ICON
A Symbol of Excellence

Adaptive Study Design for Subgroup Selection

Andy Grieve

VP Clinical Trials Methodology Innovation Centre, Icon plc, Marlow, UK

Andrew.Grieve@icon.com

What?

- ***Adaptive Design*** is one that uses accumulating data from the ongoing trial to modify aspects of the study without undermining the validity and integrity of the trial

  PhRMA ADWG, Gallo et al (2006)

- ***Adaptive design clinical study*** is defined as a study that includes a prospectively planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of data (usually interim data) from subjects in the study *FDA Guidance on AD (2010)*

- ***Validity***
  - providing correct statistical inference: adjusted p-values, estimates, confidence intervals
  - providing convincing results to a broader scientific community
  - minimizing statistical bias
- ***Integrity***
  - preplanning based on intended adaptations
  - maintaining confidentiality of data
  - assuring consistency between different stages of the study
  - minimizing operational bias

Adaptive by Design

Motivation for Adaptive Design

- Opportunity to calibrate initial assumptions used at trial design based on partial observed information
- Improved knowledge efficiency vs. conventional (non-adaptive) designs
  - Faster
  - Less expensive
  - More information for same investment
- Increase likelihood of success, or reliable early termination (e.g., futility rule)
- Improved understanding of treatment effect

Challenges for Adaptive Design

- Adaptive designs (AD) offer considerable opportunities for improving drug development, but come with risks and costs
- Industry mindset favoring traditional development approaches ⇨ change management
- Regulatory concerns with new approaches, especially in confirmatory studies: FDA draft guidance on AD quite helpful in that regard
- Need adequate operational infrastructure: recruitment, data management, drug supply, etc
- Resource needs: increased planning, more people with proper expertise; adequate commercial software for design and implementation; hardware for intensive computing

Adaptive Designs in Confirmatory Trials

Sample Size Reassessment	• Sample size adjustment based on blinded or unblinded data:
	  • Using nuisance parameter estimate
	  • Using treatment effect estimate

Adaptive Group Sequential Design	• Early stopping for efficacy, futility, harm or safety • Adjusting the number and/or timing of interim analyses • Increasing the maximum sample size
Seamless Phase II/III Design	• Design combining the objectives of Phase II dose ranging study and confirmatory Phase III trial in a single protocol • Dose selection at the interim analysis
Population Enrichment Design	• Placebo run-in; Active control run-in; Dose titration • Adaptively enrich the population at the interim analysis • Enrich based on biomarker or clinical endpoint response
Drugs with Companion Diagnostics	• Marker by Treatment Design • Targeted Design • Marker x TRT Design with Response adaptive allocation within strata

Adaptive Group Sequential Designs

- Adaptive group sequential designs generalise ordinary GSDs , where – in interim analyses - confirmatory analyses are performed under control of the Type I error rate and data dependent changes of design are allowed.
- An attractive way to derive such designs is the combination testing principle as proposed by Bauer (1989) and Bauer and Köhne (1994).

Combination Testing Principle

- Combination of *p*-values with a specific combination function, e.g., Fisher‘s combination test
- Inverse normal method: The test decision is based on

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

where the weights w_k are prefixed

- Advantage: Bounds from group sequential theory can be used
- In the predefined stages, different hypotheses can be considered, the (global) test is a test of

$$H_0 = H_0^1 \cap \ldots \cap H_0^K$$

Possible Data-Dependent Changes of Design

- Reassessment of sample size
- Adaptive choice of test statistic
- Combining Phase II/III studies (adaptive seamless phase II/III designs)
- Selection of endpoints
- Change of target parameter
- Modification of ordering of hypothesis

  - The rules for adapting the design need not be prespecified!

(Patient) Population Enrichment Designs

- Applicable where studies of unselected patients are unable to detect a drug effect and it seems necessary to "enrich" the study with potential responders (Temple, *Comm Stat Theory Meth 1994*).
- If this is done in an adaptive way (i.e., it is not clear upfront whether to use the selected population) we might use adaptive enrichment designs (Wang et al, *Biom. J.* 2009).
- Baseline characteristics that are used for patient selection are often known biomarkers, e.g., genetic.
- Proof of efficacy is done in a confirmatory sense. Hence, we use confirmatory adaptive designs that control prespecified Type I error rate.

Motivation

- The aspect of "one size fits all" surrounding the conventional design of clinical trials has been challenged, particularly
  - when the disease is considered heterogeneous
  - or the experimental therapy is tailored to a specific mechanism of action
- The development of biomarkers that define specific subsets of disease is enabling a shift from *empirical* medicine to *precision* medicine
- Next step towards the *personalized* medicine:
  - resulting in delivery of the right medicine,

  to the right patient, at the right dose, at the right time

Motivation

- Many new tailored therapies and many more potential combination therapies
- Targeted agents do not work for all patients
- Limited patient population enrolled in clinical trials
  - How to treat patients best in the trial?
  - Are there markers to guide the choice of treatments?
- How to gauge the treatment effect?
  - response rate, disease stabilization, improved survival
- How to best identify markers, gauge treatment efficacy, and treat patients in a clinical trial?

