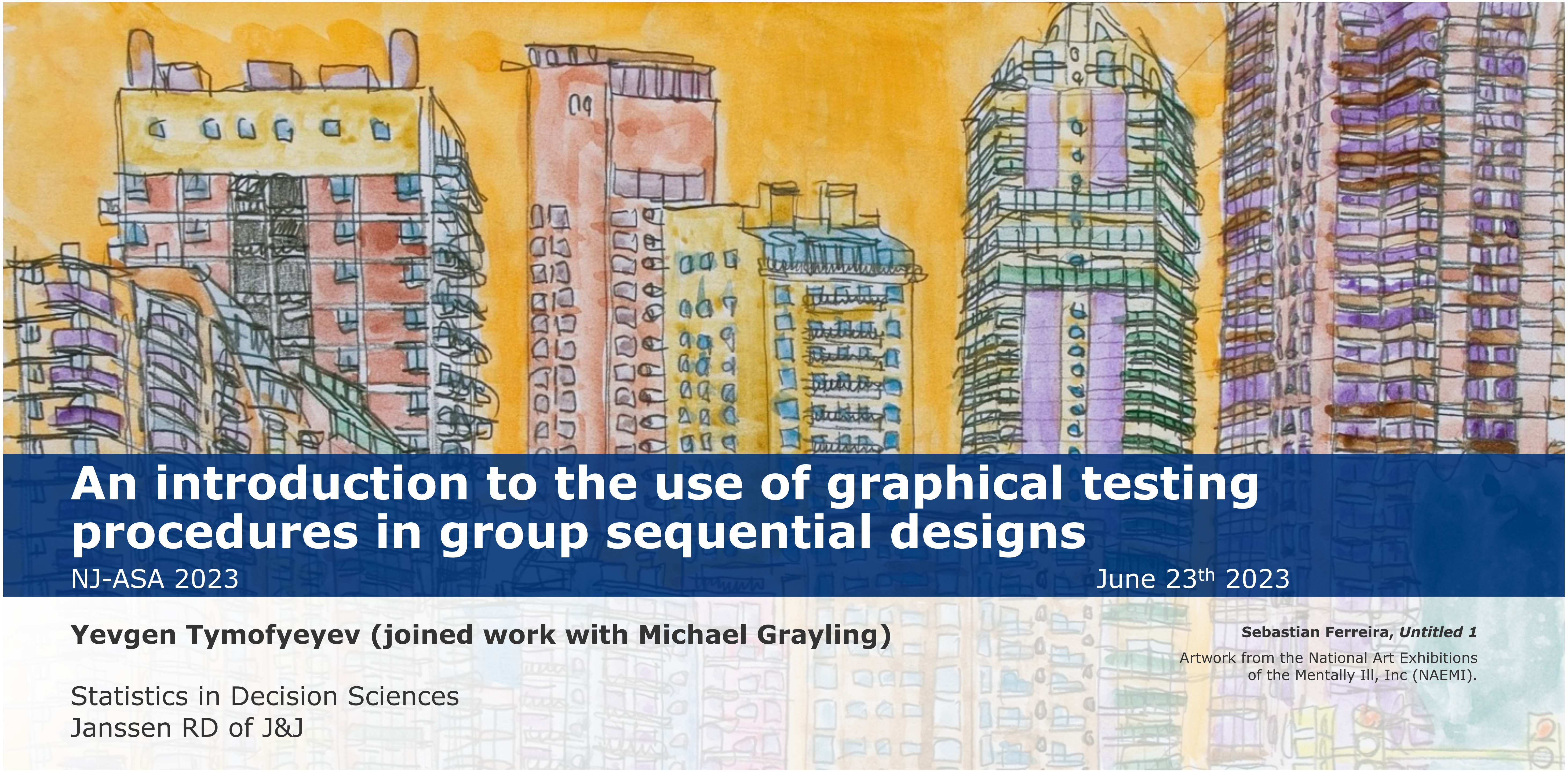


An introduction to the use of graphical testing procedures in group sequential designs

NJ-ASA 2023

June 23th 2023

Yevgen Tymofyeyev (joined work with Michael Grayling)

Statistics in Decision Sciences
Janssen RD of J&J

Sebastian Ferreira, *Untitled 1*

Artwork from the National Art Exhibitions
of the Mentally Ill, Inc (NAEMI).

Agenda

1. Group sequential design for a single endpoint

2. Graphical testing procedures in fixed-sample trials

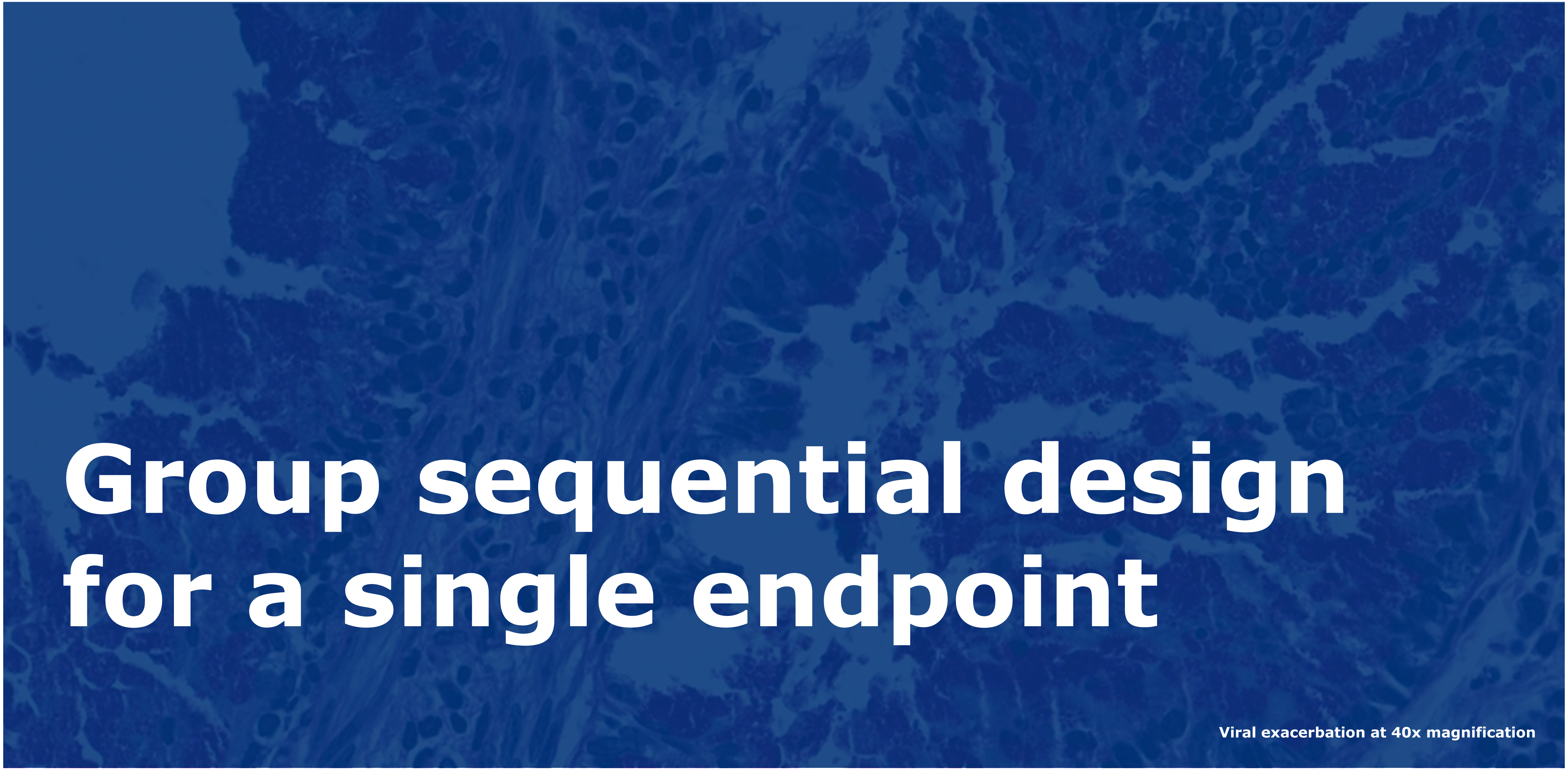
3. Graphical testing procedures in group sequential designs

4. Software

5. Discussion Q&A

Running Example 1

- Three-arm design, comparing the Experimental Treatments 1 and 2 (E1 and E2) to against Control
- 500:500:500 patient design
- Primary endpoint is progression-free survival (PFS)
  - mPFS_{Len} = 65 mo, $\text{HR}_{\text{PFS Tec}} = \text{HR}_{\text{PFS TecLen}} = 0.7$
  - Single interim analysis
- Key secondary outcome of overall survival (OS)
  - mOS_{Len} = 100 mo, $\text{HR}_{\text{OS Tec}} = \text{HR}_{\text{OS TecLen}} = 0.75$
  - Three interim analyses



Group sequential design for a single endpoint

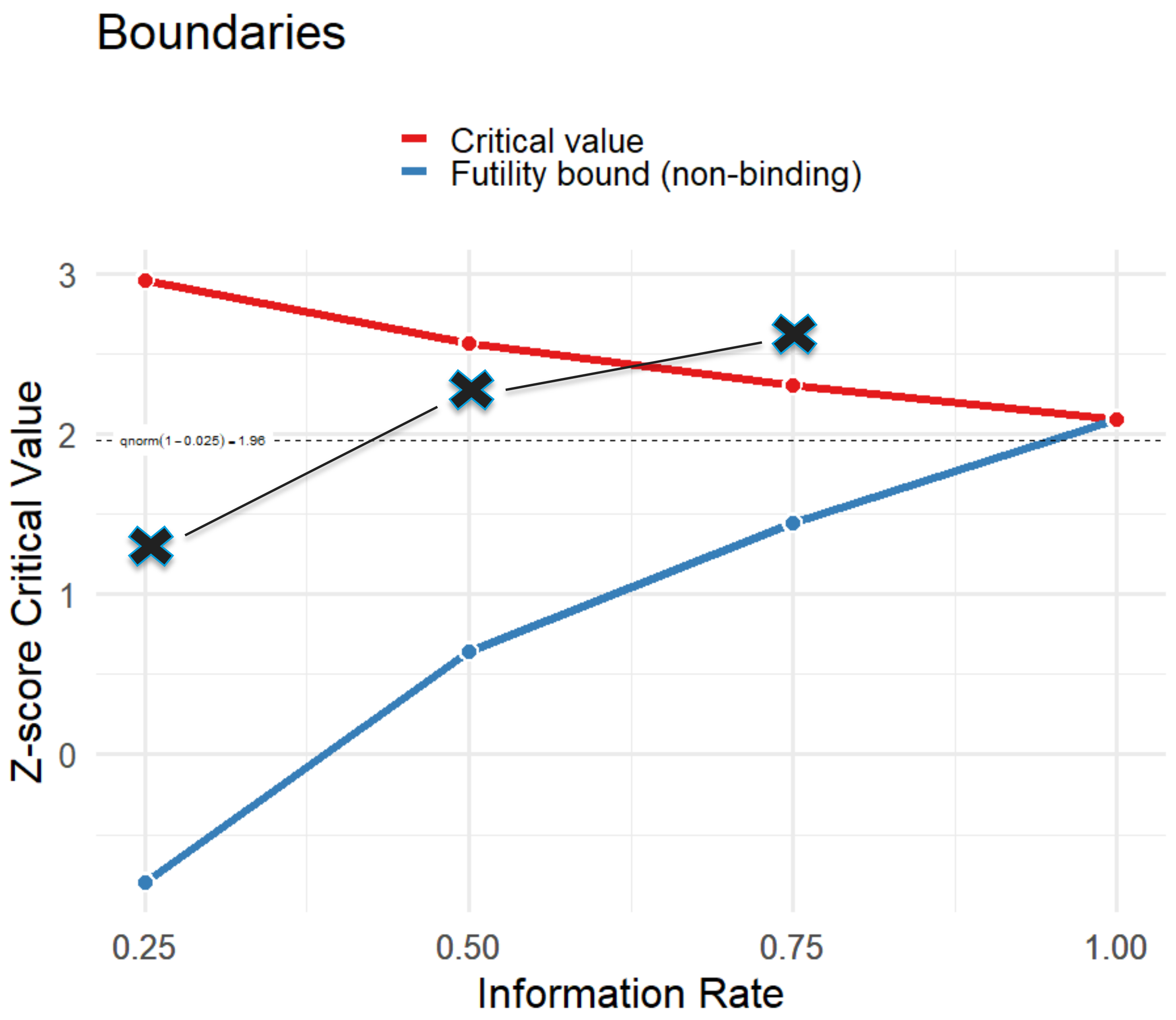
Viral exacerbation at 40x magnification

Group sequential design (GSD) introduction

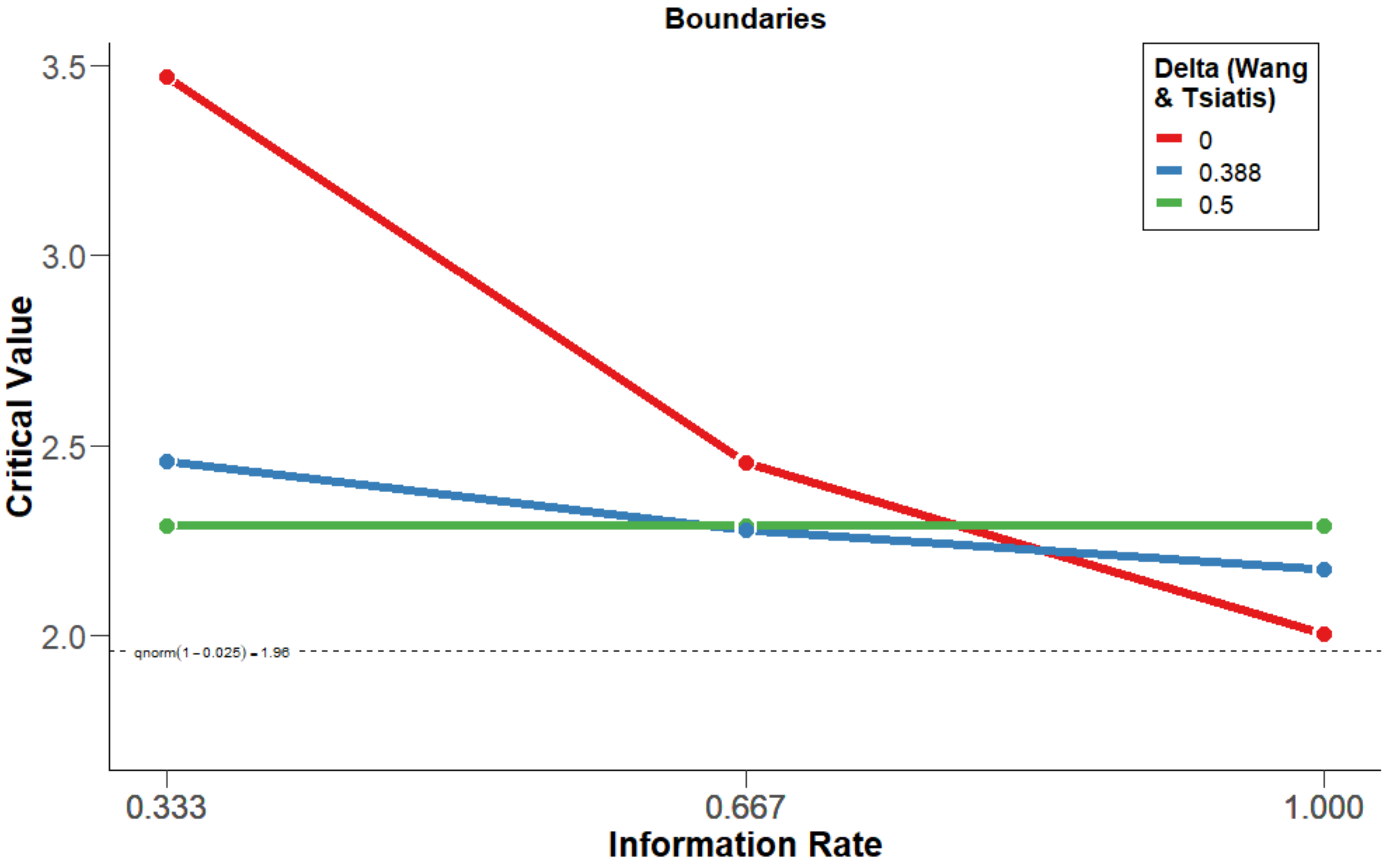
- Many clinical trials are designed considering an option of early termination
  - Overseen by Data and Safety Monitoring Board
- Reasons to conduct interim analyses as in Jennison & Turnbull (2000):
  - Ethical, administrative and economic
- Group sequential designs have been developed that avoid inflating the pre-specified type I error associated with the repeated testing of the treatment effect based on accumulating data (EMEA, 2007)
- There are also incentives to reach early decision if study is negative

