AptivSolutions℠
Accelerating the Possibilities

Enrichment Designs: Methoden und Anwendungen

Gernot Wassmer

ADDPLAN Inc, An Aptiv Solutions Company

Institutskolloquium Wien, 17.06.2013

(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 known as biomarkers, and often genetic.
- Proof of efficacy is done in a confirmatory sense. Hence, we use confirmatory adaptive designs that control prespecified Type I error rate.

Confirmatory Adaptive Designs

Confirmatory adaptive designs are a generalization of group sequential designs, where – in interim analyses - confirmatory analysis is performed under control of the Type I error rate and data dependent changes of design are allowed.

Three particular applications

- Sample size reassessment
- Treatment arm selection in multi-armed designs
- Subgroup analyses („enrichment designs“)

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

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 to be prespecified!

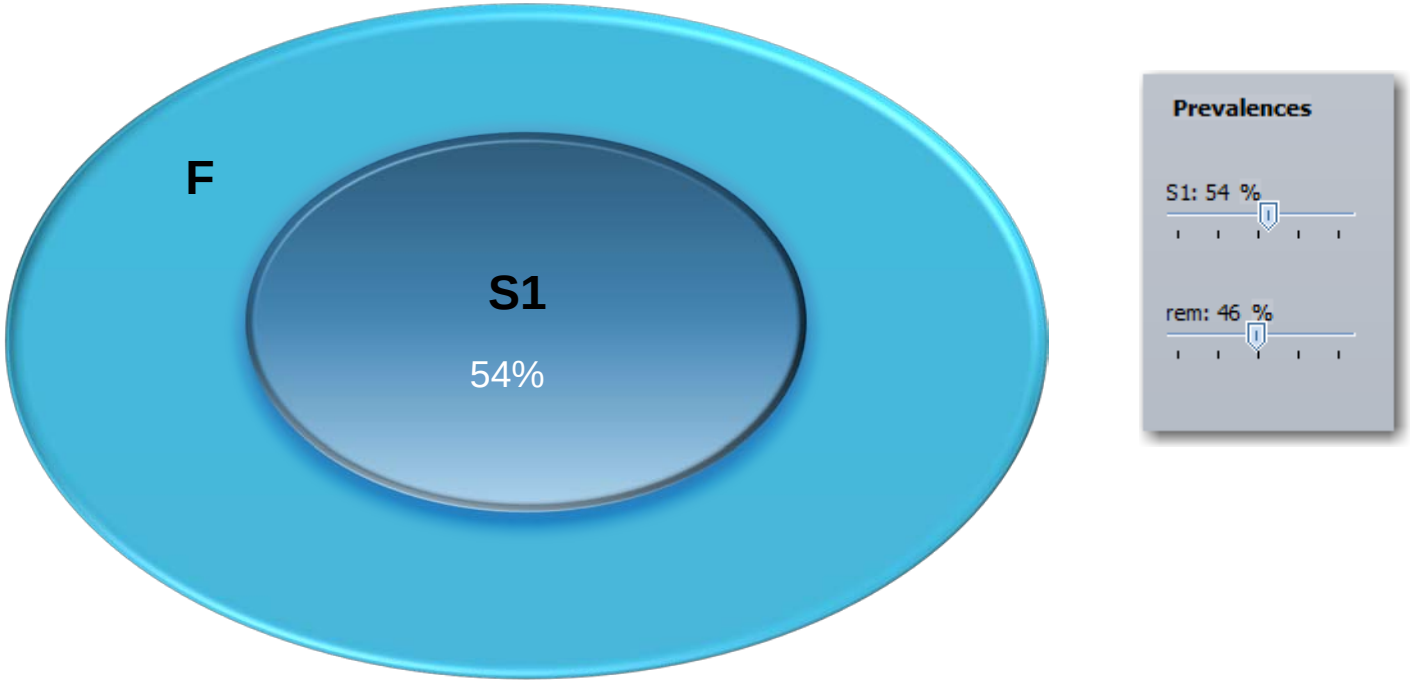
The Enrichment Test Procedure

- 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:

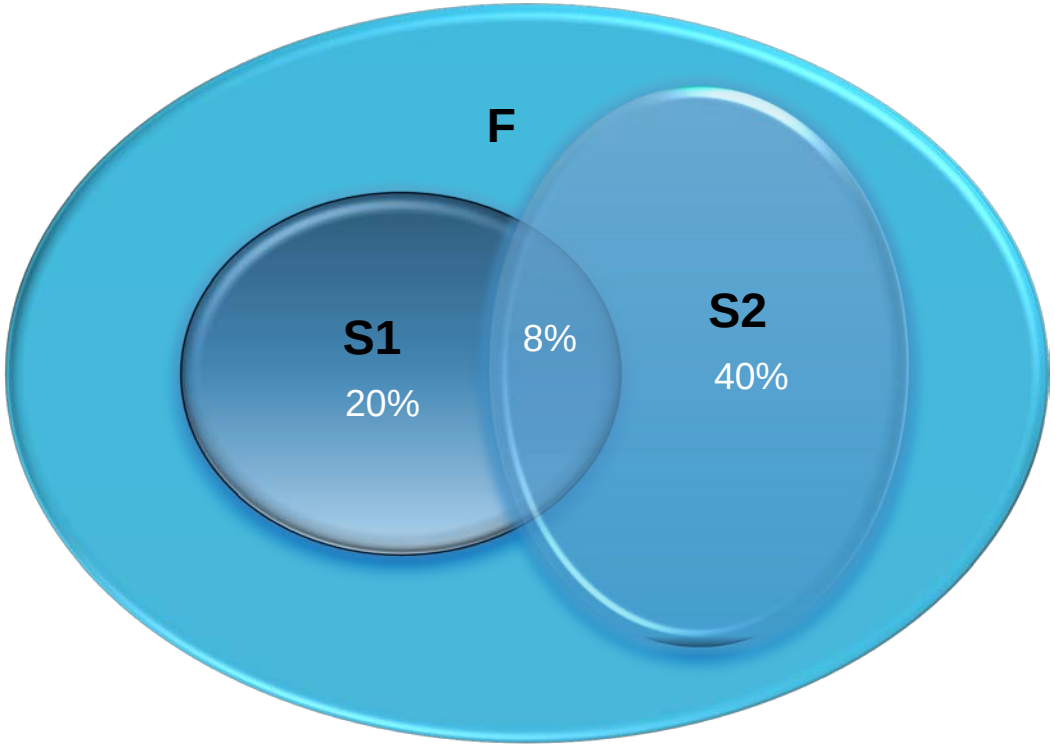
$$S_G \subset \ldots \subset S_1 \subset F$$

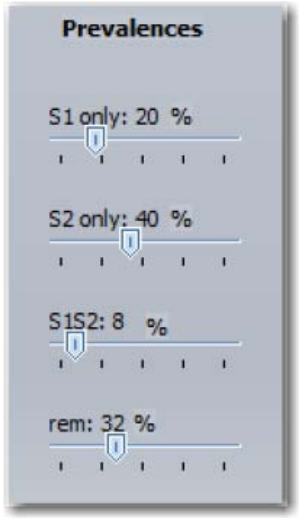
- At an interim stage it is decided which subpopulation is selected for further inference (including all subpopulations, i.e., full population).
- Not only selection procedures, but also other adaptive strategies (e.g., sample size reassessment) can be performed.

