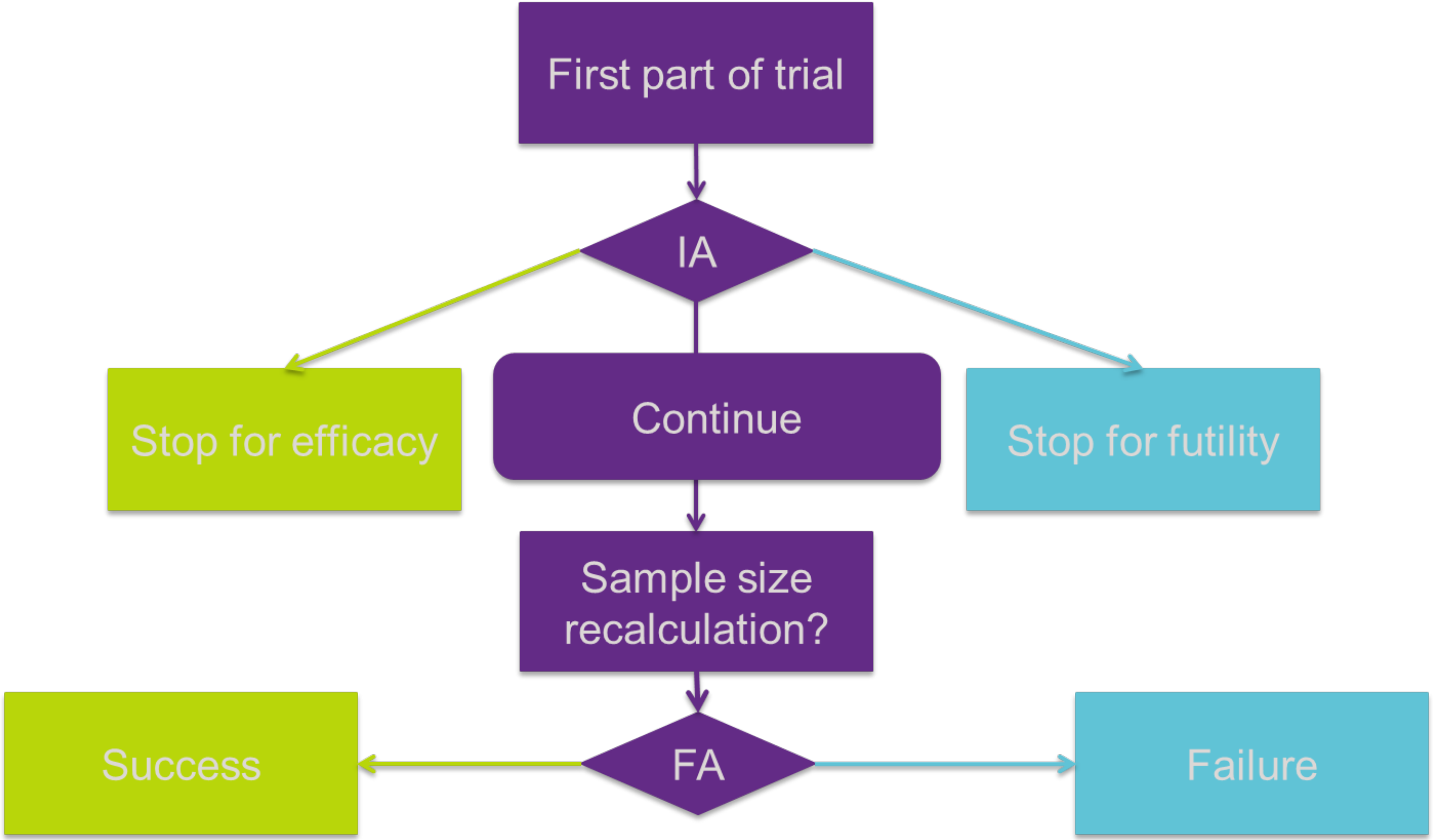
Confirmatory Adaptive Group Sequential Trials: Treatment Arm and Subpulation Selection Based on Surrogate Endpoints

Silke Jörgens

ICON Innovation Center

Gernot Wassmer

Generic Example



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)

General Methodology – Group Sequential Adaptive Design

– Throughout this presentation, consider the inverse normal method for combining *p*-values over stages:

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

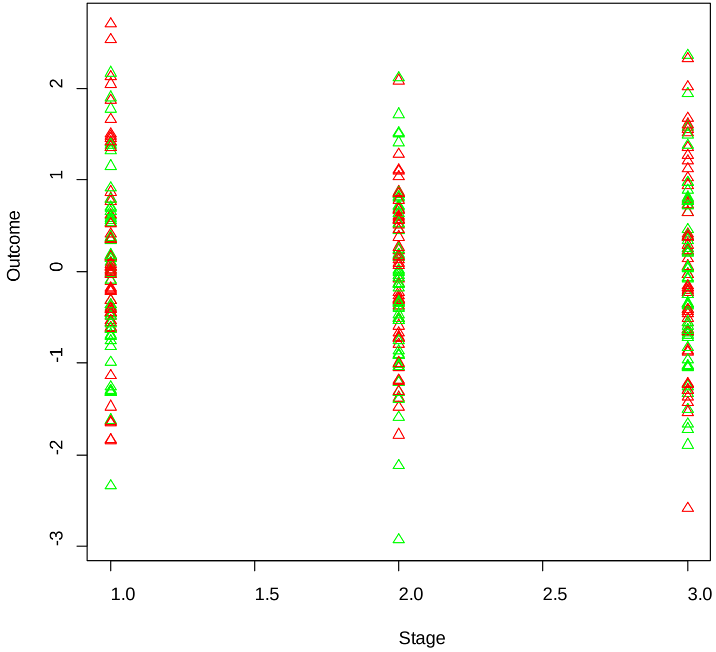
where

w_i, i = 1, …, K, denote prespecified weights for stages 1 to K

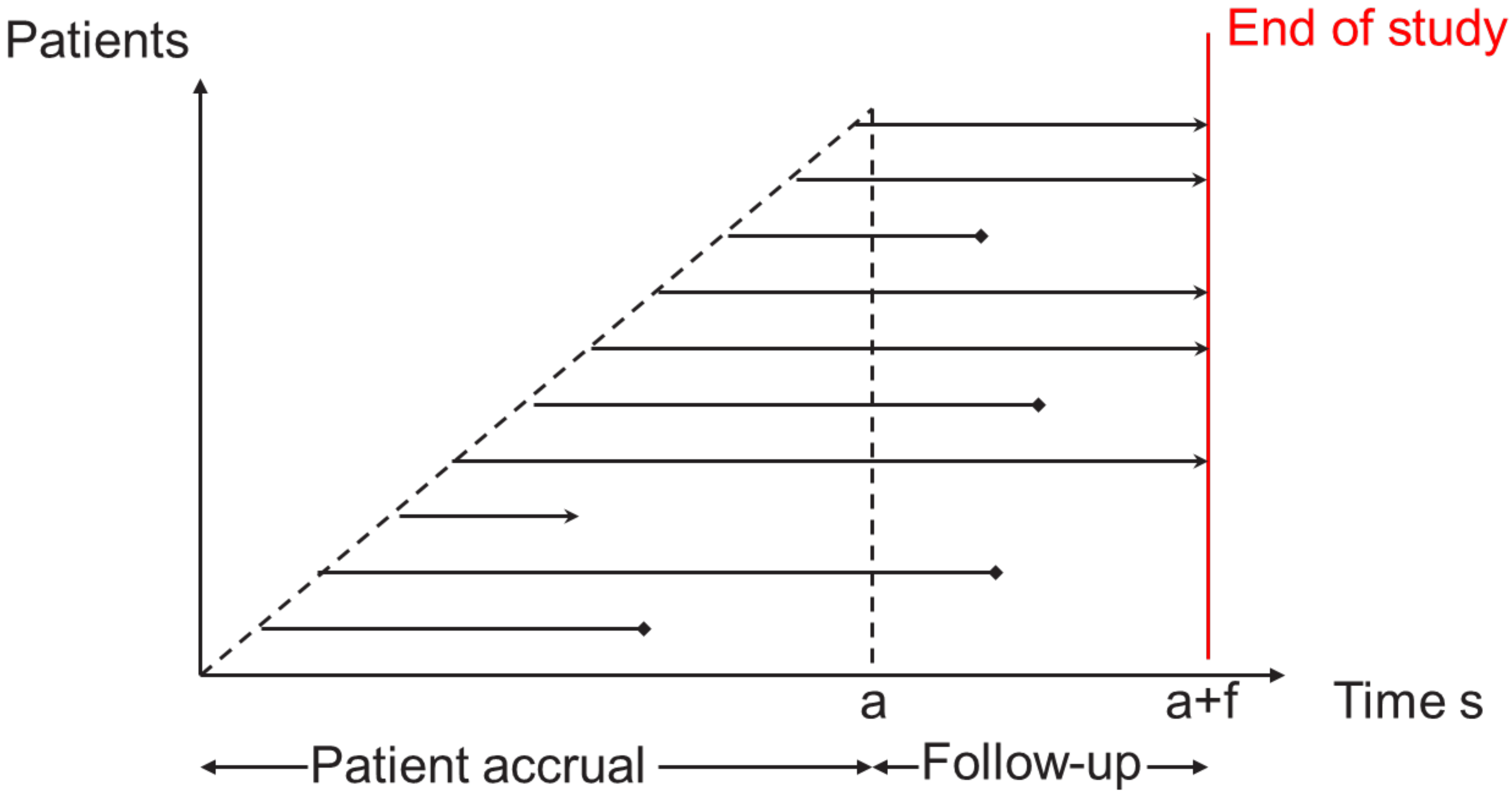
p_i, i = 1, …, K denote the stagewise *p*-values for stages i = 1, …, K

– 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



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)
  $\rightarrow$ 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

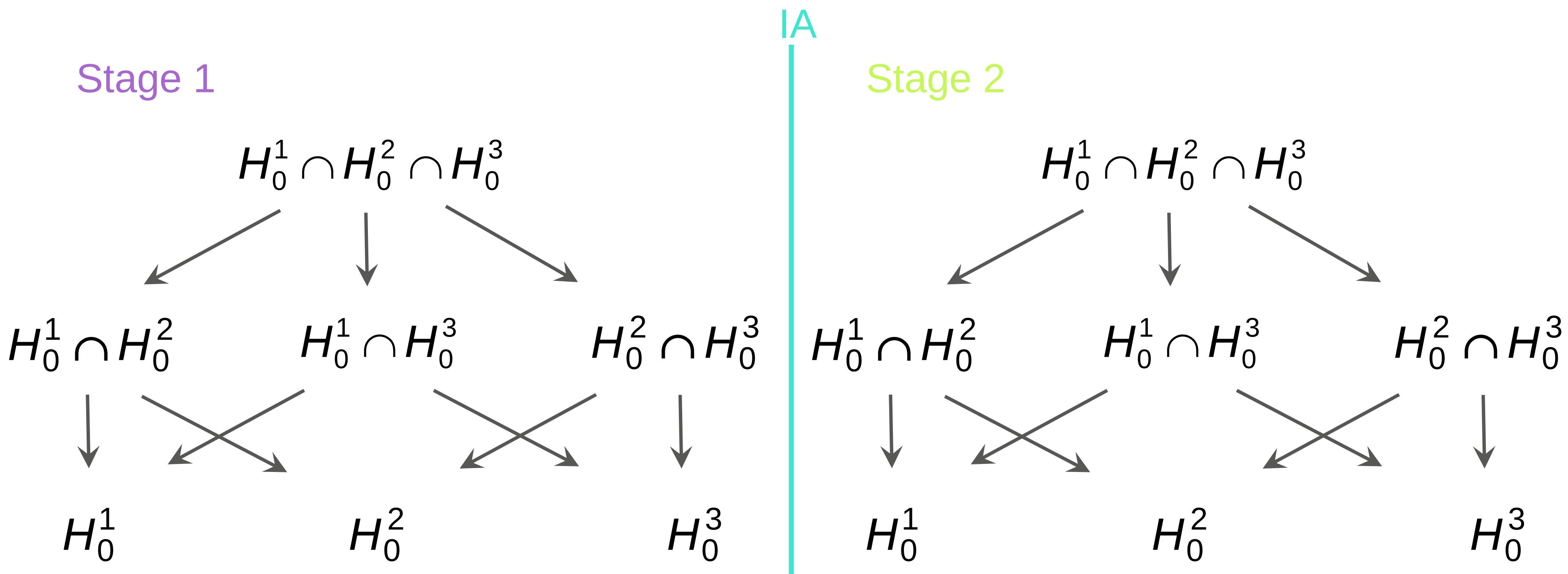
Wishlist for Multiple Treatment Arms

- 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

$$
\begin{tikzcd}
& H_0^1 \cap H_0^2 \cap H_0^3 \arrow[dl] \arrow[d] \arrow[dr] & \\
H_0^1 \cap H_0^2 \arrow[d] \arrow[dr] & H_0^1 \cap H_0^3 \arrow[dl] \arrow[dr] & H_0^2 \cap H_0^3 \arrow[dl] \arrow[d] \\
H_0^1 & H_0^2 & H_0^3
\end{tikzcd}
$$