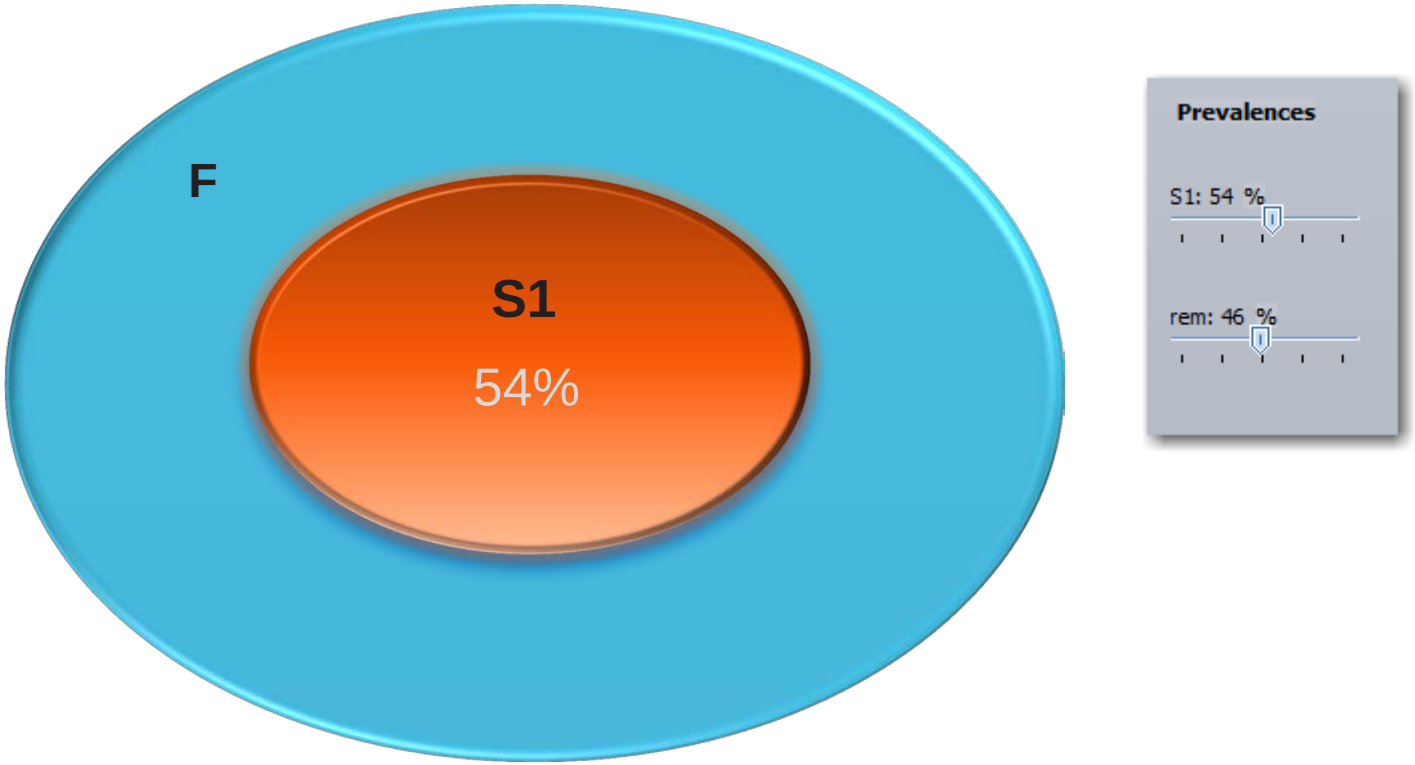
The Enrichment Test Procedure

- For simplicity, we consider a two-sample comparison case although an extension to the multi-armed case is straightforward.
- Consider prespecified subpopulation(s) $S_1,\ldots,S_G$, which can be nested, and a full population F:

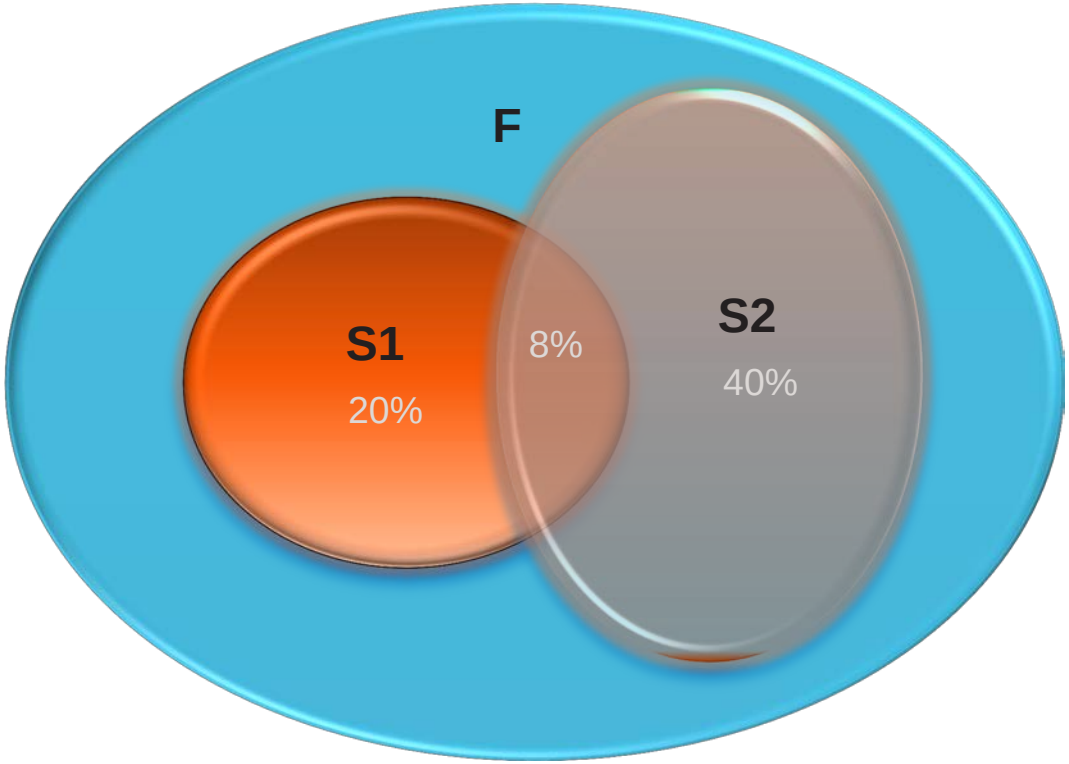
$$S_G \subset \ldots \subset S_1 \subset F$$

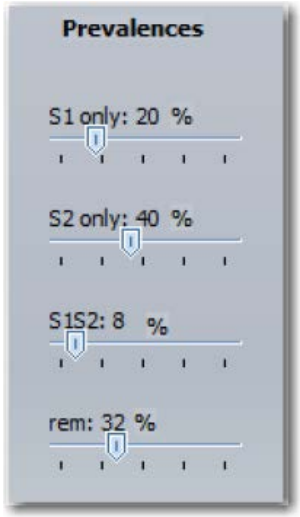
- At an interim stage it is decided which subpopulation is selected for further inference (including all subpopulations, i.e., full population).
- Other adaptive strategies (e.g., sample size reassessment) can also be performed.

