Introduction to Designing Adaptive Trials: MAMS and Platform Trials

Berry Consultants

Statistical Innovation

Outline

- The Basics
- FACTS implementation
- Advanced Topics
- Regulatory

What is a MAMS Trial?

- It would be nice if the answer was simpler and clearer cut than it is.
- The title is appealing but a bit vague, so MAMS can be taken to mean any trial with multiple stages (a.k.a. interims) and multiple arms.
  - Dose finding
  - Multi-arm phase 3
  - Umbrella trials
  - Platform trials
- BUT if we look at the original paper and the R package the term could be confined to:
  - Frequentist, Confirmatory (Type-1 FWER is controlled)
  - Multi-arm trials with multiple interims (fixed number of arms)
  - Where multiple arms might be declared successful

The start of MAMS Trial?

- The landmark MAMS / Platform Trials are:
  - STAMPEDE testing multiple treatments in prostate cancer. Started as a fixed MAMS trial in 2005 and morphed into a platform trial in 2010.
  - I-SPY 2 testing multiple treatments in neo-adjuvant breast cancer,, platform trial from the outset started in 2010.
- The first MAMS paper & tool was:
  - Patrick Royston & Daniel Bratton & Babak Choodari-Oskooei & Frederike Maria-Sophie Barthel, 2014. "**NSTAGE: Stata module for multi-arm, multi-stage (MAMS) trial designs for time-to-event outcomes**," Statistical Software Components S457931, Boston College Department of Economics, revised 08 Apr 2023. First published 2009.
- Other tools include:
  - The R Package "MAMS" released in 2019. Jaki, T., Pallmann, P., & Magirr, D. (2019). The R Package MAMS for Designing Multi-Arm Multi-Stage Clinical Trials. *Journal of Statistical Software*, *88*(4), 1–25. https://doi.org/10.18637/jss.v088.i04
  - East V6.4 (2016) had a limited MAMS option
  - ADDPLAN Classic 6.1.1 (2014) has a "Multiple Comparison" module that allows the design of MAMS trials.
  - FACTS V7.0 include a Platform Trial simulator.

Using the “MAMS” R package

- A trial to test 5 treatments against control (K=5), with 2 interims (so 3 stages, J=3)
  - One sided alpha 0.025
  - Power 0.9
  - Equal allocation to the treatment arms in each stage (r=1:3), allocate 2x to control compared to any treatment (r0=c(2,4,6))
  - Power the trial for a response to treatment of 1pt and to control of 0pt with and sd of 2. (delta=1, delta0=0, sd=2).
  - This means we have to set to null the default mechanism for defining the expected success which is to specify a target and minimum probability for a patient that their response to the treatment would be better than their response on control (p=NULL, p0=NULL)
  - We set the upper and lower boundaries to O’Brien Fleming (ushape=“obf”, lshape=“obf”)

```
m.obf <- mams(K=5, J=3, alpha=0.025, power=0.9, r=1:3,
r0=c(2,4,6), delta=1, delta0=0, sd=2, p=NULL, p0=NULL,
ushape="obf", lshape="obf")
```

Resulting Design:

- Takes a couple of minutes to complete

```
> m.obf
Design parameters for a 3 stage trial with 5 treatments

                                                Stage 1 Stage 2 Stage 3
Cumulative sample size per stage (control):        60     120     180
Cumulative sample size per stage (active):         30      60      90

Maximum total sample size:  630

             Stage 1 Stage 2 Stage 3
Upper bound:   4.444   3.142   2.566
Lower bound:  -4.444  -3.142   2.566
```

Need to simulate to get OCs

- Call mams.sim() to run the simulations
  - Run 10,000 simulations (nsim=10000)
  - Supply a matrix with the number of subjects in each arm at each stage (nMat=t(m.obf$n * m.obf$rMat))
  - Pass in the upper and lower bounds from the degign (u=m.obf$u, l=m.obf$l)
  - Define the result with a delta and sd, not probability of better response (pv=NULL, delta=c(0, 0.2, 0.4, 0.6, 1), sd=1)
  - Specify the arms to test for significance (ptest=1:5)

```
> m.obf.sim.null <- mams.sim(nsim=10000, nMat=t(m.obf$n * m.obf$rMat), u=m.obf$u,
l=m.obf$l, pv=NULL, deltav=c(0, 0, 0, 0, 0), sd=2, ptest=1:5)

> m.obf.sim.lfc <- mams.sim(nsim=10000, nMat=t(m.obf$n * m.obf$rMat), u=m.obf$u,
l=m.obf$l, pv=NULL, deltav=c(0, 0, 0, 0, 1), sd=2, ptest=1:5)

> m.obf.sim.alt <- mams.sim(nsim=10000, nMat=t(m.obf$n * m.obf$rMat), u=m.obf$u,
l=m.obf$l, pv=NULL, deltav=c(0, 0.2, 0.4, 0.6, 1), sd=2, ptest=1:5)
```

Results

```
> m.obf.sim.null

Simulated error rates based on 10000 simulations

Prop. rejecting at least 1 hypothesis:              0.025

Prop. rejecting first hypothesis (Z_1>Z_2,...,Z_K)   0.005

Prop. rejecting hypotheses 1 or 2 or 3 or 4 or 5:    0.025

Expected sample size:                             629.010

> m.obf.sim.lcf

Simulated error rates based on 10000 simulations

Prop. rejecting at least 1 hypothesis:              0.907

Prop. rejecting first hypothesis (Z_1>Z_2,...,Z_K)   0.000

Prop. rejecting hypotheses 1 or 2 or 3 or 4 or 5:    0.907

Expected sample size:                             519.708

> m.obf.sim.alt

Simulated error rates based on 10000 simulations

Prop. rejecting at least 1 hypothesis:              0.921

Prop. rejecting first hypothesis (Z_1>Z_2,...,Z_K)   0.000

Prop. rejecting hypotheses 1 or 2 or 3 or 4 or 5:    0.921

Expected sample size:                             511.689
```

Observations

- Effectively no early stopping under the Null, (maxN=630, ASNH0=629) (*This is odd, ASNH0 is hardly reduced at all, the package has an odd interpretation of OBF bounds for futility.)
- Type 1 error well controlled
- Power is ~0.90 as required
- Sample size under this alternate is 520 c/w max of 630 (*This too is odd, this is quite a reduction c/w Null, stopping whole study when one arm successful? Having one successful arm increasing type-1 error in other arms?
- Normal 2 arm GS would have maximum N of 180, ASN0 of 114.5 and ASNH1 of 140. To run 5 of these would have max N 900, ASN0 of 572.5, ASNH1 700 (all effective)

Limitations of the package

- OK computing boundaries & sample size for 3 stages (~2 minutes).
- To compute boundaries & sample size for 4 stages takes ~2 hours.
- To compute boundaries for 5 stage will take days. The documentation does warn of this. EAST uses an approximation that avoids this problem.
- The boundary options are very limited: O'Brien Fleming, Triangular and Pocock, user specified boundaries use a formulation that I find unusual and awkward to work with: supply the ratio of the test statistic at the different boundaries e.g.
  - `ushape = function(x) return(x:1)`
    `Stage 1 Stage 2 Stage 3`
  - `Upper bound: 6.125 4.084 2.042`
- Would have preferred to be able to specify Alpha and Beta spend
- No individual simulation results, and v limited OCs – in particular: no probability to reject H0 by arm & and no break down of probability of stopping at each interim by arm.