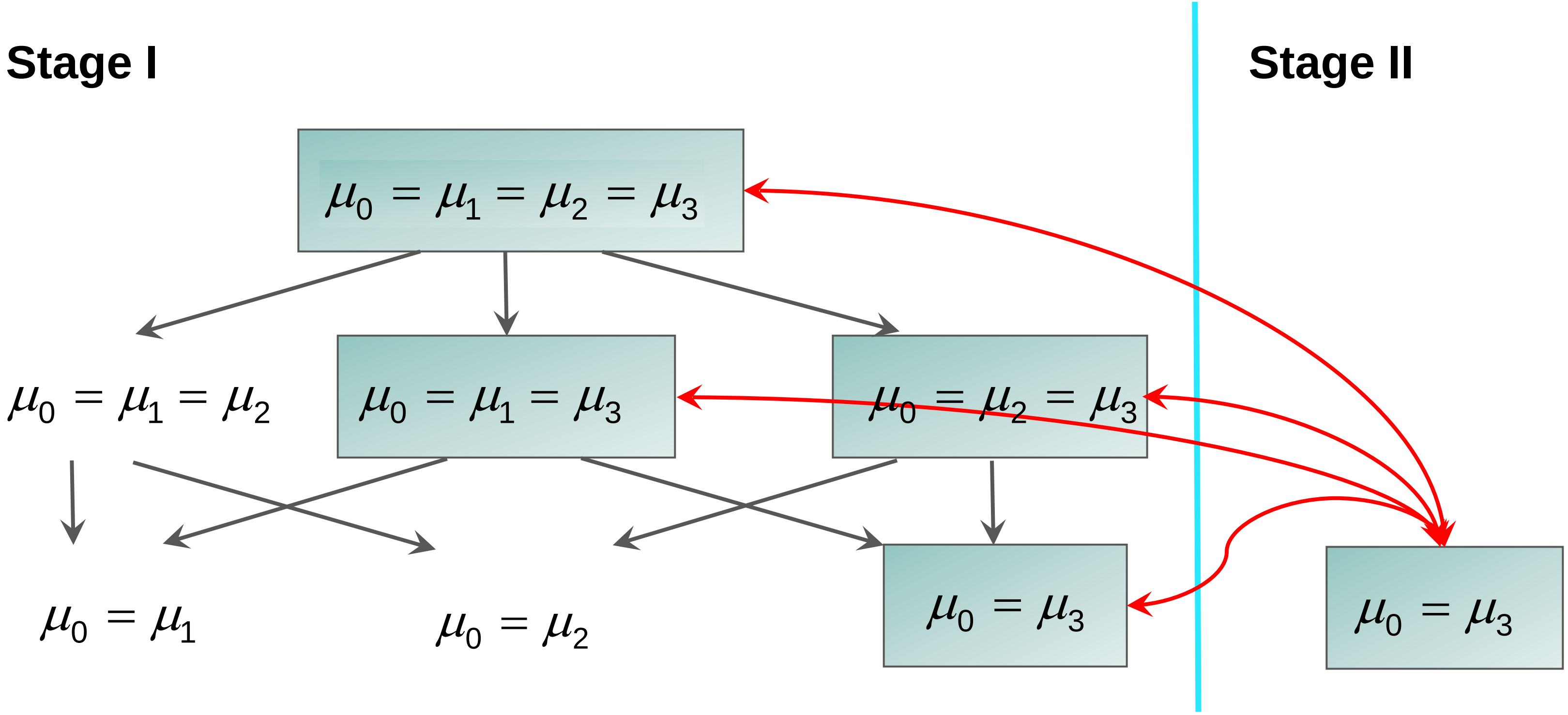
Closed Testing Procedure in Multi Stage Designs



… use inverse normal combination test for each pair of hypotheses

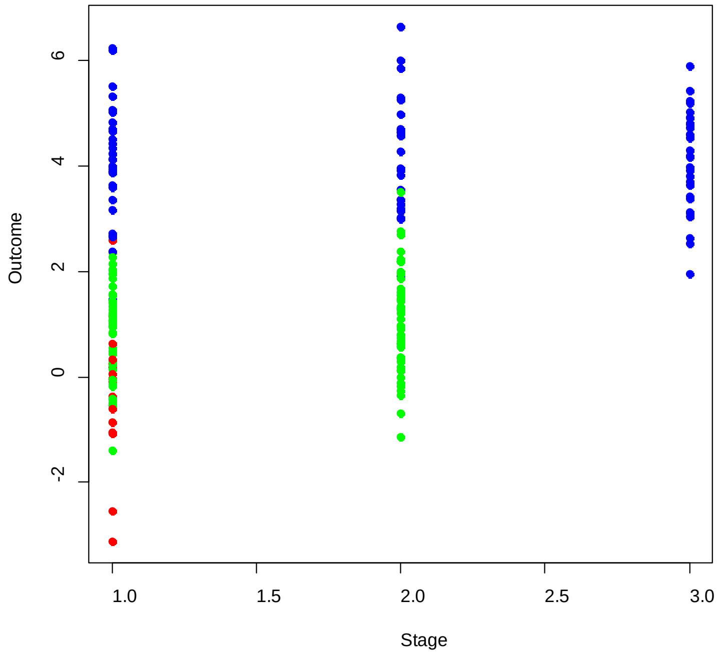
(Bauer and Kieser, 1999; Posch et al., 2005)

Treatment Selection

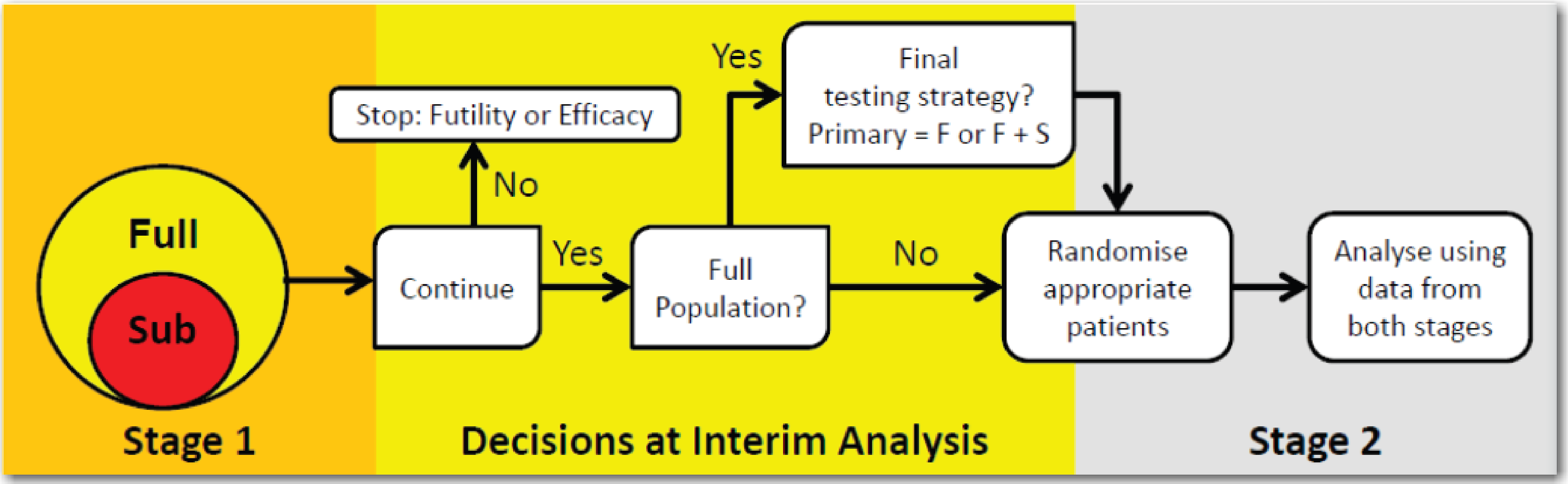


H_0^3 can be rejected if all combination tests exceed the critical value u_2

Treatment Selection



Population Selection



- 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

Population Selection

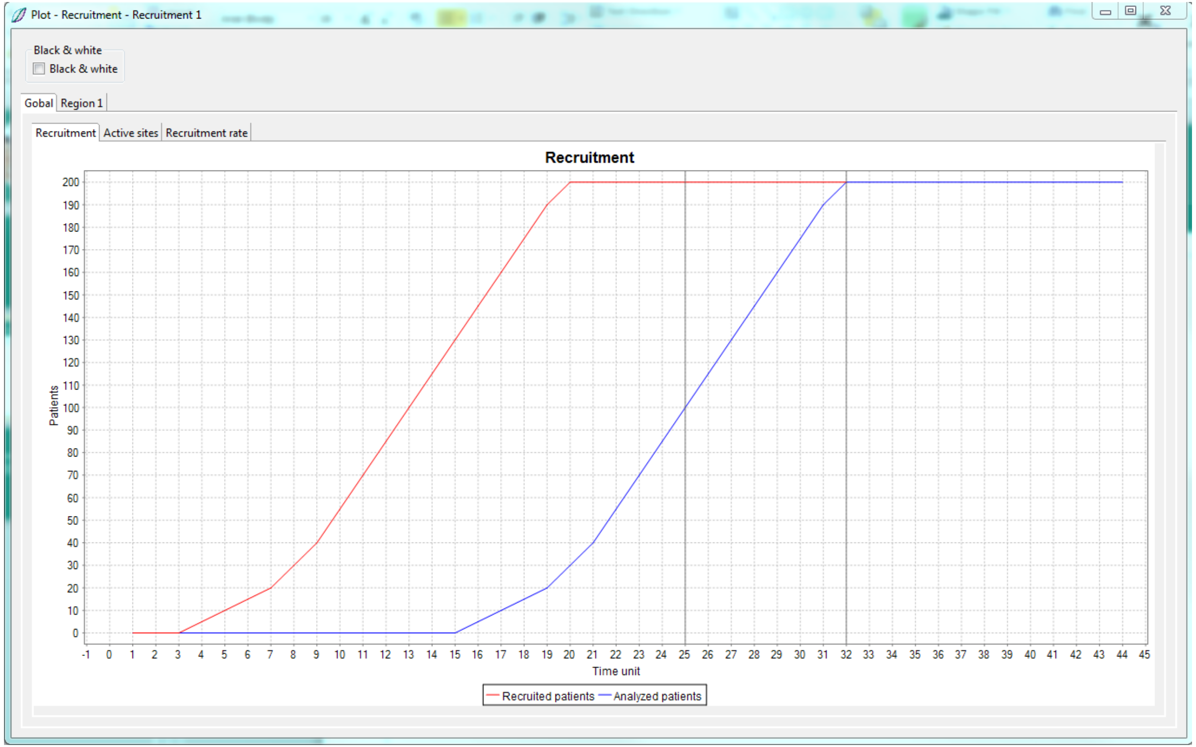
– Testing procedure works in much the same way (testing hypotheses in multiple subgroups, instead of comparing multiple treatments to one control)