One sub-population

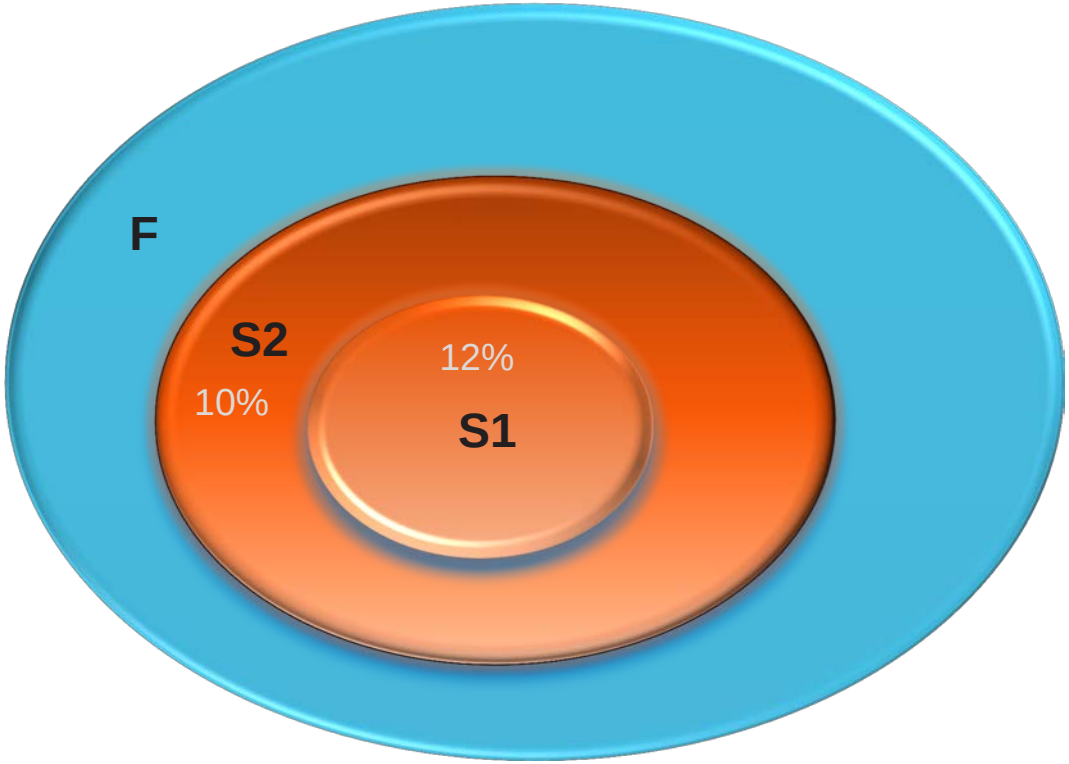


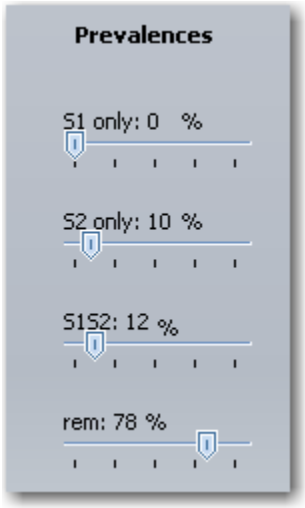
Two sub-populations of interest



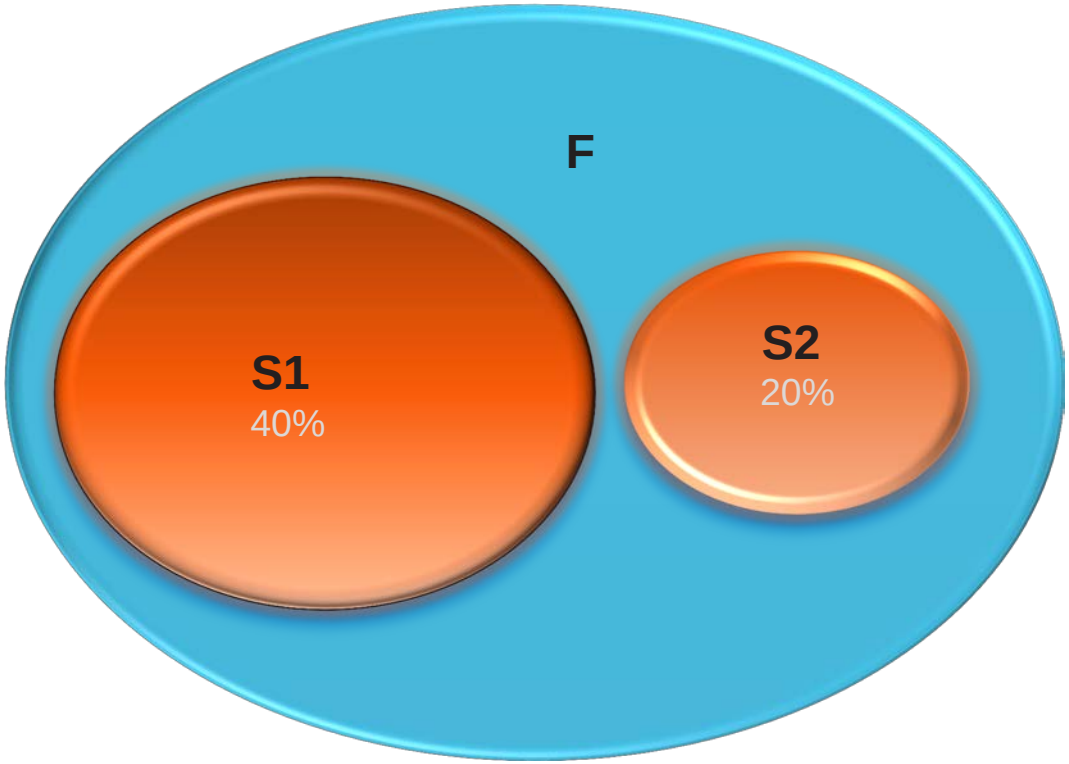


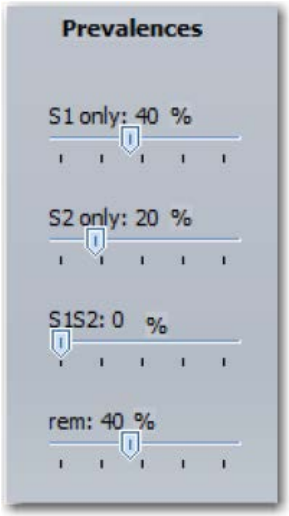
Two nested sub-populations of interest



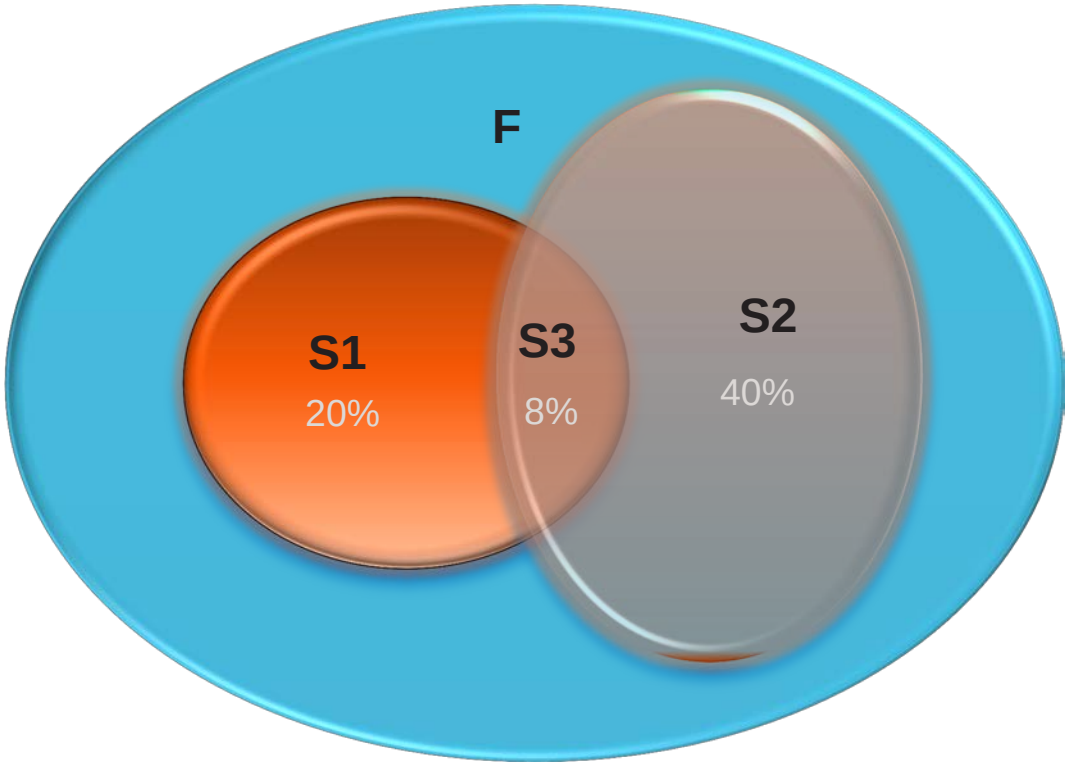


Two non-overlapping sub-populations





Three sub-populations of interest



Prevalences

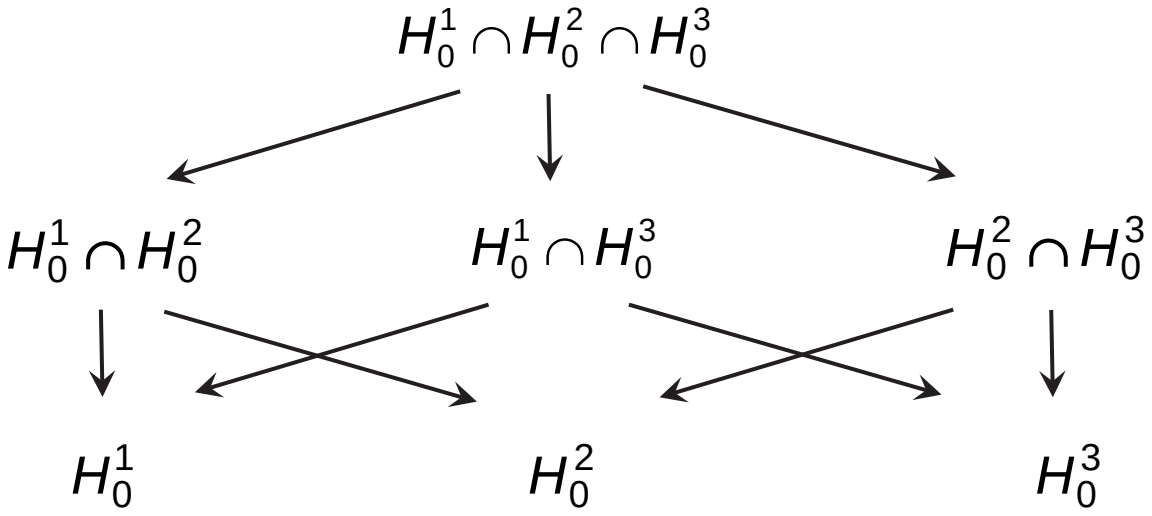
S1 only =	20	%
S2 only =	40	%
S3 only =	0	%
S1S2 only =	0	%
S1S3 only =	0	%
S2S3 only =	0	%
S1S2S3 =	8	%
rem =	32	%

Methodology

- Sources for alpha inflation
  - Interim analyses
