Testing in the Context of Group Sequential Designs

Yevgen Tymofyeyev andMichael Grayling

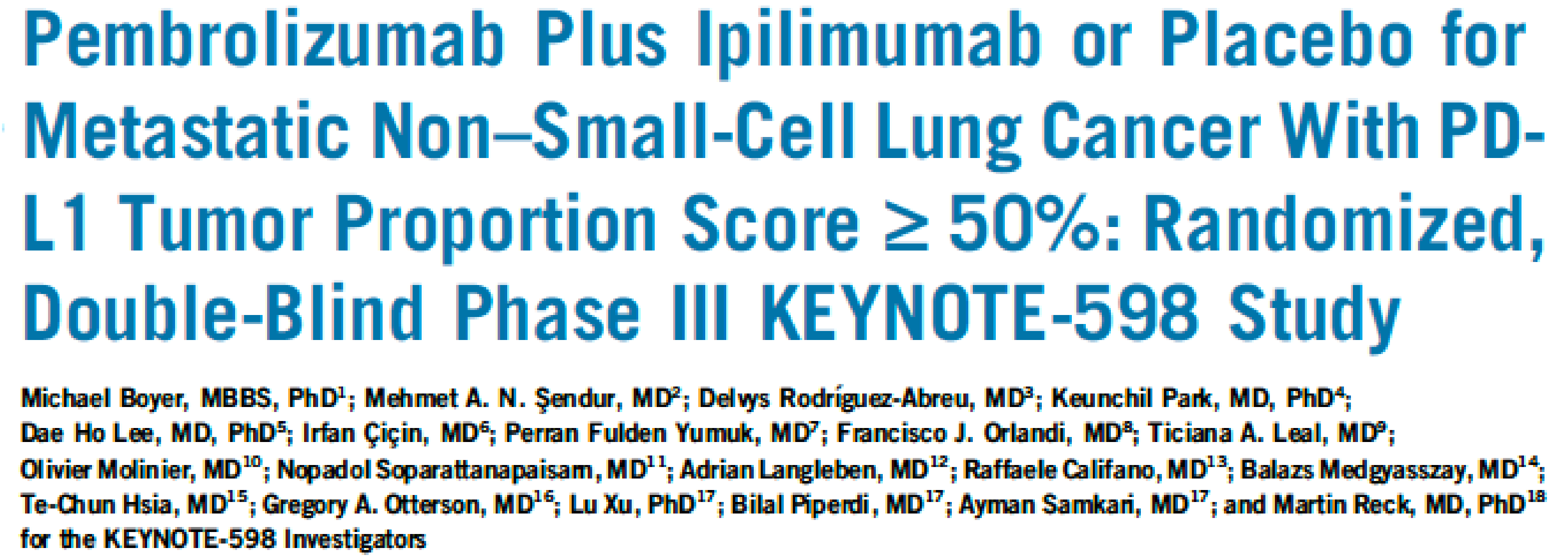
BASS- 2025

1:00-2:00pm 4 th Nov 2025

Johnson&Johnson

Running example 1: KEYNOTE-598

See /Examples/Running example 1 - KEYNOTE-598/



Pembrolizumab Plus Ipilimumab or Placebo for Metastatic Non–Small-Cell Lung Cancer With PD-L1 Tumor Proportion Score ≥ 50%: Randomized, Double-Blind Phase III KEYNOTE-598 Study

Michael Boyer, MBBS, PhD[1]; Mehmet A. N. Şendur, MD[2]; Delvys Rodríguez-Abreu, MD[3]; Keunchil Park, MD, PhD[4]; Dae Ho Lee, MD, PhD[5]; Irfan Çiçin, MD[6]; Perran Fulden Yumuk, MD[7]; Francisco J. Orlandi, MD[8]; Ticiana A. Leal, MD[9]; Olivier Molinier, MD[10]; Nopadol Soparattanapaisarn, MD[11]; Adrian Langleben, MD[12]; Raffaele Califano, MD[13]; Balazs Medgyasszay, MD[14]; Te-Chun Hsia, MD[15]; Gregory A. Otterson, MD[16]; Lu Xu, PhD[17]; Bilal Piperdi, MD[17]; Ayman Samkari, MD[17]; and Martin Reck, MD, PhD[18] for the KEYNOTE-598 Investigators