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

$\rightarrow$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

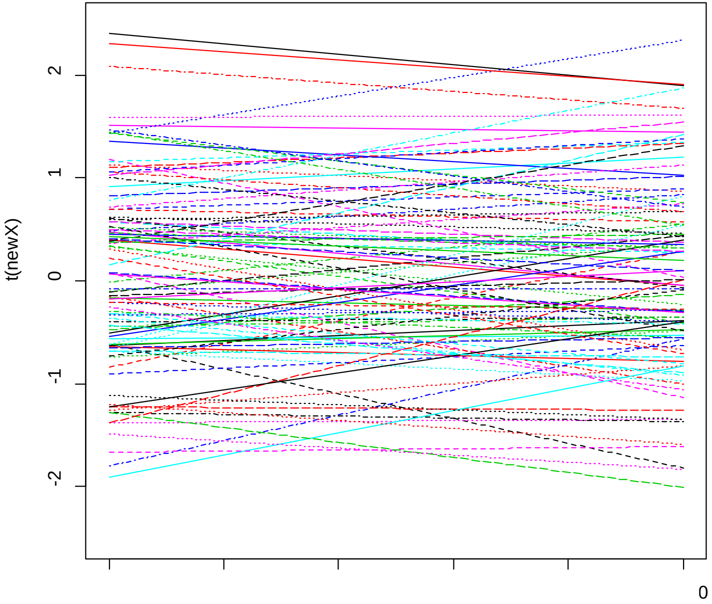
User‘s Question

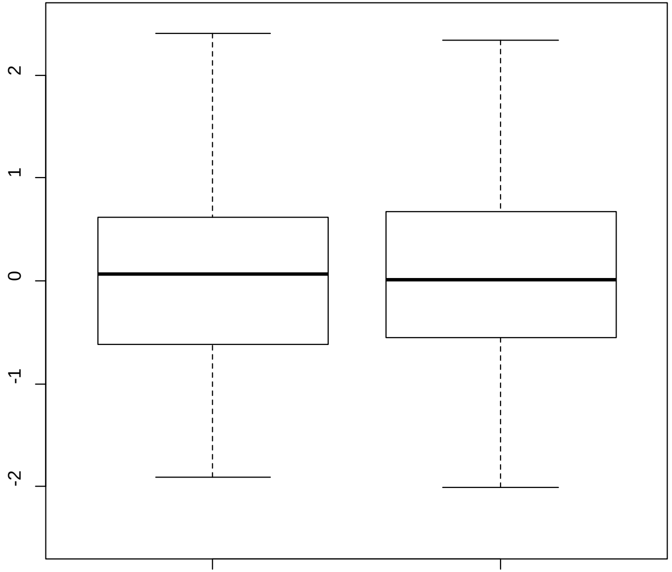
- Real life question in a project with early readout:

  „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

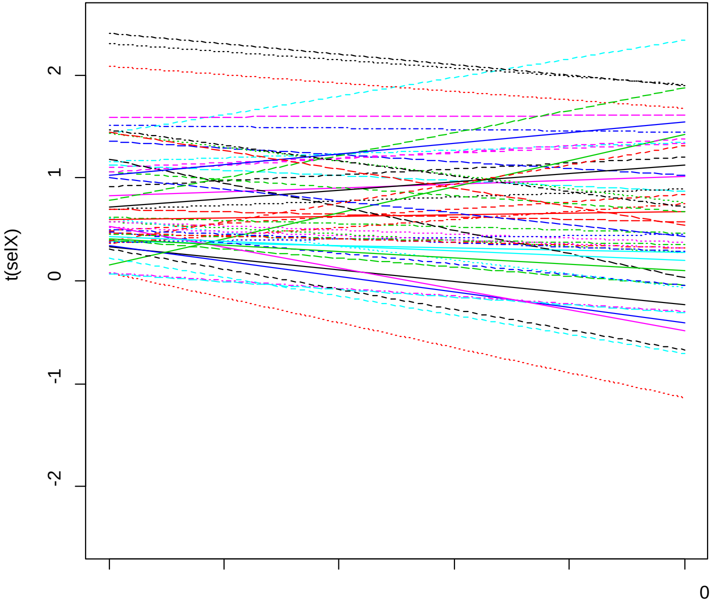
Selection Bias: Two Readouts

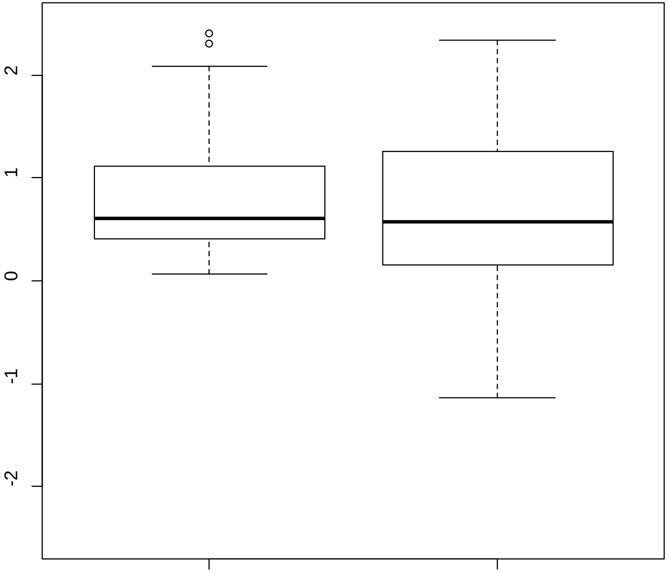




Short-term Long-term Short-term Long-term

Selection Bias: Only Patients With Positive Short-Term



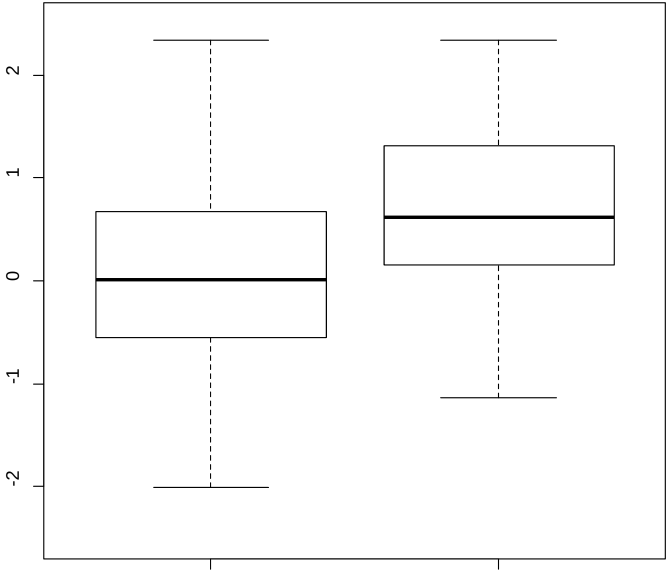


Short-term Long-term

Short-term Long-term

Selection Bias: What We Achieved …

Long-term outcome



All patients

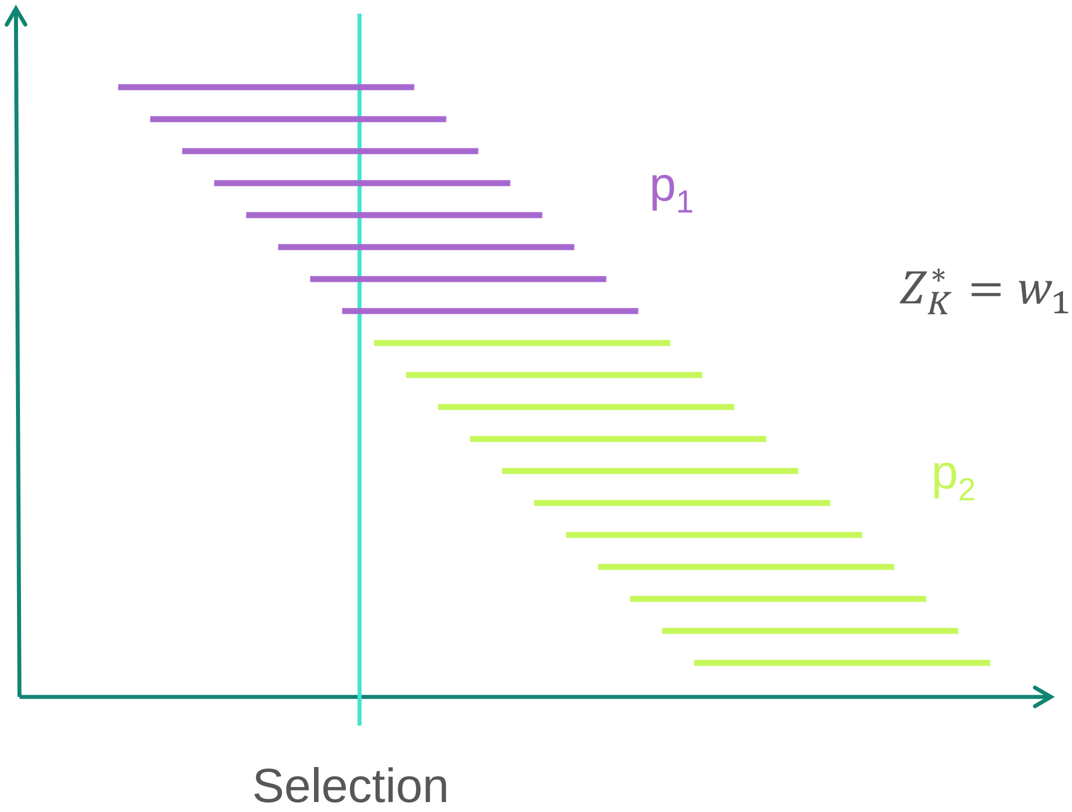
Patients with positive short-term only

… 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
  $\rightarrow$ 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



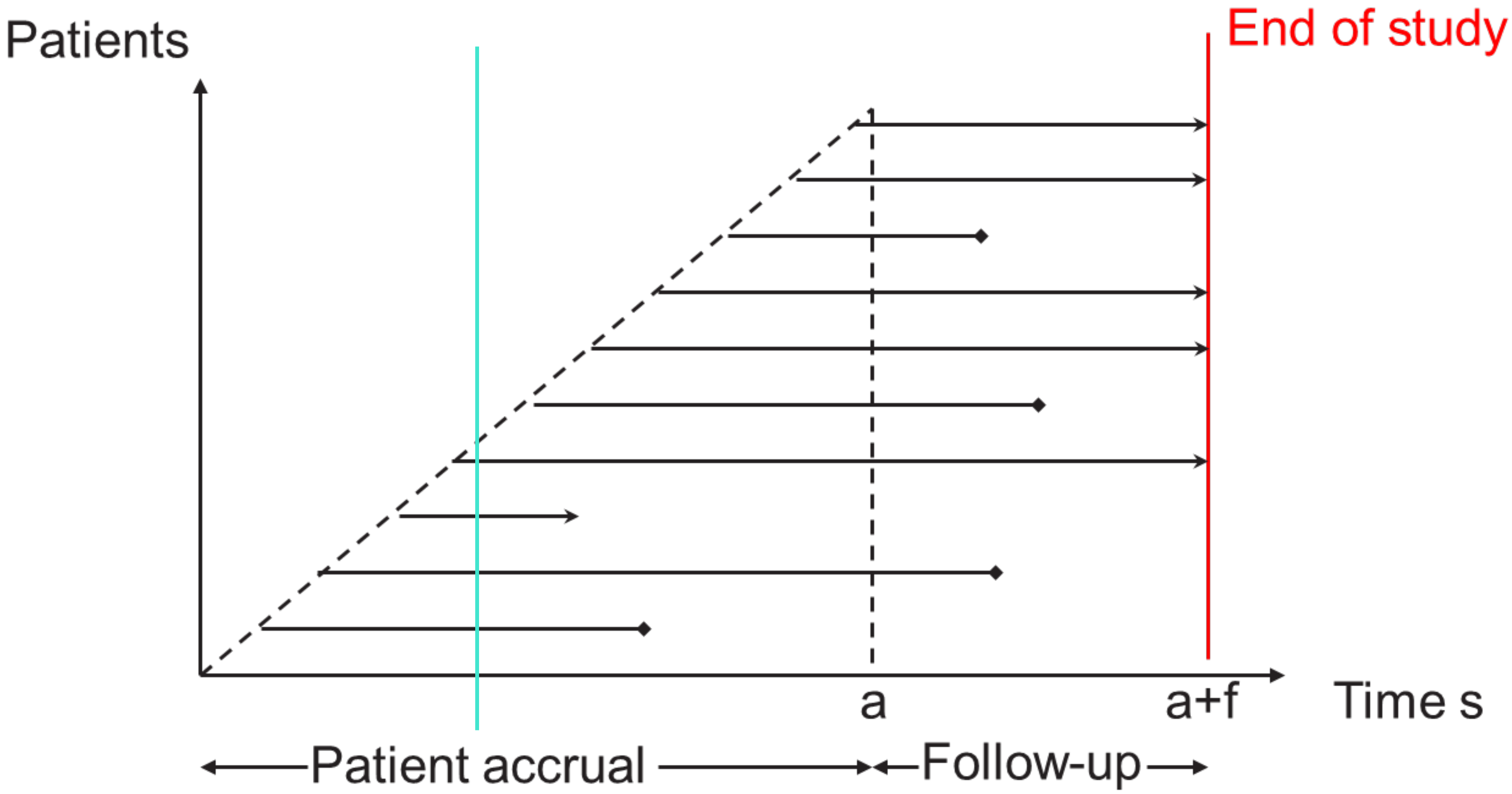
$$Z_K^* = w_1 \cdot \Phi^{-1}(1 - p_1) + w_2 \cdot \Phi^{-1}(1 - p_2)$$