One sub-population

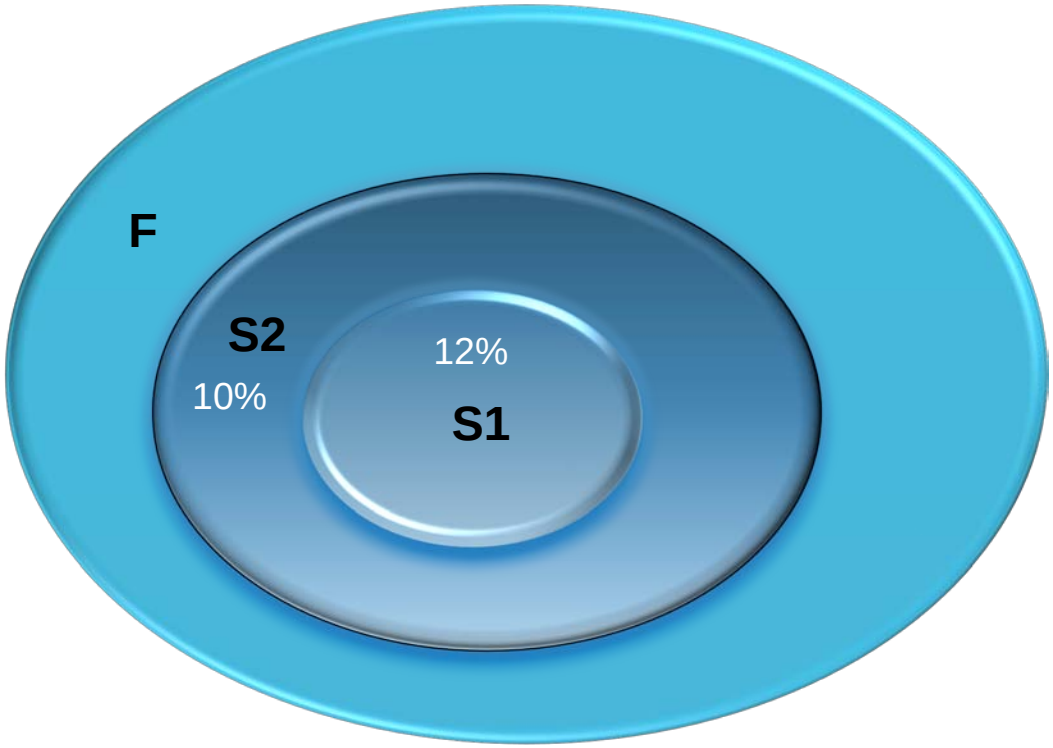


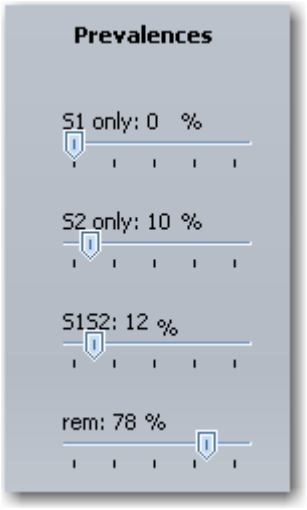
Two sub-populations of interest



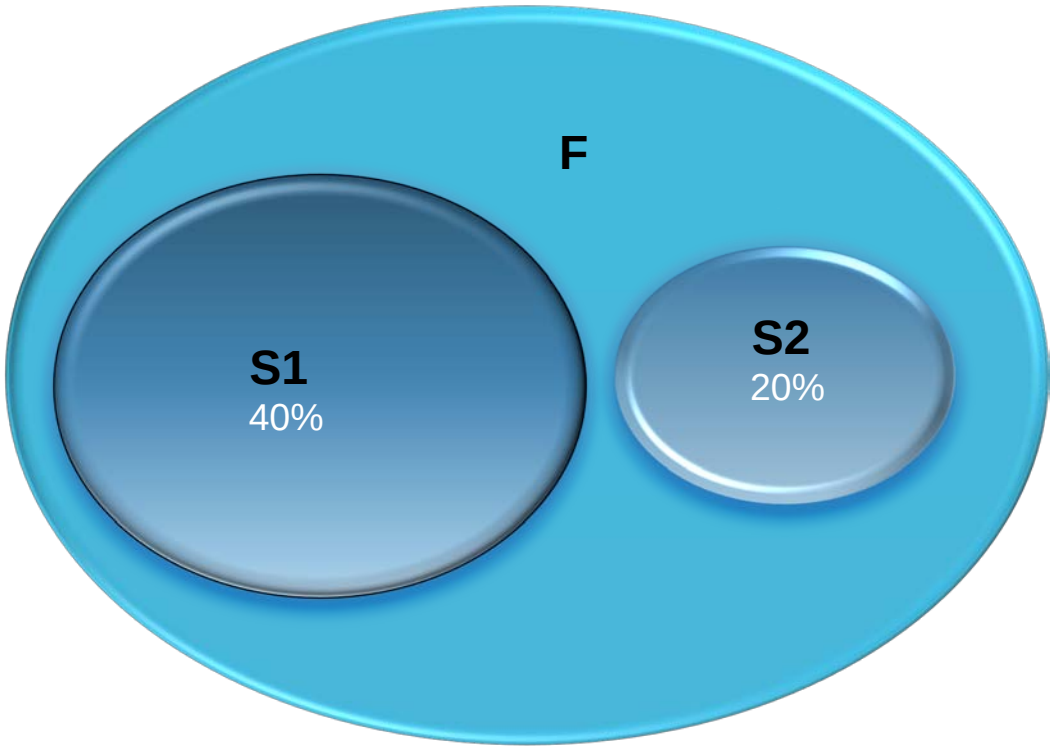


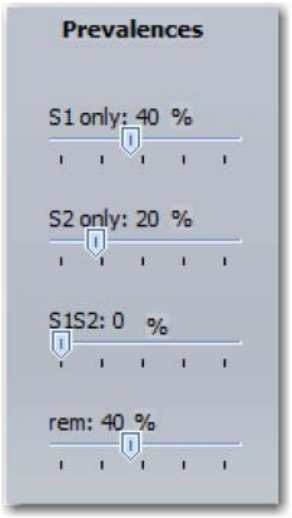
Two nested sub-populations of interest



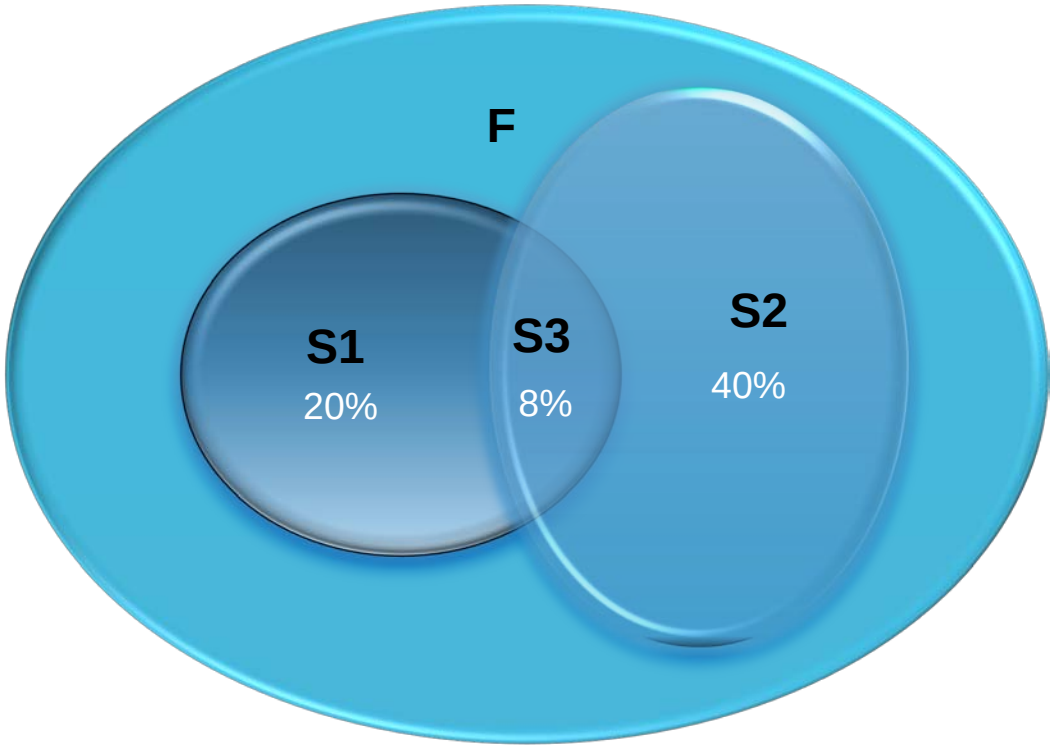


Two non-overlapping sub-populations





Three sub-populations of interest



Prevalences

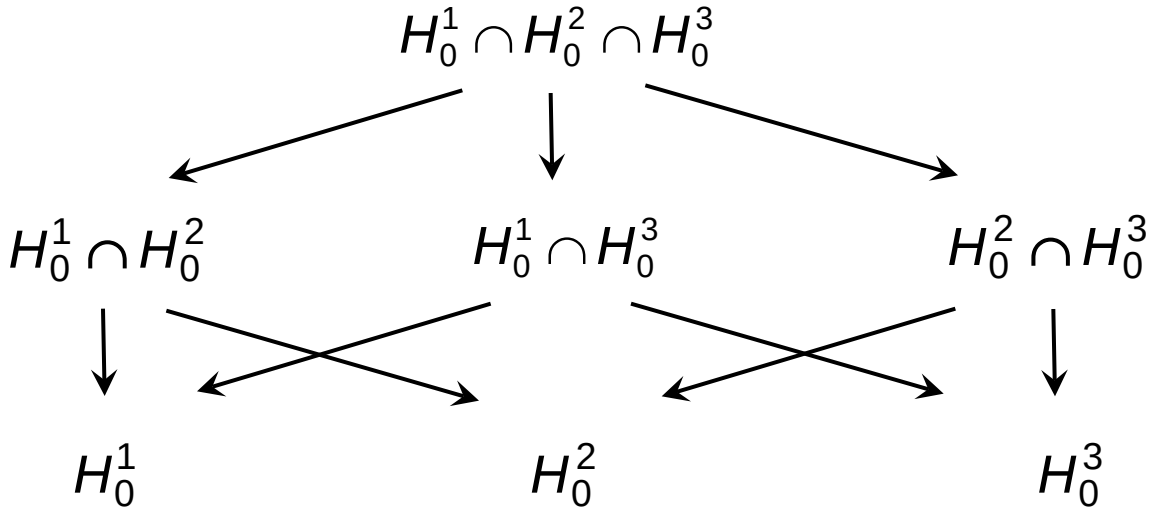
S1 only =	20	%
S2 only =	40	%
S3 only =	0	%
S1S2 only =	0	%
S1S3 only =	0	%
S2S3 only =	0	%
S1S2S3 =	8	%
rem =	32	%