Usual Complaint

- As is all too often in Group Sequential stats packages the s/w is misleading about the expected sample size because it fails to ask for time to endpoint and accrual rate and doesn't even warn of the problem.

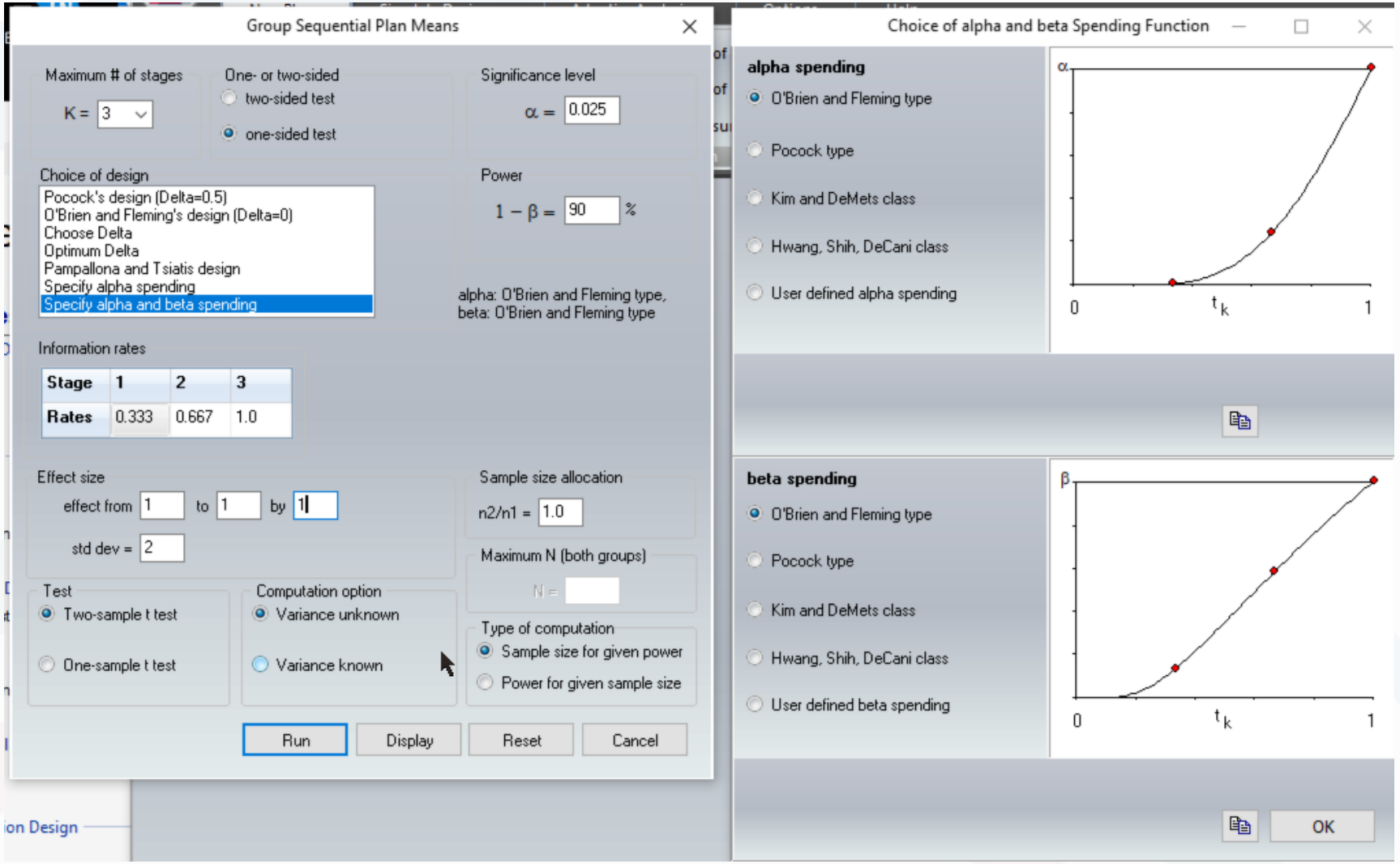
ADDPLAN

- As the MAMS paper that accompanies the MAMS R package, ADDPLNA has an “MC” module that will allow us to design mult-arm, multi-stage trials.
- How does it compare?

Sample Size

- ADDPLAN doesn’t provide a sample size calculation for the Multiple Comparison case (if we select arms at an interim – which the MC module allows - there is no closed form sample size calculation).
- So we start by sample sizing the 2 arm case…

Simple GS for initial sample size



3 stages
Alpha 0.025
Power 0.9
Alpha and beta spending stopping, using “OBF like” boundaries.
Equal allocation
Just look sample size for effect size of 1 with SD of 2.

Resulting Design and sample size

	Plan 1 Means
alpha	0.025
Futility stops	595, 1.002)
tails	1
K	3
Design	alpha: O/F type, beta: O/F
Information rates	equal
Hypothesis	diff<=0
Parameters	diff=1 std=2
Power %	90.0
Total ASN H0	114.5
Total ASN H01	147.7
Total ASN H1	139.9
Total maximum N	180.2
Allocation	1

Plan 1
Two-Sample t Test (One-Sided)
Null hypothesis H0: mu2 - mu1 <= 0
Non-binding stopping for futility bounds as specified in the table below.