Friede et al, 2011;
Kunz et al, 2015

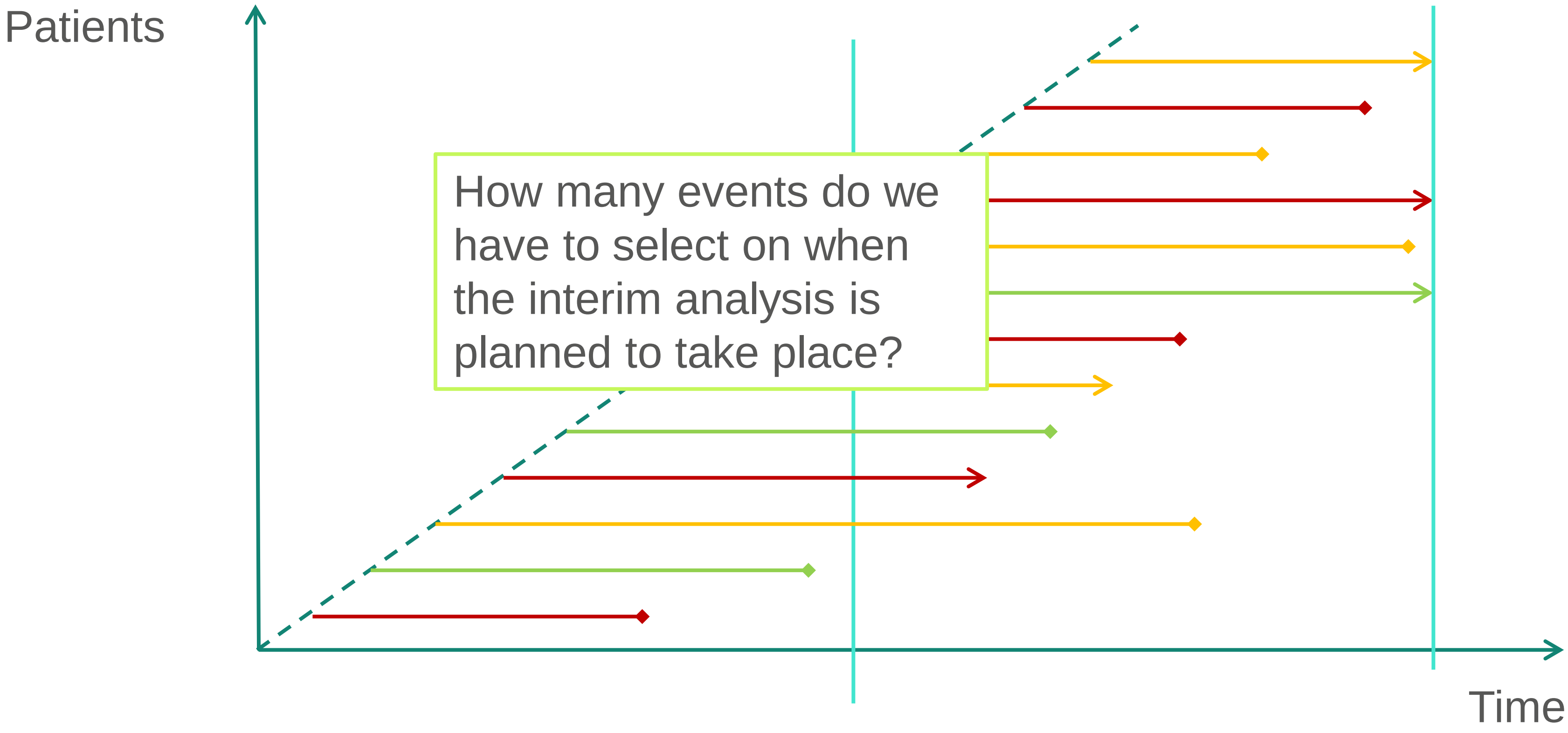
Drawbacks of This Strategy

- 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



Selection with Time to Event Data



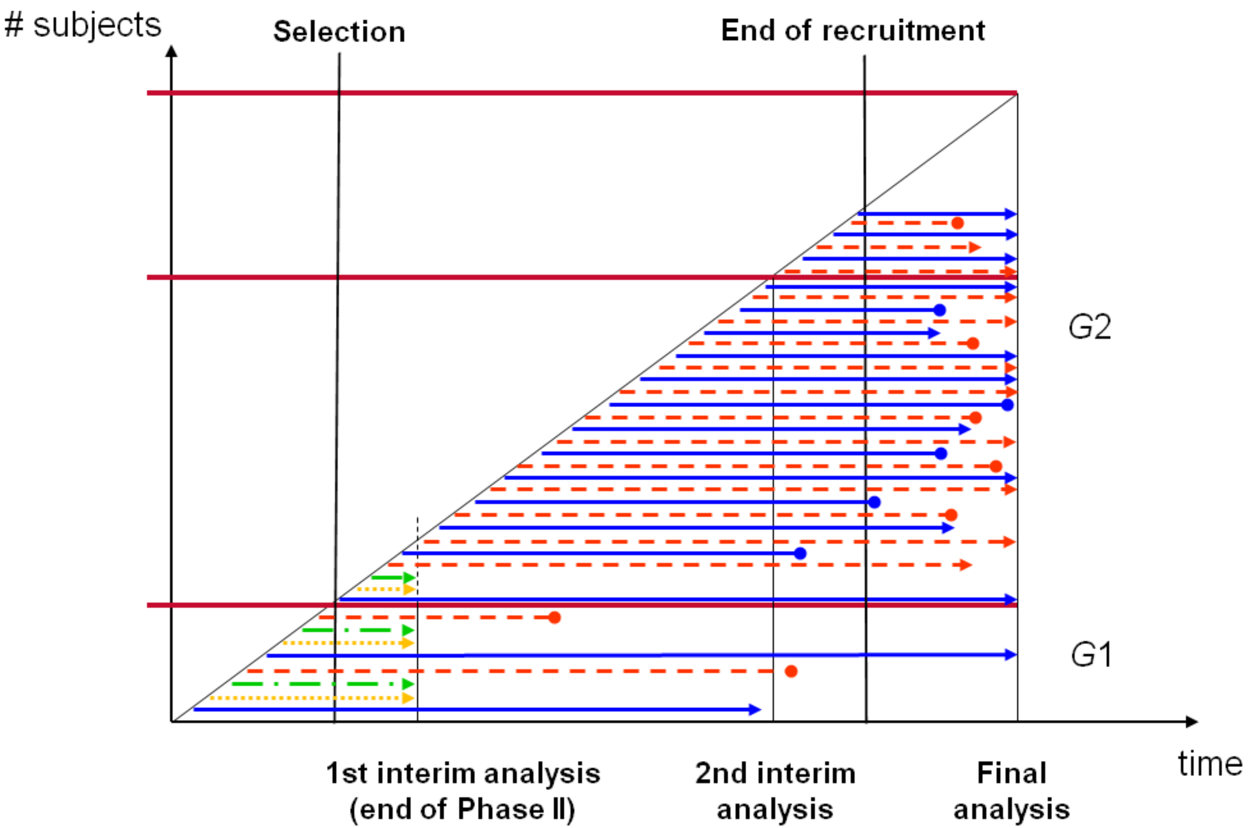
Application to Time To Event Data

– 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

$$w\Phi^{-1}\left(1-\tilde{p}_k^{G1}\right)+\sqrt{1-w^2}\,\Phi^{-1}\left(1-\tilde{p}_k^{G2}\right)$$

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)
2. Magirr et al (2014): Worst case scenario adjustment of critical level
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.
  $\rightarrow$ 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 $w \rightarrow$ often chosen based on assumed population sizes

Simulation of Such Trials

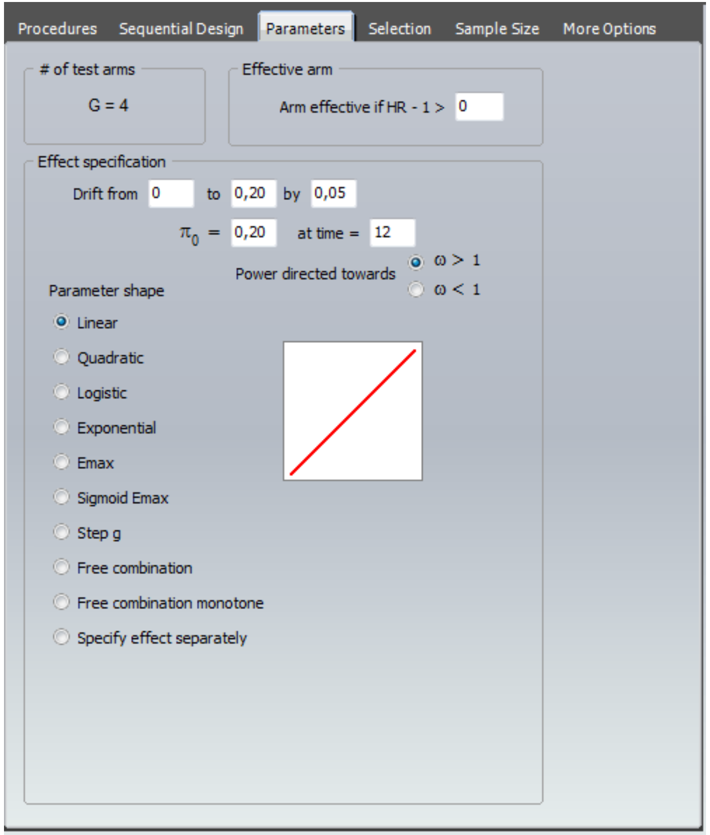
- Simulation, as always, needs a lot of setting which aim at mirroring real life

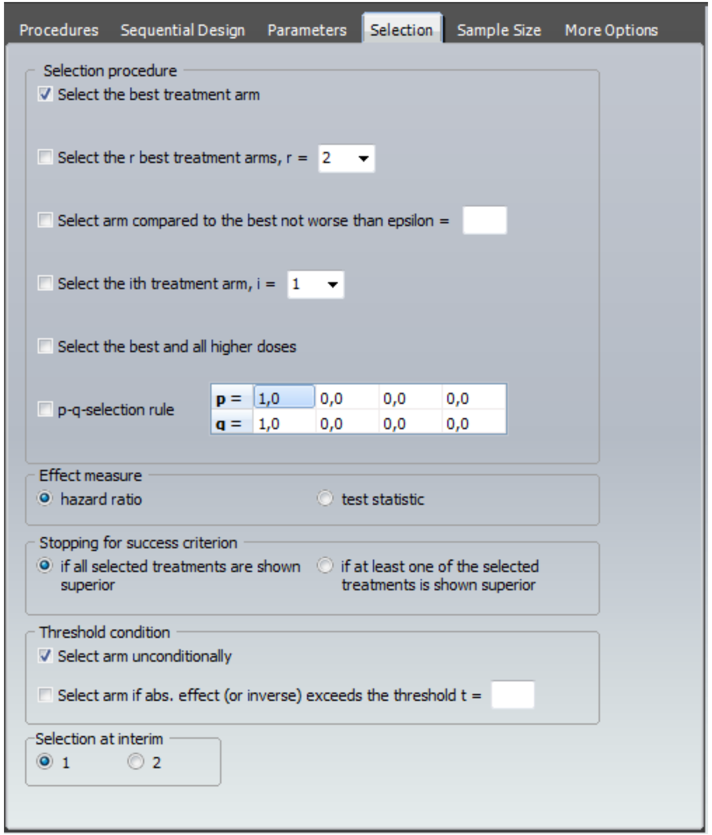
- Such trials are not automatons. Decisions will in reality made by humans, and conduct of the trial is influenced by operational constraints

- Software packages need to provide options which can cover a variety of scenarios, which makes them somewhat generic