Group Sequential Tests

- Test H_0: $\mu \leq 0$ against H_1: $\mu > 0$
- $Z_1, Z_2, \cdots, Z_J$ are standardized test statistics obtained at analyses $1, \ldots, J$
- Crossing upper boundary (denoted by b_j 's) results in early stopping for a positive outcome
  - $P_{\mu=0}\left(U_{j=1}^{J} Z_j > b_j\right) = \alpha$
- Crossing lower boundary => stopping for futility



Wang and Tsiatis Family with parameter Δ



Critical Boundary at the analysis *j*:

$$b_j = C_{WT}\, j^{\Delta - 0.5}$$

- Δ=0 O'Brien-Fleming
- Δ=0.5 Pocock boundary

Functional form efficacy bounds

I.e., the original approach

- Can easily be generalised for arbitrary information levels fixed in advance
- Small deviation from the planned information levels will not lead to substantial impact on type I / II error rates
- But a better way of designing under unpredictable information levels is…

Error spending

I.e., the approach usually used today

- Handles unpredictable information levels with strict type I error control
- Doesn't require maximum number of analyses to be pre-specified
- Use non-decreasing function $f : [0,1] \rightarrow [0,\alpha]$, that gives **cumulative α spend** at IF t_j as $f(t_j)$
  - **Information fractions (IFs)** $t_j = \frac{I_j}{I_J}$
    - Where I_j amount of statistical information at the *j-th* analysis
- Does require information level I_j **to not depend** on $\hat{\theta}_1, \dots, \hat{\theta}_{j-1}$

Error Spending Function Approach

Given Function $\boldsymbol{f}: [0,1] \rightarrow [0, \boldsymbol{\alpha}]$ non-decreasing

Fix, maximum information (N or # events) $\boldsymbol{I}_{\mathbf{max}}$

Analysis 1 get b_1:

$$P_{H_0}(Z_1 > b_1) = f(I_1/I_{max})$$

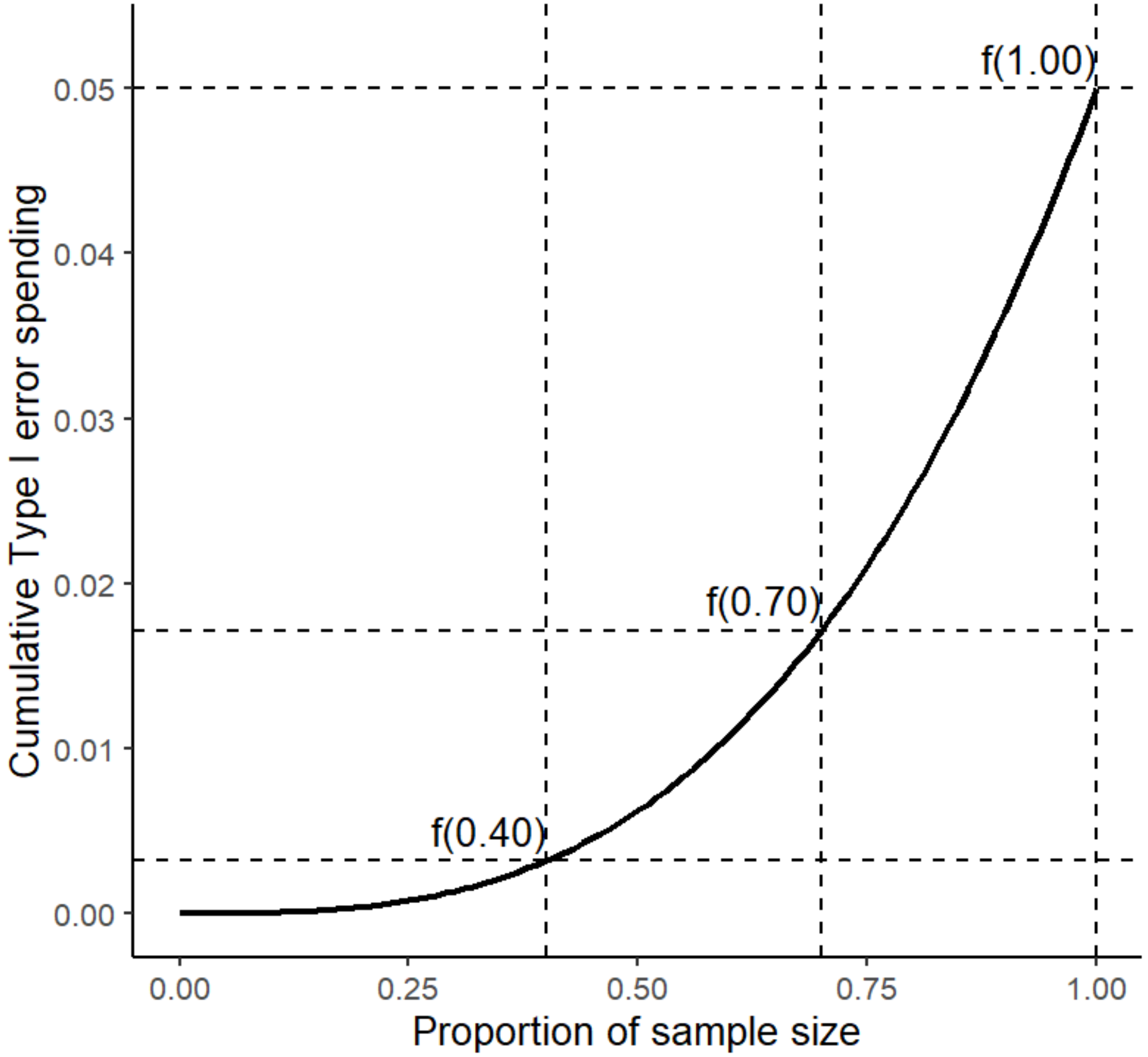
Analysis 2 get b_2:

$$P_{H_0}(Z_1 < b_1, Z_2 > b_2) = f(I_2/I_{max}) - f(I_1/I_{max})$$

…

Continue solving for b_j until reaching I_{max}, and "spend all alpha"

Also, the method accommodates "under" and "overrunning" of information scenarios



Common spending functions

- Lan and DeMets O’Brien-Fleming approximation:

$$f(t) = 2\left\{1 - \Phi\left[\Phi^{-1}(1-\alpha/2)/\sqrt{t}\,\right]\right\}$$

- Lan and DeMets Pocock approximation:

$$f(t) = \alpha \ln\{1 + (e-1)t\}$$

- Hwang, Shi and DeCani (γ-family), with $\gamma \in \mathbb{R}$:

$$f(t) = \begin{cases} \alpha(1-e^{-\gamma t})/(1-e^{-\gamma}) & \gamma \neq 0 \\ \alpha t & \gamma = 0 \end{cases}$$

$\gamma = -4$ Similar to O’Brien-Fleming

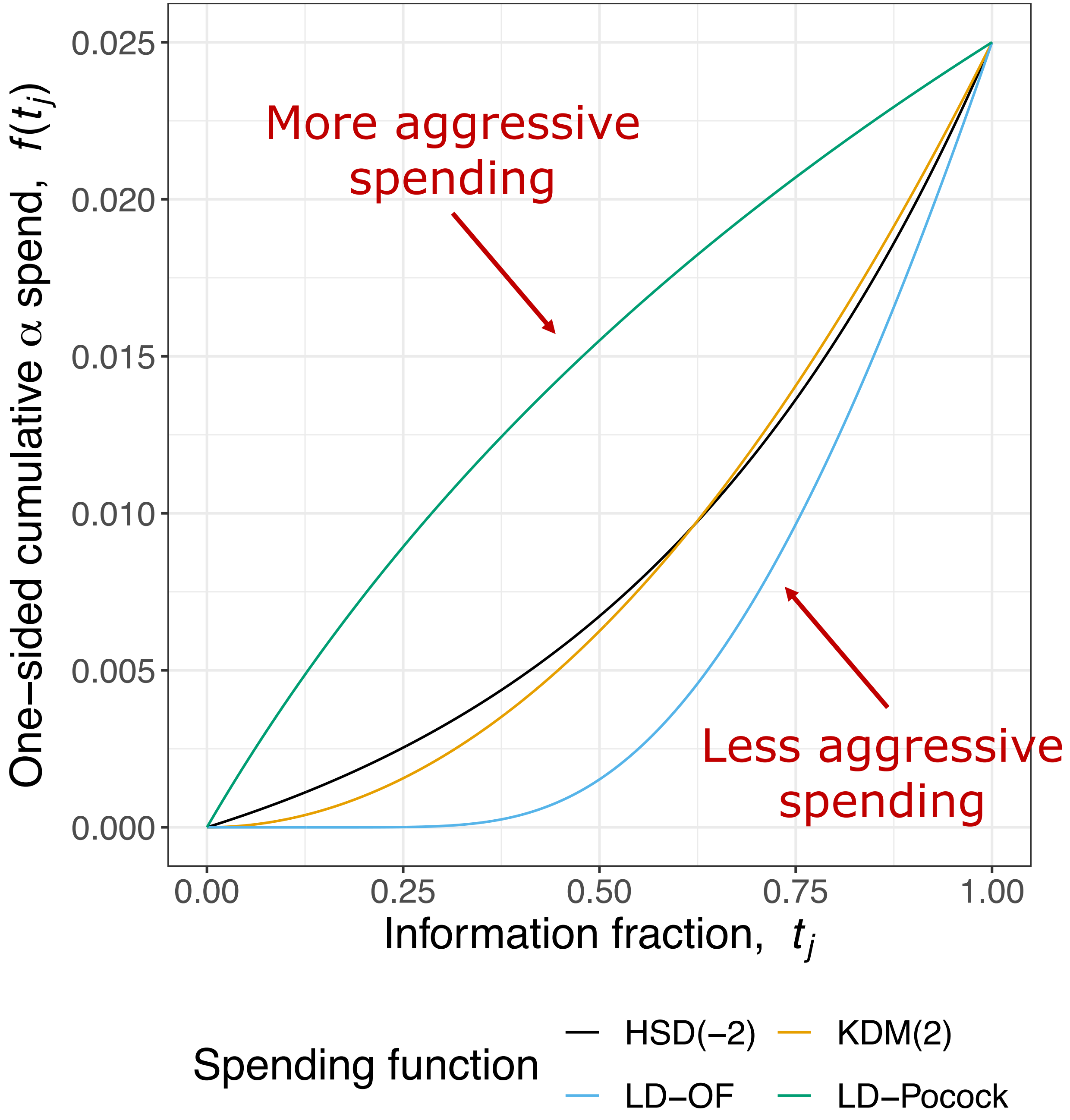
$\gamma = 1$ Similar to Pocock

- Kim and DeMets (ρ-family / power-family), with $\rho > 0$:

$$f(t) = \alpha t^{\rho}$$

$\rho = 3$ Similar to O’Brien-Fleming

$\rho = 0.75$ Similar to Pocock



Software

- EAST mycytel.cytel.com
- ADDPLAN
- SAS SEQDESIGN
- R:
  - gsDesign https://gsdesign.shinyapps.io/prod/
  - rpact (~ ADDPLAN) https://rpact.shinyapps.io/public/
  - Others too… https://cran.r-project.org/web/views/ClinicalTrials.html

Running Example 1:

- Just consider TecLen vs Len for PFS
  - mPFSLen = 65 mo, $\text{HR}_{\text{PFS TecLen}} = 0.7$
- Usual total one-sided $\alpha = 0.025$ and suppose we desire 90% power
  - More on this later though
- 5% drop-out rate for PFS
- 1000 patients with recruitment rate = 42 pts/mo
- **GSD:**
  - **Single interim analysis at 70% IF (i.e., $t_1 = 0.7$, $t_2 = 1$)**
  - **Lan and DeMets O'Brien-Fleming (LDOF) spending function**

Running Example 1:

gsDesign

```
> gsDesign::gsSurv(k         = 2,
+                 test.type = 1,
+                 alpha     = 0.025,
+                 beta      = 0.1,
+                 timing    = c(0.7, 1),
+                 sfu       = gsDesign::sfLDOF,
+                 lambdaC   = log(2)/65,
+                 hr        = 0.7,
+                 eta       = -(1/12)*log(1 - 0.05),
+                 gamma     = 42,
+                 R         = 1000/42)
Time to event group sequential design with HR= 0.7
Equal randomization:          ratio=1
One-sided group sequential design with
90 % power and 2.5 % Type I Error.

  Analysis   N    Z   Nominal p  Spend
         1 236 2.44      0.0074 0.0074
         2 337 2.00      0.0228 0.0176
     Total                      0.0250

++ alpha spending:
 Lan-DeMets O'Brien-Fleming approximation spending function with
none = 1.
```

```
Boundary crossing probabilities and expected sample size
assume any cross stops the trial

Upper boundary (power or Type I Error)
          Analysis
   Theta      1      2 Total   E{N}
  0.0000 0.0074 0.0176 0.025 335.8
  0.1779 0.6152 0.2848 0.900 274.4
               T    n   Events HR efficacy
IA 1  44.28596 1000 235.5577        0.728
Final 63.88113 1000 336.5110        0.804
Accrual rates:
        Stratum 1
0-23.81        42
Control event rates (H1):
      Stratum 1
0-Inf      0.01
Censoring rates:
      Stratum 1
0-Inf         0
```

