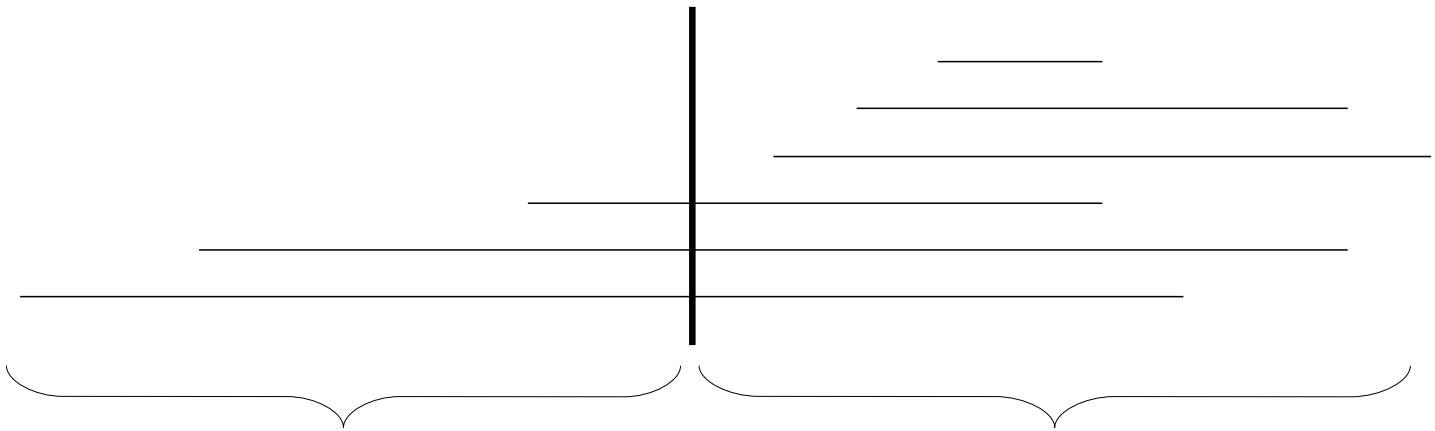
- Strict type I error rate control only if IA-decisions depend solely on OS (Bauer & Posch '01),
- Reason: When using RR or PFS, first and second stage log-rank statistics may fail to be assym. independent.

$\Longrightarrow$

No strict type I error rate control when treatment selection is based on RR or PFS!

Combination test with **stratified** log-rank test

(ZUBER ET AL. '06, BRANNATH ET AL. '08)



Stage 1: p_1, p_2, p_{12} from **stratified** (e.g. for RR) log-rank tests, follow-up **till** time of IA,

Stage 2: q_1, q_2, $q_{1,2}$ from **stratified** (e.g. for RR) log-rank tests follow-up **from** time of IA.

Combination test with **stratified** log-rank test

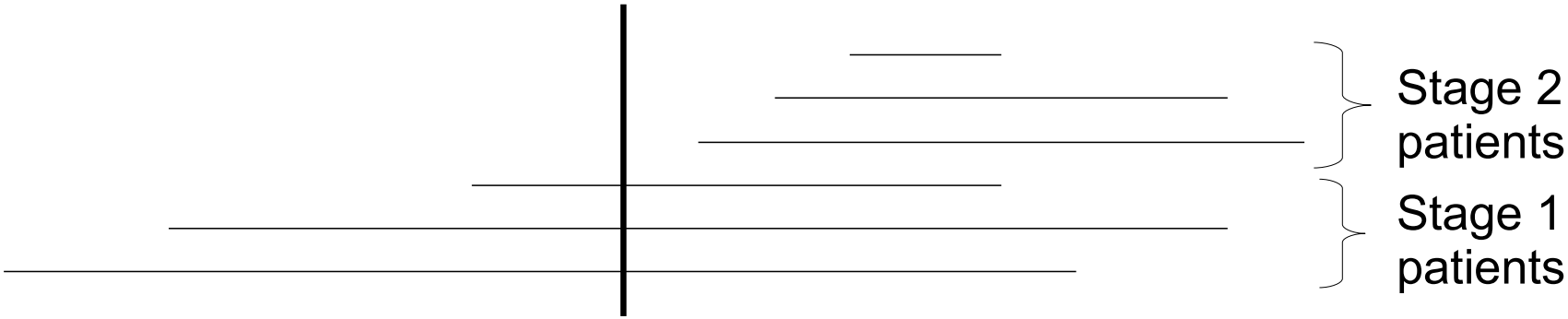
- Same type of stage-wise data splitting as in group sequential designs (Jennison and Turnbull, 2001).
- Strict type I error rate control when treatment selection is based on RR.

- Gains power only from *cytostatic* effect!
- Reason: We condition on RR. Stratified logrank test verifies (average) treatment effect only within each RR-group.

$\Longrightarrow$

Strict type I error rate control for the price of a power loss!