- Pembro-mono standard 1L therapy for mNSCLC with PD-L1 TPS ⩾50% without actionable driver mutations
- KEYNOTE-598 investigated whether addition of ipilimumab to pembro-mono improves efficacy
- See Boyer *et al.* (2021) for the primary results, where the protocol is also available
  - For further information, see NCT03302234

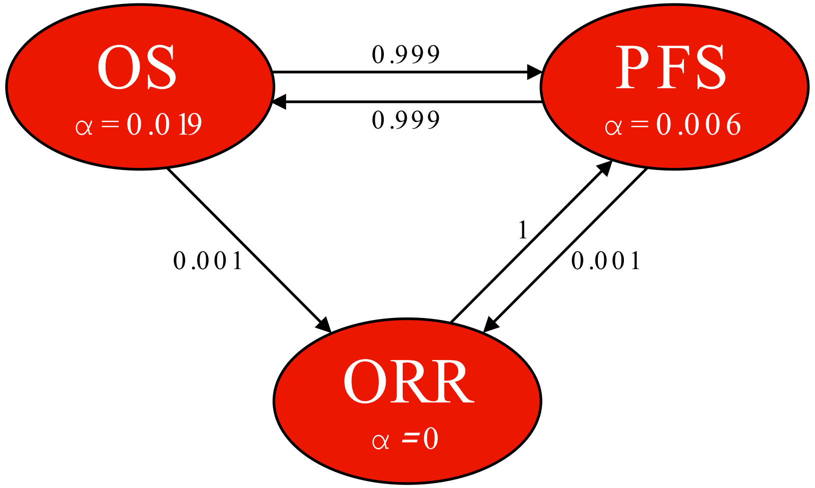
Running example 1: KEYNOTE-598

Interim analysis and multiplicity plan

- Two efficacy IAs and one FA
  - Lan-DeMets O'Brien-Fleming spending functions for OS and PFS
  - ORR matures at time of IA1

Analysis	Trigger	Primary purpose
IA1	~255 OS events	Interim PFS (~92%IF) and OS analyses (~71%IF)
IA2	~307 OS events	Final PFS analysis and interim OS analysis (~85%IF)
FA	~361OS events	Final OS analysis

- Family-wise error rate for OS, PFS, and ORR controlled in the strong sense to one-sided $\alpha = 0.025$



Running example 1: KEYNOTE-598

Efficacy boundaries and properties for PFS analyses: Table 18 in Section 8.8.2

Analysis	Value	α = 0.006	α = 0.025
IA1: 92%[a] N: 568 Events: 357 Month: ~ 32[f]	Z	2.6394	2.0667
	p (1-sided)[b]	0.0042	0.0194
	HR at bound[c]	0.7562	0.8034
	P(Cross) if HR=1[d]	0.0042	0.0194
	P(Cross) if HR=0.69[e]	0.8085	0.9251
IA2: Final PFS Analysis N: 568 Events: 389 Month: ~ 39[f]	Z	2.5869	2.0575
	p (1-sided)[b]	0.0048	0.0198
	HR at bound[c]	0.7690	0.8115
	P(Cross) if HR=1[d]	0.0060	0.0250
	P(Cross) if HR=0.69[e]	0.8692	0.9517

[a] Percentage of total planned events at each interim analysis
[b] The nominal α for testing
[c] The approximate HR required to reach an efficacy bound
[d] The probability of crossing a bound under the null hypothesis
[e] The probability of crossing a bound under the alternative hypothesis
[f] The approximate number of months since first subject randomized