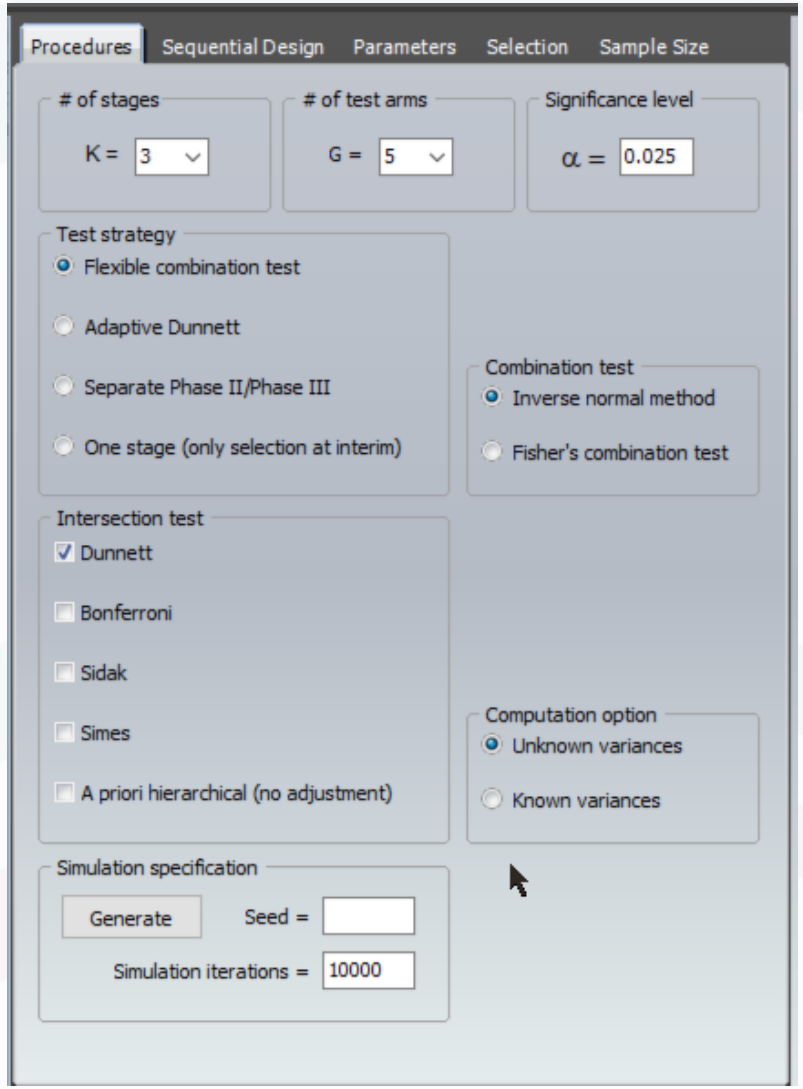
A design with a maximum of K = 3 stages was chosen.
The critical values and the test characteristics of the group sequential test design were calculated for an O'Brien and Fleming type alpha spending function, and an O'Brien and Fleming type beta spending function.
For specified alpha = 0.025, effect mu2 - mu1 = 1 and standard deviation sigma = 2 the power (1 - beta) is 90.0% if the test stages consist of the sample sizes given in the last two columns of the table.
The computation assumes an allocation ratio (n2/n1) = 1.0.
This yields a total of 90.1 + 90.1 = 180.2 observations.
For comparison, the sample size in a fixed sample size design is n1 = 85.0, n2 = 85.0.
Thus, the maximum sample size in the group sequential test design is 1.059 times the sample size in a fixed sample size design.
The expected (average) total sample size under the alternative hypothesis is 139.9, under a value midway between H0 and H1 it is 147.7, and under the null hypothesis it is 114.5.

Inform. rate	bounds accept H0	bounds reject H0	sign.level one-sided	alpha spent	beta spent	power achieved	stage n1	sizes n2
0.333	-0.695	3.710	0.0001	0.0001	0.0044	0.0372	30.0	30.0
0.667	1.002	2.511	0.0060	0.0060	0.0440	0.5845	30.0	30.0
1.0	1.993	1.993	0.0231	0.0250	0.1000	0.9000	30.0	30.0

The “OBF – like” alpha and beta bounds are quite different from MAMS.