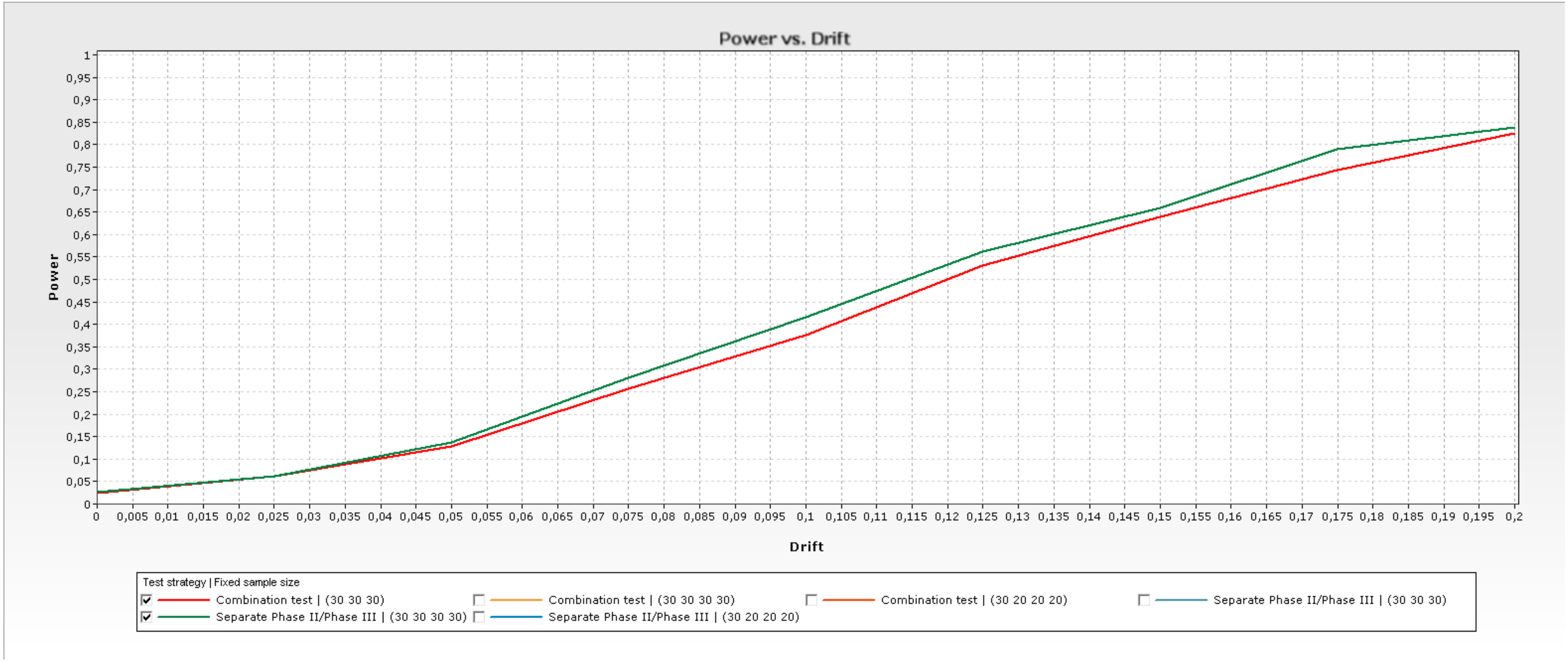
Simulation of Such Trials

ADDPLAN – Multi Armed Time To Event Trials

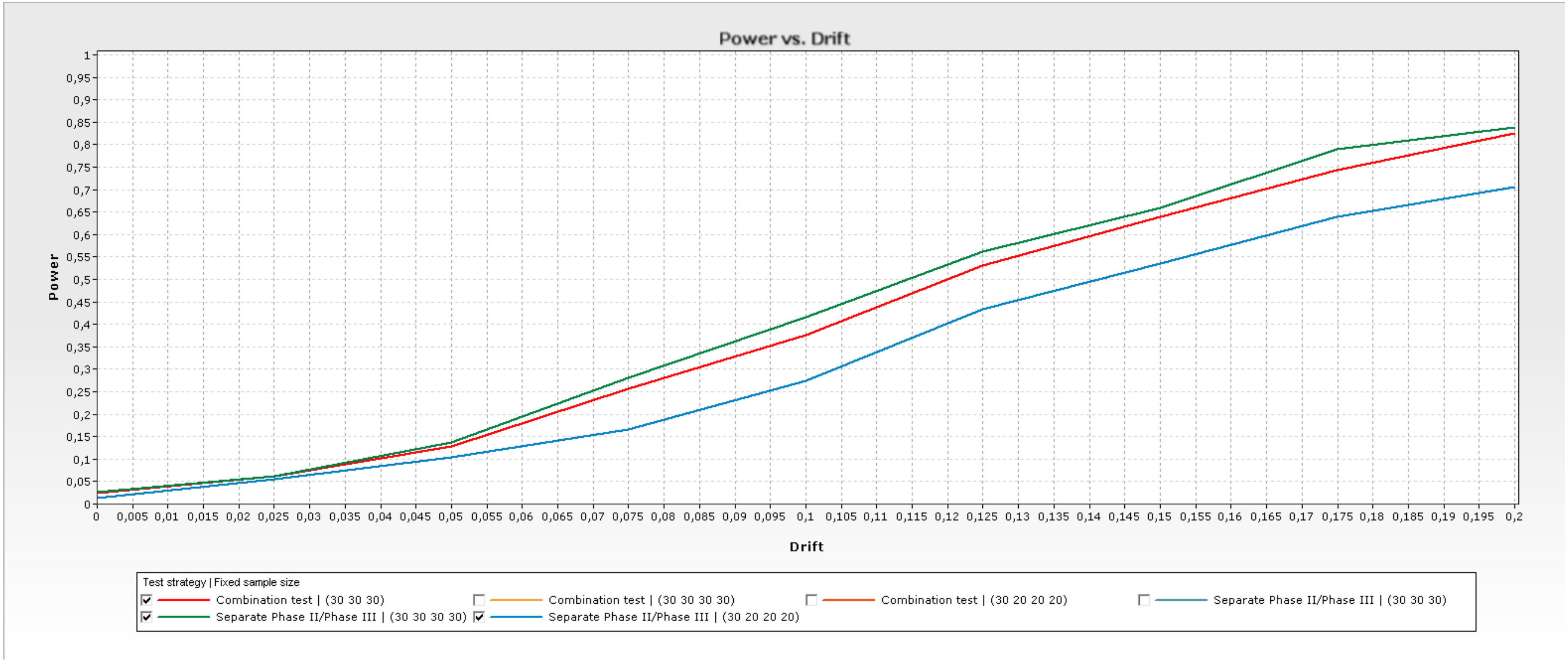




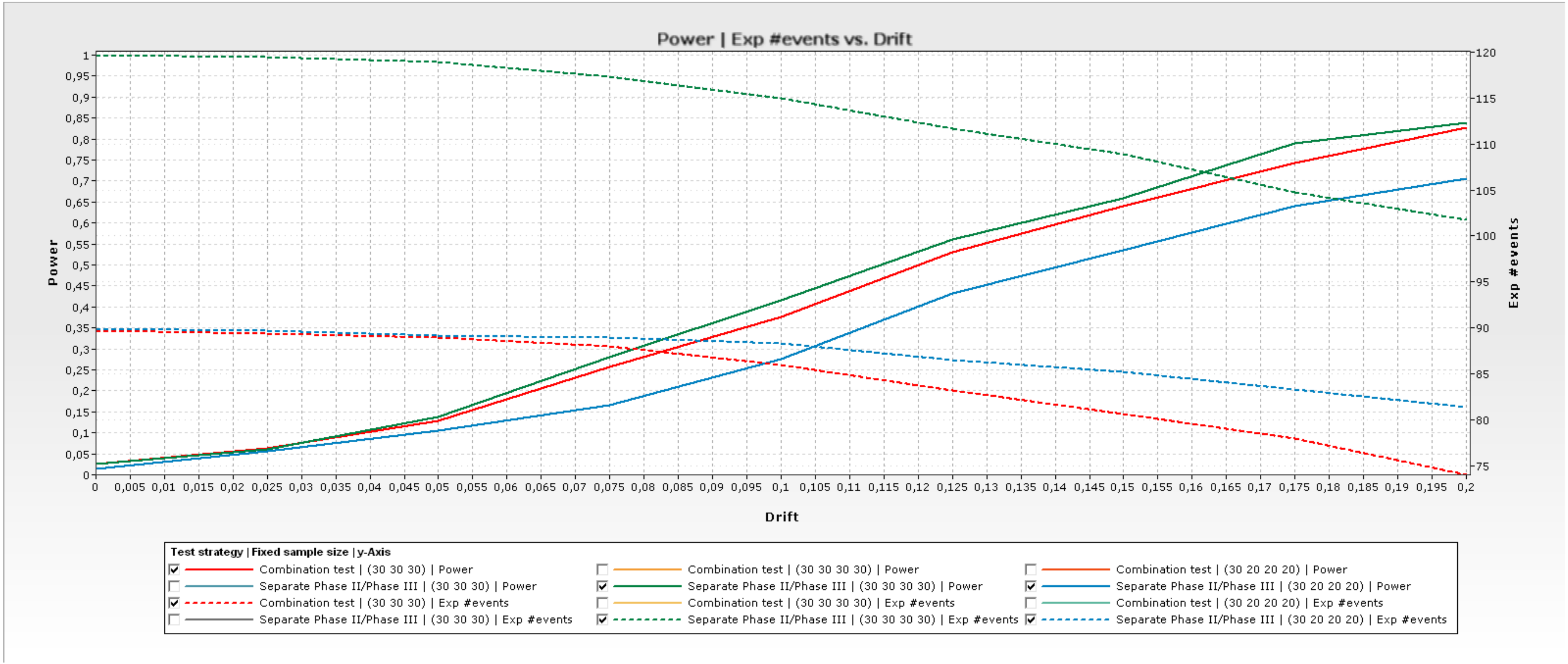
Simple Simulation Result (Seamless Phase II/III vs Separate)



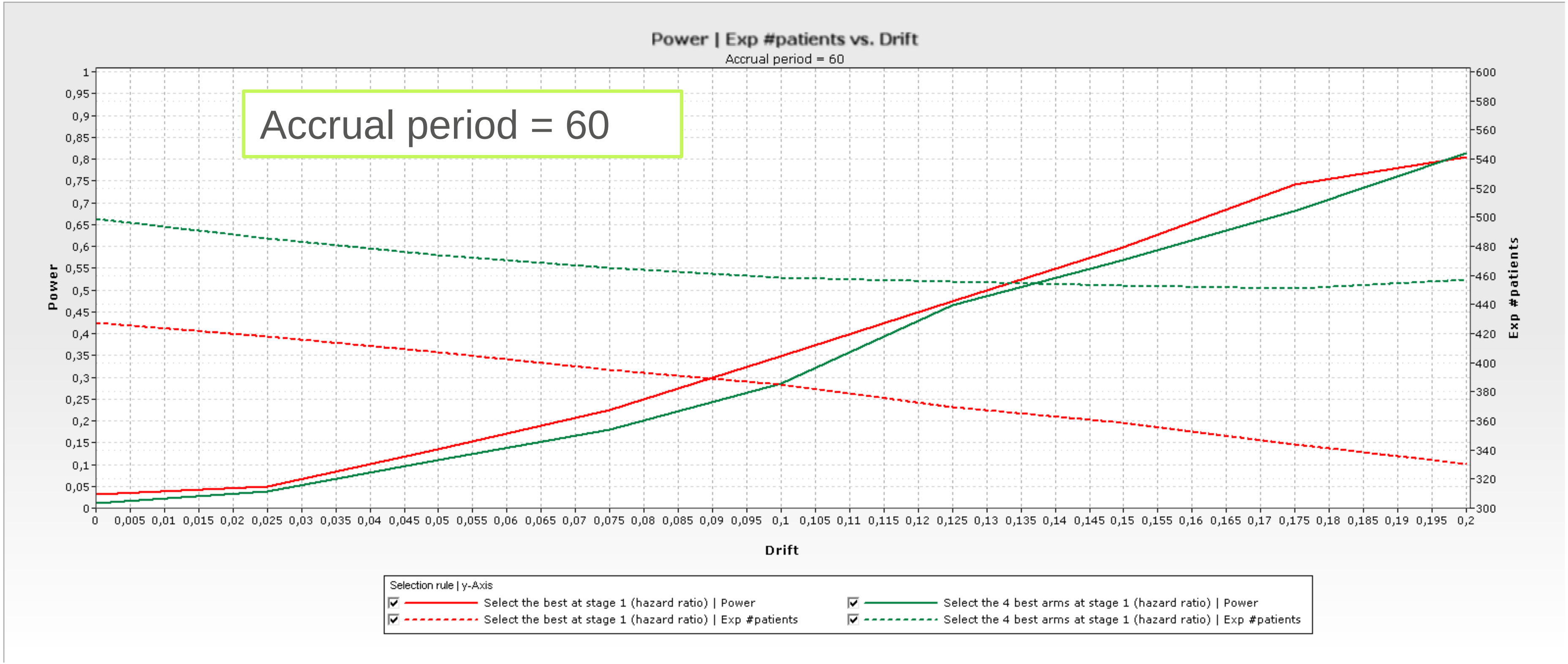
Simple Simulation Result (Seamless Phase II/III vs Separate)