Running example 2

See /Examples/Running example 2/

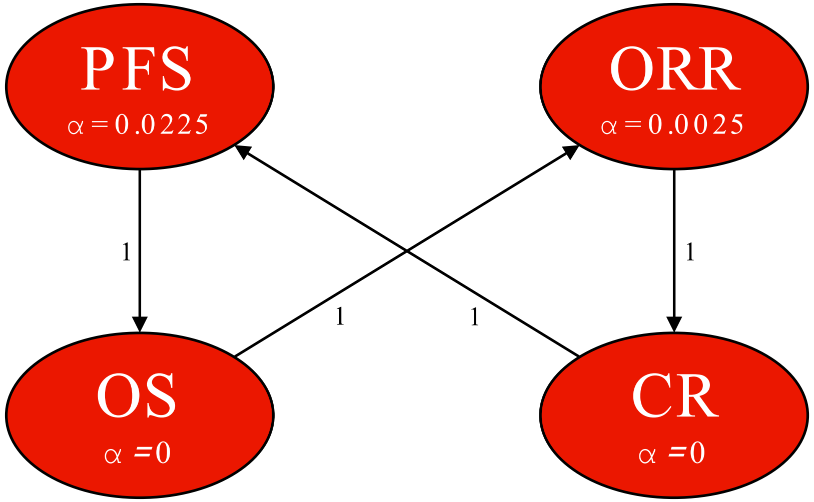
- Ph3 oncology trial, comparing CON vs TRT
- Enrollment
  - N = 450 pts, randomized 1:1
  - 10 pts/mo for months 1-2; 15 pts/mo for months 3-4; 25 pts/mo thereafter
- Four endpoints for which strong control to one-sided $\alpha = 0.025$ is ensured
  - ORR and PFS as dual primary endpoints
  - CR and OS as key secondary endpoints
- Three efficacy IAs and one FA
  - Lan-DeMets O'Brien-Fleming spending for PFS, Kim-DeMets-2 spending for OS

Analysis	Trigger	Primary purpose
IA1	6 mo after LPR	Final ORR analysis
IA2	~254 PFS events	Final PFS analysis
IA3	~211 OS events	Interim OS analysis
FA	~243 OS events	Final OS analysis

Running example 2

See /Examples/Running example 2/

- PFS
  - Analysis method: Log-rank
  - Control arm: Exponential with a median of 15 mo
  - Treatment arm: HR = 0.66
  - Drop-out: 5% per year
- ORR
  - Analysis method: Pooled comparison of proportions
  - Control arm: 50%
  - Treatment arm: Δ = 20%
- CR
  - Analysis method: Pooled comparison of proportions
  - Control arm: 30%
  - Treatment arm: Δ = 20%
- OS
  - Analysis method: Log-rank
  - Control arm: Exponential with a median of 27 mo
  - Treatment arm: HR = 0.69
  - Drop-out: 2% per year



2. Refresher on group-sequential design for a single endpoint

- Stopping rules
- Information fractions
- Choice of spending function
- {gsDesign} and {rpact}