  - Sample size reassessment
  - Multiple sub-populations
- The proposed adaptive procedure strongly controls the pre-specified family-wise Type I error rate
- The procedure is based on the application of the closed test procedure together with combination tests (e.g., Bauer & Kieser, *Statistics in Medicine*,1999)

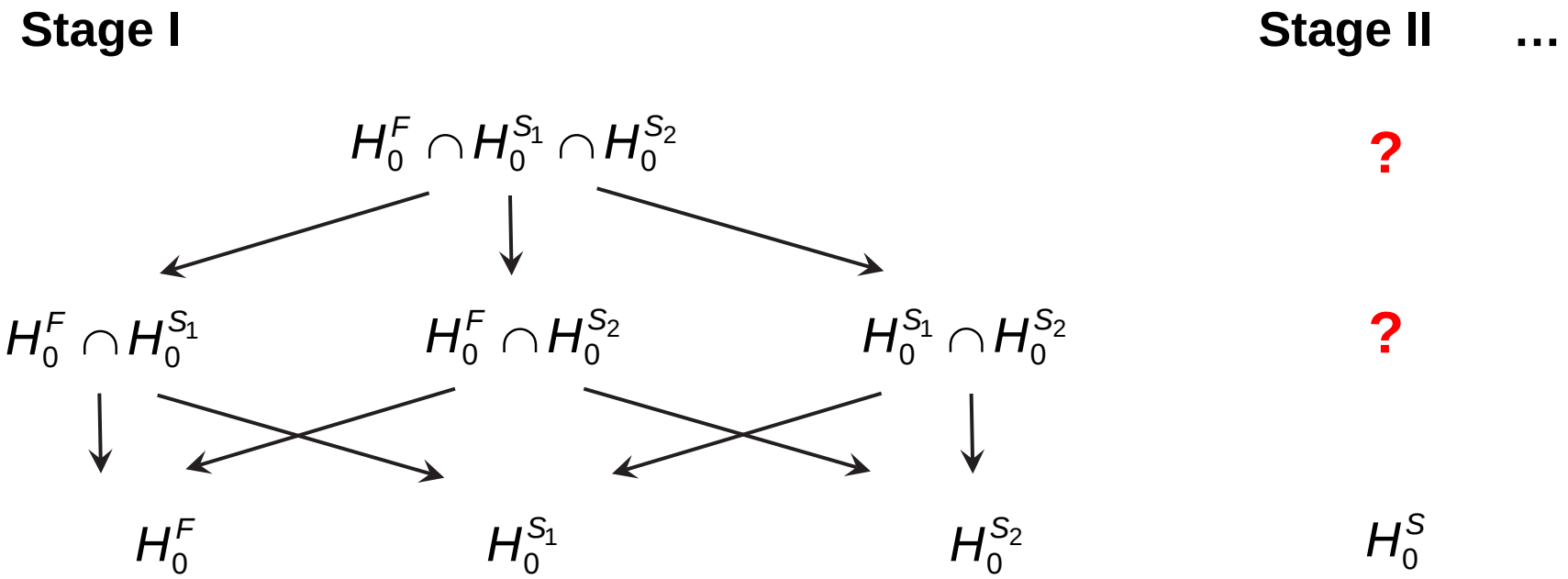
Closed Testing Procedure

- Closed system of hypotheses:



- A hypothesis $H_0^i, i = 1, 2, 3,$ is rejected if H_0^i itself and all hypotheses which are a subset of H_0^i are rejected at level α.
- This procedure controls the expiremententwise error rate α in a strong sense.