Running Example 1:

rpact

```
> design            <-
+   rpact::getDesignGroupSequential(kMax             = 2,
+                                   alpha            = 0.025,
+                                   beta             = 0.1,
+                                   sided            = 1,
+                                   informationRates = c(0.7, 1),
+                                   typeOfDesign     = "asOF")
> sampleSizeResult <-
+   rpact::getSampleSizeSurvival(design           = design,
+                                lambda2          = log(2)/65,
+                                hazardRatio      = 0.7,
+                                dropoutRate1     = 0.05,
+                                dropoutRate2     = 0.05,
+                                dropoutTime      = 12,
+                                accrualTime      = c(0, 1000/42),
+                                accrualIntensity = 42)
> summary(sampleSizeResult)
```

```
Sample size calculation for a survival endpoint

Sequential analysis with a maximum of 2 looks (group sequential design),
overall
significance level 2.5% (one-sided).
The sample size was calculated for a two-sample logrank test,
H0: hazard ratio = 1, H1: hazard ratio = 0.7, control lambda(2) = 0.011,
accrual time = 23.81, accrual intensity = 42, dropout rate(1) = 0.05,
dropout rate(2) = 0.05, dropout time = 12, power 90%.

Stage                                         1      2
Information rate                            70%   100%
Efficacy boundary (z-value scale)         2.438  2.000
Overall power                            0.6152 0.9000
Expected number of subjects              1000.0
Number of subjects                       1000.0 1000.0
Cumulative number of events               234.5  335.0
Analysis time                              44.1   63.5
Expected study duration                    51.6
Cumulative alpha spent                   0.0074 0.0250
One-sided local significance level       0.0074 0.0228
Efficacy boundary (t)                     0.727  0.804
Exit probability for efficacy (under H0) 0.0074
Exit probability for efficacy (under H1) 0.6152

Legend:
  (t): treatment effect scale
```

Summary

- GSDs seek to reduce the expected time to a significant result
- Easy to control type I error rate using **error spending** approach
- On top of usual requirements for sample size calculation, specify:
  - **IFs at the interim analyses**
  - **Spending function**

Graphical testing procedures in fixed-sample trials

Viral exacerbation at 40x magnification

Multiple testing procedures

- Most clinical trials evaluate significance for multiple important outcomes
- Some evaluate significance for multiple treatment arms
- In either case, we then typically need to control the probability of committing one or more type I errors across the analyses
  - **Family-wise error rate** (FWER) control
- **Multiple testing procedures** are methods for achieving such FWER control

Graphical testing procedures (GTPs)

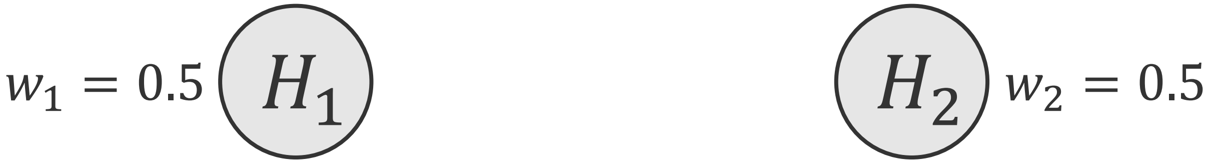
- Flexible multiple testing framework that can be **tailored to reflect the relative importance of hypotheses**
  - I.e., can deal with complex trial objectives and multiple structured hypotheses
- Built on the principle of **closed testing**
  - I.e., they can be thought of as a shortcut to specifying a closed testing procedure
  - Ensures strong FWER control
- Very visual technique
  - **Easily and efficiently communicable**
- Includes many common multiple testing procedures as special cases
  - Fixed sequence, Bonferroni, Holm, …

The graph

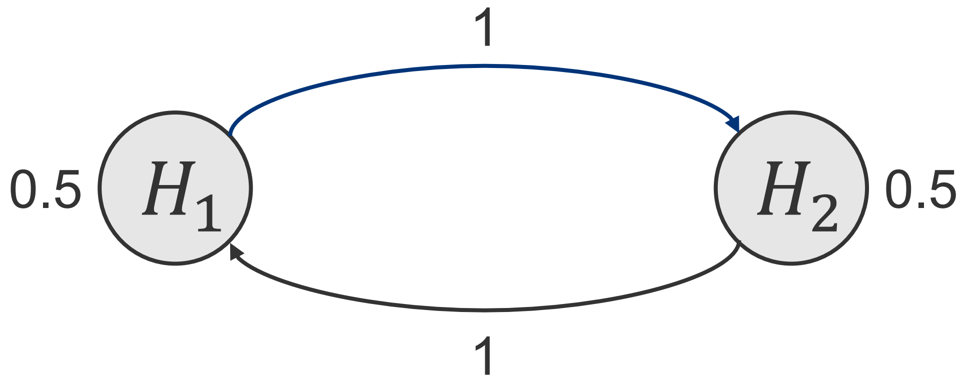
Specification

1. Hypotheses $H_1, \dots, H_K$ represented as **nodes**

2. (Initial) split of significance level represented by **weights** $w_1, \dots, w_K$



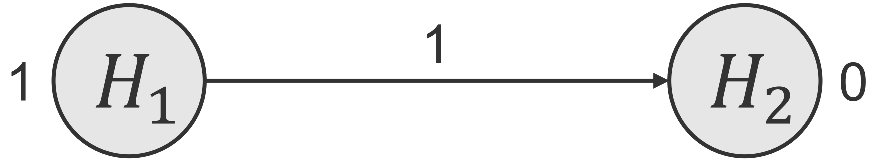
3. 'α-recycling' through **weighted directed edges**



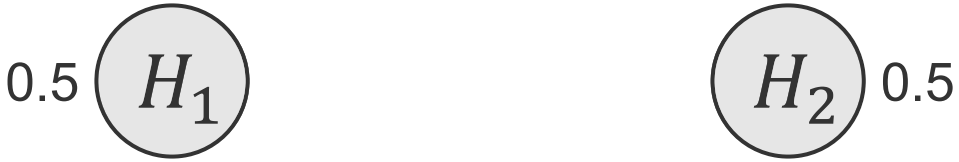
Examples

$K = 2$

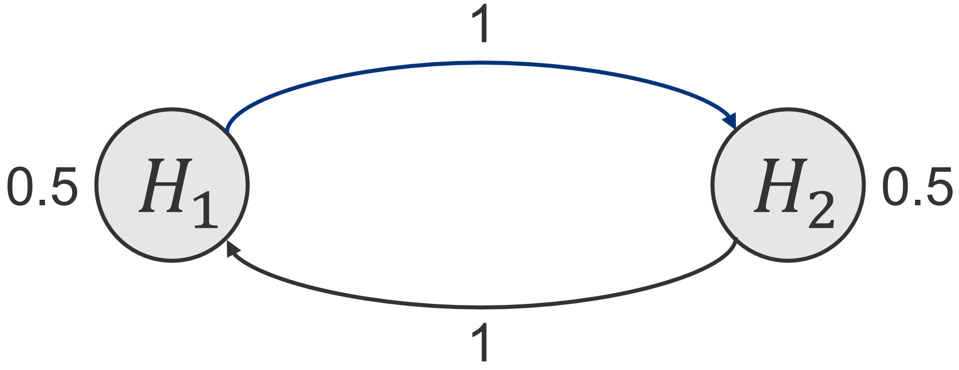
- **Fixed sequence:** Maximizes power if previous hypotheses rejected as all tests performed at level α



- **Bonferroni:** No α-recycling



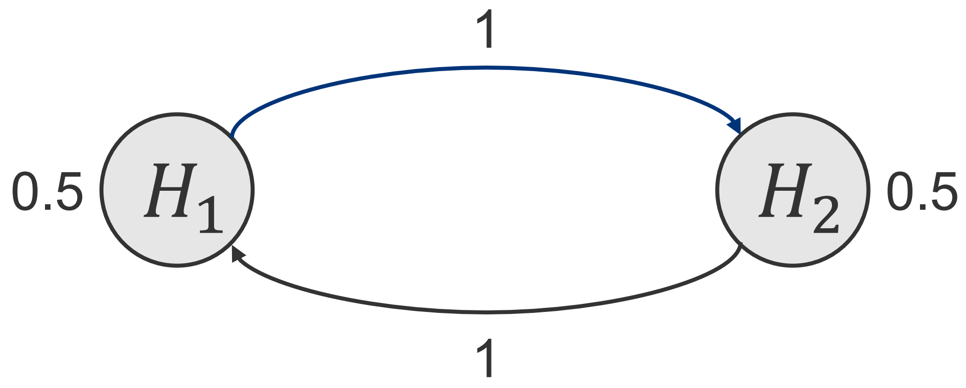
- **Holm:** Everything in Bonferroni + more → more powerful



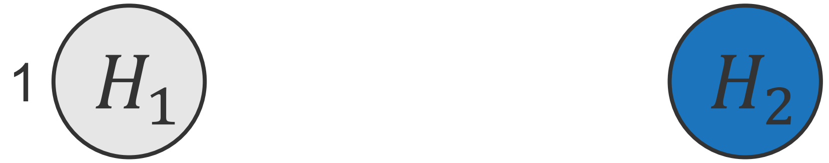
Example: Holm

$K = 2$ *and* $\alpha = 0.025$

- Suppose that $p_1 = 0.02$ and $p_2 = 0.01$ are the p-values for H_1 and H_2



- As $p_2 = 0.01 \leq 0.0125 = 0.5(0.025) = w_2\alpha$, reject H_2 and update the graph



- As $p_1 = 0.02 \leq 0.025 = 1(0.025) = w_1\alpha$, we can now also reject H_1

Technical basis

- The graph defines a closed testing procedure with **weighted tests** (e.g., weighted Bonferroni) for each intersection hypothesis
- If a hypothesis H_k can be rejected at level $w_k\alpha$ (i.e., $p_k \leq w_k\alpha$), recycle its level $w_k\alpha$ to the remaining (not yet tested) hypotheses, according to a prefixed rule, and continue testing with the updated α levels
- Can be shown that the order you test in does not matter
  - I.e., would always end with the same hypotheses being rejected

Technical basis

Graph update algorithm

- Transition matrix $G = \{g_{ij}\}$, where g_{ij} is the fraction of w_i allocated to H_j if H_i is rejected
- Require $0 \le g_{ij} \le 1$, $g_{ii} = 0$ and $\sum_{k=1}^{K} g_{ik} = 1$ for $i, j = 1, \dots, K$

0. Set $\mathcal{K} = \{1, \dots, K\}$

1. Select a $k \in \mathcal{K}$ such that $p_k \le w_k \alpha$ and reject H_k; otherwise stop

2. Update the graph:

$$\mathcal{K} \to \mathcal{K} \backslash \{k\}$$

$$w_l \to \begin{cases} w_l + w_k g_{kl} & : l \in \mathcal{K} \\ 0 & : \text{otherwise} \end{cases}$$

$$g_{lm} \to \begin{cases} \frac{g_{lm} + g_{lk} g_{km}}{1 - g_{lk} g_{kl}} & : \text{for } l, m \in \mathcal{K}, l \neq m, g_{lk} g_{kl} < 1 \\ 0 & : \text{otherwise} \end{cases}$$

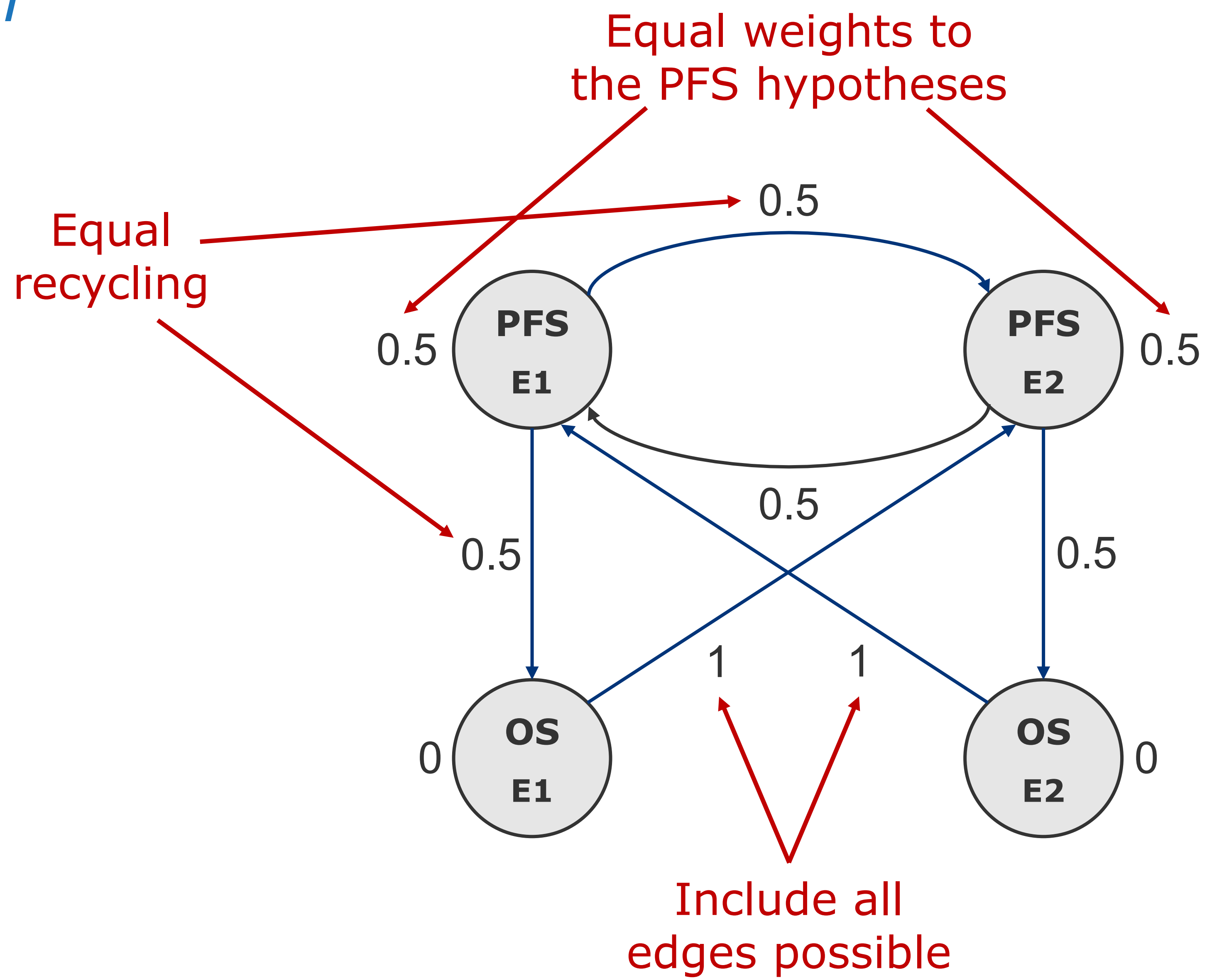
3. If $|\mathcal{K}| \ge 1$, go to Step 1; otherwise stop

Running Example 1:

- Four hypotheses
  - PFS and OS for Experimental 1 and Experimental 2
- PFS hypotheses have all α initially as the primary endpoint
- Equal priority to both comparisons
- Recycle to corresponding OS and other PFS hypothesis

Running Example 1:

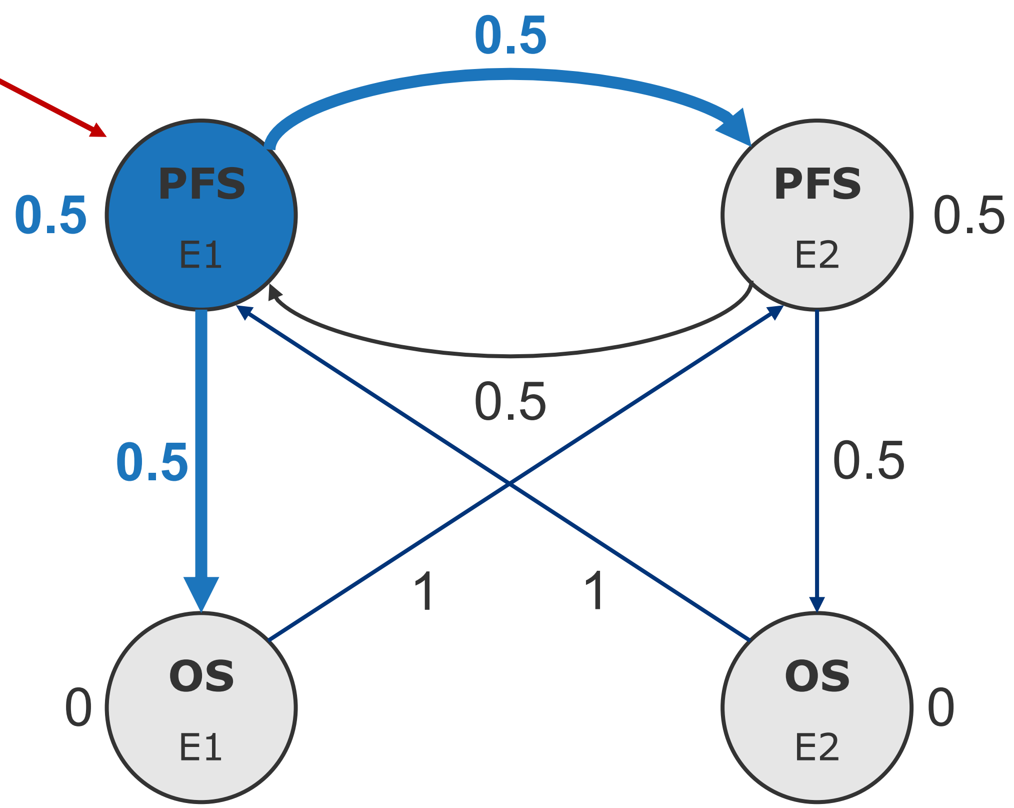
Initial graph



Running Example 1:

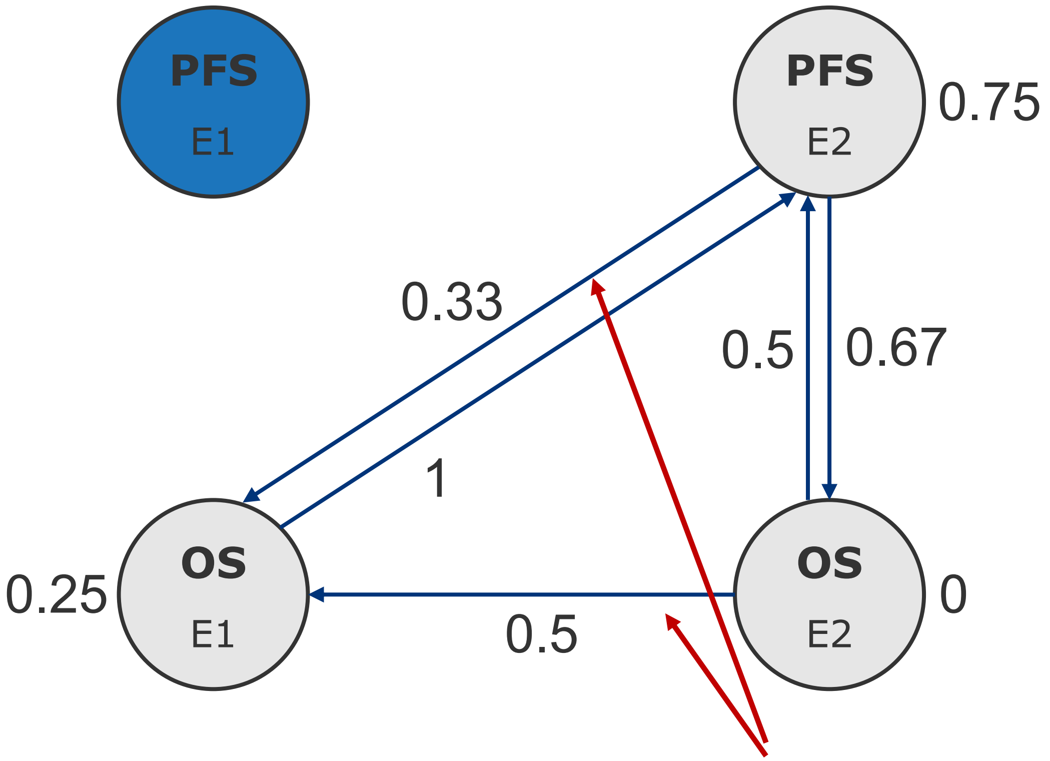
Sequential updating

Suppose we achieve significance for PFS on E1 vs Cntrl



Running Example 1:

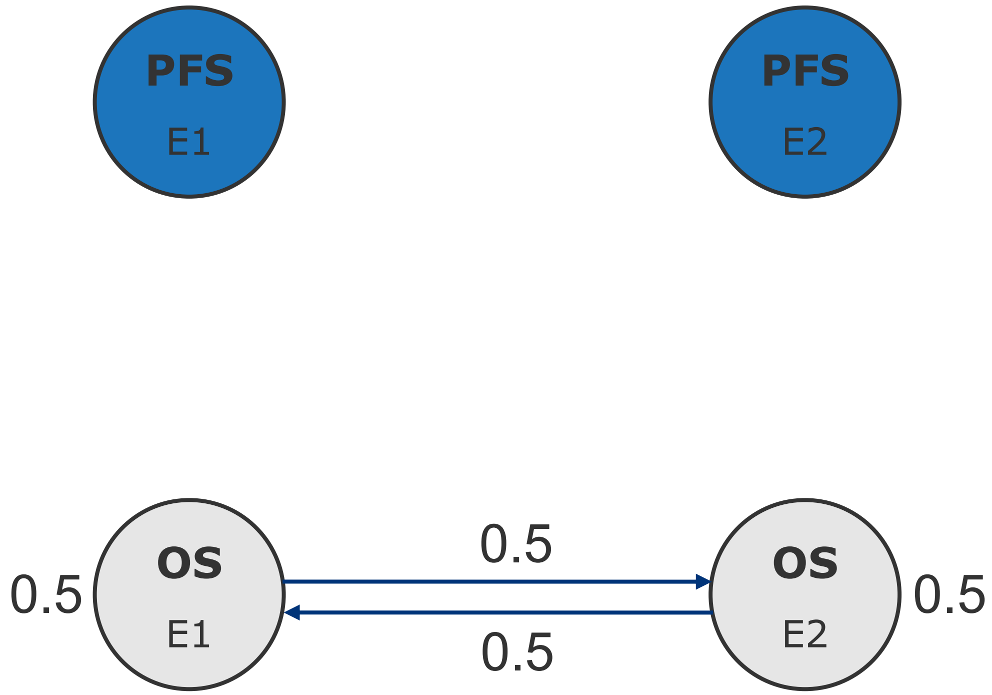
Sequential updating



Note: There are now edges that weren’t previously in the graph

Running Example 1:

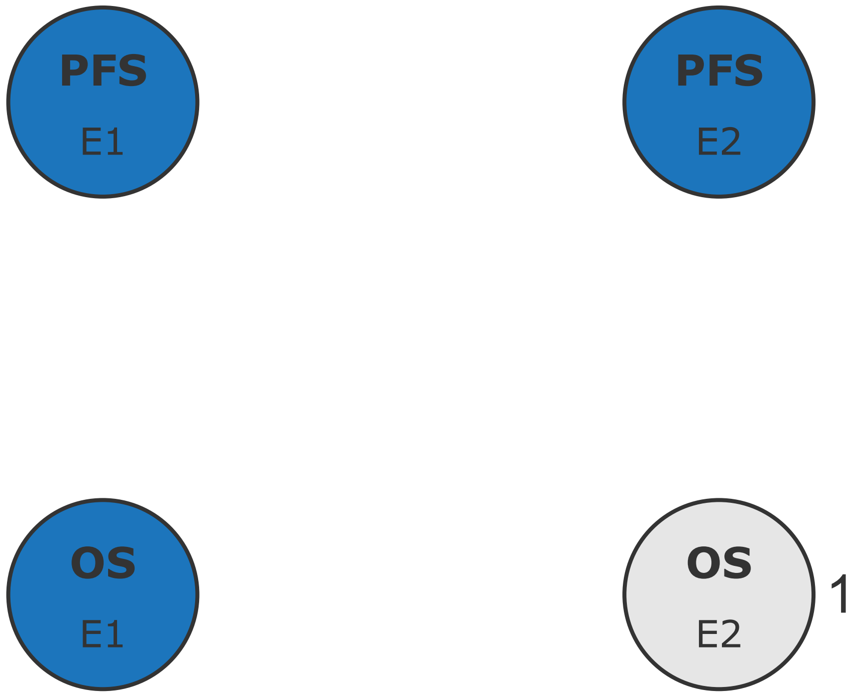
Sequential updating



It’s now symmetric again: the graph would look like this regardless of which of the PFS hypotheses was rejected first

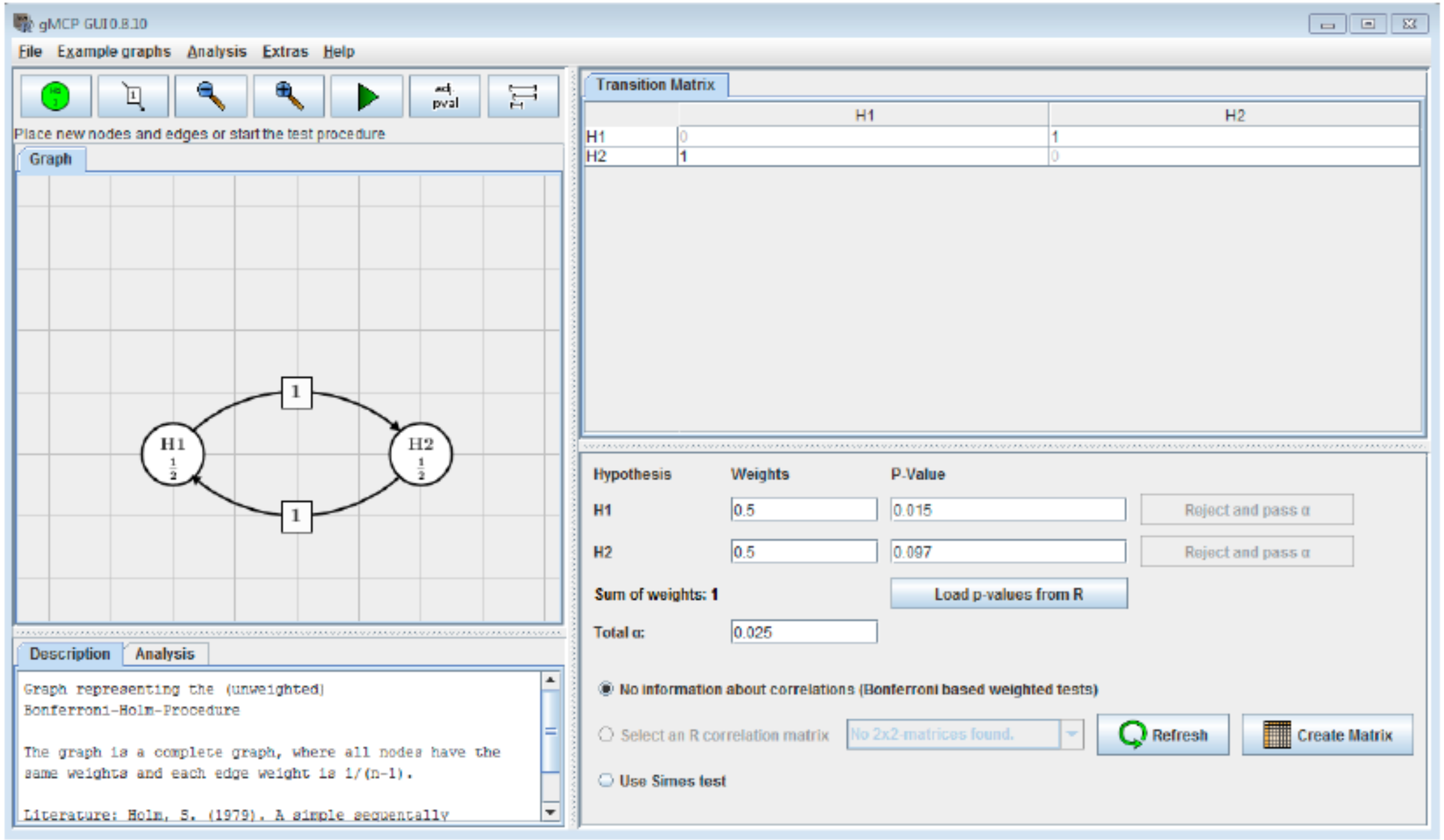
Running Example 1:

Sequential updating



Software

- R:
  - gMCP
  - gsDesign
  - gMCPLite



Place new nodes and edges or start the test procedure

	H1	H2
H1	0	1
H2	1	0

Hypothesis	Weights	P-Value
H1	0.5	0.015
H2	0.5	0.097

Sum of weights: 1

Total α: 0.025

● No information about correlations (Bonferroni based weighted tests)

○ Select an R correlation matrix No 2x2-matrices found.