15 min

Introduction to group-sequential design

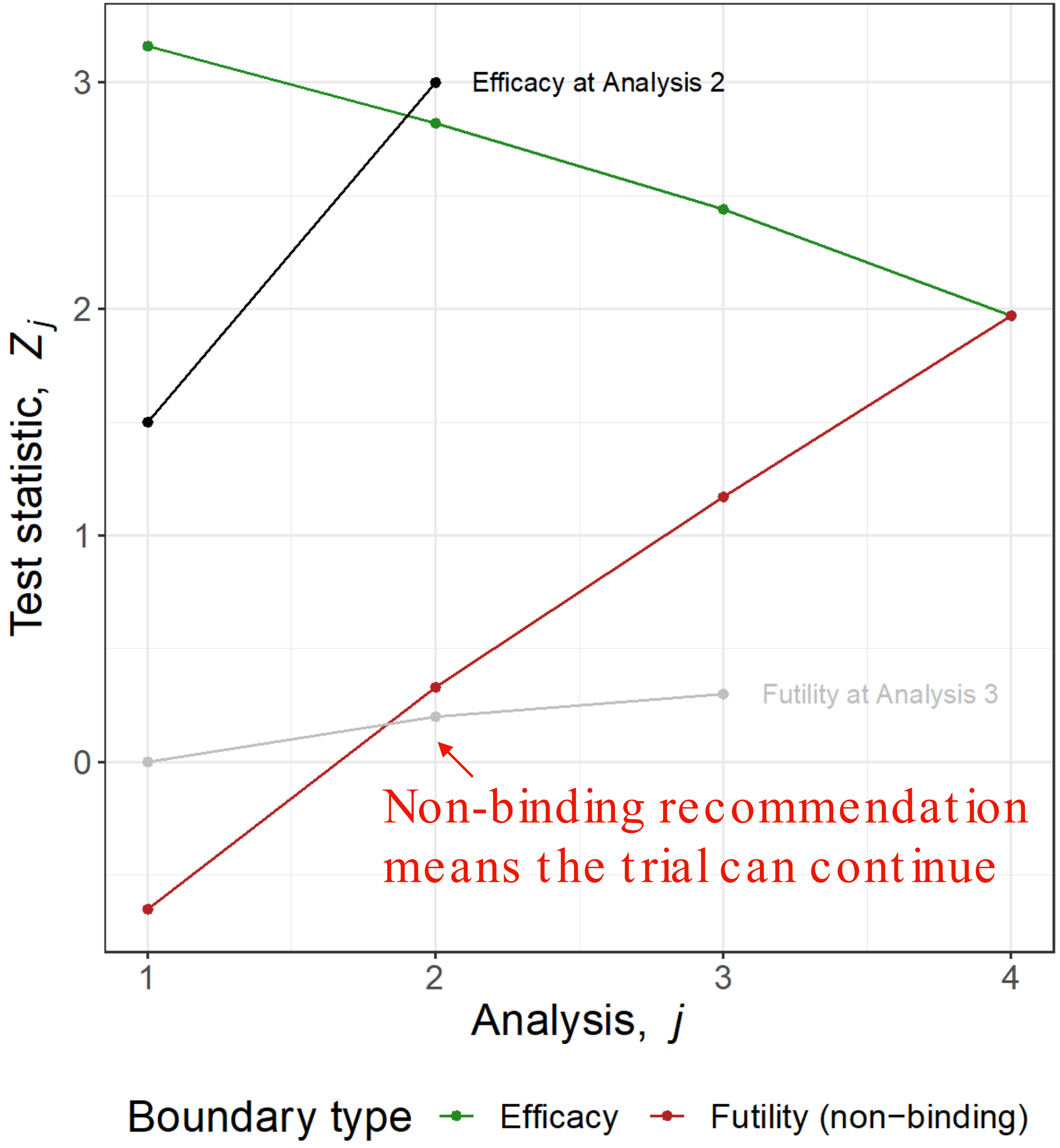
What, why, and how?

- Over-arching idea is that in many trials its likely an early decision can be made about hypotheses
  - I.e., without the maximal amount of planned data
  - GSD can therefore **reduce the expected sample size (ESS) and expected study duration (ESD)**
- Repeated testing of hypotheses will **inflate the study-wide type I and/or type II error rate** unless care is taken in the approach to assessing significance
- Extensive literature on how to choose testing rules

Stopping rules

Example for $J = 4$

- Testing rules depend on **futility bounds** $\boldsymbol{f_1, \ldots, f_J}$ **and efficacy bounds** $\boldsymbol{e_1, \ldots, e_J}$. At analysis j:
  - If $Z_j \geq e_j$, stop the trial and reject H_0
  - If $Z_j < f_j$, stop the trial and do not reject H_0 (typically a non-binding recommendation)
  - If $Z_j \in [f_j, e_j)$, continue to analysis $j + 1$
  - Common to have $f_J = e_J$, so there is a recommendation either way at the final analysis