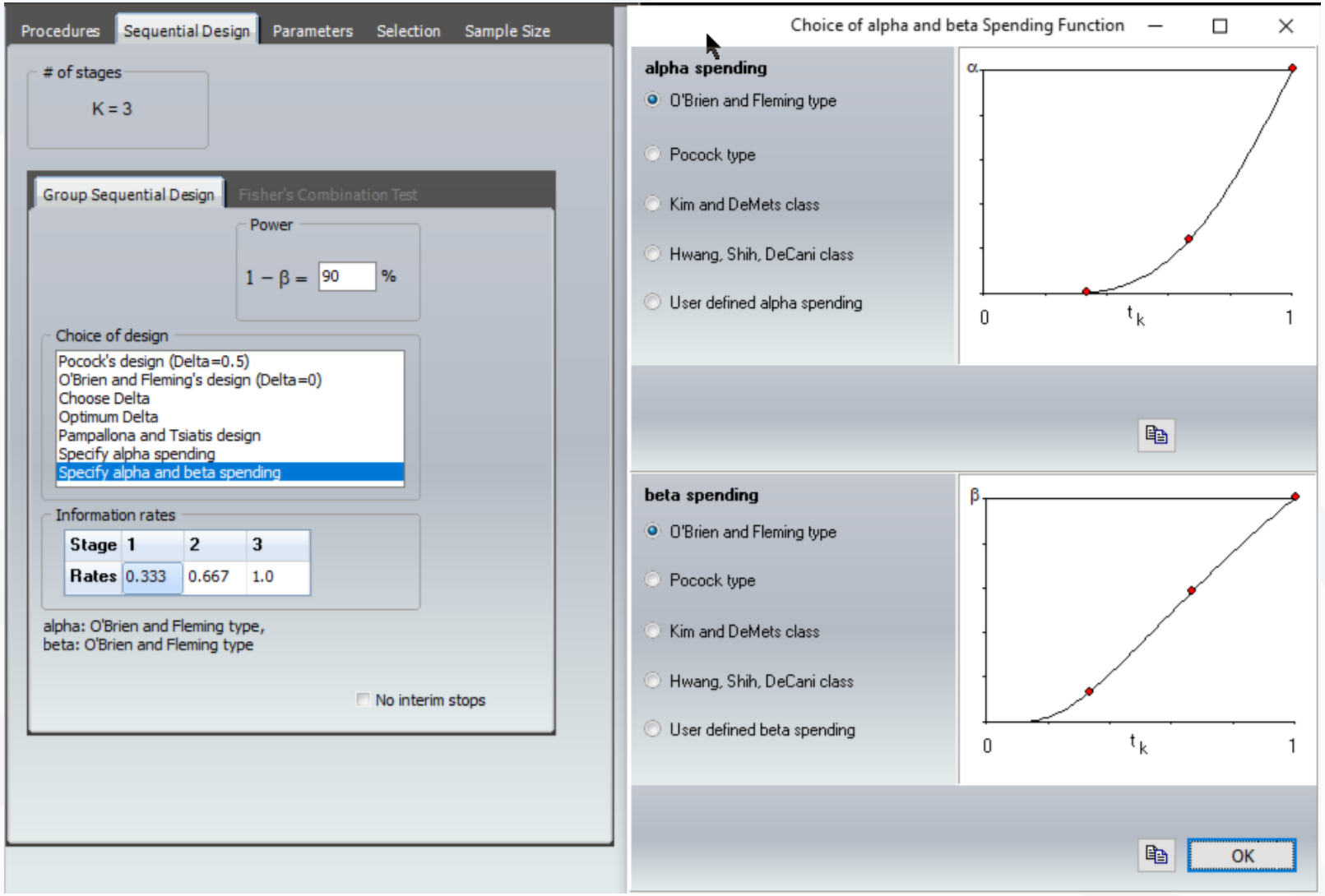
Sample size is 30 per arm per stage

Now we can enter our initial MC design



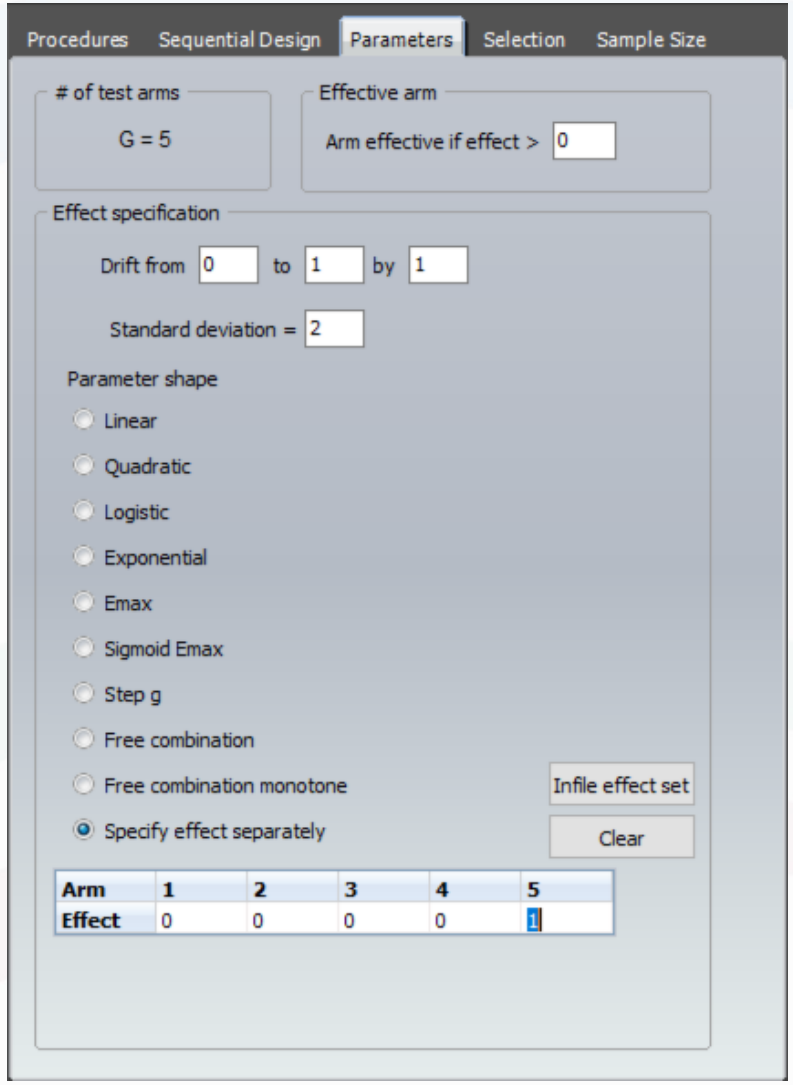
3 stages
5 treatment arms
Alpha 0.025
Use Flexible combination test using inverse Normal across the stages
Use Dunnett adjustment for multiplicity from multiple arms.
Unknown variance

Boundaries



We retain the "OBF like" Alpha and Beta spending boundaries.

Simulation parameters

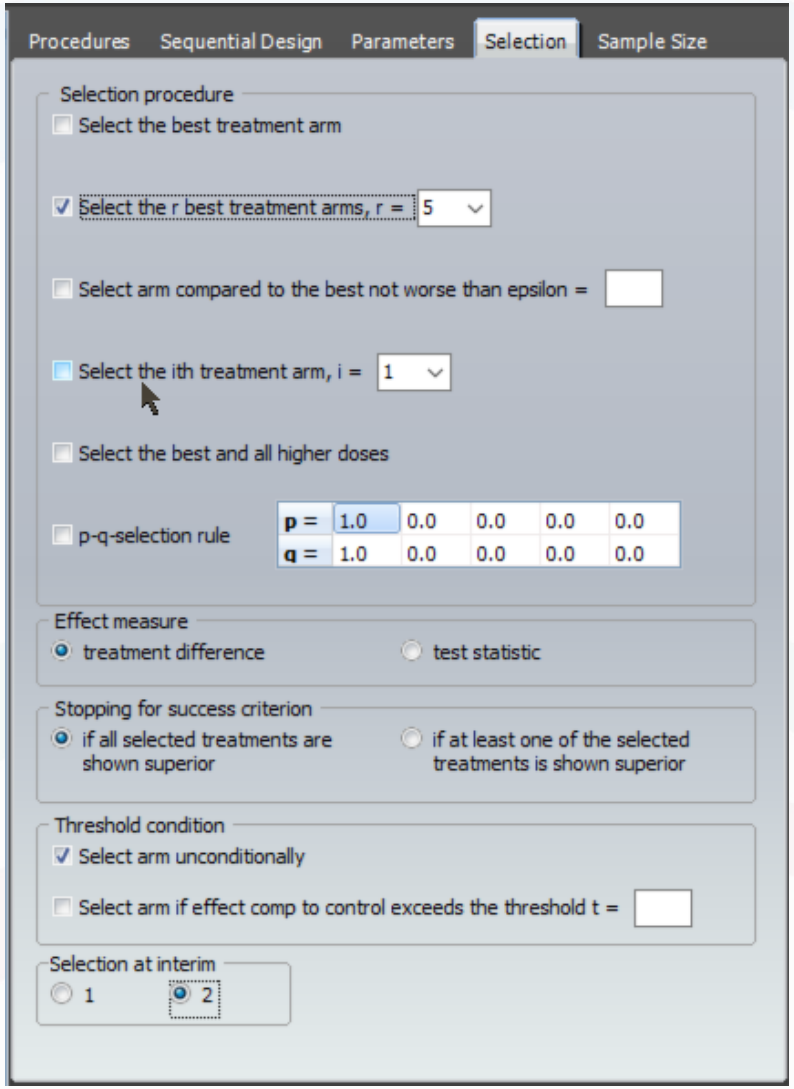


Rather than use a model, we specify the response to simulate per arm, and variously specify:

0 0 0 0 0
0 0 0 0 1
0 0.2 0.4 0.6 1

As our Null, LFC and Alt scenarios

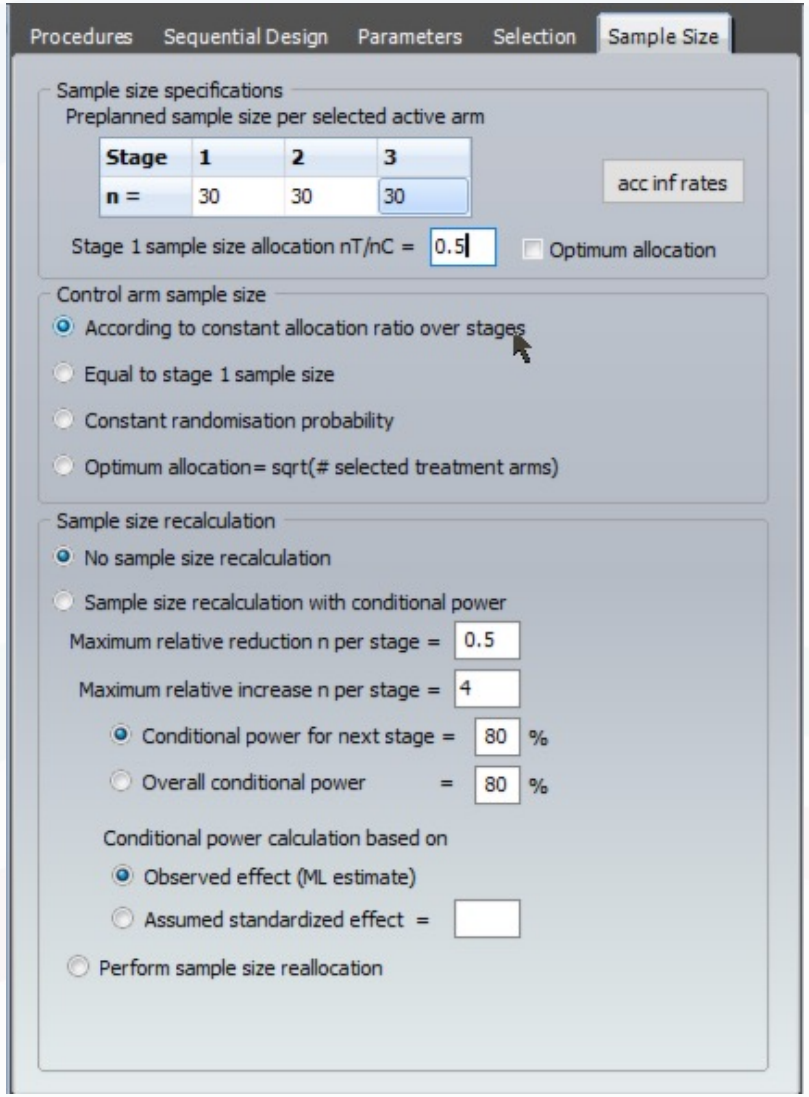
Selection Rule



We don't want to select arms, only stop them early individually for success or futility.

So we simply set the "select the best" rule to select the 5 best, i.e. all arms.

Sample size



Set the sample size to 30 per active arm per stage.
Set the treatment/control allocation ratio to 0.5
So we get 2:1:1:1:1:1 allocation
No sample size re-assessment

Results – power.

Max size stage 3	P_RejectatleastOne	P_RejectEffective
30	0.009	0.000
30	0.737	0.737
30	0.813	0.813
45	0.921	0.921
40	0.879	0.879
44	0.918	0.918
43	0.905	0.905

Top line is Null
Second Line is LFC “Least Favourable Configuration” (Dunnett) one arm has a response of the alternate (1 in our case) and all the others are Null.
Third line is mixed alternate (0, 0.2, 0.4, 0.6, 1)
Type-1 error is over controlled at 0.009.
Power in LFC is 0.737.

Try simulating LFC at sample sizes of 45, 40, 44, 43.

At 43 per active arm, per stage we get power of ~0.9

Simulating at new sample size

	P RejectatleastOne	P RejectEffective	P RejectIneffective
1	0.009	0.000	0.009
2	0.907	0.907	0.015
3	0.950	0.950	0.013
4	0.999	0.999	0.025

Scenarios are: Null, LFC, Alt, All effective but 1.

Power in the LFC is ~0.9

Type-1 error of the 1 Null arm is 0.025

	P Reject arm 2	P Reject arm 3	P Reject arm 4	P Reject arm 5
1	0.002	0.003	0.002	0.002
2	0.004	0.004	0.003	0.907
3	0.064	0.228	0.511	0.941
4	0.977	0.980	0.980	0.978

The closed testing procedure has highest FWER and highest power per arm when most arms are successful.

Frequentist borrowing?

	P Stop stage 1	P Stop stage 2	P Stop stage 3
1	0.245	0.591	0.164
2	0.014	0.564	0.422
3	0.002	0.167	0.831
4	0.000	0.421	0.578

	P Futility stage 1	P Futility stage 2	Total ASN
1	0.245	0.590	576.8
2	0.009	0.040	701.4
3	0.002	0.015	766.5
4	0.000	0.000	677.9

We can see that the procedure is not as powerful as the MAMS procedure and has required larger sample sizes (smaller in the Null, but only because of the odd interpretation of OBF futility boundary by the MAMS package). And higher MaxN of 803

Simulating in FACTS

- Can we import the MAMS design into FACTS and get a better understanding of it?
- Initially we use FACTS Core…
- We need to translate the boundaries reported as test statistic, into p-values:

```
> 1-pt(4.444, 29)
[1] 5.925855e-05
> 1-pt(3.142, 29)
[1] 0.001923288
> 1-pt(2.566, 29)
[1] 0.007859529
```

FACTS Core simulation

- Study:
  - Adaptive Design
  - Continuous recruitment
  - Max subjects: 630
  - Higher response is subject improvement
  - Time to endpoint 0.1 weeks
- Treatment arms: control + 5 treatment arms
- Virtual Subject Response: Null, LFC, Alt
- Execution:
  - median accrual 5 per week,
  - no dropouts

- QOIs:
  - Bayesian Posterior $Pr(\theta_d > \theta_{control})$
  - P-value LOCF, unadjusted
  - Decision P-value at min p-value
  - $Pr(\theta_d > \theta_{control})$ at max prob
- Design:
  - Independent Dose Model
  - Bayes prior for all doses N(0,10)
  - Prior for sigma IG(2, 1)
- Allocation 2:1:1:1:1:1
- Interims at 210 and 420 opportunity to complete

Stopping thresholds from MAMS

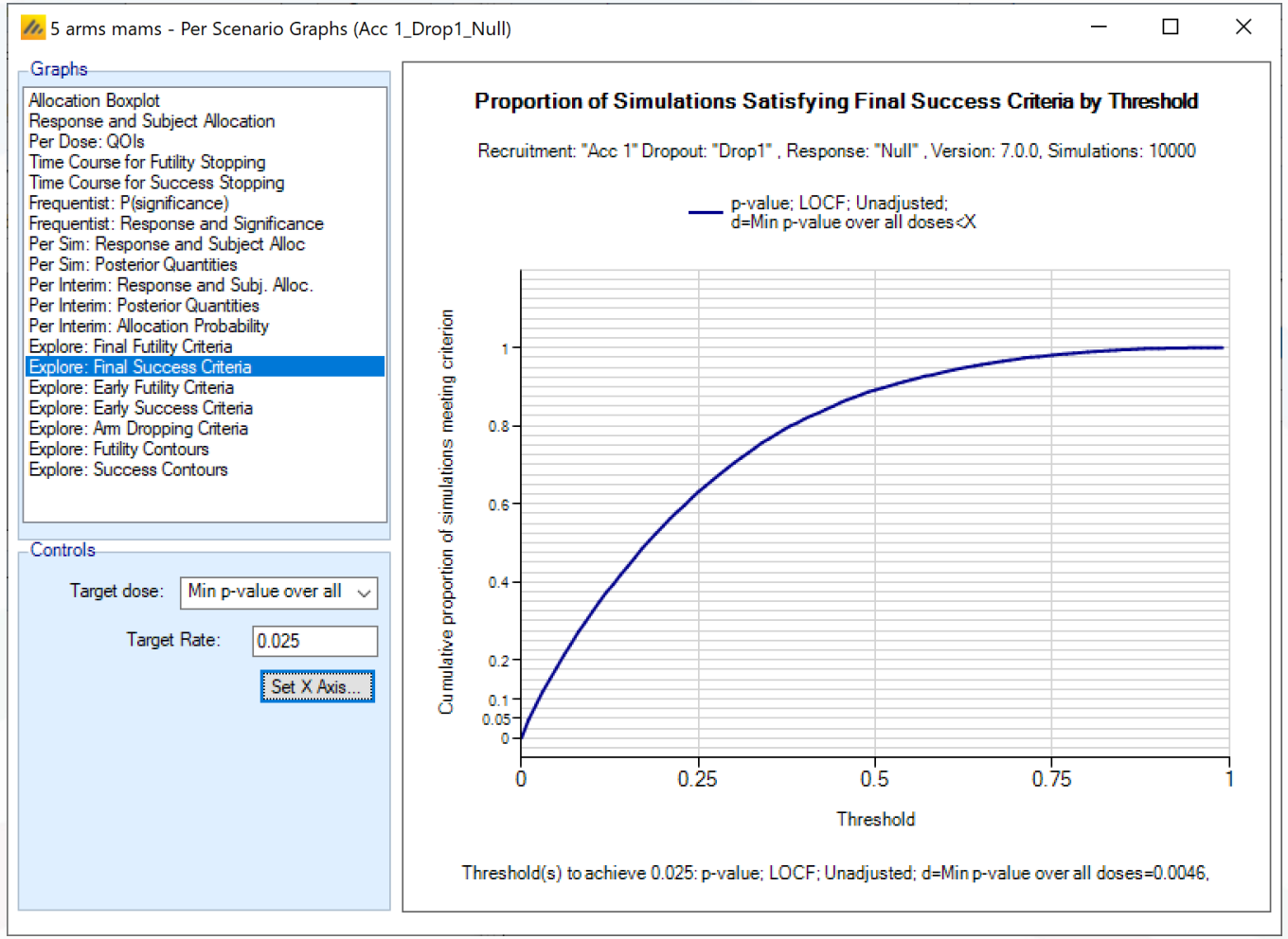
- We'll start using the MAMS thresholds converted to p-values