Closed Testing Procedure

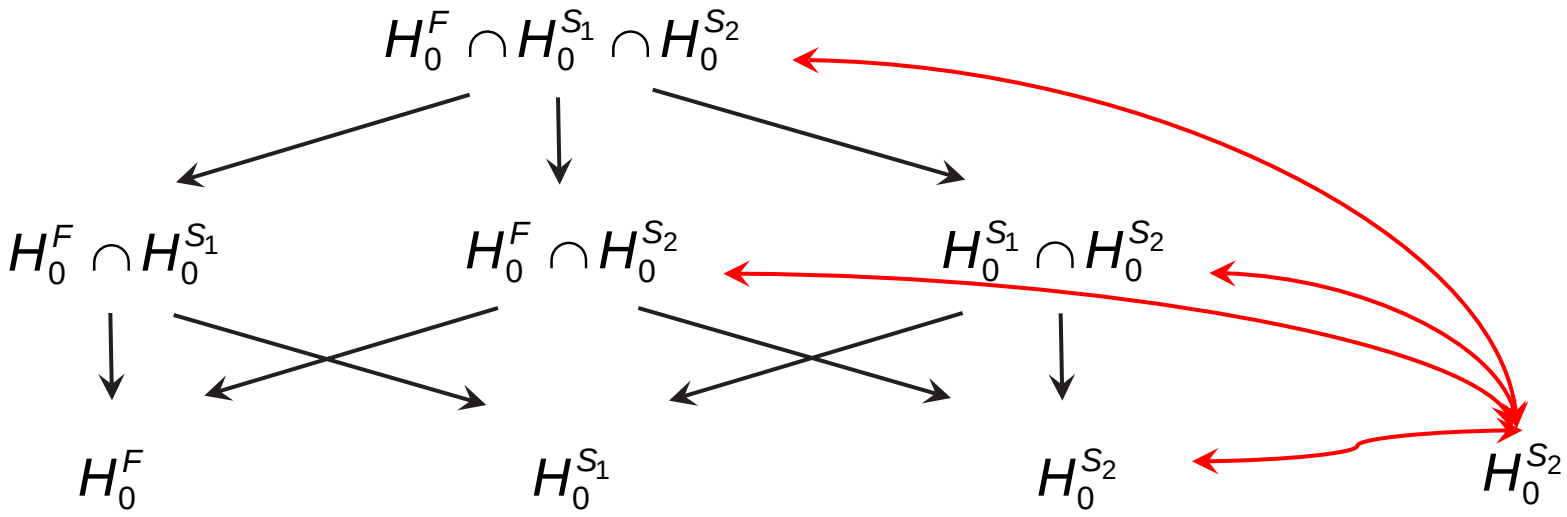


- H_0^S can be rejected if all combination tests exceed the critical value u_2.

Closed Testing Procedure

Example: 2 stages $S = S_2$

Stage I **Stage II** **…**



$H_0^{S_2}$ can be rejected if all combination tests exceed the critical value u_2.

Closed Testing Procedure

- The choice of combination tests is free - you might use inverse normal or Fisher‘s combination test.
- The choice of tests for intersection hypotheses is free. - you might use Bonferroni, Simes or Sidak tests.
- For one subgroup also Dunnett‘s test can be applied
- You might also use the CRP principle. i.e., perform conditional Dunnett test (Friede et al., *Stat Med*)

- Calculation of RCIs and overall p-values straightforward
- Except conditional Dunnett, all procedures available in ADDPLAN PE, Version 6.0

Questions To Be Answered by Simulation
(Maurer, Branson, Posch, 2010)

ICON
A Symbol of Excellence

- What is the influence of the unknown prevalence of *S* on the power?
- How powerful is the design under different assumptions about treatment differences in *S* and S^c?
- How does this strategy compare to other strategies (e.g., group sequential designs)?
- How robust is the design to selecting the correct subpopulation at the first stage, i.e., how often is the wrong group selected for continuation?
- It is even possible to select an unspecified subpopulation (i.e., *adding* a hypothesis acc. Hommel, 2001). What is the effect of prespecification?

Questions on Analysis

- Since we have positive correlation between the test statistics the usual intersection tests (Sidak, Simes) can be used.
- What are the criteria for selecting a population or sticking to the full population?
- What are the criteria for assessing sample size of full and subset population?
- Are estimation procedures and procedures for the calculation of overall *p*-values available?
- Is a user friendly and illustrative assessment of study results available?

Overall *p*-Values

- Defined as smallest significance level for which the test results yield rejection of the considered (single) hypothesis
- Overall *p*-value can be calculated at any stage of the trial („Repeated *p*-value“).
- That is, $p_k^g \leq \alpha \quad \Leftrightarrow \quad H_0^g$ can be rejected at stage k

- *p*-values account for the step-down nature of the closed testing principle and are completely consistent with the test decision.

Overall Confidence Intervals

- Confidence intervals based on stepwise testing are difficult to construct. This is a specific feature of multiple testing procedures and not of adaptive testing.
- Posch et al. (2005) proposed to construct confidence intervals based on the single step adjusted overall *p*-values. These can also be applied for the conditional Dunnett test.
- The RCIs are not, in general, consistent with the test decision. It might happen that, e.g., a hypothesis is rejected but the lower bound of the CI is smaller 0.
- They can be provided for each step of the trial.