Non-binding recommendation means the trial can continue

Error rate requirements

- We typically want to identify a design that the correct type I error rate and power
- The probability we reject H_0 for general θ is:

$$\mathbb{P}_\theta(\text{Reject } H_0) = \underbrace{\mathbb{P}_\theta(Z_1 > e_1)}_{\text{Reject at analysis 1}} + \overbrace{\mathbb{P}_\theta(Z_1 \le e_1, Z_2 > e_2)}^{\text{Reject at analysis 2}} + \cdots + \underbrace{\mathbb{P}_\theta\big(Z_1 \le e_1, Z_2 \le e_2, \ldots, Z_J > e_J\big)}_{\text{Reject at analysis } J}$$

- We therefore desire $e_1, \ldots, e_J$ and $I_1, \ldots, I_J$ such that:

$$\mathbb{P}_0(\text{Reject } H_0) \le \alpha$$

Futility rules ignored when calculating type I error rate

$$\mathbb{P}_\delta(\text{Reject } H_0) \ge 1 - \beta$$

Sometimes futility rules treated as binding when computing power; ignored here

Computing a design

- Recall $Z_j = \hat{\theta}_j\sqrt{I_j}$. Then, in a very broad range of settings
  - $(Z_1, \ldots, Z_J)$ has (approximately) a multivariate normal (MVN) distribution with
  - $\mathbb{E}(Z_j) = \theta\sqrt{I_j}$ for $j = 1, \ldots, J$
  - $\mathrm{Cov}(Z_{j_1}, Z_{j_2}) = \mathrm{Cov}(Z_{j_2}, Z_{j_1}) = \sqrt{I_{j_1}/I_{j_2}}$ for $j_1, j_2 = 1, \ldots, J,\ j_2 \geq j_1$

- We can compute $\mathbb{P}_\theta(Z_1 \leq e_1, Z_2 \leq e_2, \ldots, Z_j > e_j)$ for each j using MVN distribution function integration
- But we still need a method to set the $2J$ parameters $e_1, \ldots, e_J$ and $I_1, \ldots, I_J$
- Approach to this has evolved over time…

Functional form efficacy bounds

I.e., the original approach

- Early literature on GSDs assumes ‘equally spaced’ analyses, such that

$$I_j = \frac{jI_J}{J}$$

- Also assumed a simple form for the efficacy boundaries. Wang and Tsiatis give a unified approach
  - $e_j = C_{WT}\left(\frac{j}{J}\right)^{\Delta - 0.5}$
  - $\Delta = 0$ O’Brien and Fleming (1979)
  - $\Delta = 0.5$ Pocock (1977)
  - 1d search gives C_{WT} that controls the type I error rate
  - Additional 1d search for sample size / events to control type II error rate

J&J

Functional form efficacy bounds

Wang and Tsiatis family with parameter Δ

O'Brien-Fleming boundaries required higher evidence of efficacy to terminate earlier in the trial

Pocock boundaries are constant across analyses

Note: $\Delta = 0.389$ provides minimum ESS among 3-stage designs for 5% two-sided alpha and 80% power

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 prespecified
- Based on **information fractions (IFs)** $t_j = \frac{I_j}{I_J}$
- And non-decreasing function $f : [0,1] \rightarrow [0, \alpha]$, that gives **cumulative α spend** at IF t_j as $f\left(t_j\right)$
- Does require information level I_j **to not depend** on $\hat{\theta}_1, \dots, \hat{\theta}_{j-1}$
  - Use of interim data to update information levels requires more general methodology (p-value combination / conditional error) for strict type I error control