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

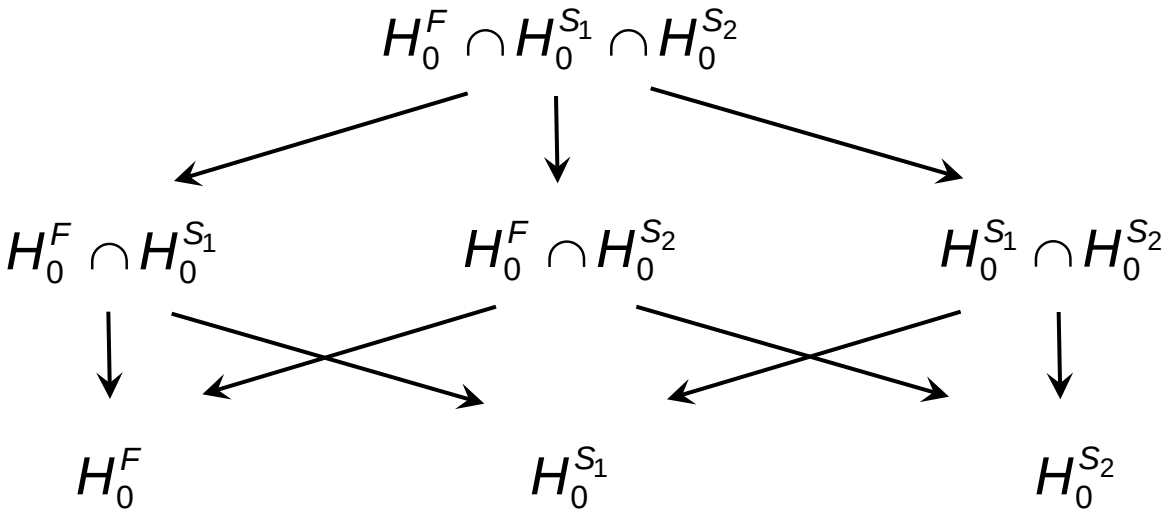
Closed system of hypotheses:



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 expirementwise error rate α in a strong sense.

Closed Testing Procedure

Stage I **Stage II** **…**



?

?

H_0^S

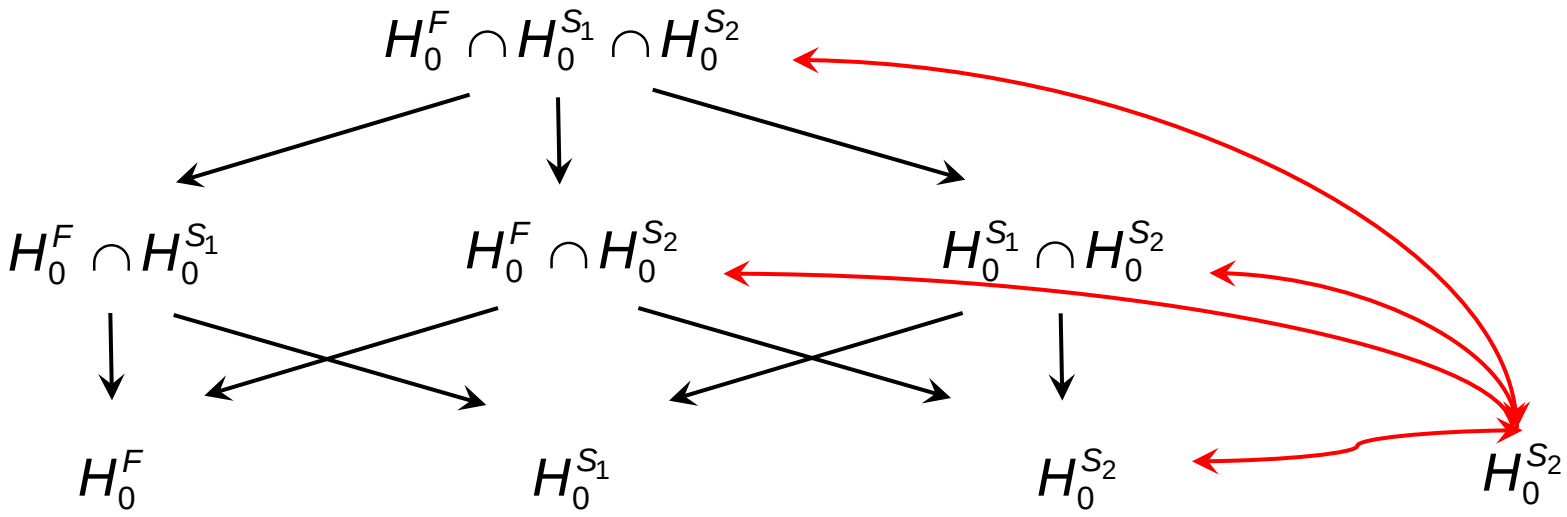
Simple “trick”: Test of intersection hypotheses are formally performed as tests for H_0^S.

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. E.g., you might use inverse normal or Fisher‘s combination test.
- The choice of tests for intersection hypotheses is free. E.g., 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, 2012*)
- Calculation of RCIs and overall p-values straightforward
- Except conditional Dunnett, all procedures available in ADDPLAN PE, Version 6.0

Questions on Simulation
(Maurer, Branson, Posch, 2010)

1. What is the influence of the unknown prevalence of S on the power?
2. How powerful is the design under different assumptions about treatment differences in S and S^c?
3. How does this strategy compare to other strategies (e.g., group sequential designs)?
4. 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?
5. 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

1. Since we have positive correlation between the test statistics the usual intersection tests (Sidak, Simes) can be used.
2. What are the criteria for selecting a population or sticking to the full population?
3. What are the criteria for assessing sample size of full and subset population?
4. Are estimation procedures and procedures for the calculation of overall *p*-values available?
5. 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,

$$p_k^g \leq \alpha \quad \Leftrightarrow \quad H_0^g \text{ can be rejected at stage } k$$

- *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

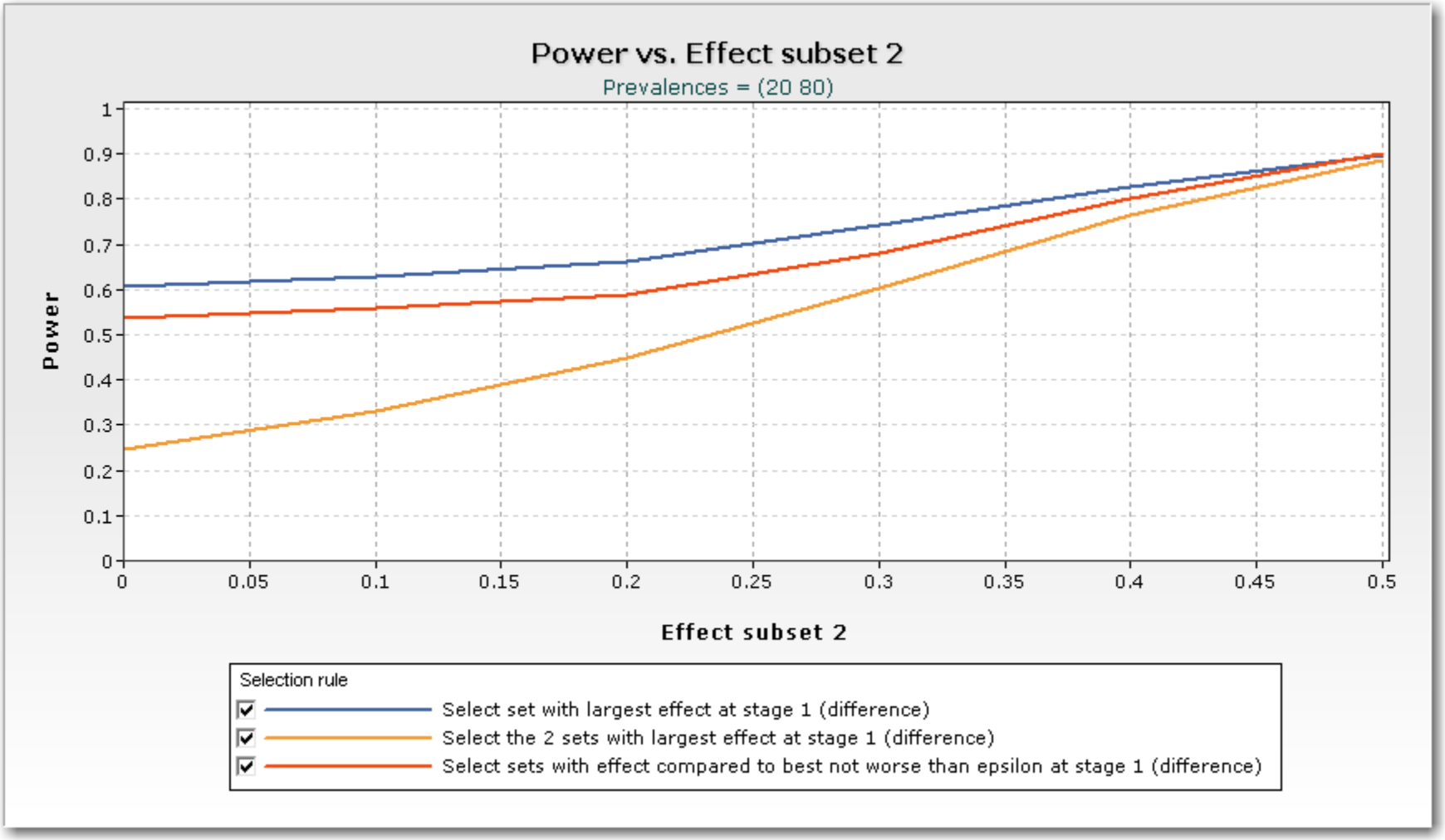
Adaptive Designs – Plans and Analyses®

- 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 (free of charge)
  - 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