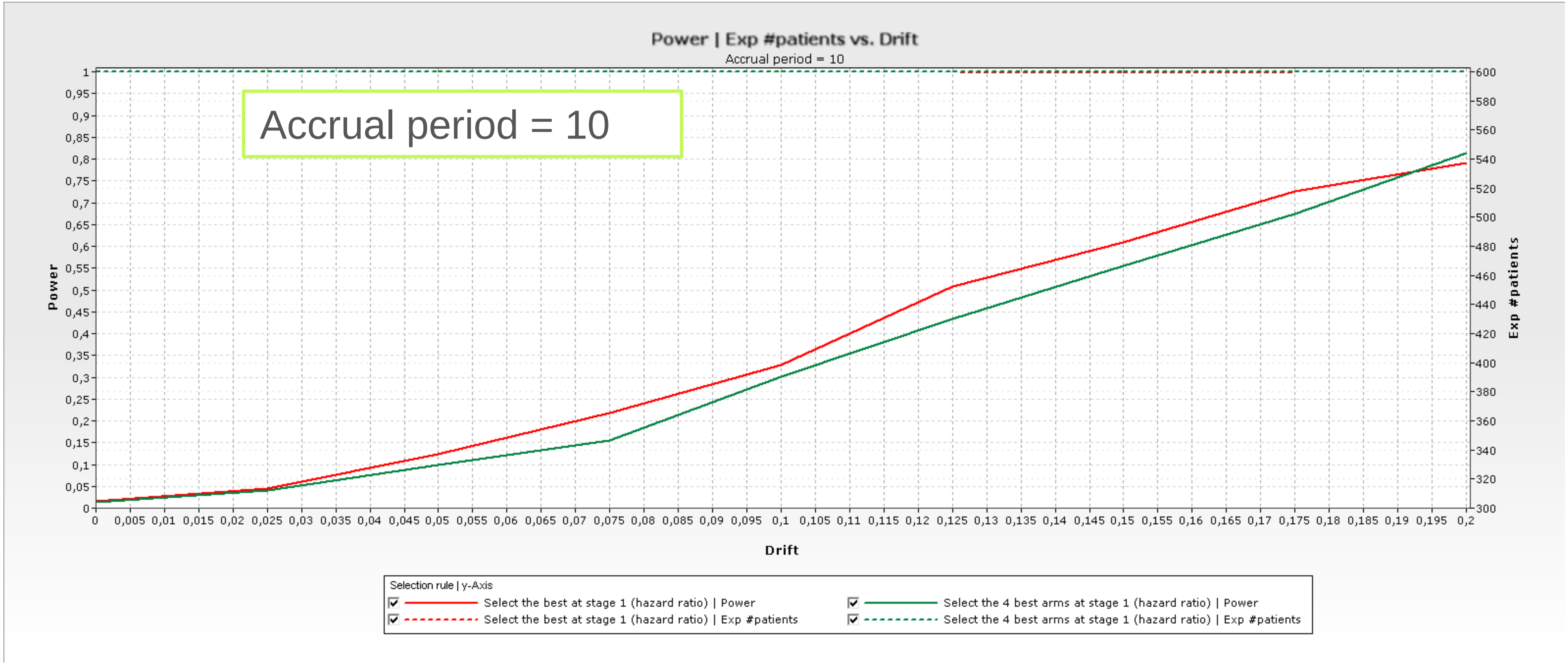
Expected Number of Events (Seamless Phase II/III vs Separate)



Saving Patients by Selection (Seamless Phase II/III)



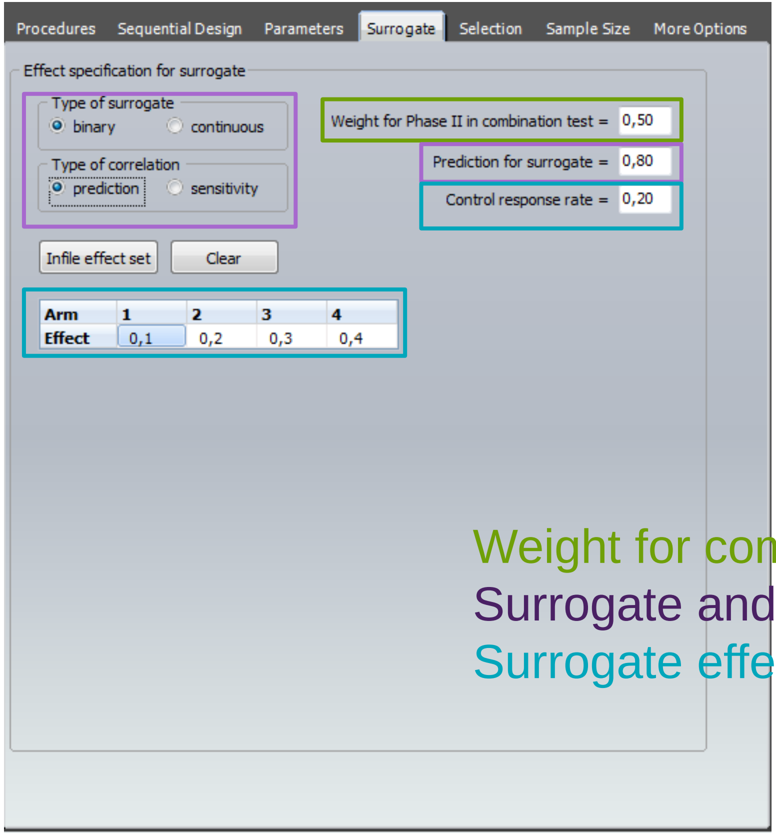
Saving Patients By Selection?