Combination test with patient wise stages



Stage 1: p_1, p_2, p_{12} from — unstratified log-rank tests from patients recruited **before** the IA,

Stage 2: q_1, q_2, $q_{1,2}$ from — unstratified log-rank tests from patients recruited **after** the IA.

Combination test with patient wise stages

- Strict type I error rate control when treatment selection is based on PFS, RR, OS or any other information.
- Gains power from both, *cytotoxic* and *cytostatic* effect.
- First stage p-value can only be computed at end of trial and is not available at the IA

  $\Rightarrow$ early efficacy testing impossible (often anyhow not anticipated).

Combination test with patient wise stages

- Design for first stage patients must be fixed and remain unaltered.
- Consequently:

  - Stage 1 patients must stay in trial and be treated and followed-up as pre-planned also when treatment arm is terminated.
  - Patients of terminated treatment arm could be switched to selected treatment because only used for $H_{0,1} \cap H_{0,2}$.
  - IA-decision should not influence follow-up time of stage 1 patients (e.g. avoid change of overall event/patient number by event/sample size reshuffling).
  - Only a minor problem when proportion of censoring among first stage patients is small.

Simulation (10^5 runs)

Two doses (E_1, E_2) and standard treatment S,

305 patients per treatment group,

uniform recruitment within 35 month.

exponential survival, no censoring

$\lambda_{S,N} = 0.0462$ (hazard in S and NR),

$HR_{R/N} = 0.7$ in S (hazard rate of RR vs. NR)

IA at 50% patients per arm; we select dose with higher RR.

Simulation (10^5 runs)

Fisher's product test.

null hyp.: $\theta_S = \theta_{E_1} = \theta_{E_2} = 0.4$, $c_1 = c_2 = 1$
alternative 1: $\theta_S = \theta_{E_1} = \theta_{E_2} = 0.4$, $c_1 = c_2 = 0.8$
alternative 2: $\theta_S = 0.2$, $\theta_{E_1} = 0.3$, $\theta_{E_2} = 0.4$, $c_1 = 0.9$, $c_2 = 0.8$

STAGES TEST	follow-up wise unstratified	follow-up wise stratified	pat. wise (unstratified)
null hyp.	0.019	0.017	0.019
alternative 1	0.83	0.70	0.85
alternative 2	0.89	0.65	0.89

Further considerations

- What information should be used for IA-decision (RR only, RR and OS, RR and PFS and OS)? How to use this information? Answer requires modeling and extensive simulations.
- When (at which proportion of patients) should the interim analysis be done?
- Comparison of Phase II/III design to separate phase II and phase III trials.
- Combination test with follow-up wise stages with p-values from a joint model of RR and OS.

Discussion

- There is no universally applicable method. All methods require specific additional assumptions or restrictions.
- Usual combination test (follow-up wise stages) with survival data and treatment selection based on surrogate endpoints may not control type I error rates.
- Stratification for IA-information (e.g. RR) allows control of type I error rate but will be inefficient.
- Combi. tests with patient wise stages, seems best and rigorously valid when follow-up times of stage 1 patients remain unaltered or are complete (at end of study).
- Phase II/III designs in oncology are an interesting option, but require modeling and extensive simulations.

Selected References

- Schäfer H., Müller H.-H. Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections *Statistics in Medicine* 2001, **20**:3741–3751
- Bauer B., Posch M. Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections by H. Schäfer and H.-H. Müller, Statistics in Medicine 2001; 20: 3741-3751. *Statistics in Medicine*, **23**:1333–1334
- Wassmer G. Planning and Analyzing Adaptive Group Sequential Survival Trials *Biometrical Journal* 2006, **48**:714–729
- Brannath W., Zuber E., Branson M., Bretz F.,Gallo P., Posch M., Racine-Poon A. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine* 2009, **28**:14451463.
- Zuber E., Brannath W., Branson M., Bretz F., Gallo P., Posch M., Racine-Poon A. Phase II/III seamless adaptive designs with bayesian decision tools for an efficient development of a targeted therapy in oncology. Technical Report 2006-05, Department of Statistics and Decision Support Systems, University of Vienna.

Simulation (10^5 runs)

Inverse normal combination function with equal weights.

null hyp.: $\theta_S = \theta_{E_1} = \theta_{E_2} = 0.4$, $c_1 = c_2 = 1$
alternative 1: $\theta_S = \theta_{E_1} = \theta_{E_2} = 0.4$, $c_1 = c_2 = 0.8$
alternative 2: $\theta_S = 0.2$, $\theta_{E_1} = 0.3$, $\theta_{E_2} = 0.4$, $c_1 = 0.9$, $c_2 = 0.8$

STAGES TEST	follow-up wise unstratified	follow-up wise stratified	pat. wise (unstratified)
null hyp.	0.017	0.022	0.018
alternative 1	0.80	0.75	0.85
alternative 2	0.89	0.71	0.85