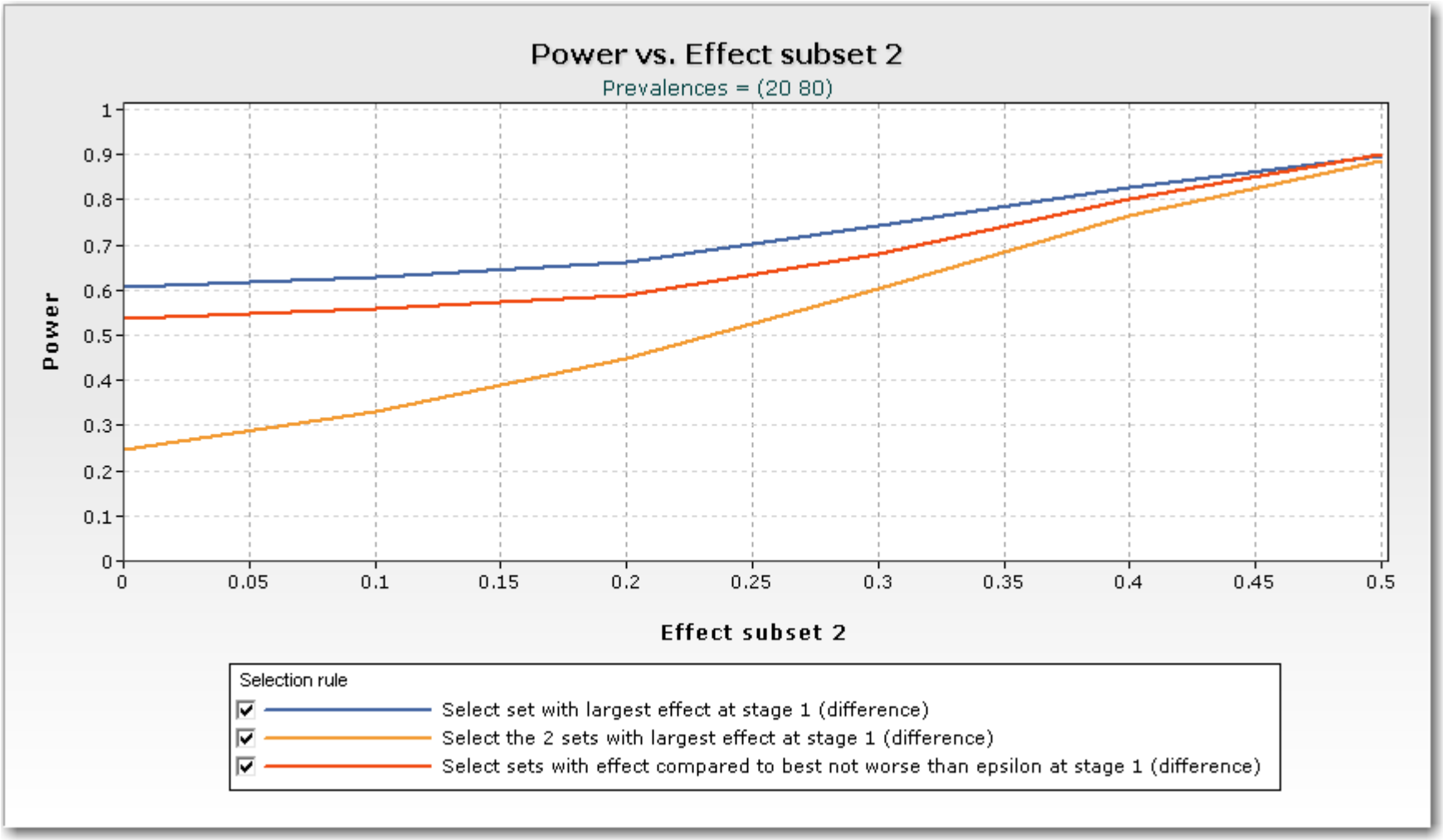
ADDPLAN Software

- ADDPLAN combines traditional designing and sample size calculation features with a very powerful simulation engine, and with extensive adaptive analysis tools.
- Traditional non-flexible designs are just a special case, which makes it an all-purpose tool for study designers.
- BASE module
  - Planning, simulation and adaptive analysis for up to two treatment arms for continuous, binary, and survival endpoints
- MC Module
  - Additional multiple comparison features for more than two treatment arms in simulation and analysis
- PE Module
  - Additional features for patient enrichment designs in simulation and analysis

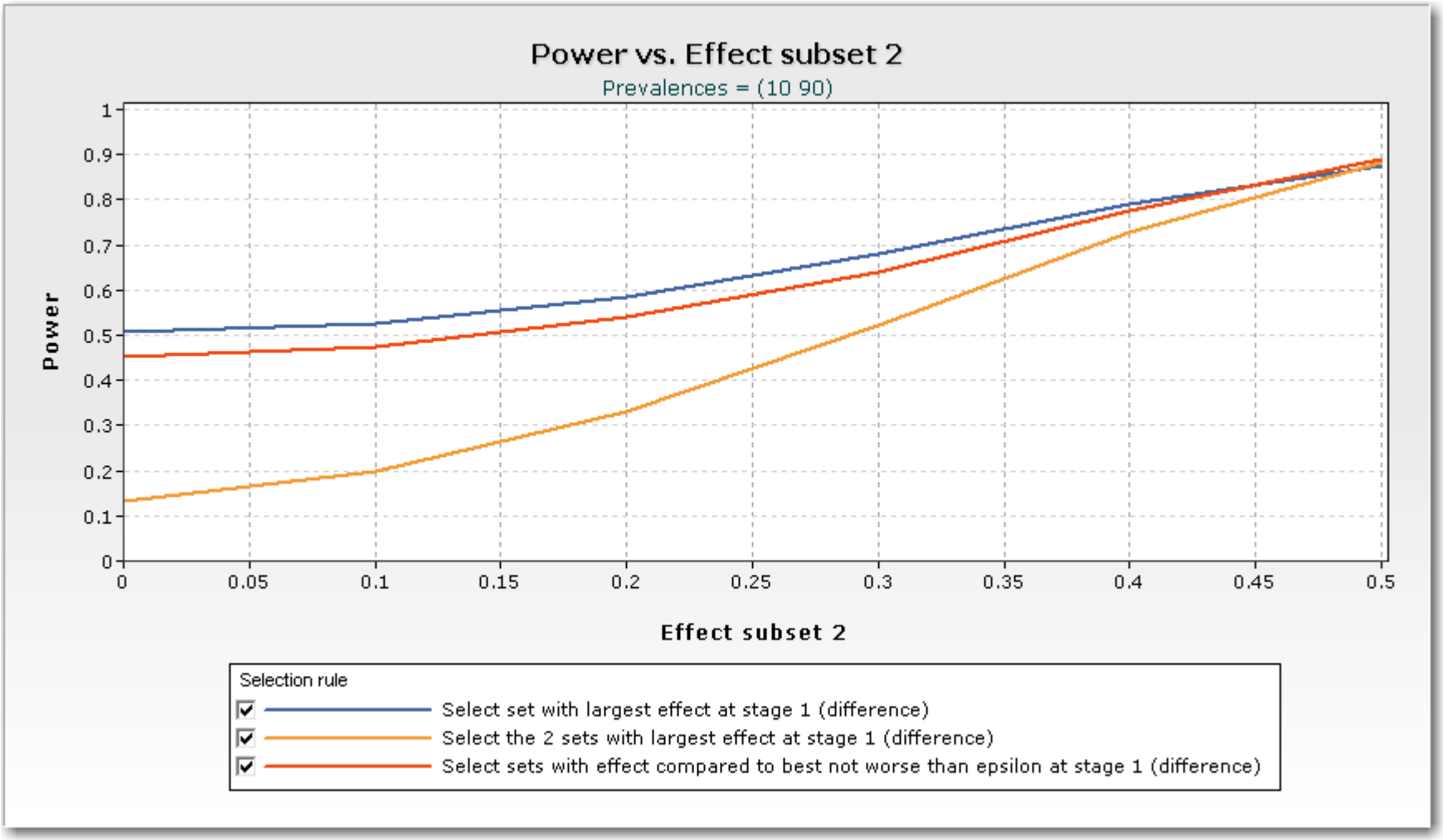
Simulation Example

- Two-stage design with no early stopping, one sub-population
- In the biomarker positive population a standardized effect of 0.5 is assumed, biomarker negative population has effect sizes ranging from 0.0 to 0.5
- Selection rules
  - Select the population with highest effect size
  - Select the population with effect size compared to the better not worse than 0.25 (say)
  - Never select
- Prevalance of biomarker +ve population: 5%, 10%, 20%.
- Sample sizes 100 patients per stage
- Simes‘ test is used for testing intersection hypotheses.