	Interim 1	Interim 2	Final
Futility	0.9999	0.998	0.0079
Success	0.000059	0.0019	0.0079

Combined success 0.0381 in the Null, too high! And as already noted there is no early futility stopping.
Also ASN is low in the alternate - in FACTS Core when one arm is successful, the trial stops. We'll need to switch to FACTS Platform Trial sim, but we'll use FACTS Core to determine our success/futility thresholds first.

	Num Sims	Mean Subj.	Ppn Early Success	Ppn Late Success	Ppn Late Futility	Ppn Early
_Null	10000	628.0088	0.0092	0.0289	0.9619	0
_LCF	10000	492.9851	0.6076	0.3225	0.0699	0
_Alt	10000	483.7211	0.6467	0.3002	0.0531	0

Adjusting Final Success Criteria

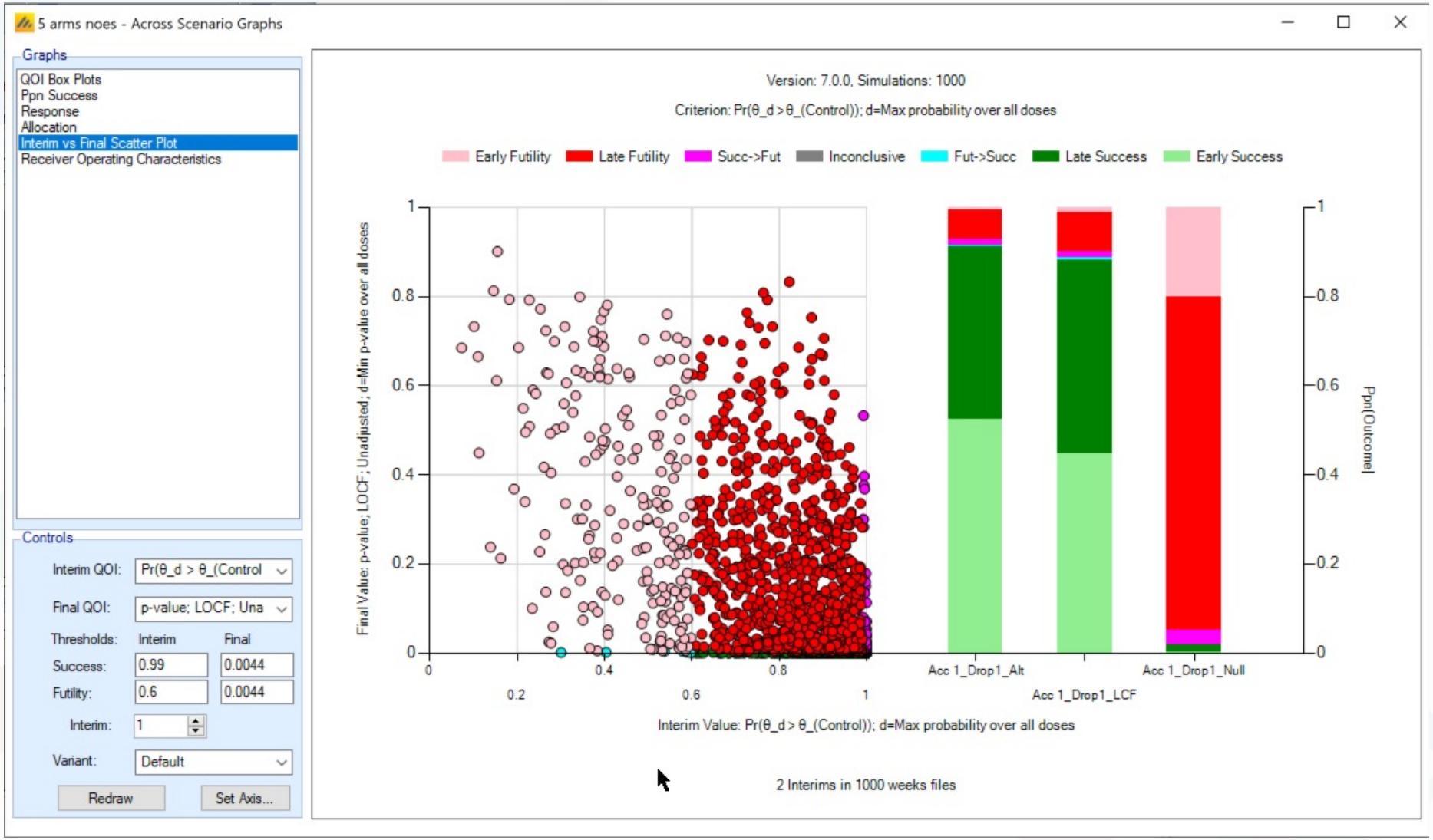


Explore: Final Success Criteria showing the Null scenario can be used to show the threshold that would limit successes to some number – here an alpha level of 0.0046 would limit (trial) success to 0.025