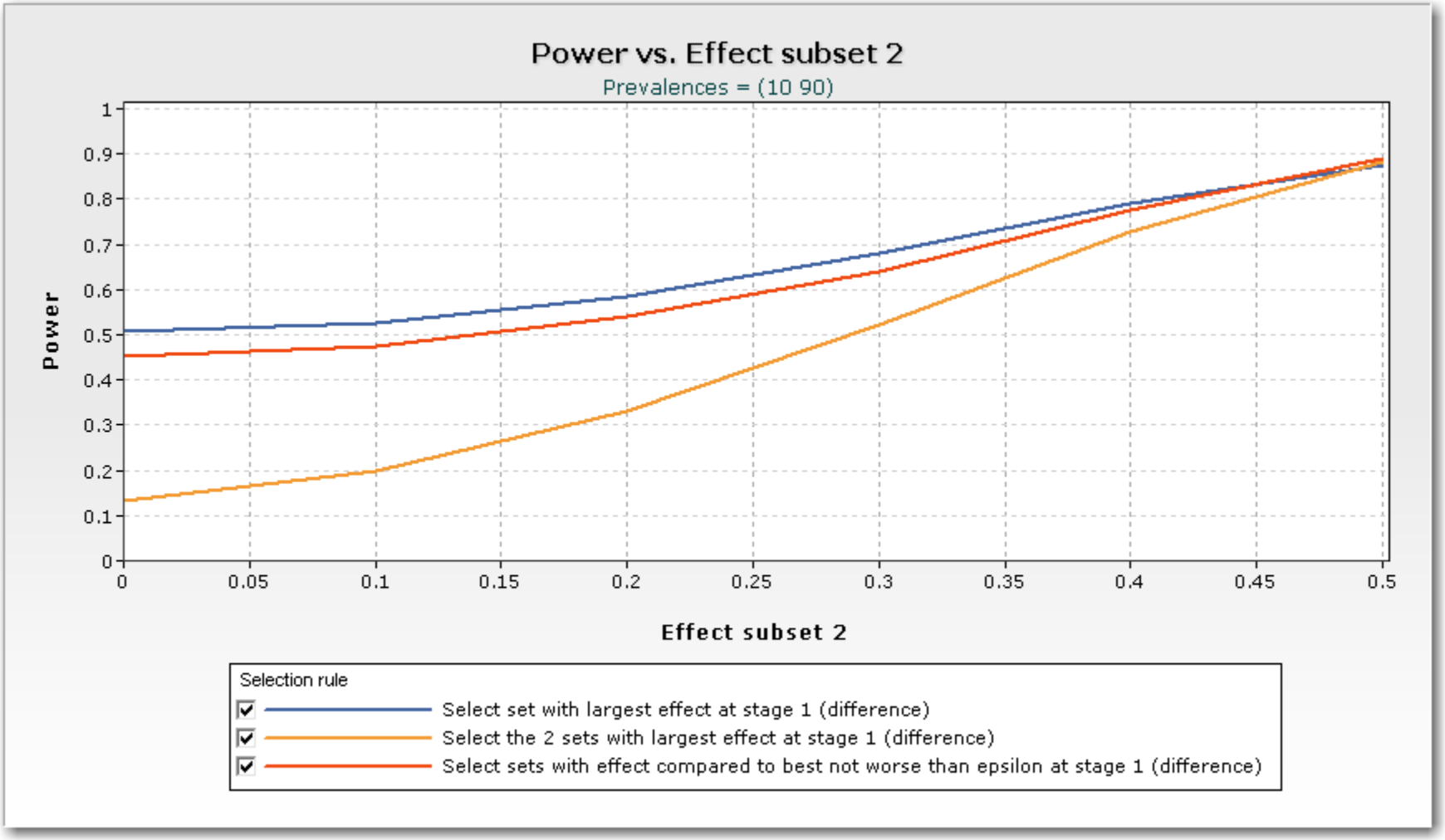
Simulation Example

- 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 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
- Prevalances of biomarker positive population is 5%, 10%, 20%.
- Sample sizes 100 patients per stage
- Simes‘ test is used for testing intersection hypotheses.