○ Use Simes test

```
Graph representing the (unweighted)
Bonferroni-Holm-Procedure

The graph is a complete graph, where all nodes have the
same weights and each edge weight is 1/(n-1).

Literature: Holm, S. (1979). A simple sequentally
```

Summary

- GTPs are a **flexible and powerful** method of strongly controlling the FWER across multiple hypotheses
- Completely defined by the initial graph, which contains:
  - **Nodes defining hypotheses**
  - **Weights defining initial α split**
  - **Edges defining how to recycle α**

Graphical testing procedures in group sequential designs

Viral exacerbation at 40x magnification

History

- Long history of methods / application of GSDs to clinical trials
- Similar is true of GTPs
- But development of methods for use of GTPs in GSDs has occurred mostly over last 10-15 years
- Much was motivated by…

Hierarchical testing of a primary and one secondary endpoint

- Hung *et al* (2007) considered a two-stage GSD with a primary and one key secondary endpoint

- The primary endpoint tested according to some GSD with cumulative one-sided type I error of $\alpha = 0.025$

- **Question:** How should we test the secondary endpoint after the primary endpoint achieves significance (either at the IA or FA)?
  - ***Assuming that Secondary EP data accumulates from Interim to Final***

- Investigated **naïve strategy** for secondary endpoint:
  - Since the secondary endpoint is tested at most once, when the primary endpoint is significant, it seems reasonable to use the **whole** α (regardless of IA or FA)

Hierarchical testing of a primary and one secondary endpoint

- Demonstrated that this approach does not control the FWER
- Depending on the correlation between the endpoints, FWER could be as much as 4.1%
- So specialist methodology required for FWER control

Naïve strategy

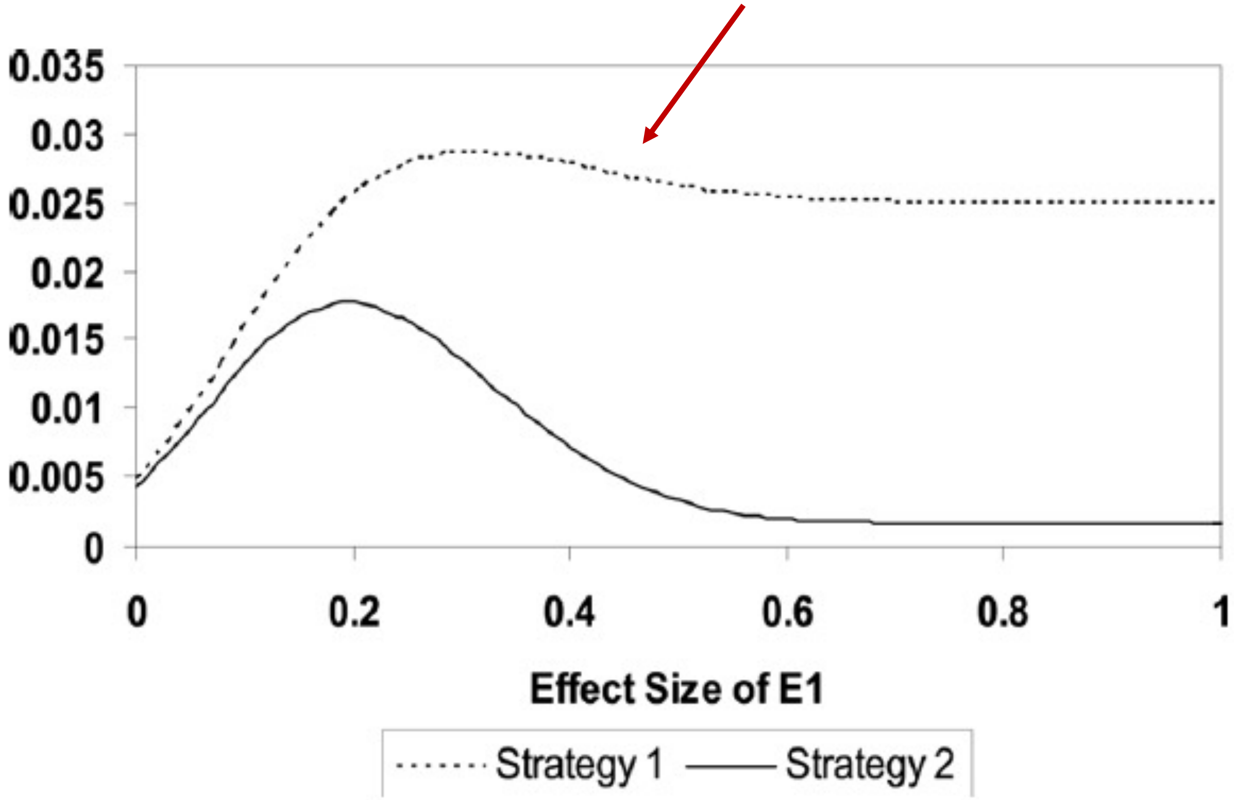


Figure 1 Type I error rate of E2 ($\rho = 0.5$).

Hung *et al* (2007)

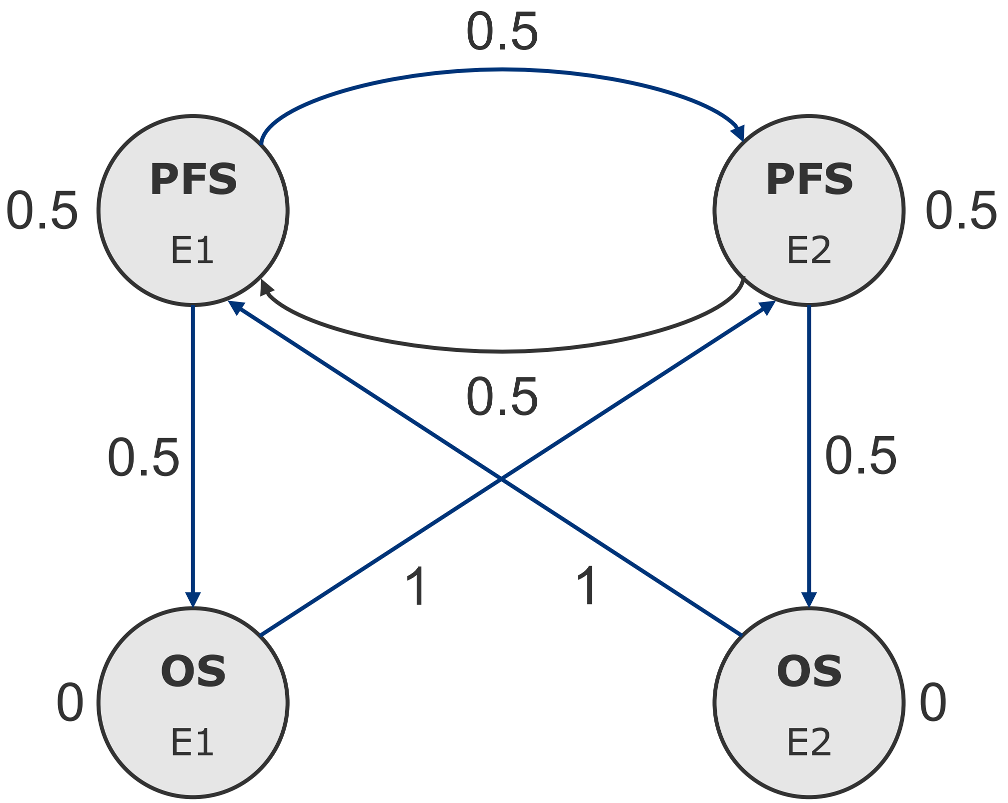
GTPs for GSDs

- Maurer and Bretz (2013), amongst others, provide highly general methodology for testing primary and secondary endpoints in GSD setting with strong control of the FWER

- **Take home message: Essentially all you have to do is specify your initial GTP and your GSD for each hypothesis**
  - I.e., think of it as the union of two more familiar steps: specifying a GTP and specifying GSDs
  - There are some finer points, but this gets you the majority of the way there

Running Example 1:

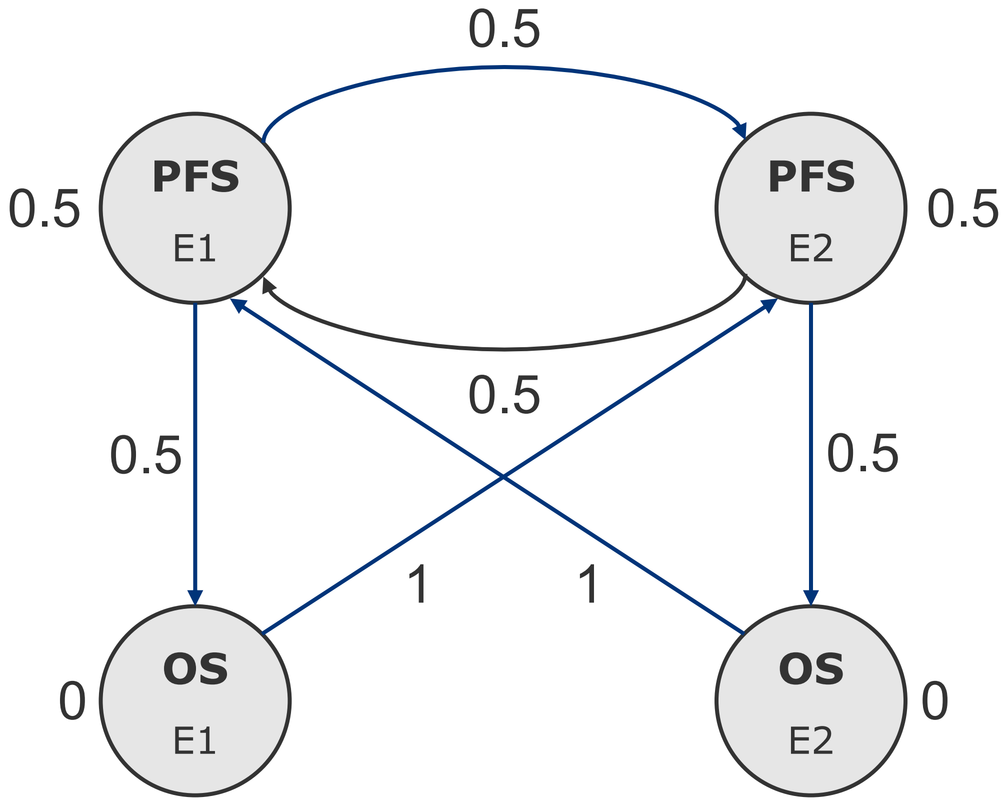
Focus on PFS for E1 vs Cntrl

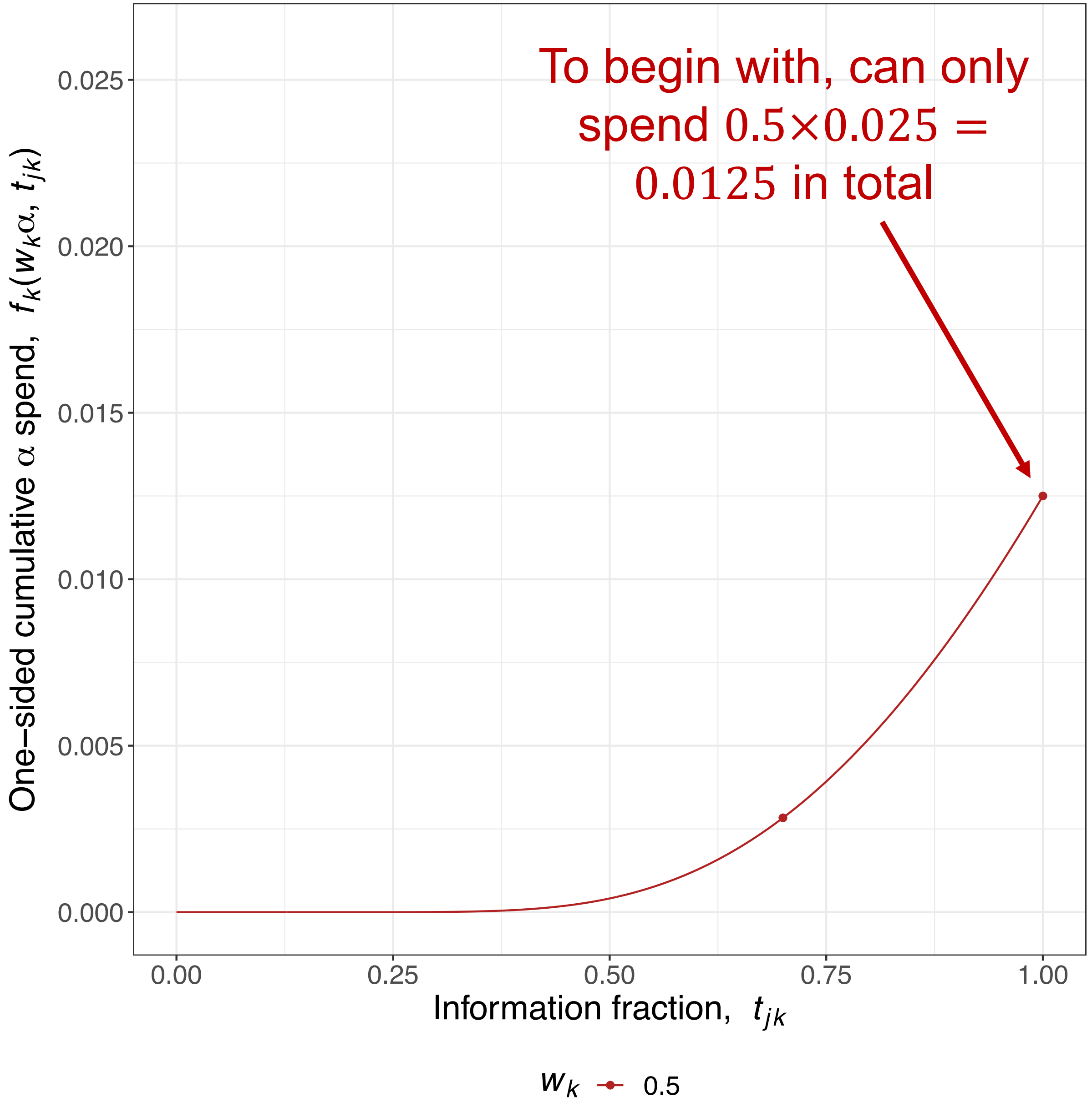


- Single IA at ~70% IF
- LDOF spending function
- Initially it has weight of 0.5
- Overall one-sided $\alpha = 0.025$

Running Example 1:

Focus on PFS for E1 vs Cntrl

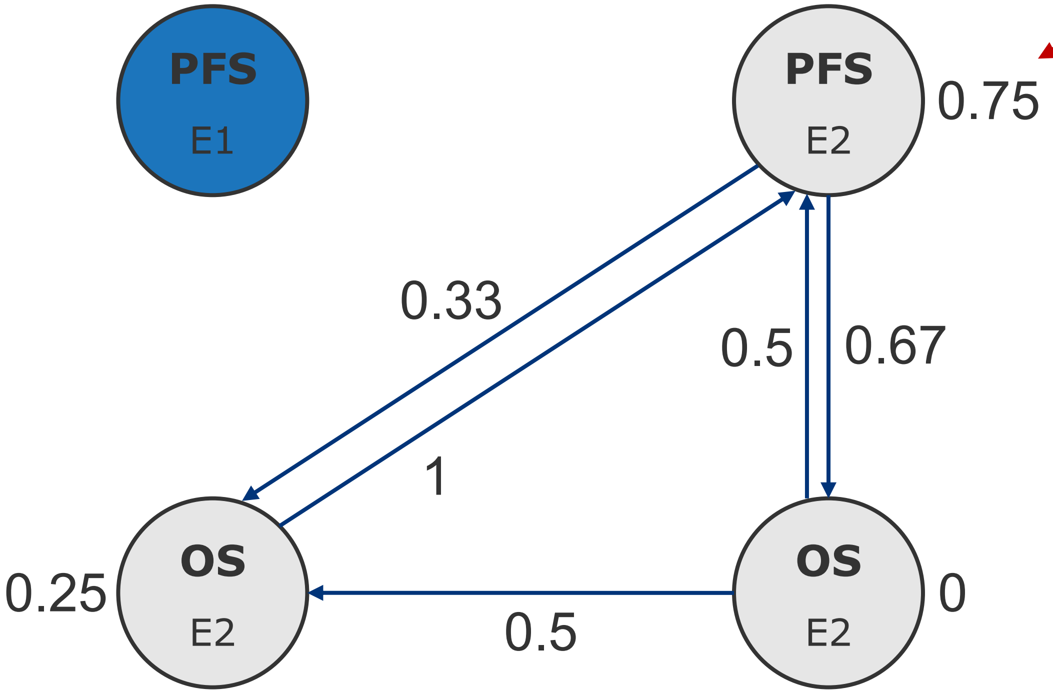




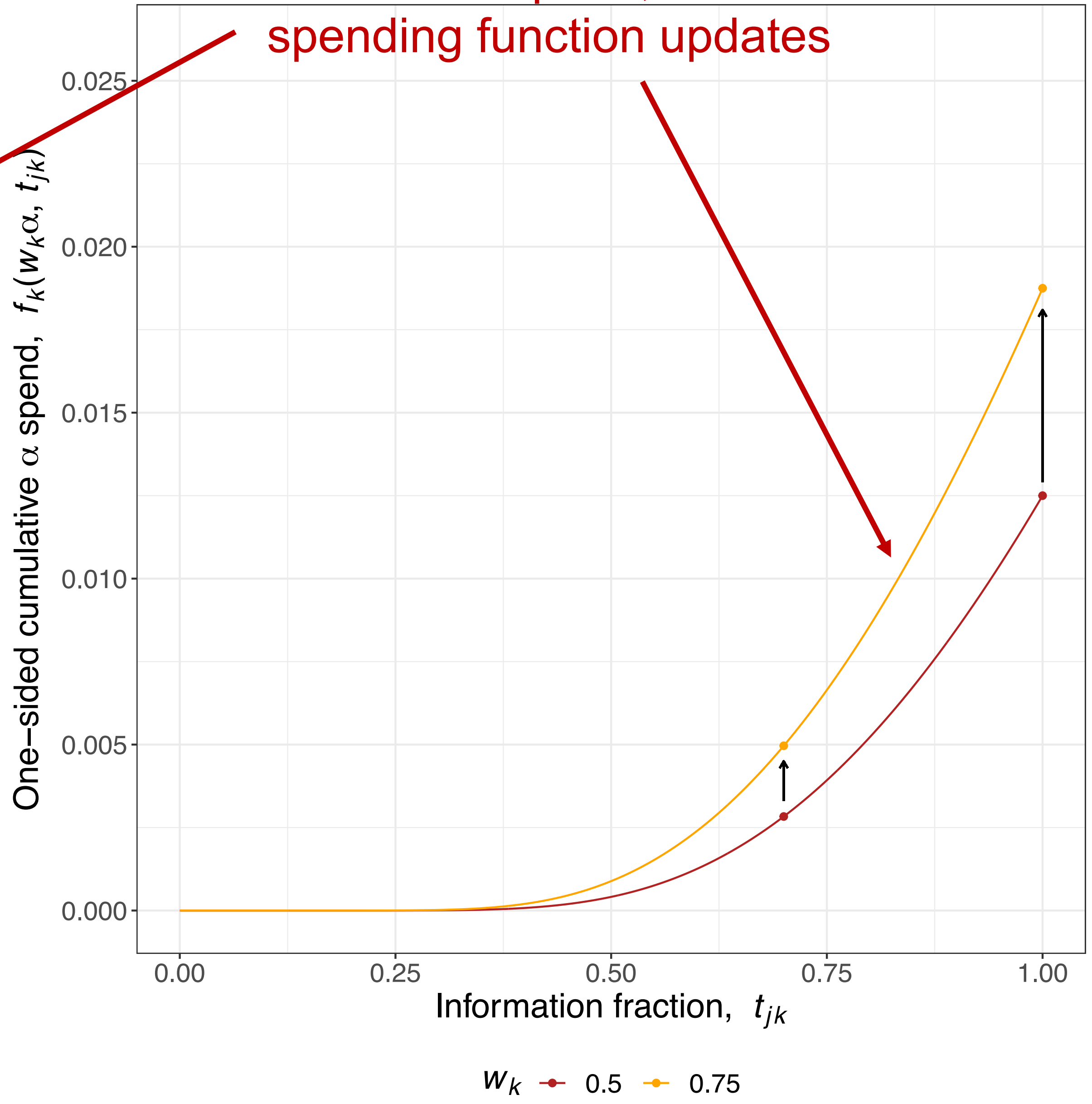
To begin with, can only spend $0.5 \times 0.025 = 0.0125$ in total

Running Example 1:

Focus on PFS for E1 vs Cntrl

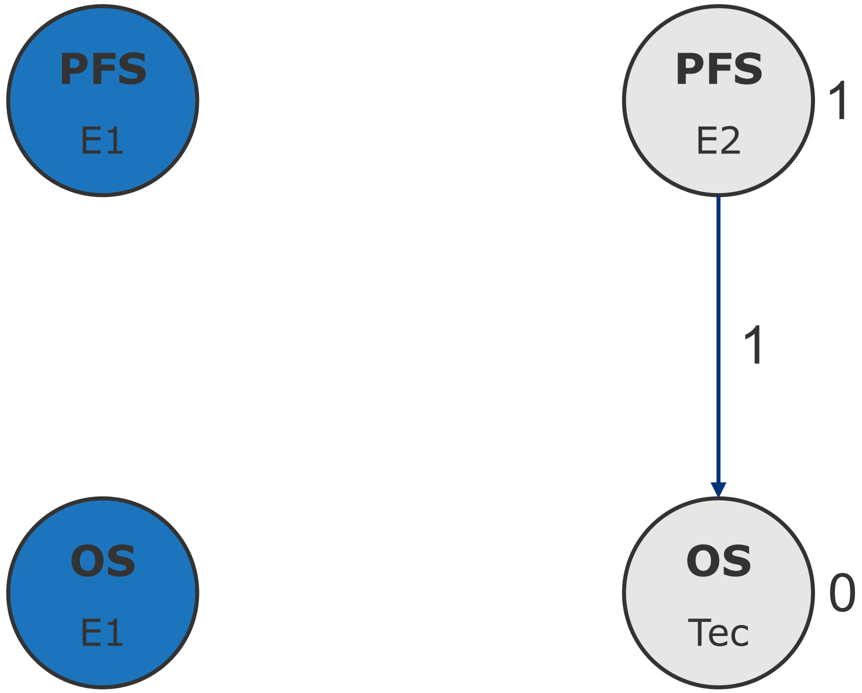


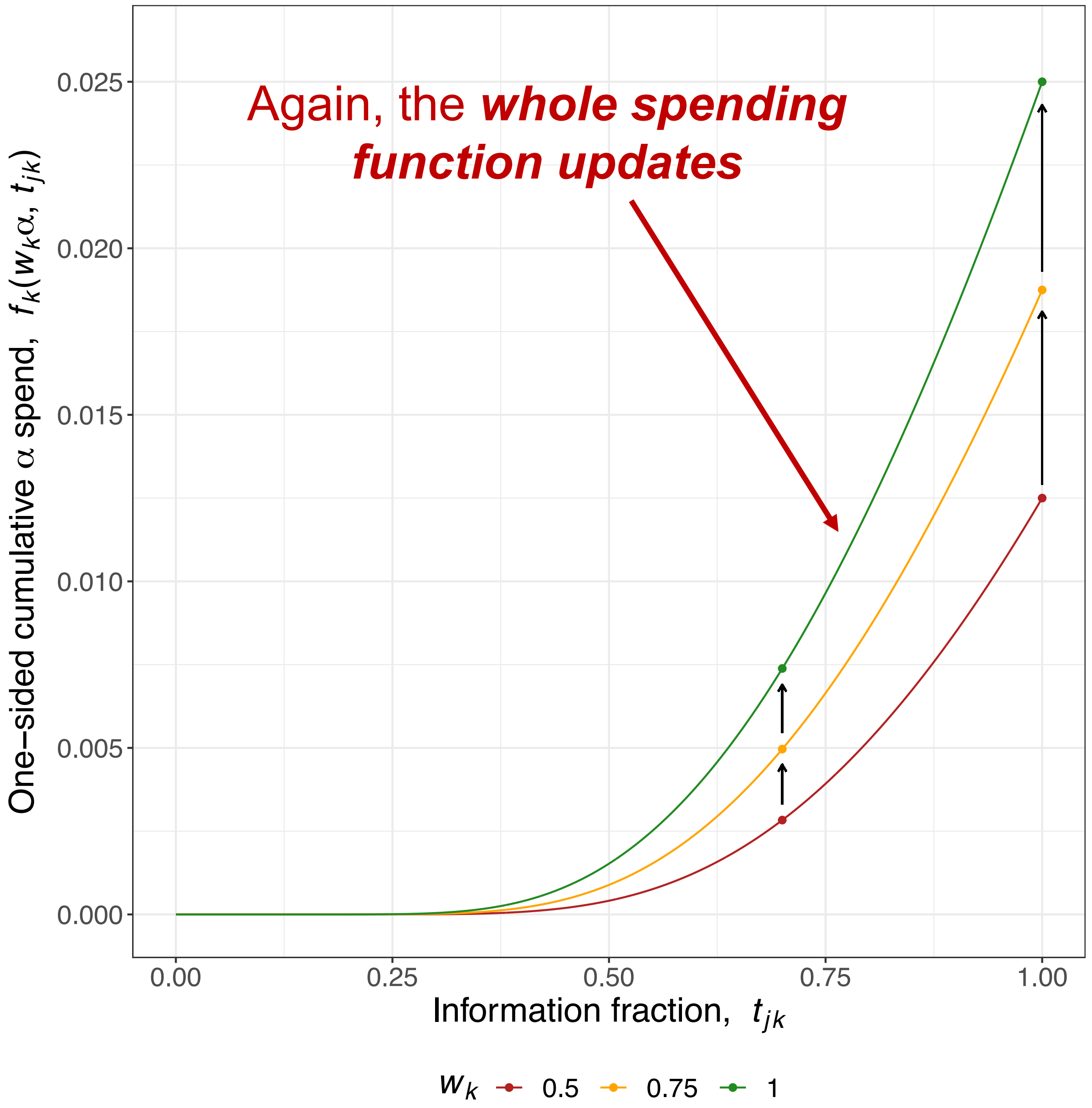
As the graph updates the allowed total spend, the whole spending function updates



Running Example 1:

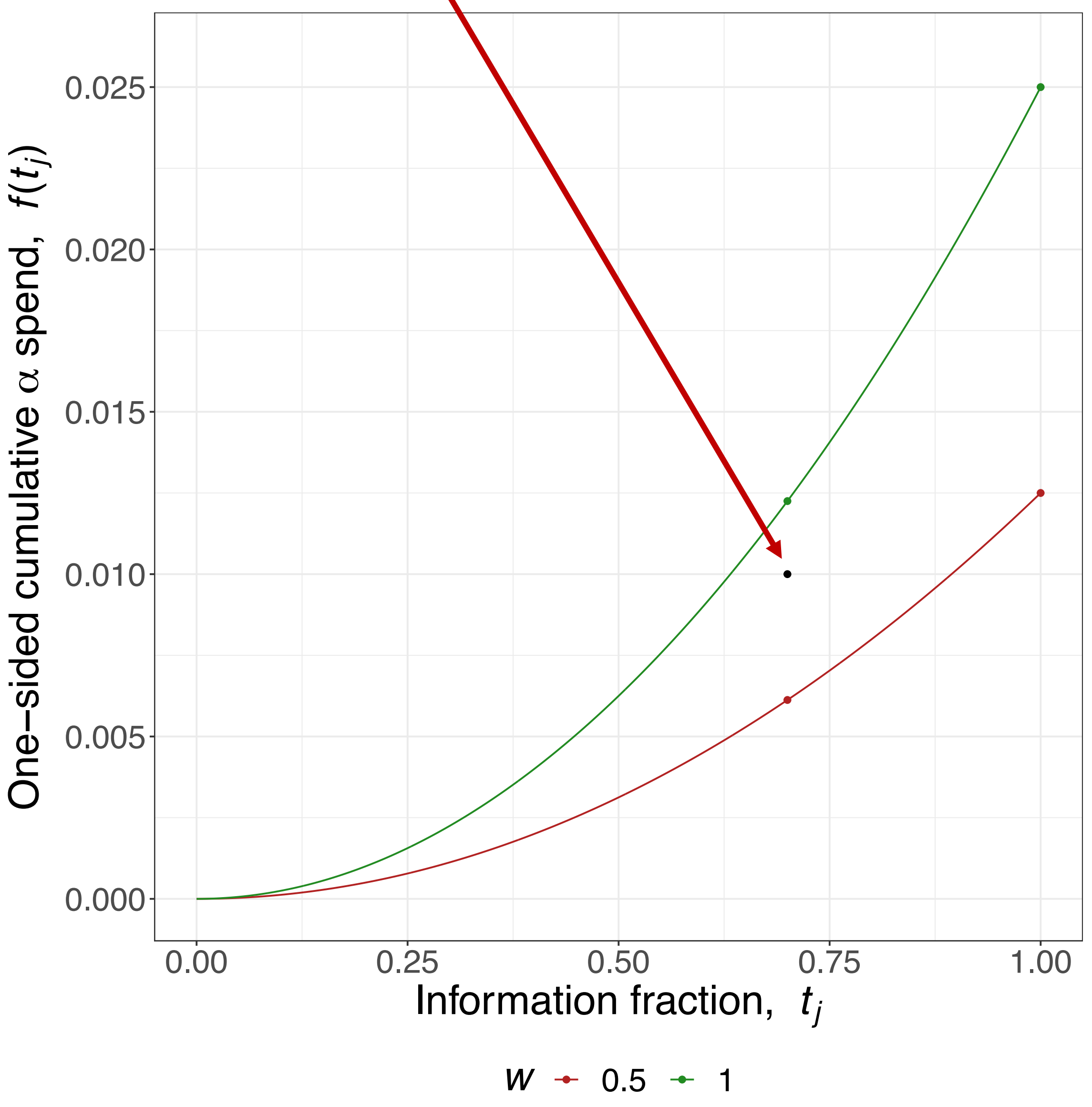
Focus on PFS for Tec vs Len




Again, the ***whole spending function updates***

‘Look back’ analyses

- The algorithm stated earlier allows for what has been termed ’look back’ analyses
- E.g., consider a simple case where there’s two possible spending function shapes, based on $w = 0.5$ or $w = 1$, and a single IA
- Suppose that at the IA we have to stay at $w = 0.5$ and so we aren’t able to reject the null based on the black dot in the plot

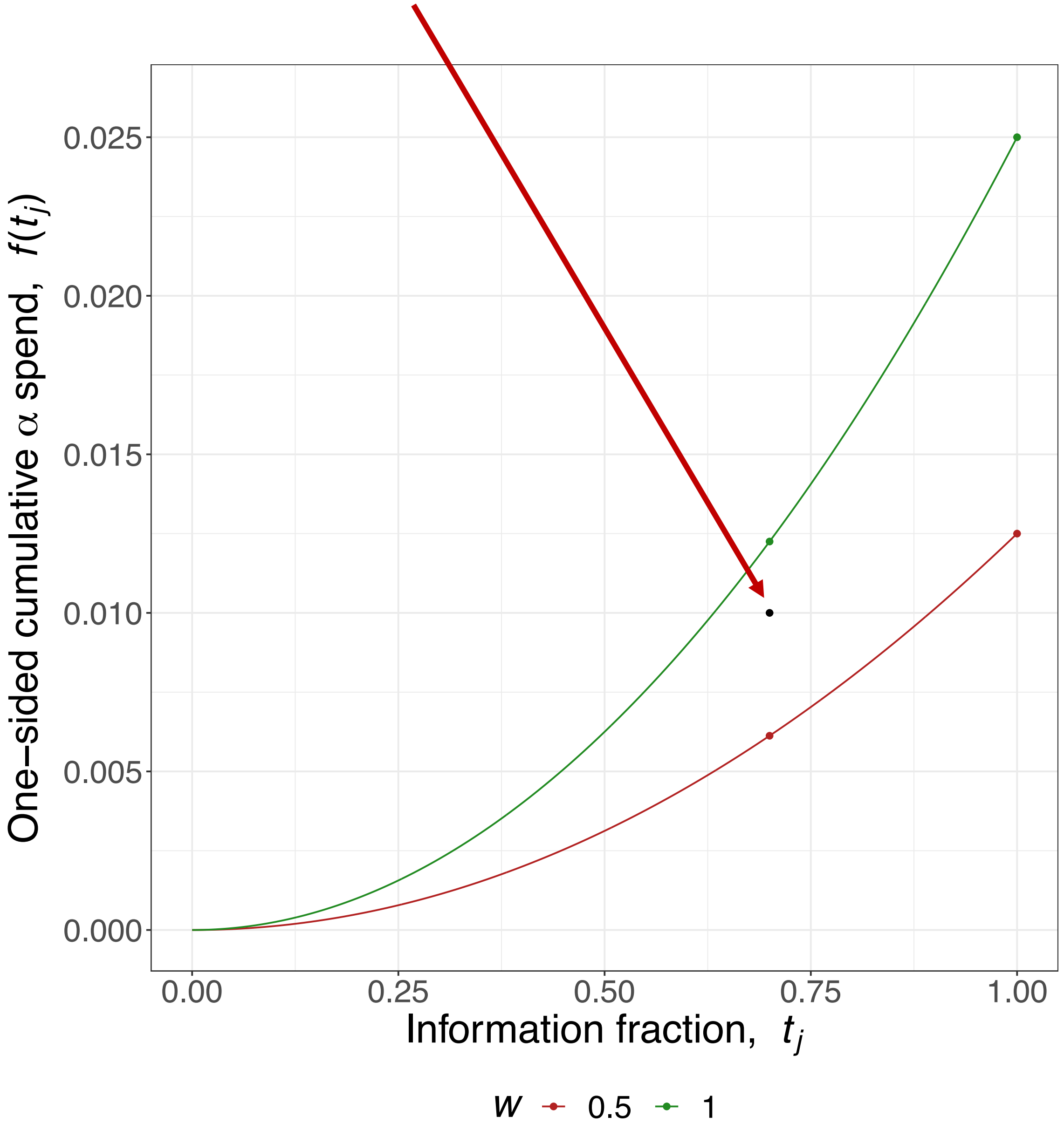

Significant if *w* = 1, but not if *w* = 0.5

‘Look back’ analyses

- If we reach $w = 1$ at the FA, we are technically allowed to ‘look back’ and claim significance for this hypothesis based on the IA p-value