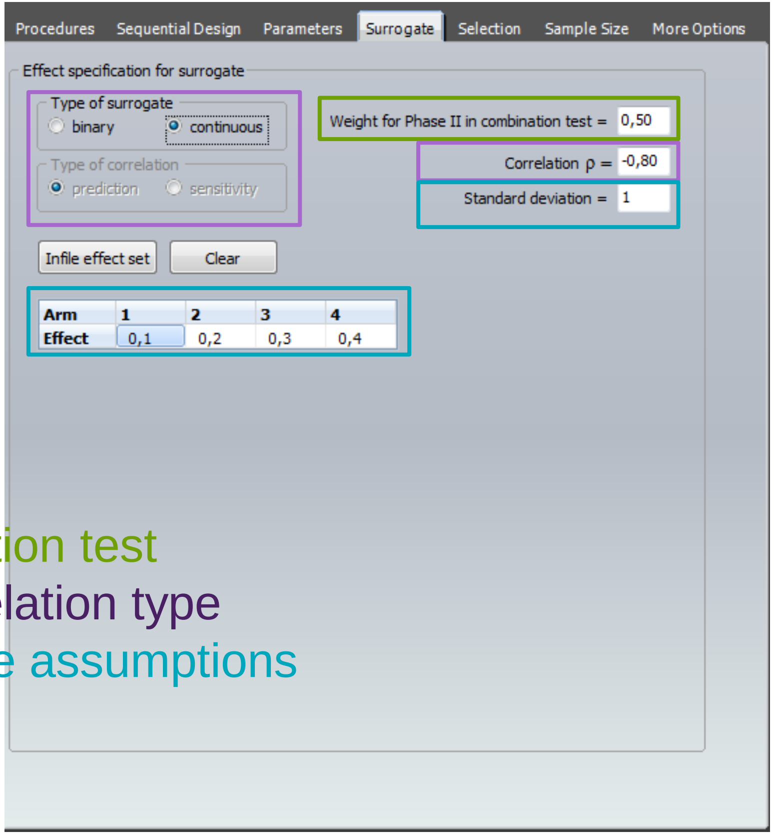


Introducing Surrogate For Selection

– 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



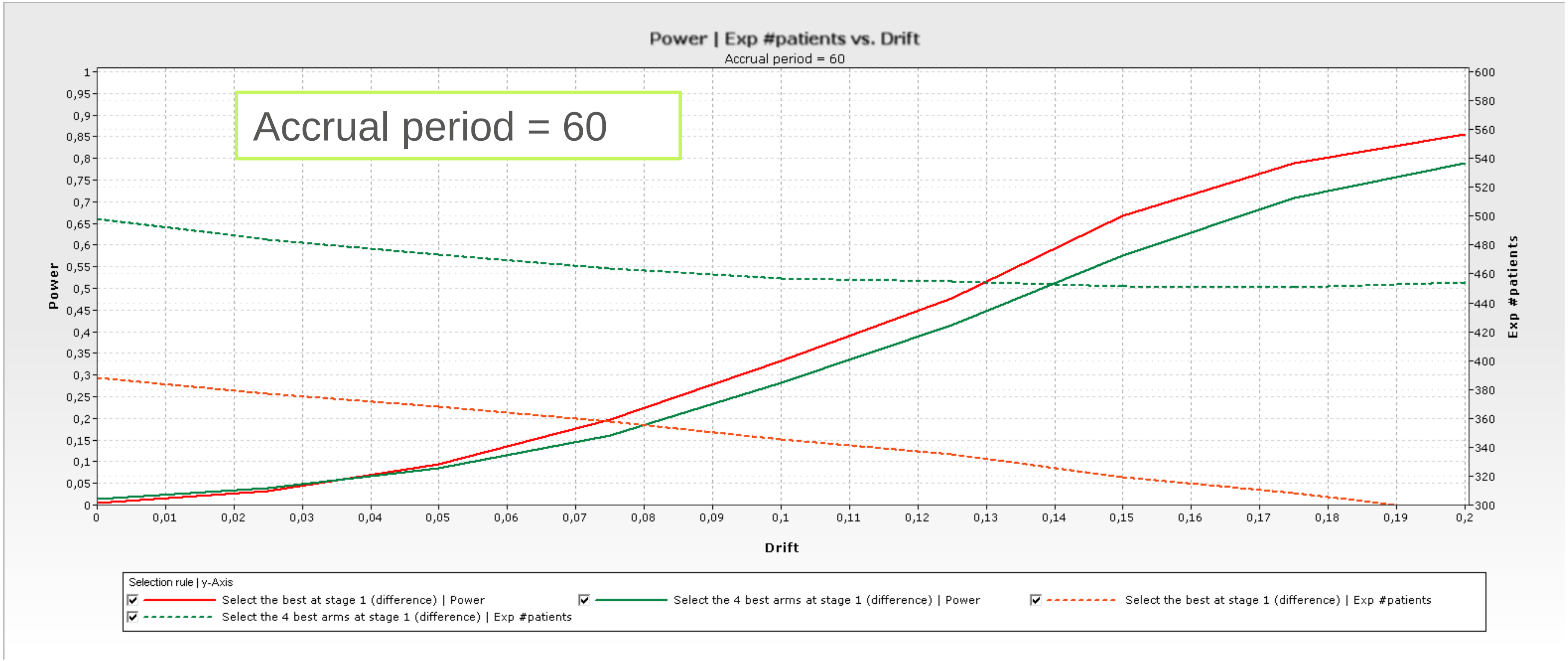


Weight for combination test

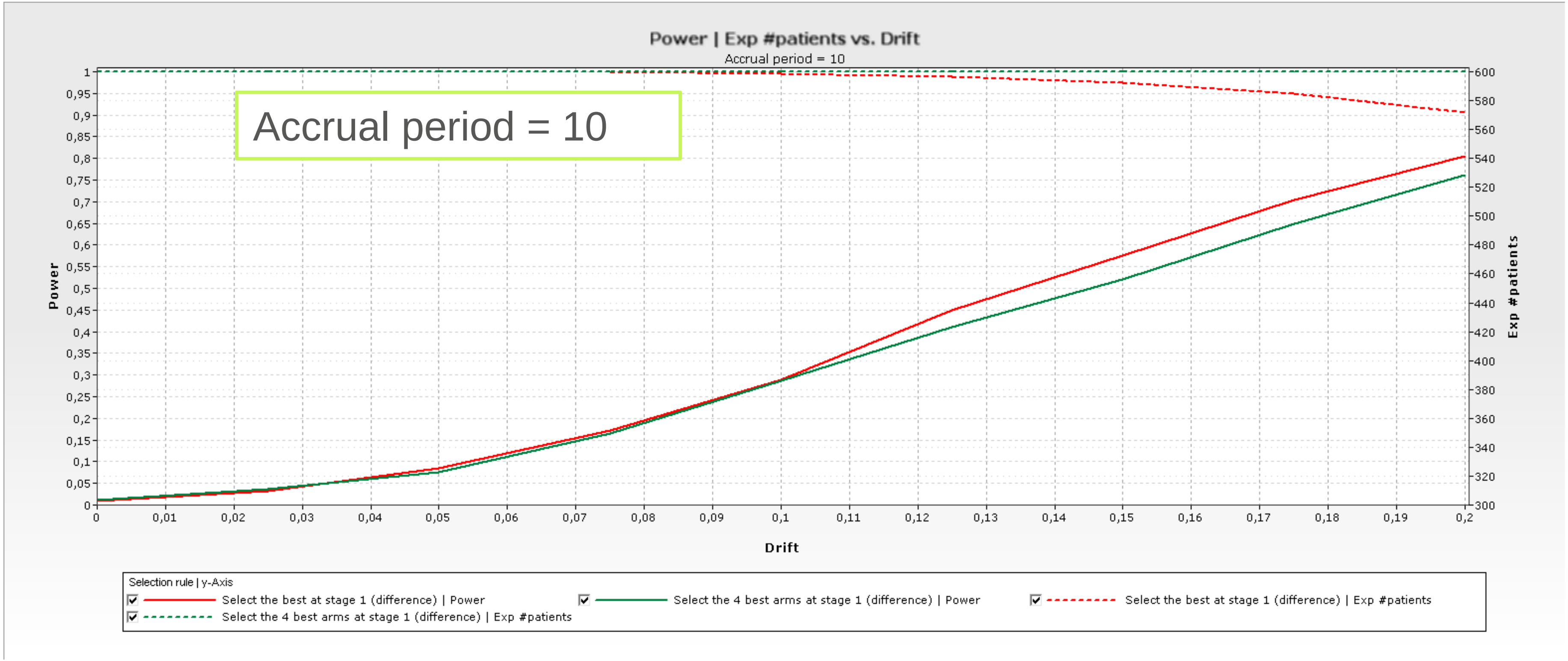
Surrogate and correlation type

Surrogate effect size assumptions

Saving Patients By Selection



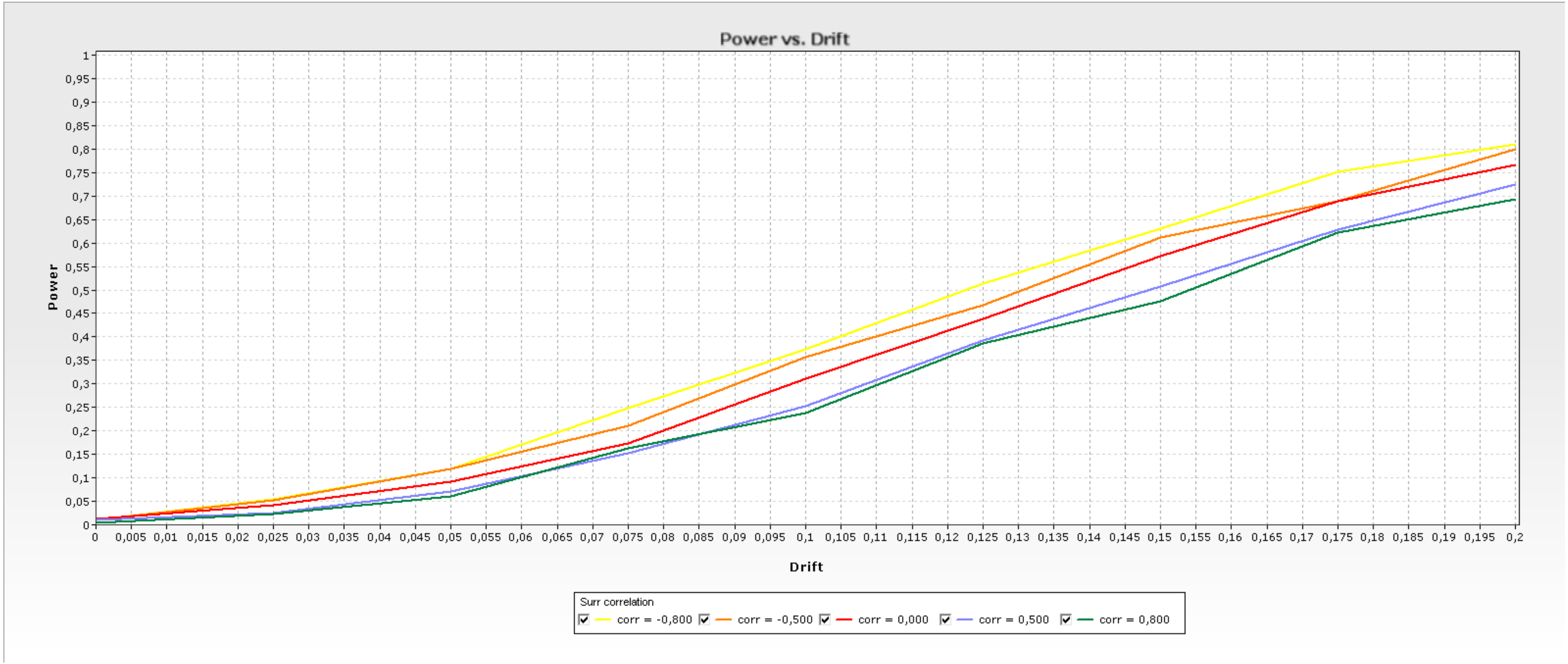
Saving Patients By Selection!



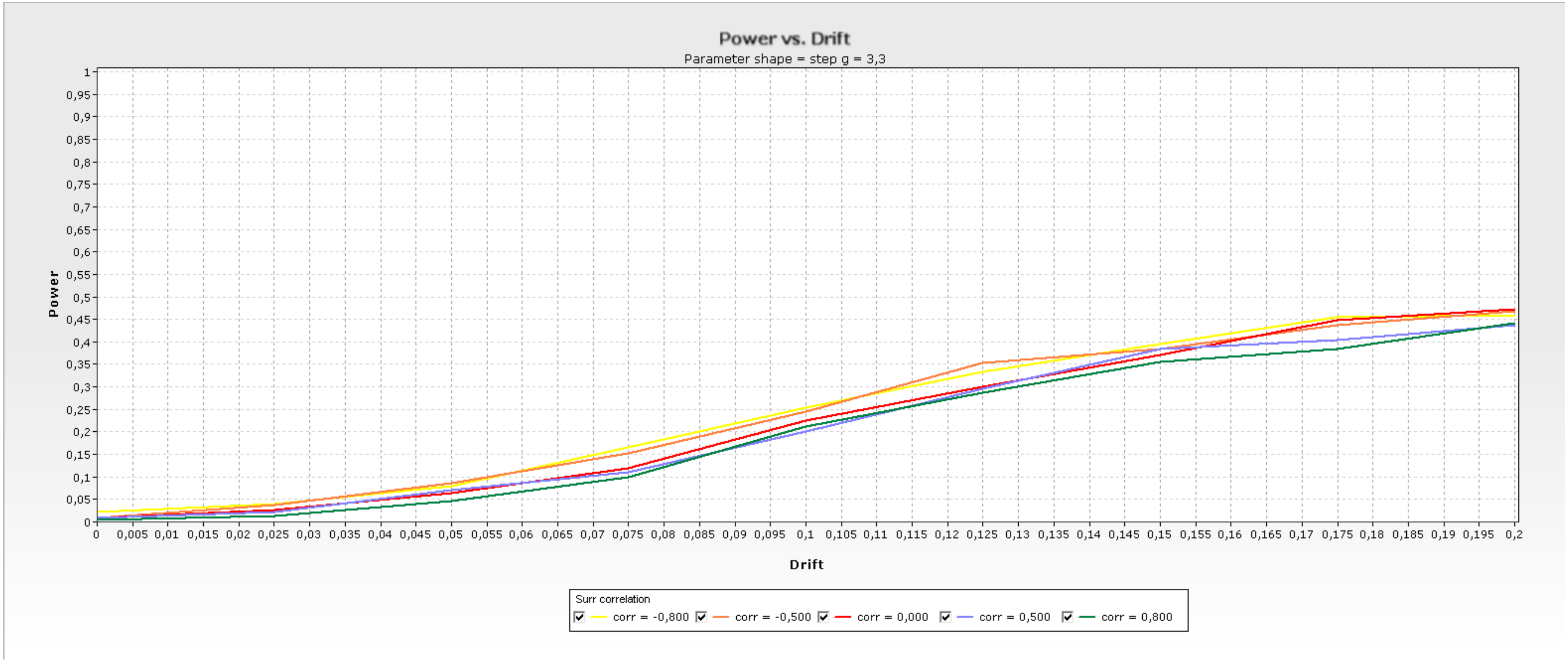
Continuous Surrogate

- 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)

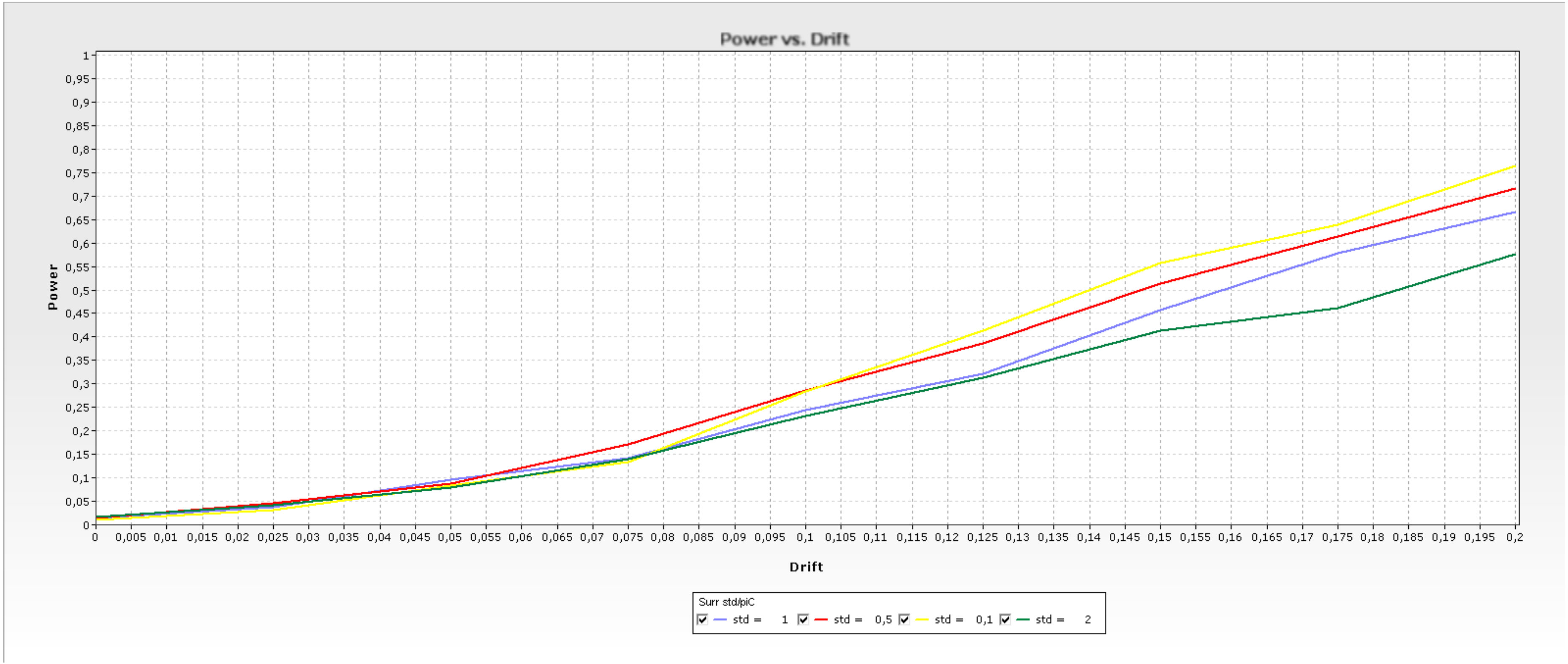
Impact Of The Correlation



… 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

Binary Surrogates – Example (Enrichment Design)