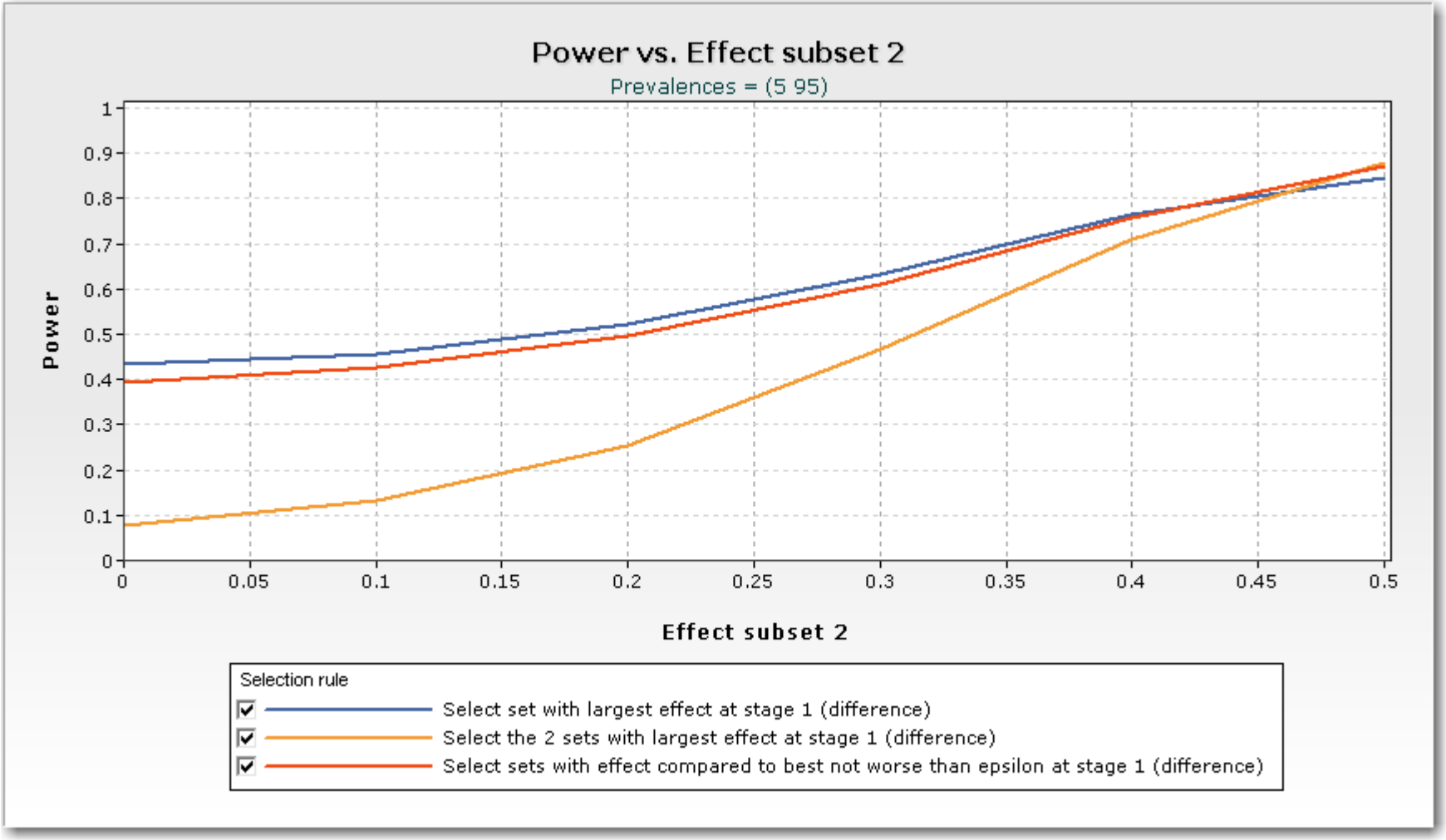
Results


Power vs. Effect subset 2
Prevalences = (20 80)
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Effect subset 2
Selection rule
Select set with largest effect at stage 1 (difference)
Select the 2 sets with largest effect at stage 1 (difference)
Select sets with effect compared to best not worse than epsilon at stage 1 (difference)

Results


Power vs. Effect subset 2
Prevalences = (10 90)
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Effect subset 2
Selection rule
Select set with largest effect at stage 1 (difference)
Select the 2 sets with largest effect at stage 1 (difference)
Select sets with effect compared to best not worse than epsilon at stage 1 (difference)

Results


Power vs. Effect subset 2
Prevalences = (5 95)
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Effect subset 2
Selection rule
Select set with largest effect at stage 1 (difference)
Select the 2 sets with largest effect at stage 1 (difference)
Select sets with effect compared to best not worse than epsilon at stage 1 (difference)

Results

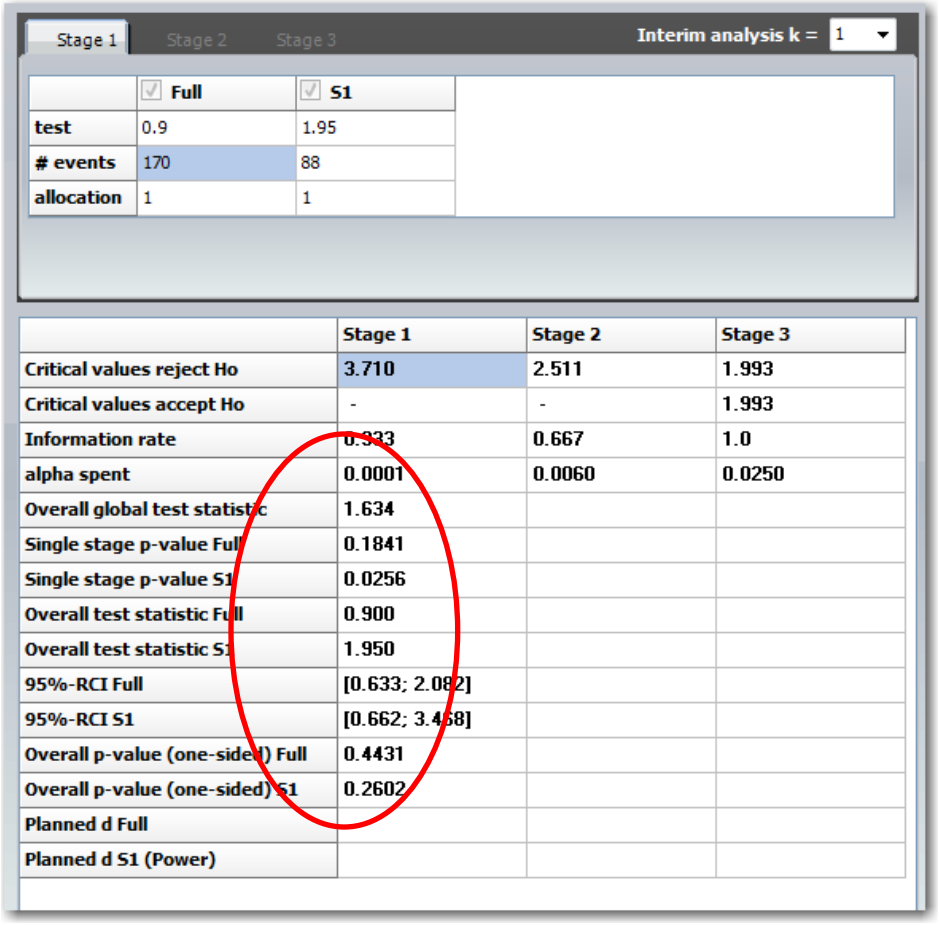
- Clear power disadvantage for procedure that never selects a sub-population
- No clear advantage of selecting always (and only) the best population
- For small prevalences, always selecting the best can even provide a small loss in power