- In practice, this might be a hard sell to regulators as at the FA we have more data available and still have α available for retesting this hypothesis

Significant if *w* = 1, but not if *w* = 0.5



‘Look back’ analyses

- Where this ‘look back’ is useful is if we have data that matures at different rates

- E.g., suppose there’s two hypotheses with expected IFs at three analyses of:
  - H_1: 50%, 100%, 100%
  - H_2: 33%, 67%, 100%

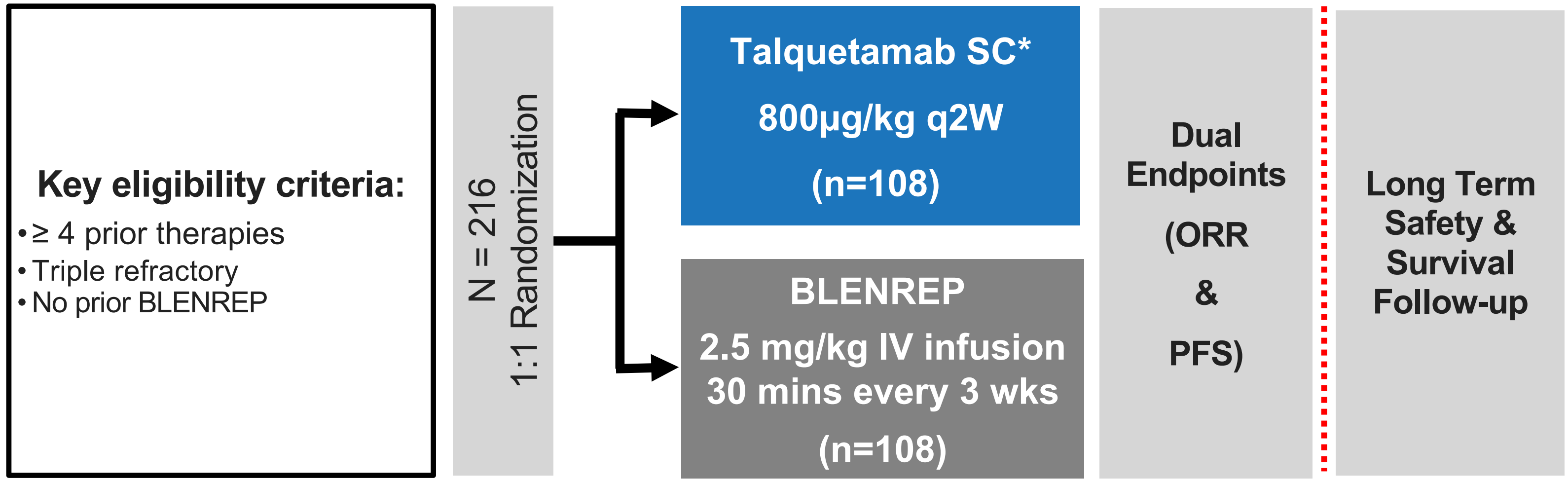
- Suppose we don’t manage to reject H_1 at IA2, and eventually reject H_2 at the FA

- Then we are allowed to retest H_1 using its IA2 p-value with the recycled α

Example: MonumenTAL-5

Tal vs Belamaf

- Phase 3 study in subjects with relapsed/refractory multiple myeloma who have received at least 4 prior lines of therapy



Key eligibility criteria:

- ≥ 4 prior therapies
- Triple refractory
- No prior BLENREP

N = 216
1:1 Randomization

Talquetamab SC*
800µg/kg q2W
(n=108)

BLENREP
2.5 mg/kg IV infusion 30 mins every 3 wks
(n=108)

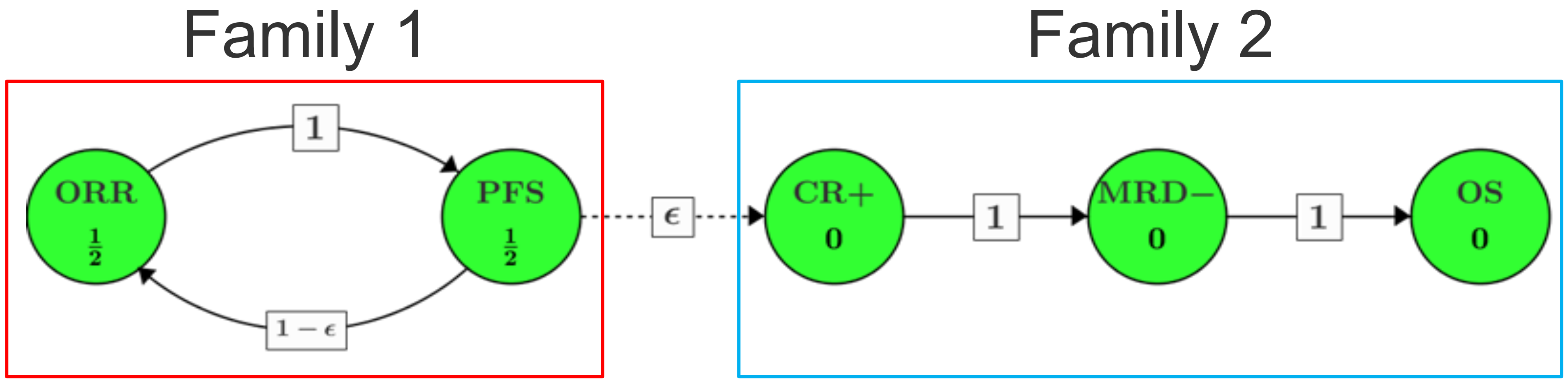
Dual Endpoints
(ORR
&
PFS)

Long Term Safety & Survival Follow-up

Example: MonumenTAL-5

Tal vs Belamaf

- Dual primary endpoints of ORR and PFS are grouped into a primary family, which serves as a gatekeeper for the second family (CR+, MRD-, OS)



Example: MonumenTAL-5

Tal vs Belamaf

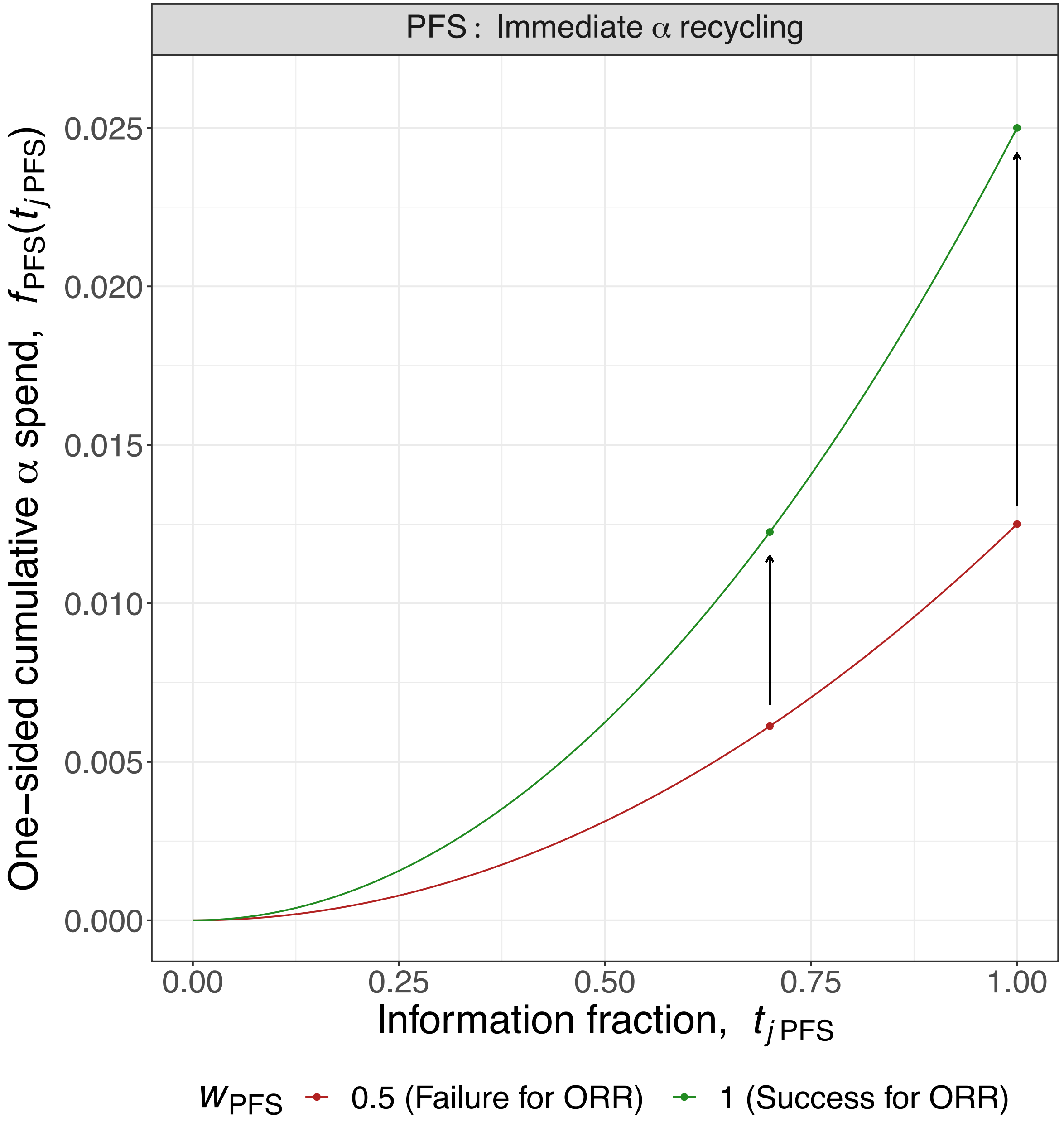
- Single IA, 3 months after the 140th participant (n = 216) is randomized
  - So ORR tested with ~140 subjects included
  - PFS expected to be tested with ~114 events

- FA for PFS when 163 events have occurred
  - At this point, ORR may be retested with ~216 subjects included

- PFS designed using KDM(2) spending function, but ORR uses a different approach to α-recycling

PFS uses immediate recycling

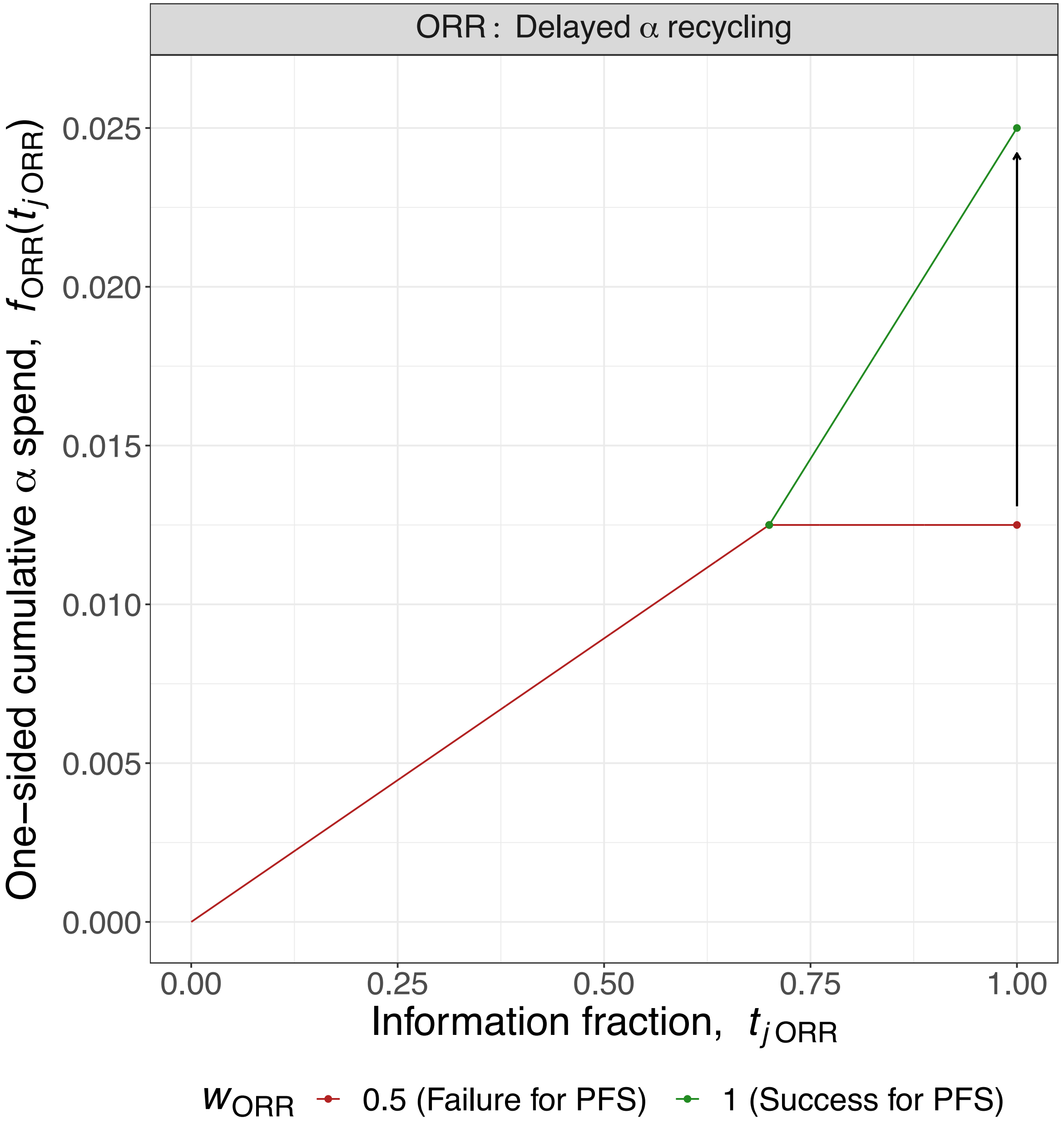
- This means that the entire spending function trajectory updates when a larger weight becomes available to PFS
- Creates an ‘issue’ that some α may be wasted if we only recycle at the FA



ORR uses delayed recycling

- Alternative, can prospectively say that additional α will only be used at the FA if more weight becomes available

- Can think of this a little bit like changing the spending function
  - vs. immediate recycling which keeps the same spending function, but just updates how much can be spent



Immediate vs. delayed recycling

Which is best?