Running example 1: KEYNOTE-598

E.g., OS spending for $\boldsymbol{\alpha = 0.025}$

Ends at α: $f(1) = 0.025$

Three analyses

Spending functions always non-decreasing

Starts at 0: $f(0) = 0$

Error spending

Technical approach

- Iterative approach used to determine e_1, then e_2, then e_3, etc.
- Analysis 1 choose e_1 such that

$$\mathbb{P}_0(Z_1 > e_1) = f\left(I_1/I_J\right) = f(t_1)$$

- Analysis 2 choose

$$\begin{aligned}\mathbb{P}_0(Z_1 \leq e_1, Z_2 > e_2) &= f\left(I_2/I_J\right) - f\left(I_1/I_J\right)\\ &= f(t_2) - f(t_1)\end{aligned}$$

- Continue solving until reach final analysis, spending all alpha
- Method accommodates under- and over-running

Common spending functions

- Lan-DeMets O'Brien-Fleming approximation: "LDOF"

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

- Lan-DeMets Pocock approximation: "Pocock"

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

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

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

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

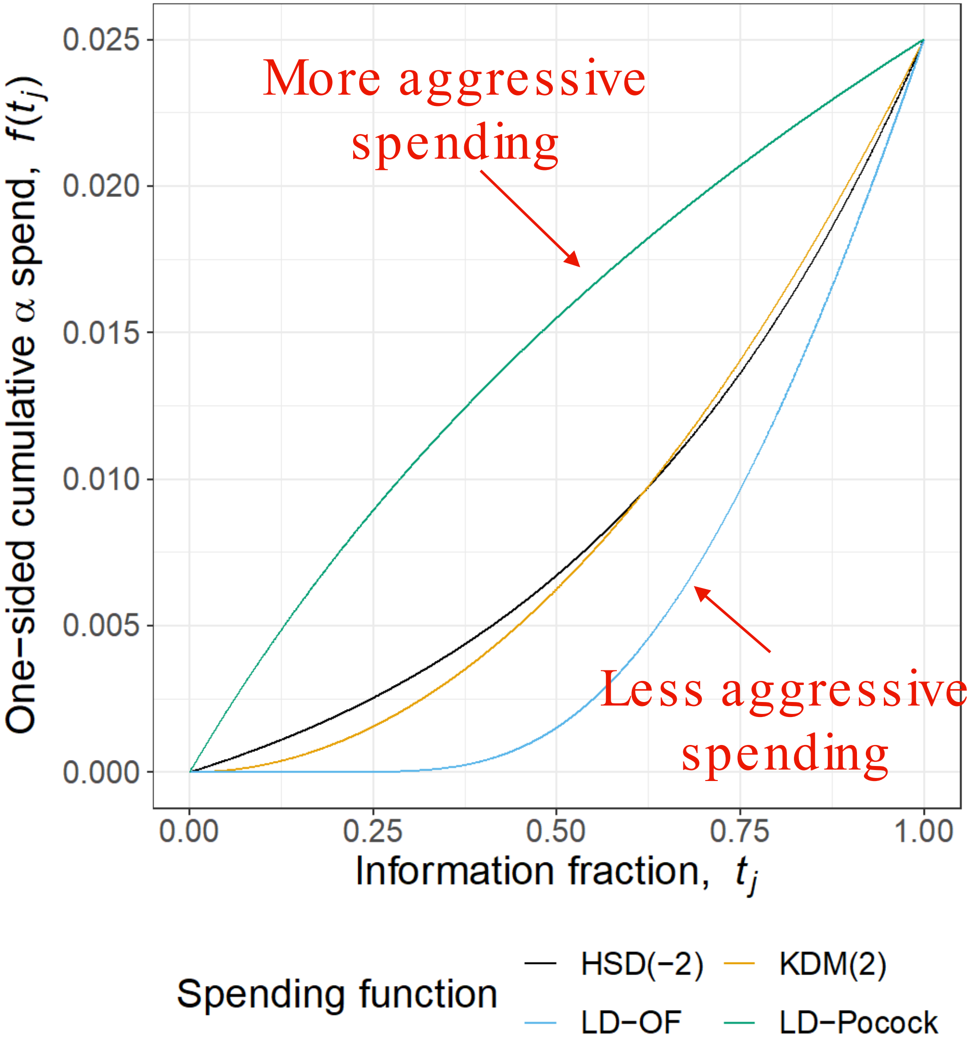
$\gamma = 1$ Similar to Pocock

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

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

$\rho = 3$ Similar to O'Brien-Fleming

$\rho = 1$ Similar to Pocock



J&J

Spending options

Speed of spending trades reduction in expected sample size (ESS) or events for lower power