Example by Brannath et al. (2009)

- They considered a three-stage design with inverse normal combination test using fixed weights 0.43, 0.64, and 0.63 according to planned information rates 170/918 = 0.185, 551/918 = 0.60, and 1.
- They show how Bayesian decision tools can be used for the population selection decision making.
- Critical values according to an O‘Brien and Fleming α-spending function approach are chosen.
- Simes‘ test is used for testing intersection hypotheses.
- Increments of logrank statistics from the right-censored event times are used for decision making (e.g., Wassmer, 2006)

ADDPLAN Output

- Test decision and overall statistical inference, incl. confidence intervals and overall test results, performed with ADDPLAN:



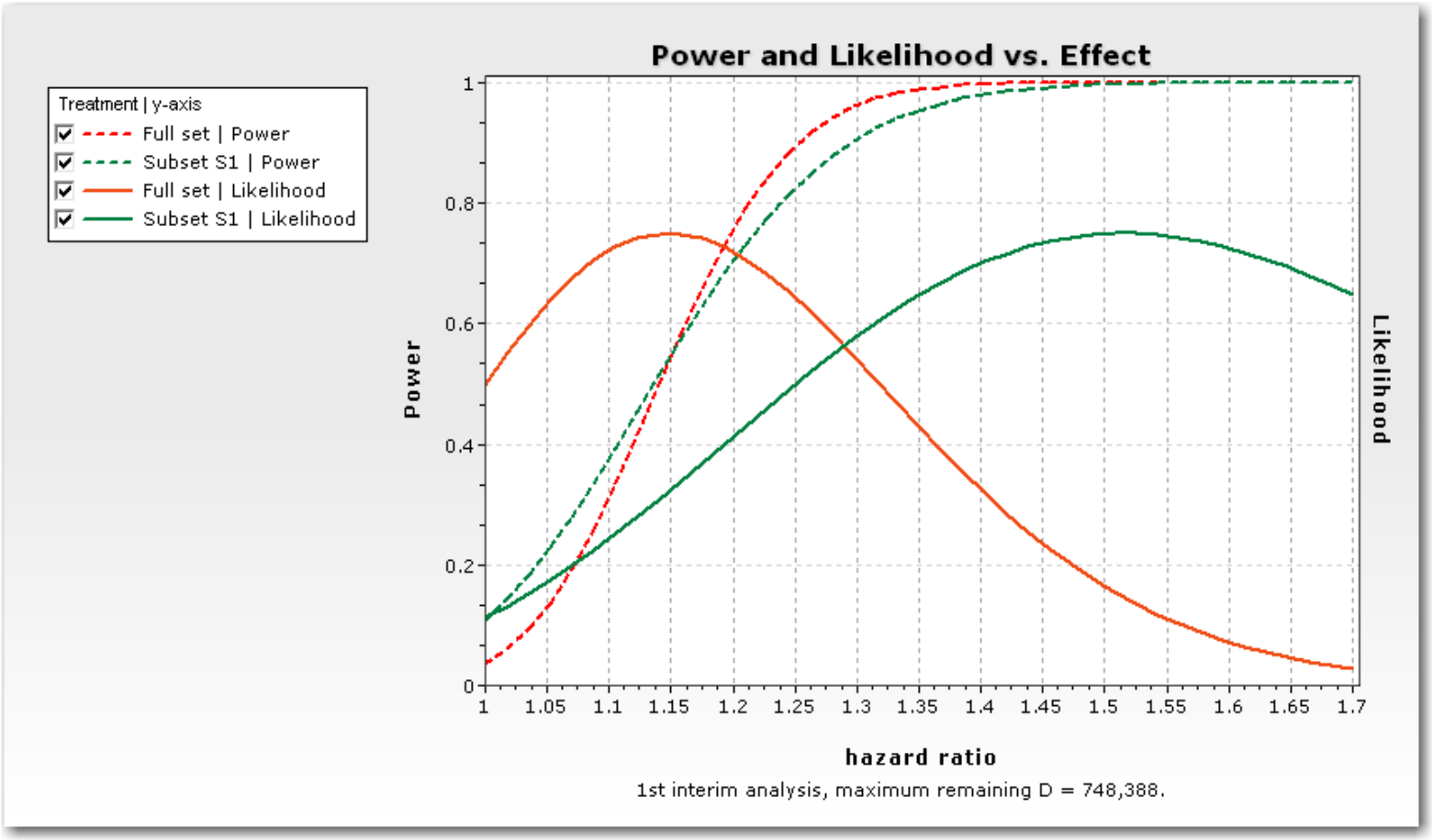
Stage 1 | Stage 2 | Stage 3 — Interim analysis k = 1

	Full	S1
test	0.9	1.95
# events	170	88
allocation	1	1

	Stage 1	Stage 2	Stage 3
Critical values reject Ho	3.710	2.511	1.993
Critical values accept Ho	-	-	1.993
Information rate	0.333	0.667	1.0
alpha spent	0.0001	0.0060	0.0250
Overall global test statistic	1.634		
Single stage p-value Full	0.1841		
Single stage p-value S1	0.0256		
Overall test statistic Full	0.900		
Overall test statistic S1	1.950		
95%-RCI Full	[0.633; 2.082]		
95%-RCI S1	[0.662; 3.468]		
Overall p-value (one-sided) Full	0.4431		
Overall p-value (one-sided) S1	0.2602		
Planned d Full			
Planned d S1 (Power)			

ADDPLAN Output

- A convenient way to find the sub-group to be selected through the conditional power plot together with likelihood (similar to Bayesian predictive power):



Summary

- Attractive and general procedure for adaptive confirmatory design that controls Type I error rate
- The “rules” for adaptation and stopping for futility
  - not need be pre-specified
  - Adaptations may depend on all interim data including secondary and safety endpoints.
  - can make use of Bayesian principles integrating all information available, also external to the study
  - should be evaluated (e.g. via simulations) and preferred version recommended, e.g., in DMC charter
- Comparison of different strategies and options for analyses is mandatory. The role of simulation becomes increasingly important.