- Depends on study specifics and objectives
- Usually, immediate recycling will be the preferred approach
  - Corresponds to the usual reason for doing a GSD: trying to increase the chance of an earlier significant result
- In the given example, delayed recycling kind of maximize currently available alpha at IA
- Also, delayed recycling may make more sense for outcomes around which there is more uncertainty about the effect or for which an early significant result is unlikely
- It’s also possible to defines recycling to begin at a certain analysis
  - E.g., recycling from analysis 3 in a trial with up to 5 analyses
  - But you cannot choose the time from which you recycle adaptively: it has to be prespecified

Protocol / SAP

What to include?

- Important to make the problem clearly defined
- So definitely specify exactly what we've discussed:
  - Initial graph
  - Spending functions / expected IFs for each GSD
    - Approach to α-recycling (immediate vs delayed)
- May also be helpful to list all associated nominal p-values based on the possible weights that the hypotheses could have
  - Becomes totally transparent what the thresholds for significance should be at each analysis

Summary

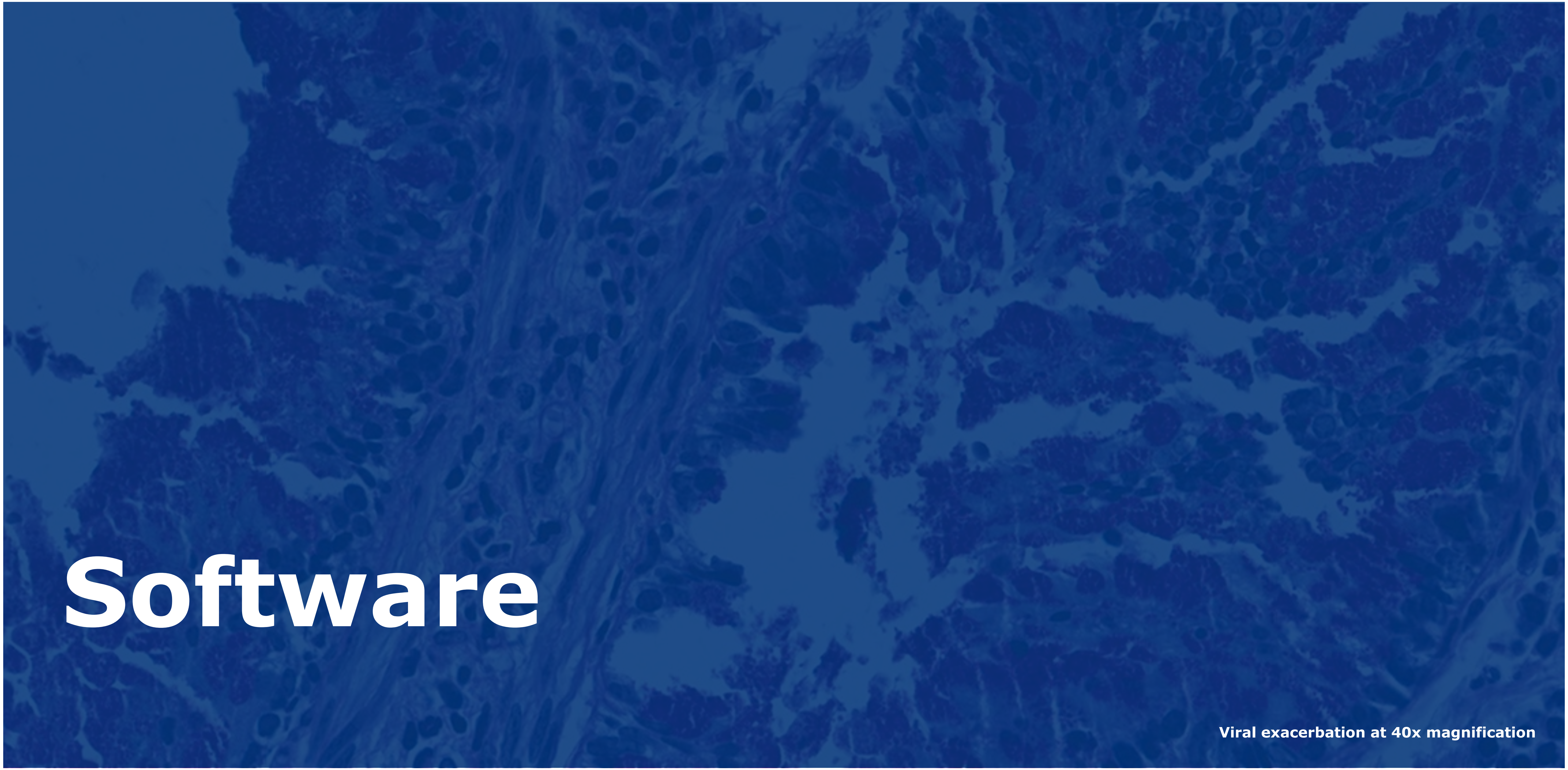
- GTPs can easily be incorporated in a GSD framework

- Specify:
  - Initial graph
  - Spending function and IFs for each hypothesis

- Tip: decouple the graph and the spending in your mind
  - The graph only tells you how much α, in total, you have to spend on a hypothesis. It tells you nothing about how it will be spent

- **I.e., it involves specifying what you would for a GTP in a fixed-sample trial and what you would for each hypothesis in a GSD**

Software



Viral exacerbation at 40x magnification

Derivation of testing boundaries

- For a simple graph, it is easy to determine all possible α levels a given hypothesis can be tested
- Becomes labor intensive / more challenging as graph complexity increases
- Tools for automation become more helpful…
- gMCPLite includes some useful functions, but has a steep learning curve
  - https://merck.github.io/gMCPLite/articles/GraphicalMultiplicity.html
- We will use some R Markdown

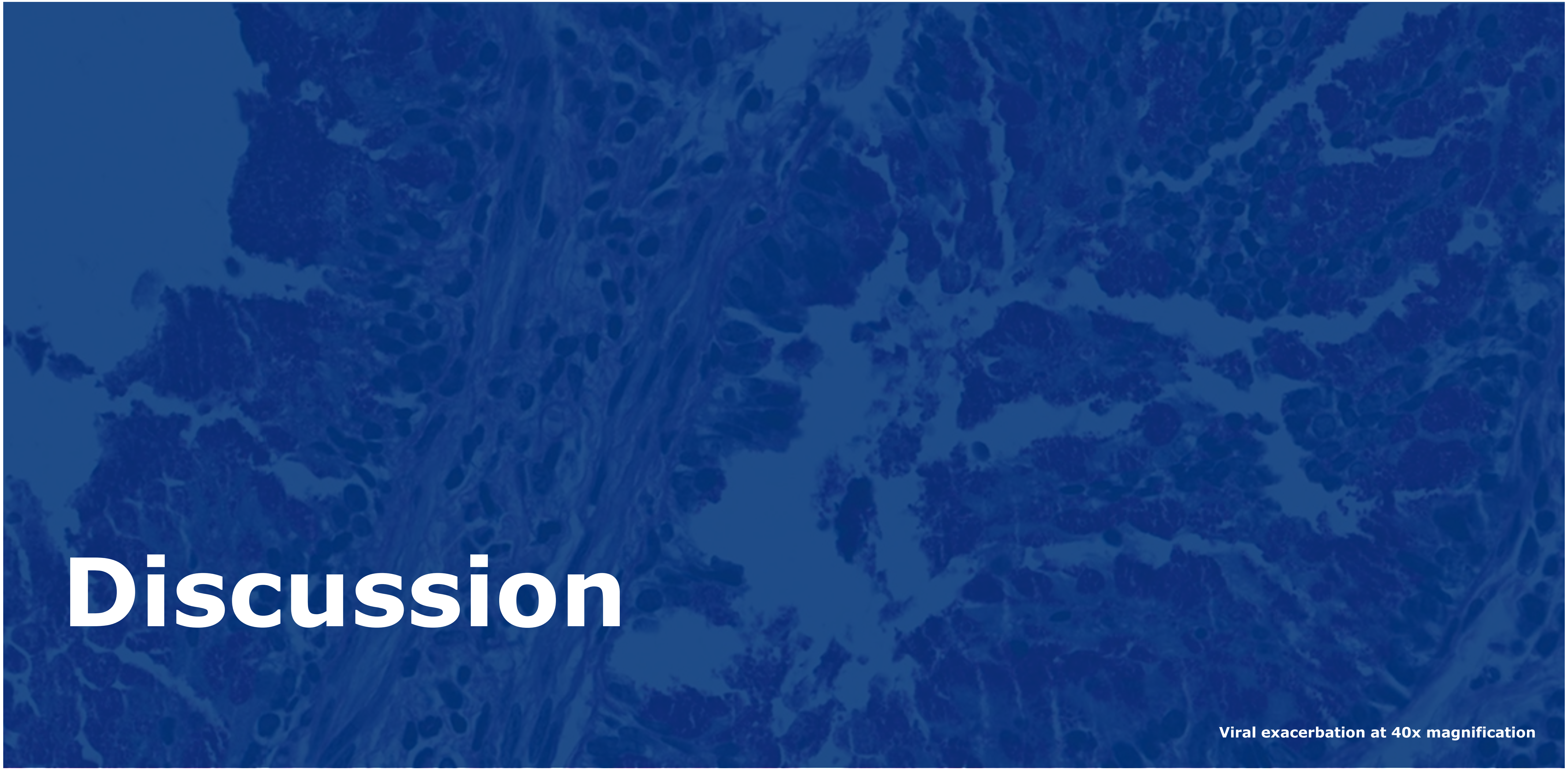
R Markdown

- Created a template that shows how we can use gsDesign and gMCP to find all possible nominal p-values
- Can download the underlying .Rmd file and the .html output
- Link here

Summary

- You can easily use standard software for computing the stopping rules under a simple graph
- For more complex graphs, if you need all the possible stopping rules then using available tools for automation can expedite things substantially
- For all graphs, certain ‘conditional powers’ are easy to get: if you need **unconditional powers**, you likely need **simulation**

Discussion



Viral exacerbation at 40x magnification

Summary

- **Approaches to testing multiple hypotheses in a GSD** framework that may seem reasonable **can inflate the FWER**

- Specialist methodology is therefore required: GTPs are such an approach, that can be readily used in a GSD setting

- We must specify:
  - **The initial graph**
  - **The GSD for each of the hypotheses in the graph**
  - (And the approach to using recycled α: immediate vs delayed)

References

Viral exacerbation at 40x magnification

References

Multiple testing procedures for GSDs

De S, Baron M (2012) Step-up and step-down methods for testing multiple hypotheses in sequential experiments. *J Stat Plan Infer* **142:**2059-70

Fu Y (2018) Step-down parametric procedures for testing correlated endpoints in a group-sequential trial. *Stat Biopharm Res* **10:**18-25

Glimm E, Maurer W, Bretz F (2010) Hierarchical testing of multiple endpoints in group-sequential trials. *Stat Med* **29:**219-28

Gou J (2020) Sample size optimization and initial allocation of the significance levels in group sequential trials with multiple endpoints. *Biom J* **64:**301-11

Hung H, Wang S, O'Neill R (2007) Statistical considerations for testing multiple endpoints in group sequential or adaptive clinical trials. *J Biopharm Stat* **17:**1201-10

Kosorok M, Yuanjun S, DeMets D (2004) Design and analysis of group sequential clinical trials with multiple primary endpoints. *Biometrics* **60:**134-45

Li H, Wang J, Luo X, Grechko J, Jennison C (2018) Improved two-stage group sequential procedures for testing a secondary endpoint after the primary endpoint achieves significance. *Biom J* **60:**893-902

Li X, Wulfsohn M, Koch G (2017) Considerations on testing secondary endpoints in group sequential design. *Stat Biopharm Res* **9:**333-7

Maurer W, Bretz F (2013) Multiple testing in group sequential trials using graphical approaches. *Stat Biopharm Res* **5:**311-20

Maurer W, Glimm E, Bretz F (2011) Multiple and repeated testing of primary, coprimary, and secondary hypotheses. *Stat Biopharm Res* **3:**336-52

Ohrn F, Niewczas J, Burman CF (2021) Improved group sequential Holm procedures for testing multiple correlated hypotheses over time. *J Biopharm Stat* **32:**230-46

Proschan M, Follmann D (2022) A note on familywise error rate for a primary and secondary endpoint. *Biometrics*

Tamhane A, Gou J, Jennison C, Mehta C, Curto T (2018) A gatekeeping procedure to test a primary and a secondary endpoint in a group sequential design with multiple interim looks. *Biometrics* **74:**40-8

Tamhane A, Mehta C, Liu L (2010) Testing a primary and a secondary endpoint in a group sequential design. *Biometrics* **66:**1174-84

Tamhane A, Xi D, Gou J (2021) Group sequential Holm and Hochberg procedures. *Stat Med* **40:**5333-50

Tang D, Gnecco C, Geller N (1989) Design of group sequential clinical trials with multiple endpoints. *J Am Stat Assoc* **84:**775-9

Xi D, Tamhane A (2015) Allocating recycled significance levels in group sequential procedures for multiple endpoints. *Biom J* **57:**90-107

Ye Y, Li A, Liu L, Yao B (2013) A group sequential Holm procedure with multiple primary endpoints. *Stat Med* **32:**1112-24

Other

Bretz F, Maurer W, Brannath W, Posch M (2009) A graphical approach to sequentially rejective multiple test procedures. *Stat Med* **28:**586-604

Hwang IK, Shih WJ, DeCani JS (1990) Group sequential designs using a family of type I error proability spending functions. *Stat Med* **9:**1439-45

Jennison C, Turnbull BW (2000) *Group sequential methods with applications to clinical trials*. Chapman & Hall: Boca Raton, FL

Kim K, DeMets DL (1987) Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* **74:**149-54

Lan KKG, DeMets DL (1983) Discrete sequential boundaries for clinical trials. *Biometrika* **70:**659-63

Marcus R, Peritz E, Gabriel KR (1976) On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63:**655-60

O'Brien PC, Fleming TR (1979) A multiple testing procedure for clinical trials. *Biometrics* **35:**549-56

Pocock SJ (1977) Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64:**191-99

Wang SK, Tsiatis AA (1987) Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43:**193-200

Extensions

- GTPs (typically) **do not make use of correlation** between test statistics
- Generally speaking we can’t use estimates of unknown correlations / it often isn’t a great idea to pre-specify guesses for unknown correlations
  - E.g., the correlation between endpoints like PFS and OS
- But using known correlations can make things more efficient
  - E.g., the correlation induced by a **shared control arm in a multi-arm trial**
- There are extensions to what’s been discussed to use such correlations
- In fact, if we need a very general testing approach, any closed testing procedure can be incorporated into a GSD framework

janssen

PHARMACEUTICAL COMPANIES OF

Johnson & Johnson