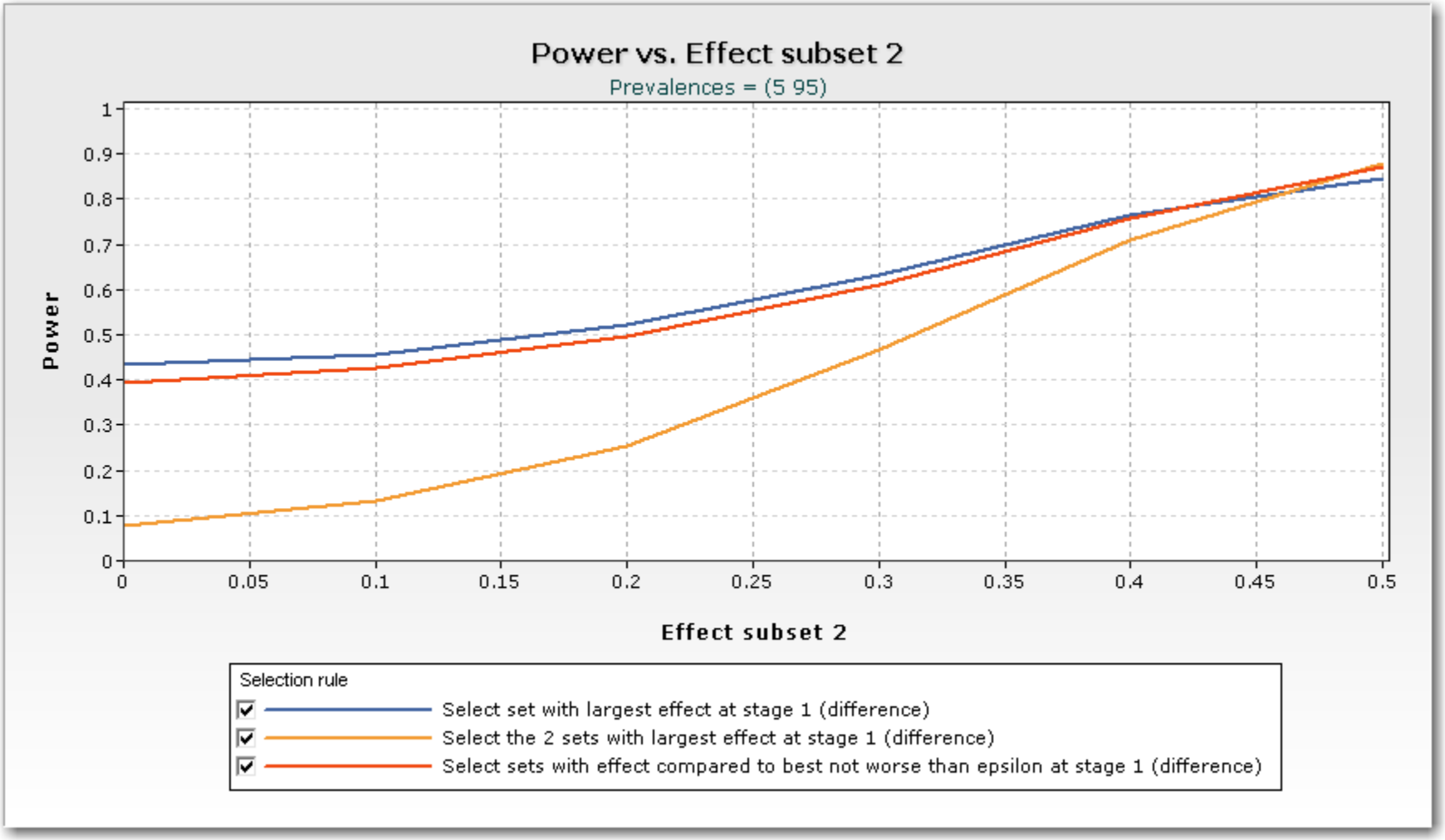
Results



Results



Results



Results

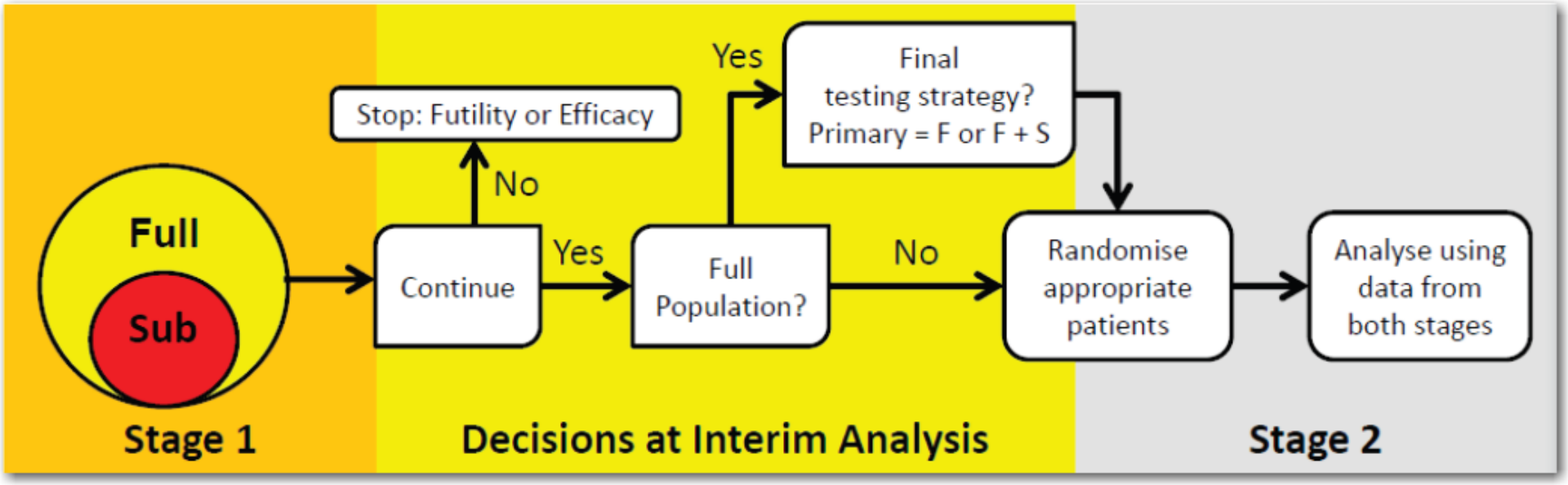
- Clear power disadvantage for procedure that never selects a sub-population
- No clear advantage of selecting always (and only) the best population
- For small prevalences, always selecting the best can even provide a small loss in power

A Case Study Example

Simulation Example
Case Study: Phase 3 Trial in HER2- MBC Patients

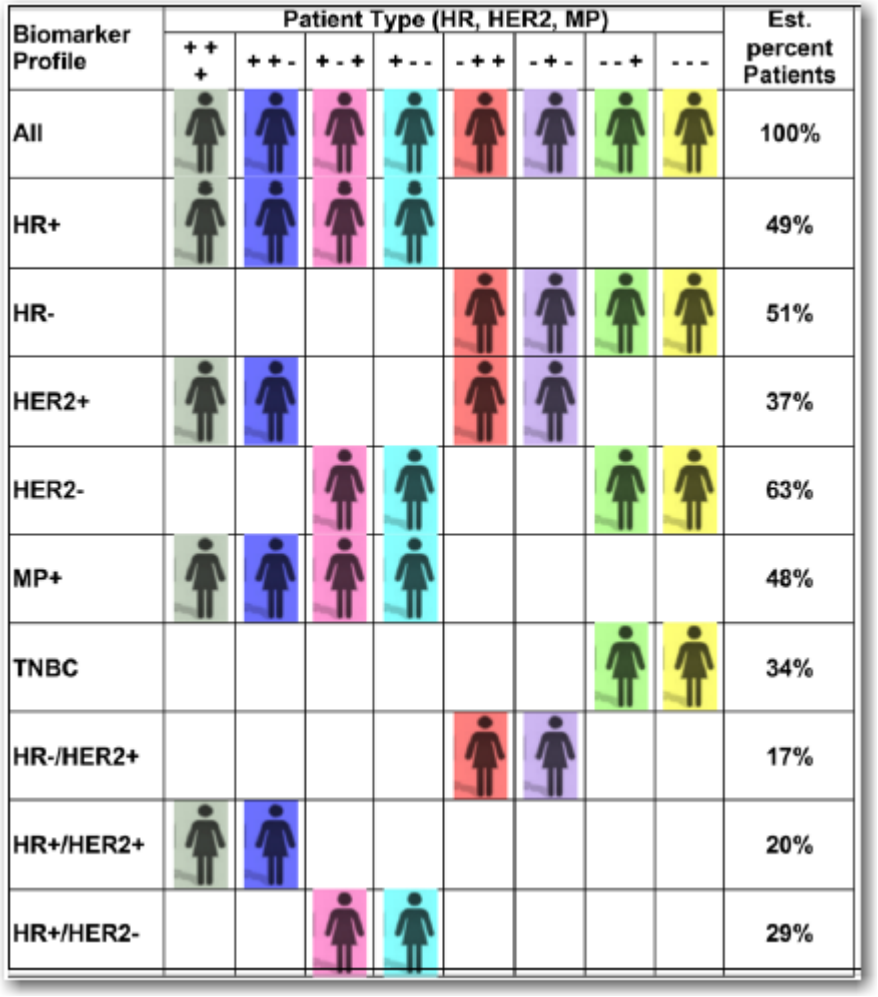
- Assume that one of the experimental drugs has been graduated from the I-SPY 2 trial with the biomarker signature of triple negative breast cancer (TNBC) but also with some promising effect in HER2- biomarker signature.
- **Option 1**: a confirmatory Phase 3 trial in TNBC patients only
  - prevalence of TNBC is only about 34%
- **Option 2**: a confirmatory Phase 3 trial in HER2- patients
  - prevalence of HER2- is about 63%
- **Option 3**: Adaptive enrichment design
  - run a confirmatory trial with a two-stage enrichment design
  - starting with the full population (HER2- patients),
  - but with the preplanned option of selecting only the TNBC patients after the 1st stage in case the observed effect is not promising in the HER2- patients with positive hormone-receptor status HR+

Adaptive Population Enrichment Design



- Stage1 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

Case Study: Phase 3 Trial in HER2- MBC Patients



Biomarker Profile	Patient Type (HR, HER2, MP)								Est. percent Patients
	+ + +	+ + -	+ - +	+ - -	- + +	- + -	- - +	- - -	
All	●	●	●	●	●	●	●	●	100%
HR+	●	●	●	●					49%
HR-					●	●	●	●	51%
HER2+	●	●			●	●			37%
HER2-			●	●			●	●	63%
MP+	●	●	●	●					48%
TNBC							●	●	34%
HR-/HER2+					●	●			17%
HR+/HER2+	●	●							20%
HR+/HER2-			●	●					29%

Planning the Trial

- Primary Endpoint : pathologic complete response (pCR) at surgery
- Power: 90%
- Sign. Level: 0.025
- Control Rate: pCR=0.3
- TRT Effect: 0.2
- Apply Bonferroni correction

	Plan 1 Rates	Plan 2 Rates	Plan 3 Rates	Plan 4 Rates
alpha	0.0125	0.0125	0.0125	0.0125
Futility stops	-	-	-	-
tails	1	1	1	1
K	1	1	1	1
Design	-	-	-	-
Information rates	-	-	-	-
Hypothesis	diff<=0	diff<=0	diff<=0	diff<=0
Parameters	pi1=0.3 pi2=0.4	pi1=0.3 pi2=0.45	pi1=0.3 pi2=0.5	pi1=0.3 pi2=0.55
Power %	90.0	90.0	90.0	90.0
Total ASN H0	-	-	-	-
Total ASN H01	-	-	-	-
Total ASN H1	-	-	-	-
Total maximum N	1124.9	512.6	293.3	189.5
Allocation	1	1	1	1