- 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

Control

		Surrogate		
		No	Yes	
Primary	No			0.47
	Yes		0.08	0.53
		0.90	0.10	

Binary Surrogates - Example

Control

		Surrogate		
		No	Yes	
Primary	No	0.45	0.02	0.47
	Yes	0.45	0.08	0.53
		0.90	0.10	

Binary Surrogates - Example

Active in S

		Surrogate		
		No	Yes	
Primary	No			0.30
	Yes		0.40	0.70
		0.50	0.50	

Binary Surrogates - Example

Active in S

		Surrogate		
		No	Yes	
Primary	No	0.20	0.10	0.30
	Yes	0.30	0.40	0.70
		0.50	0.50	

Binary Surrogates - Example

Active in S^c

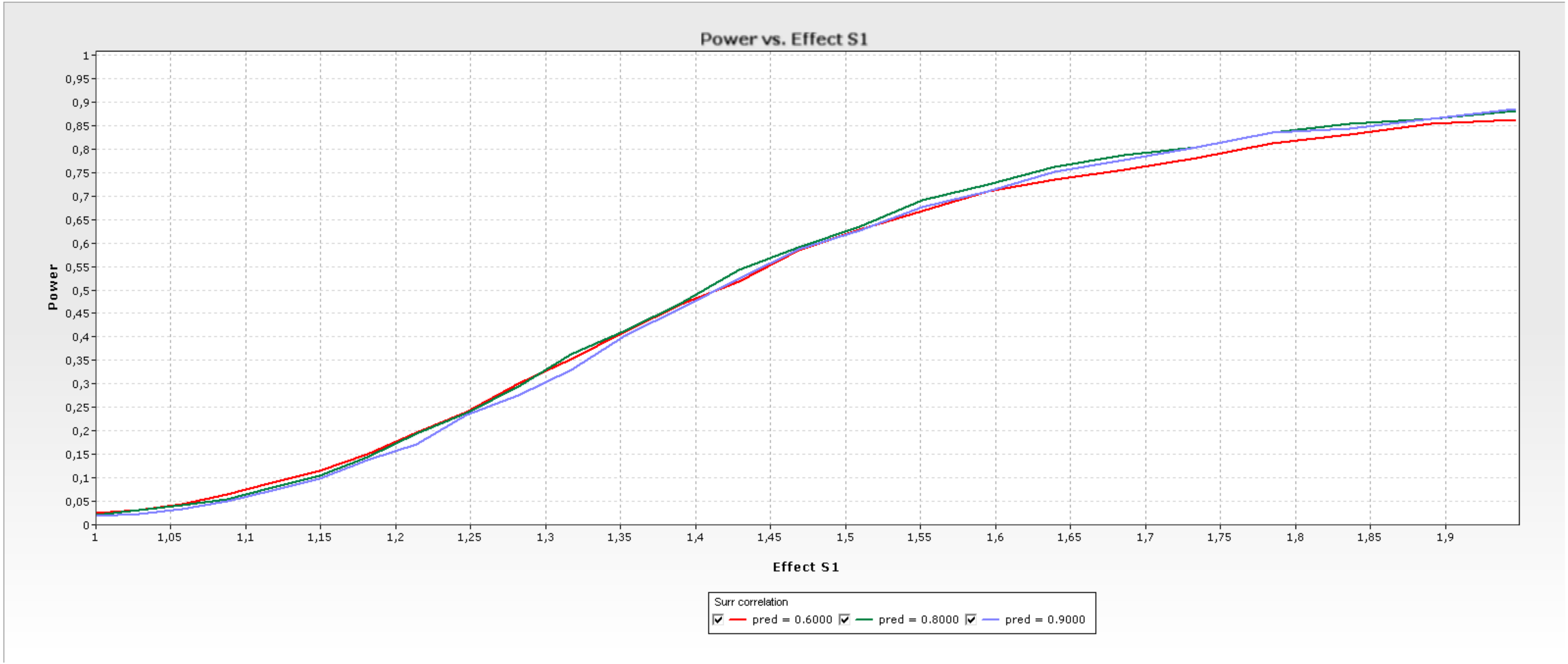
		Surrogate		
		No	Yes	
Primary	No			0.41
	Yes		0.24	0.59
		0.70	0.30	

Binary Surrogates - Example

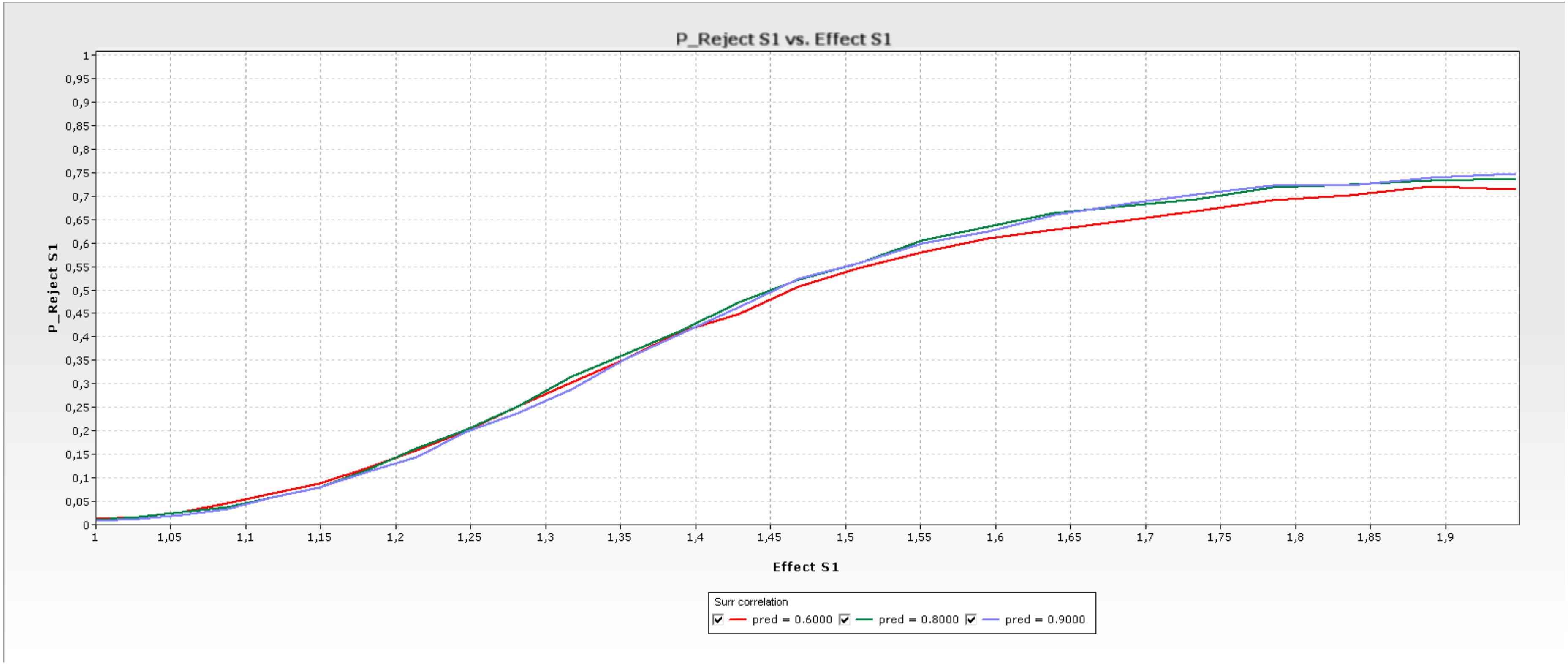
Active in S^c

		Surrogate		
		No	Yes	
Primary	No	0.35	0.06	0.41
	Yes	0.35	0.24	0.59
		0.70	0.30	

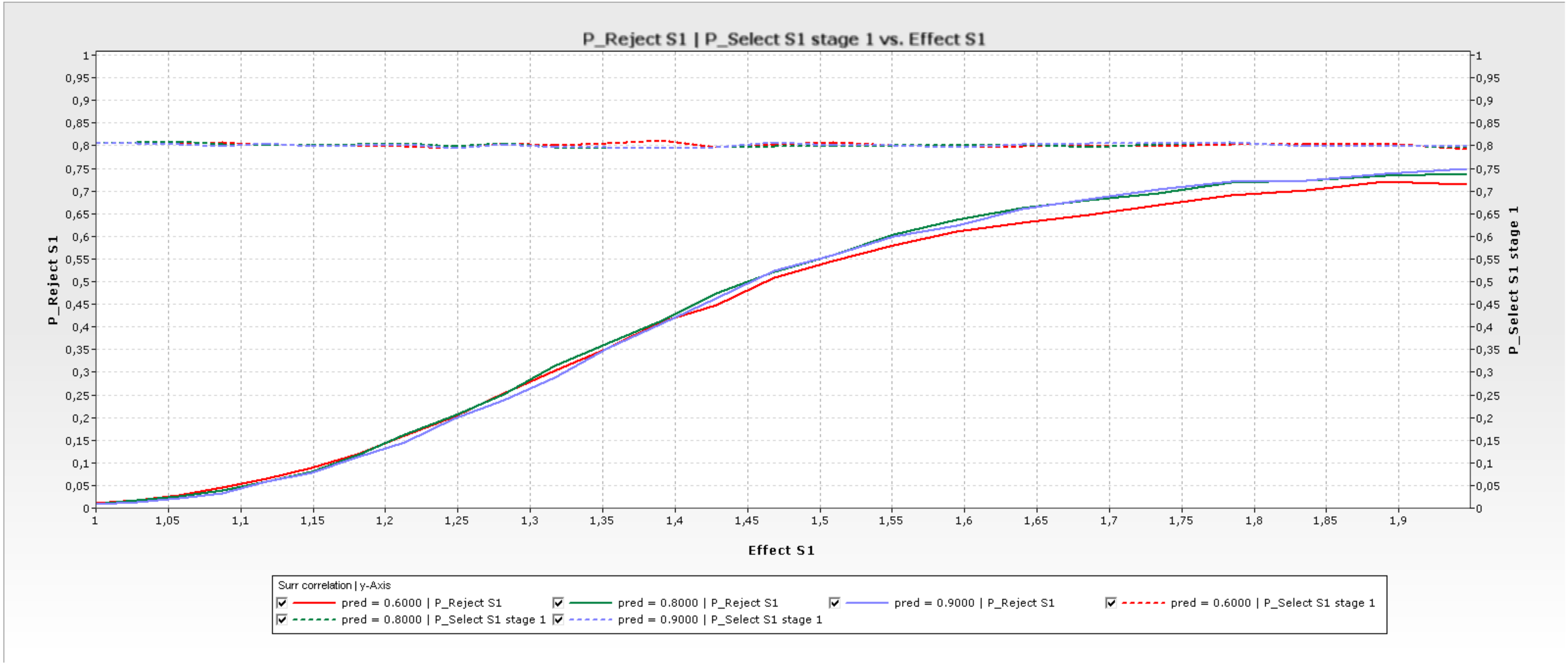
Overall Power



Power for Subgroup



Power for Subgroup



References

- Bauer, P., Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 18, 1833-1848.
- Bauer, P. and Posch, M. (2004). Letter to the Editor: Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in Medicine* 23, 1333–1335.
- Posch M., König, F., Branson M., Brannath W., Dunger-Baldauf C., Bauer P. (2005): Testing and estimation in group sequential designs with adaptive treatment selection. *Statistics in Medicine* 2443: 3697-3714.
- 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.
- Friede, T., Parsons, N., Stallard, N., Todd, S., Valdes Marquez, E., Chataway, J., Nicholas, R. (2011). Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: an application in multiple sclerosis. *Statistics in Medicine* 30: 1528-1540.
- Kunz U., Friede T., Parsons N., Todd S., Stallard S. (2015): A comparison of methods for treatment selection in Seamless Phase II/III clinical trials incorporating information on short-term endpoints. Journal of Biopharmaceutical Statistics 25: 170-189.

References

– Friede, T., Parsons, N., & Stallard, N. (2012). A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine*, *31:* 4309-4320 (correction in Statistics in Medicine 32).

– Irle, S., Schäfer, H. (2012). Interim design modifications in time-to-event studies. *Journal of the American Statistical Association* 107: 341-348.

– Jenkins, M., Stone, A., Jennison, C. (2011). An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics* 10: 347–356.

– Magirr, D., Jaki, T., König, F., Posch, M. (2014). Adaptive designs for time-to-event trials. PLoS ONE **11**, e0146465.

– Mehta, C., Schäfer, H., Daniel, H., Irle, S. (2014). Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. *Statistics in Medicine*, *33*(26), 4515-4531.

– Wassmer, G., Dragalin, V. (2014). Designing issues in confirmatory adaptive population enrichment trials. *J Biopharm Stat* 25: 651-669.

– Wassmer, G. (2006): Planning and analyzing adaptive group sequential survival trials. *Biometrical Journal* 48: 714-729.

Silke.Joergens@iconplc.com

+49 221 5999 233
+49 1761 5999 278

iconplc.com