Adjusting Early stopping criteria

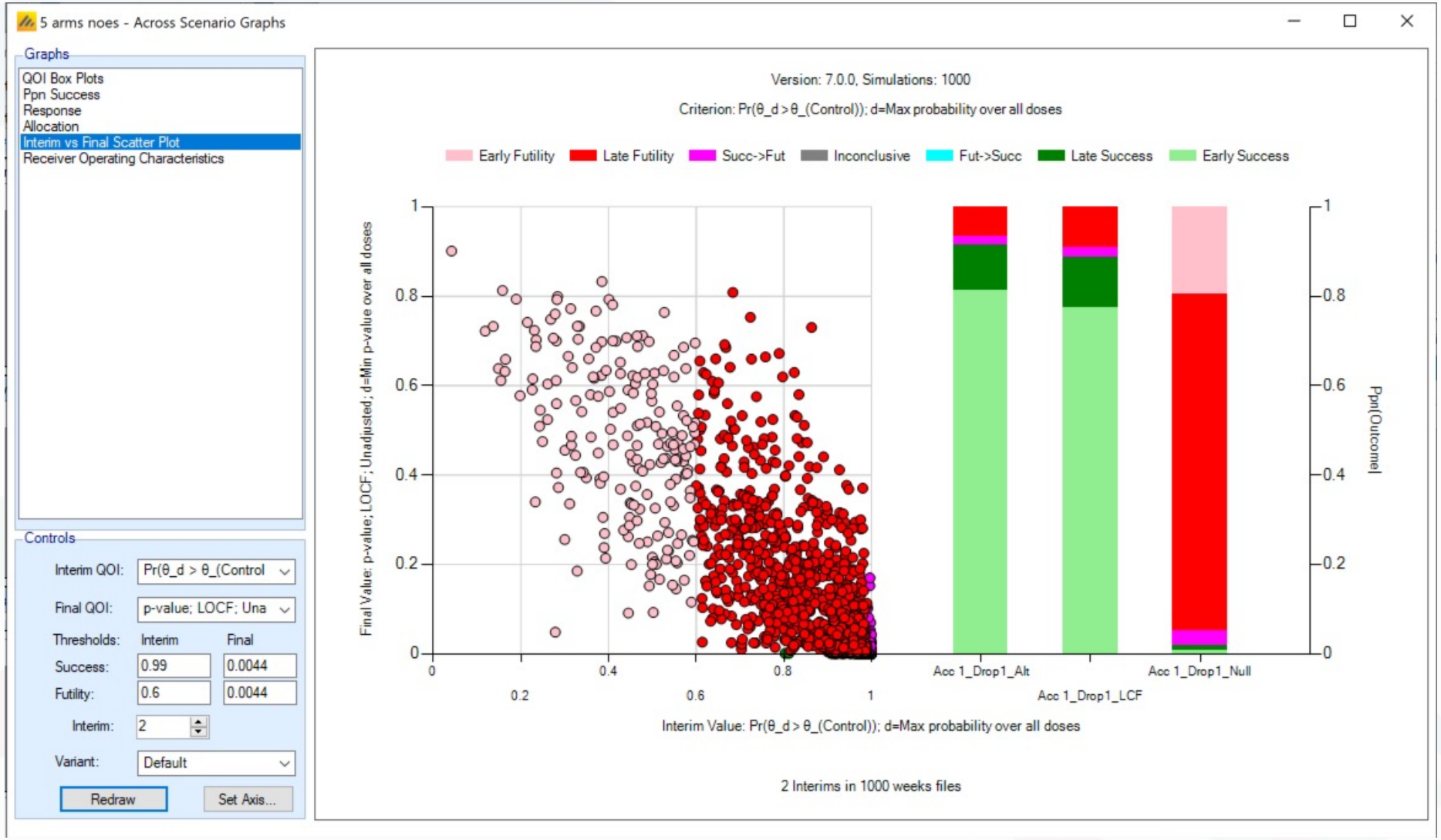
- We will use Bayesian posterior $Pr(\theta_d > \theta_{control})$
- Would like to use Bayesian Predictive Probability of success (Goldilocks trail design) but these are not currently available in FACTS PT

Futility boundary interim 1



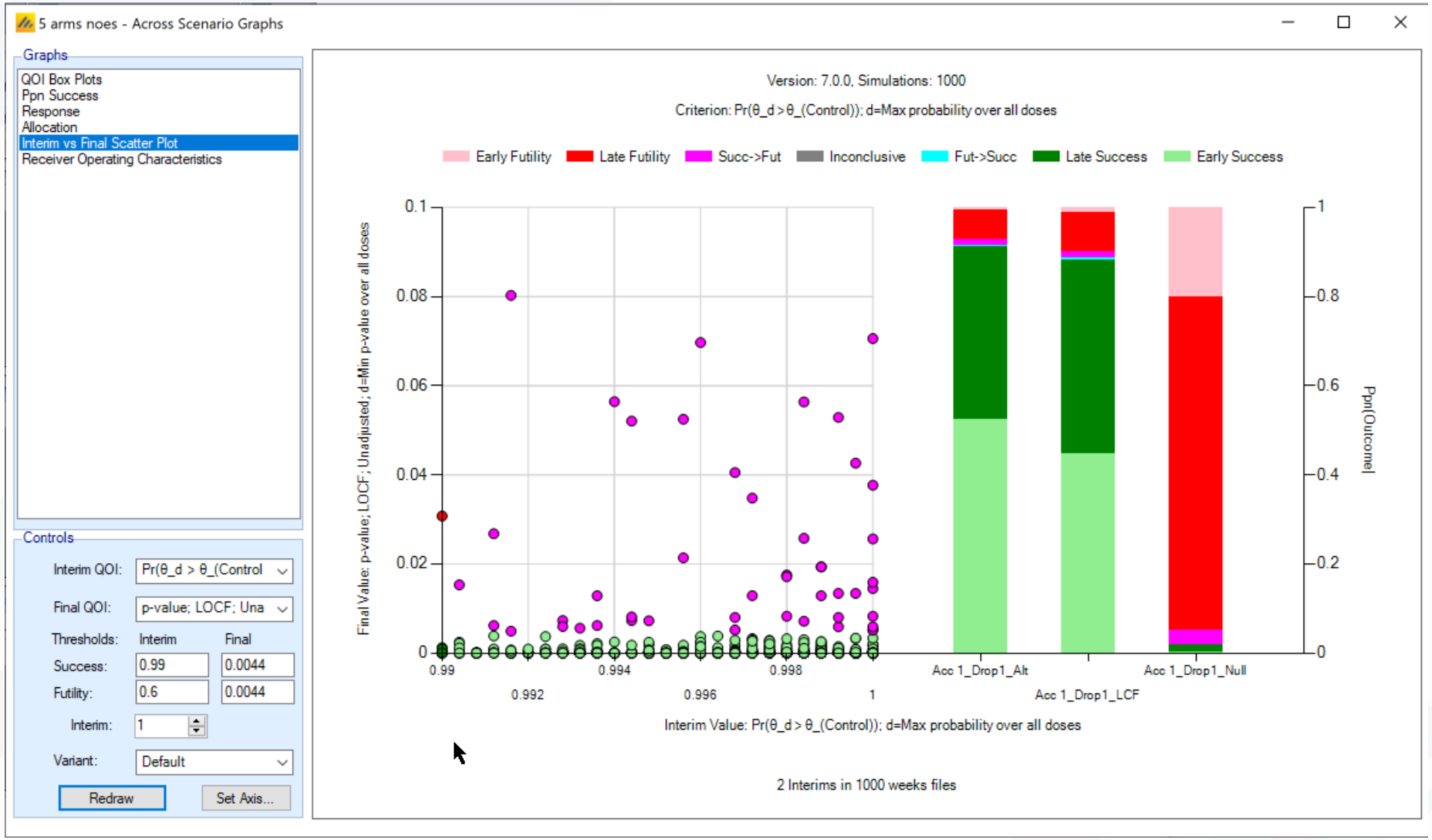
Stopping for futility at the first interim if $Pr(\theta_d > \theta_{control}) < 0.6$ Introduces only about 0.002-3 type-2 error