Adaptive PE Simulation

Prevalences

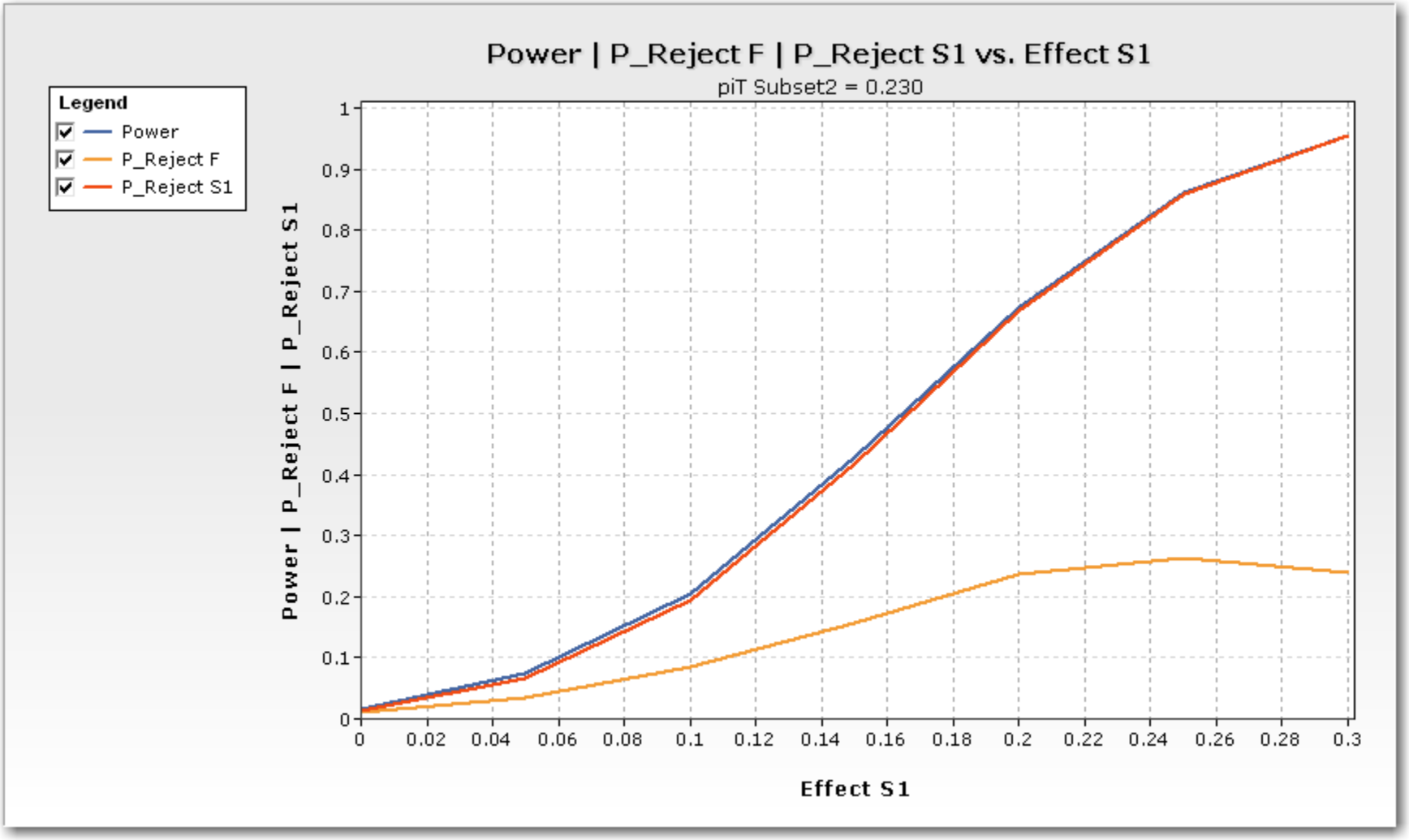
	MP Hi-1		MP Hi-2	
	HR +	HR-	HR+	HR-
HER2+	16%	7%	4%	10%
HER2-	23%	6%	6%	28%

pCR rates

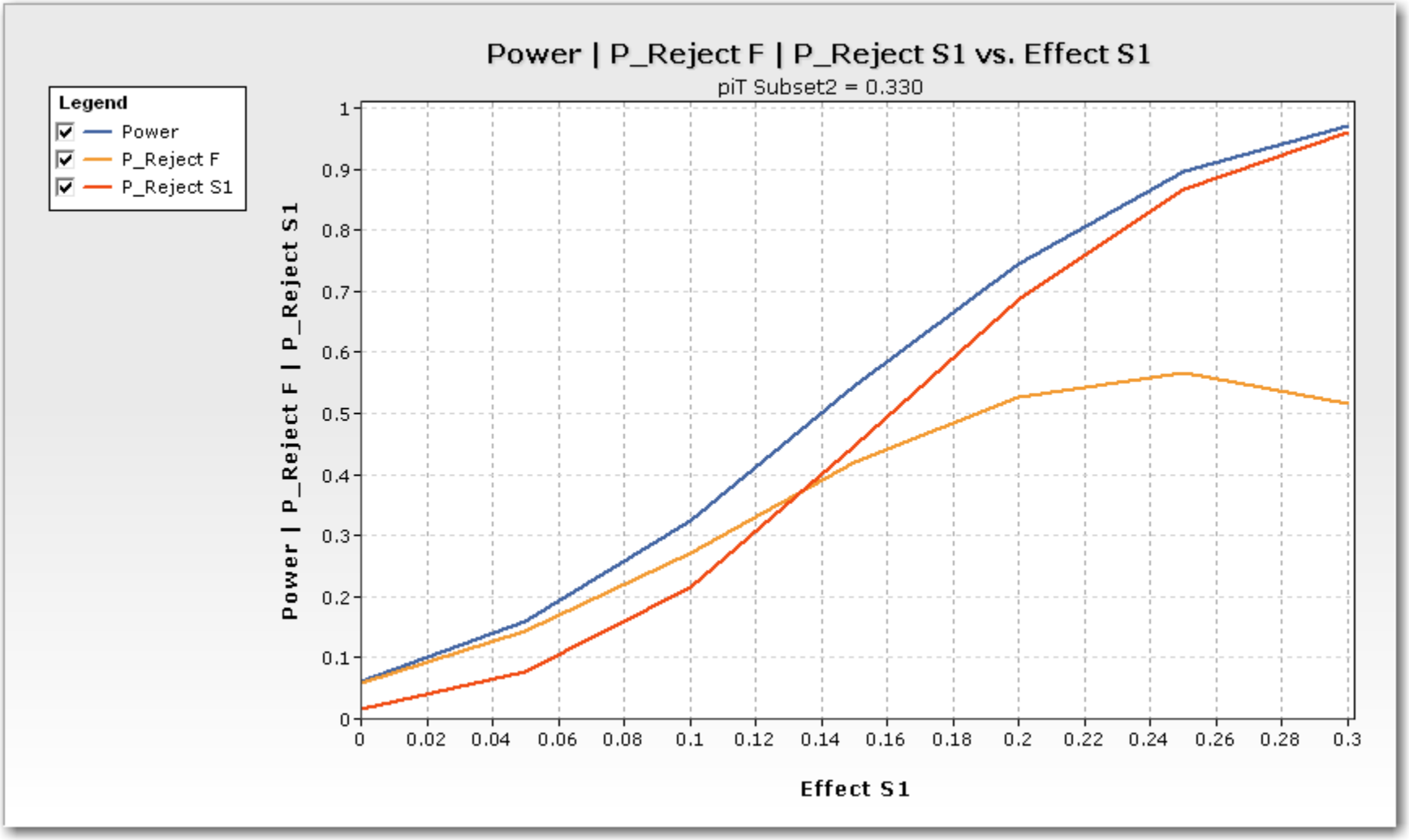
	MP Hi-1		MP Hi-2	
	HR +	HR-	HR+	HR-
HER2+	47%	67%	35%	55%
HER2-	25%	43%	17%	32%

- Prevalence of TNBC in HER2- : 54% (= (6 + 28)/63)
- Control pCR Rate in TNBC: 0.34 (= (6*0.43+28*0.32)/ 34))
- Control pCR Rate in HER2- ∩ HR+: 0.23 (= (23*0.25+6*0.17)/29)
- Total of 21 Simulation Scenarios:
  - TRT effect in TNBC: 0 to 0.3 by 0.05
  - TRT effect in HER2- ∩ HR+: 0, 0.1, 0.2
- Selection rule: Select set with effect compared to best not worse than $\varepsilon = 0.1$
- Total sample size: 300 patients, stage 1 sample size:150 patients

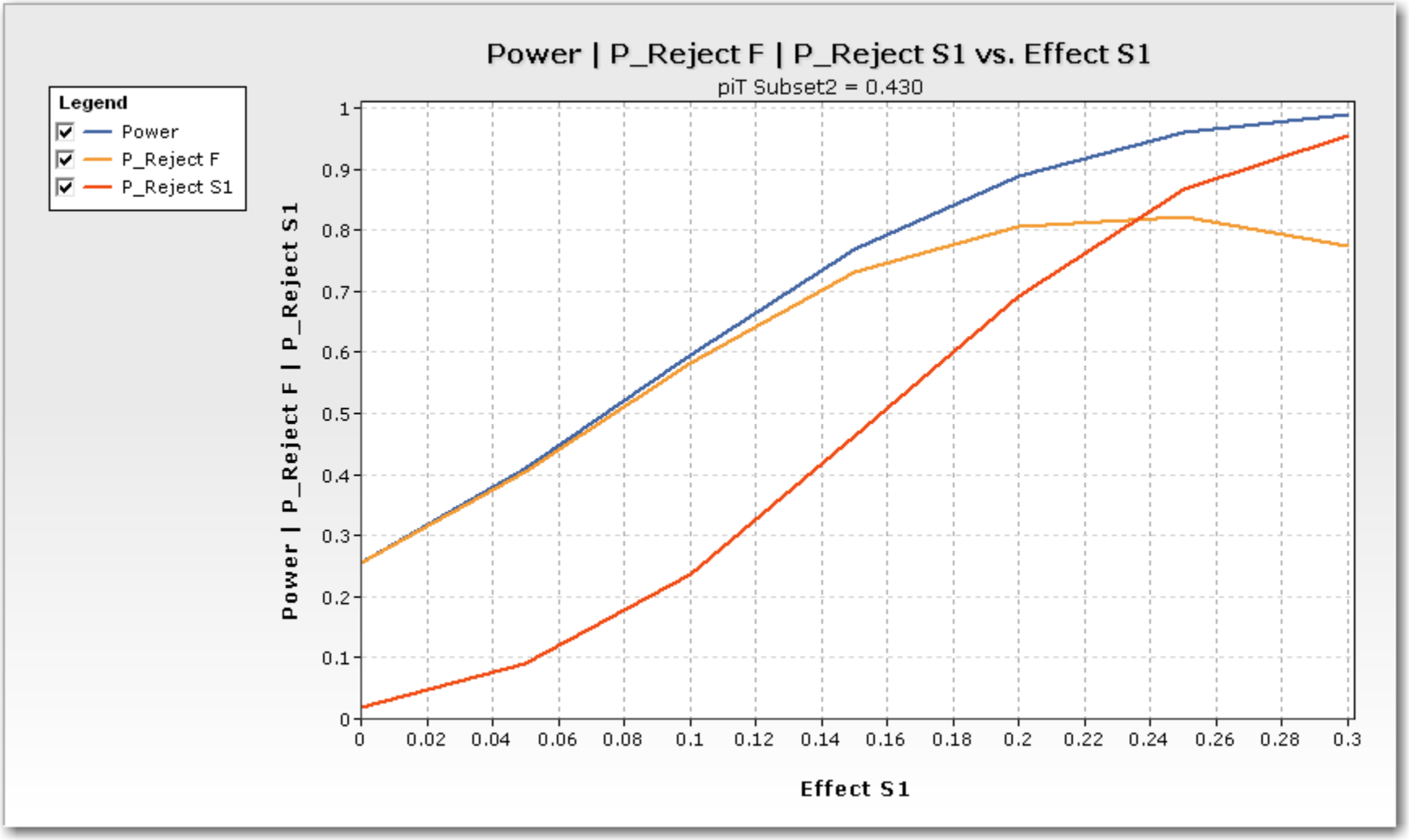
Operating Characteristics:



Operating Characteristics:



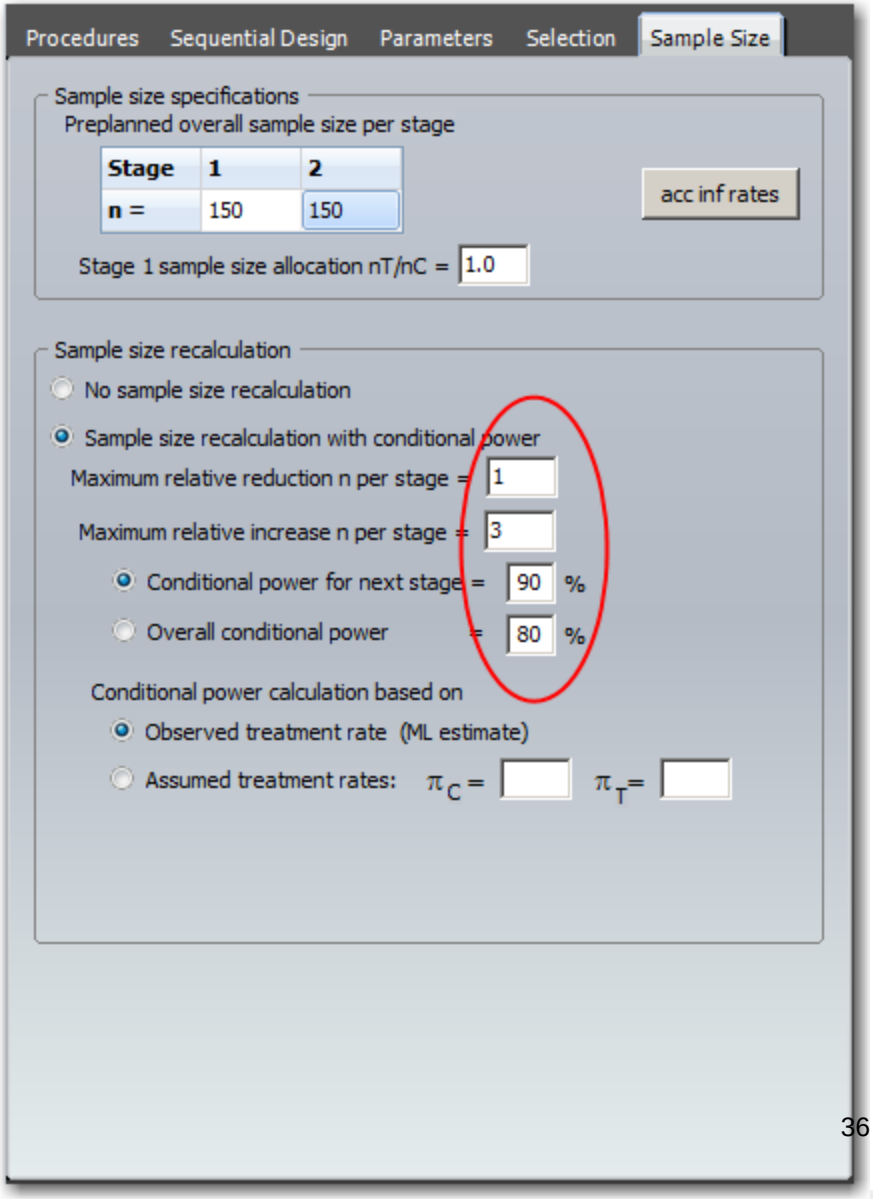
Operating Characteristics:



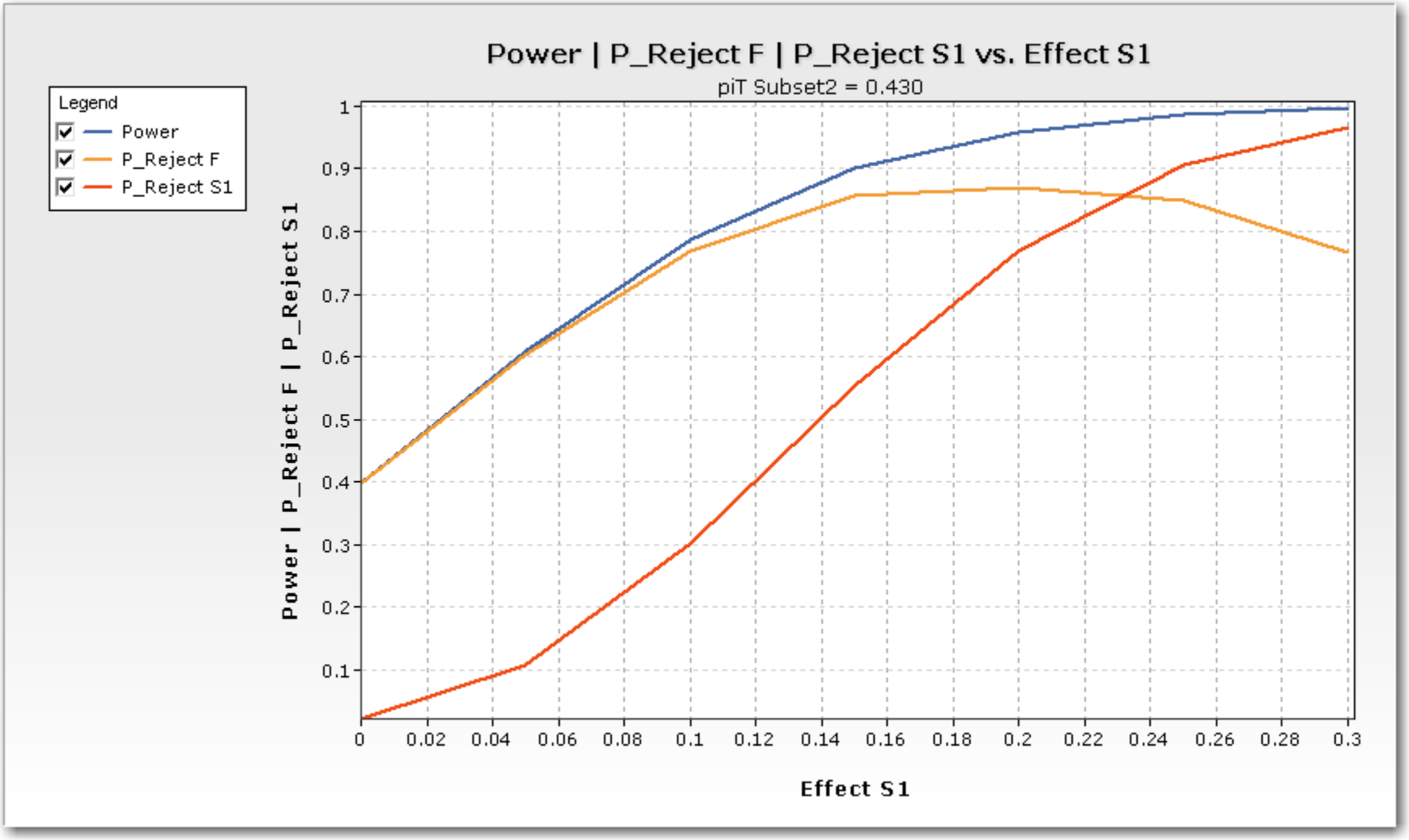
Sample Size Reestimation

- Allow up to a 3-fold sample size increase for Stage 2
- 90% Conditional Power based on observed TRT effect
- Additional futility rule: Select only if effect exceeds 0.1
- Total Sample Size:

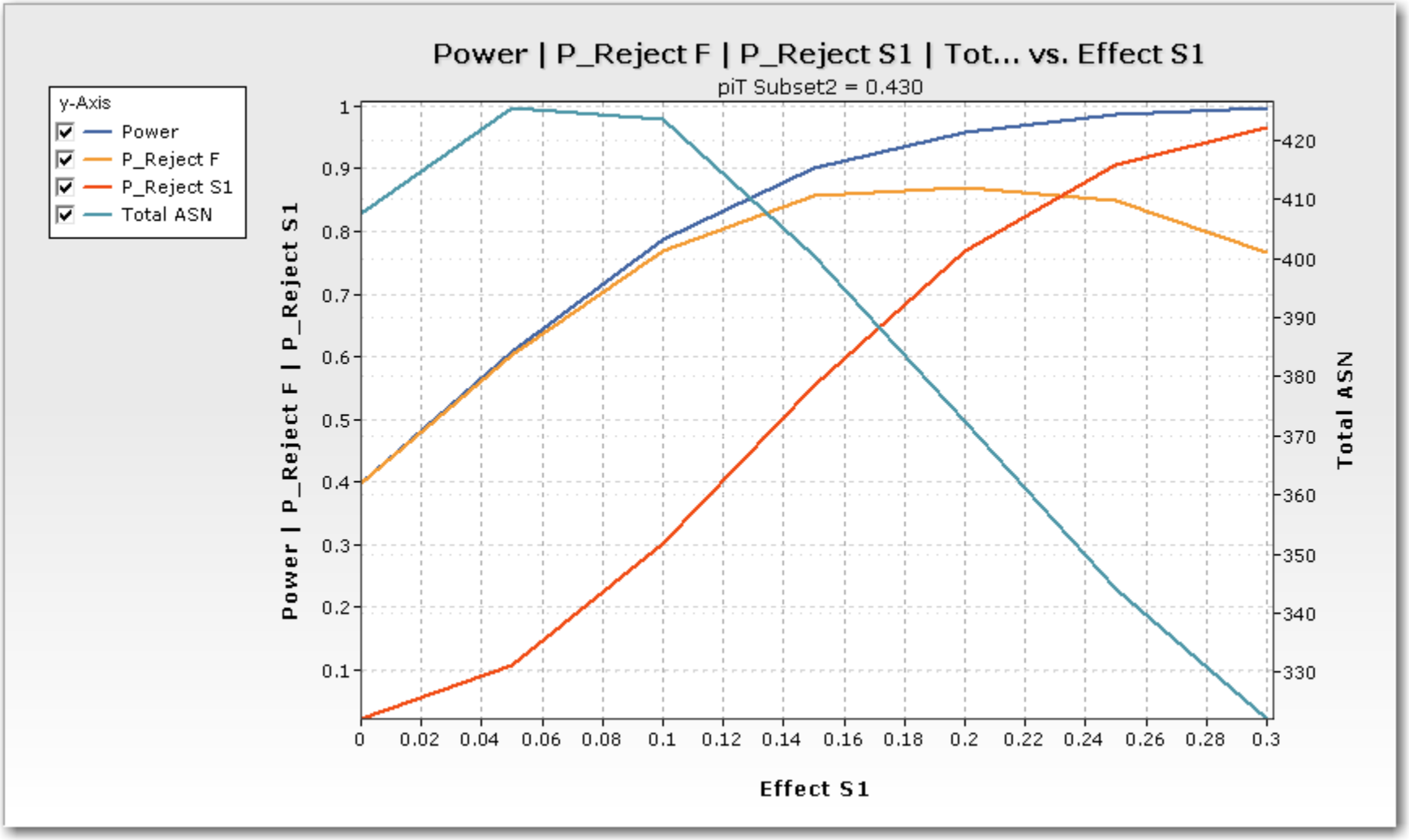
300 - 600



Operating Characteristics



Operating Characteristics



Patient Enrichment Designs: Analysis

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:

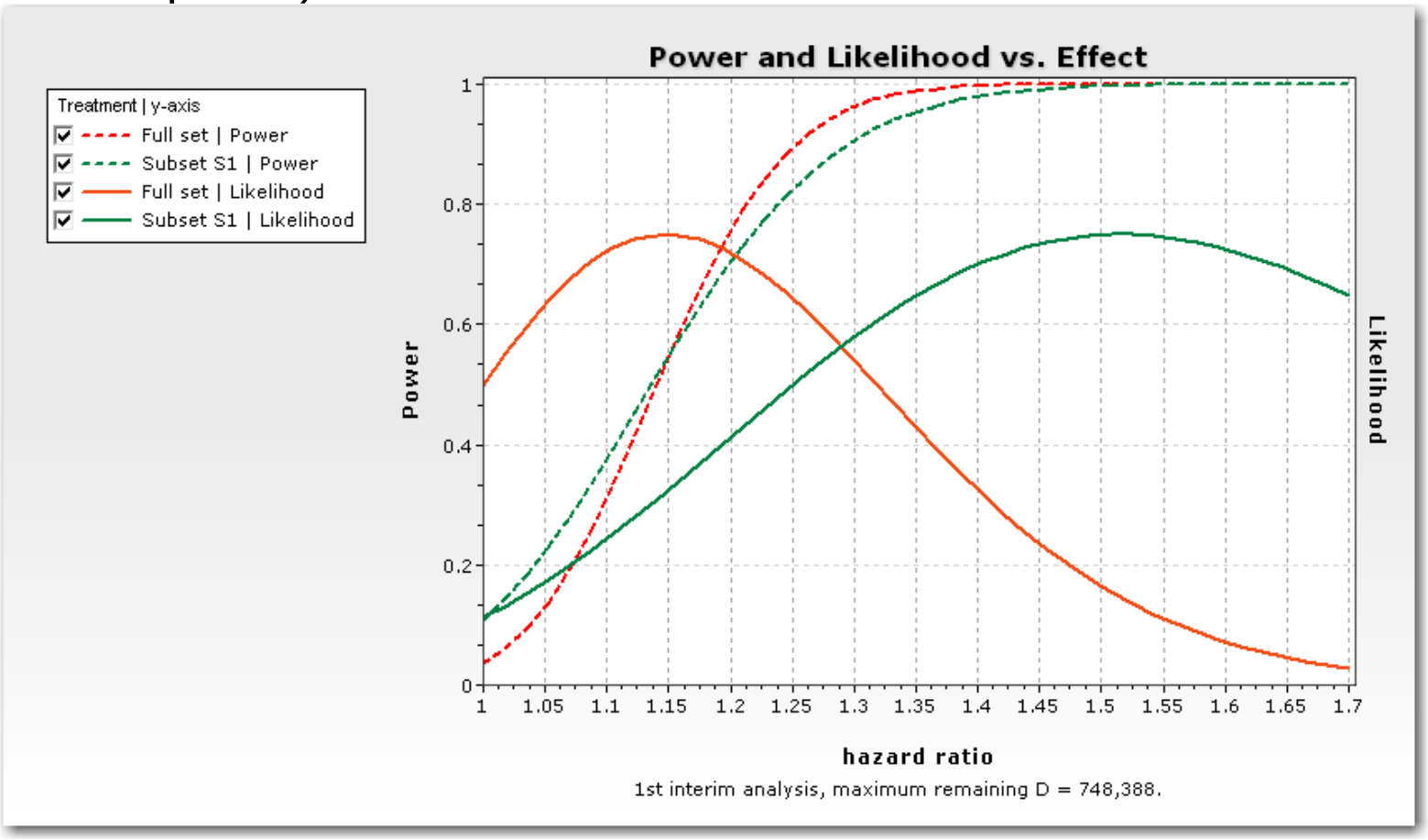
Stage 1 | Stage 2 | Stage 3 — Interim analysis k = 1

	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

- Consider 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 to 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.

References

- Bauer, P. (1989). Multistage testing with adaptive designs. Biom. und Inform. in Med. und Biol. 20, 130-148.
- Bauer, P., Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. Biometrics 50, 1029–1041.
- Bauer, P., Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. Statistics in Medicine 18,1833–1848.
- Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M., Racine-Poon, A., 2009: Confirmatory adaptive designs with with Bayesian decision tools for a targeted therapy in oncology. Statistics in Medicine 28:1445-1463.
- Bretz, F., König, F., Brannath, W., Glimm, E., Posch, M.: Tutorial in Biostatistics: Adaptive designs for confirmatory clinical trials. Statistics in Medicine 28: 1181-1217, 2009.
- Friede, T., Parsons, N., Stallard, N.: A conditional error function approach for subgroup selection in adaptive clinical trials. Statistics in Medicine **31**:4309–4320.
- Hommel, G., 2001: Adaptive modifications of hypotheses after an interim analysis. Biometrical J. 43, 581–589.
- Maurer, W., Branson, M., Posch, M. (2010). Adaptive designs and confirmatory hypothesis testing. In: Multiple Testing Problems in Pharmaceutical Statistics. CRC Press Chapman & Hall.
- Posch, M., König, F., Branson, M., Brannath, W., Dunger-Baldauf, C., Bauer, P. (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. Statistics in Medicine 24, 3697–3714.
- Temple, R. J. (1994). Special study designs: early escape, enrichment, studies in non-responders. Comm. Stat. - Theory Meth., 23, 499 - 531
- Wang, S.-J., Hung, H. M. J., O'Neill, R. T. (2009). Adaptive patient enrichment designs in therapeutic trials. Biometrical J. 51, 358–374.
- Wassmer, G. (2006). Planning and analyzing adaptive group sequential survival trials. Biometrical J. 48 714-729.
- Wassmer, G. (2011). On Sample Size Determination in Multi-Armed Confirmatory Adaptive Designs. J. Biopharm. Stat., 21, 802-817