Futility boundary interim 2



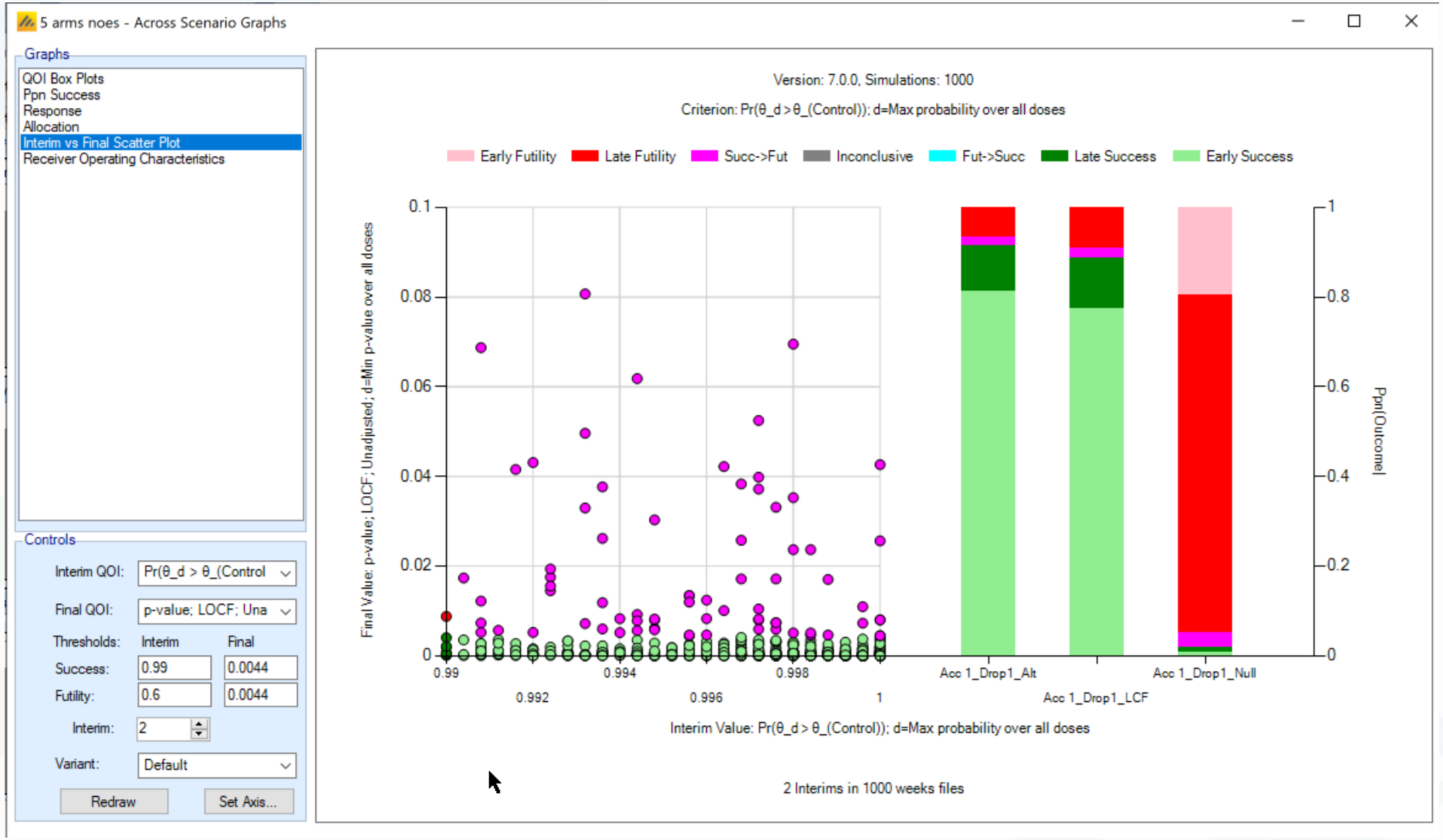
Stopping for futility at the second interim if $Pr(\theta_d > \theta_{control}) < 0.8$ Introduces only about 0.001 type-2 error

Success boundary interim 1



Stopping for success at the first interim if $Pr(\theta_d > \theta_{control}) > 0.9995$
Introduces about 0.007 type-1 error.
This is more than I'd have liked.

Success boundary interim 2



Stopping for success at the second interim if $Pr(\theta_d > \theta_{control}) > 0.999$
Introduces about 0.006 type-1 error.
Hopefully some of these are already type-1 errors at interim 1.

Using FACTS PT Simulator

- Study:
  - Enable adaptive
  - Max participants: 636
  - Max per treatment 90
  - Max concurrent treatments
  - Time to final endpoint 0.1
- Trial arms:
  - Control plus 5 treatment arms
  - All available at time 0
- Virtual Response:
  - Treatment “Good” if > 0.5
  - Treatment “Unacceptable if <= 0.3
  - Null, LFC, Alt scenarios as before
- Execution:
  - Mean accrual 5 per week
  - No Dropouts
- QOIs
  - $Pr(\theta_d > \theta_{control})$
  - P-value, LOCF Unadjusted
- Design
  - Prior for control and treatments N(0,10)
  - Prior for Sigma IG(2,1)
- Allocation 2:1:1:1:1:1
- Trial updates
  - First at 210 complete
  - At 210 complete thereafter
  - Treatment milestones at 29 & 59 subjects
- Initial success/futility
  - $Pr(\theta_d > \theta_{control}) > 0.9995$, 0.999 p-value < 0.0044
  - $Pr(\theta_d > \theta_{control}) < 0.6$, 0.8 p-value < 0.0044

Initial results

- Type-1 error too high
- ~5% early failures for successful arms

- 285 success in 10,000 sims of the Null, 110 early, 174 late
- Increased early thresholds to 0.9999 and 0.9995 now 86 early
- Reduced final success from 0.44, to 0.42, 0.4 then 0.38 now total success 245.
- Power in LFC 0.9 (but power with successful arm 0.88)
- Expected sample sizes: 449, 470, 528

BUT

- Is FWER type-1 error control really necessary?
- If difference versions (strength, regimen, combinations) of a treatment then yes
- But if all v different treatments, from different sources, then these could have been 5 separate trials each with there own 0.025 type-1 error.
- In this case it is surely inconsistent to require overall FWER type-1 rate of 0.025, 0.025 per arm would be consistent
  - Of course there is some correlation due to shared control this reduces type-1 error marginally and increases the probability of multiple type-1 error (these still v v small)

Platform Trial with 10 arms over time

- In FACTS …

To Conclude

- Thank You for attending
- Link to Recording will be sent out tomorrow
- Slides will be available via our website at the end of the series
- Any questions please contact us:
  - tom@berryconsultants.com
  - Kert@berryconsultants.com
  - facts@berryconsultants.com
- If you would like a demo and/or a free evaluation copy of FACTS – just ask.