- For fixed sample size / events, more aggressive spending ofα typically results in lower power
  - Maximal power = spend all α at a single final analysis
- But it will typically reduce the ESS, also expected study duration
- Alternatively, for fixed power, more aggressive spending results in larger required sample size / events
  - Often see this reflected in ‘inflation factors’ that give the ratio of the maximal information required by a GSD compared to a corresponding fixed-sample trial

Spending options

Inflation factors and ESS reduction for Wang-Tsiatis bounds

- 3-stage ($J = 3$) equally spaced analyses ($t_1 = 1/3, t_2 = 2/3$) GSD
- $\alpha = 0.025, \beta = 0.2$

Δ	Inflation factor	ESS reduction under H_1
(OF) 0.000	1.017	0.856
0.100	1.027	0.840
0.200	1.045	0.826
(Optimal) 0.389	1.103	0.811
0.400	1.105	0.811
(Pocock) 0.500	1.166	0.818

Minimizes the ESS reduction under H_1

Increasing in Δ

Spending options

Design considerations

- From a purely statistically perspective, selecting a spending function could be viewed as a multi-parameter optimization task of a multi -valued function
  - max N, expected N, power at IA, expected duration, …
  - Such globally 'optimal' designs can be a useful benchmarking approach
- In practice, truly optimal GSD rarely/never used (because of clinical/regulatory requirements), but this doesn't cost too much in terms of efficiency loss
  - Early stopping for efficacy at IA should ensure that it provides adequate evidence of the treatment effect to warrant such action
  - Regulatory agencies often discourage analyses that are "too early" and/or spend "too much" alpha
  - Under most realistic pragmatic scenarios, α spending that is more aggressive than the O'Brien-Fleming type approaches an optimal design
  - Some moderate alpha spending strategies tend to be quite robust in terms of having good operating characteristics

Software

- EAST
- 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

Summary

- GSDs seek to reduce the expected sample size / 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**
- Machinery now well established to support design
  - {gsDesign} and {rpact} in R cover most scenarios.

3. Refresher on graphical testing procedures in fixed sample designs

15 mins

With thanks to David Robertson (MRC Biostatistics Unit, University of Cambridge)

Multiple testing procedures

Why?

- Most clinical trials need to address the problem of multiple testing
- Happens because trials evaluate significance for multiple important outcomes
- Some evaluate significance for multiple treatment arms
- In any 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
  - Otherwise the probability of committing a type I error rises rapidly in the number of tests
- **Multiple testing procedures** are methods for achieving such FWER control

Running example 1: KEYNOTE-598

- Ignore the presence of interim analyses for now, and assume that

$$p_{\mathrm{OS}} = 0.0249, \qquad p_{\mathrm{PFS}} = 0.004, \qquad p_{\mathrm{ORR}} = 0.001$$

- Compare Bonferroni, Holm, Fixed sequence and Fallback
  - Fixed sequence: OS → PFS → ORR
  - Fallback: Sequence as above, with $\alpha_{\mathrm{OS}} = \alpha_{\mathrm{PFS}} = 0.012$ and $\alpha_{\mathrm{ORR}} = 0.001$

Hypothesis	Bonferroni	Holm	Fixed sequence	Fallback
OS	Not rejected	Rejected	Not rejected	Not rejected
PFS	Rejected	Rejected	Not rejected	Rejected
ORR	Rejected	Rejected	Not rejected	Rejected

Graphical testing procedures (GTPs)

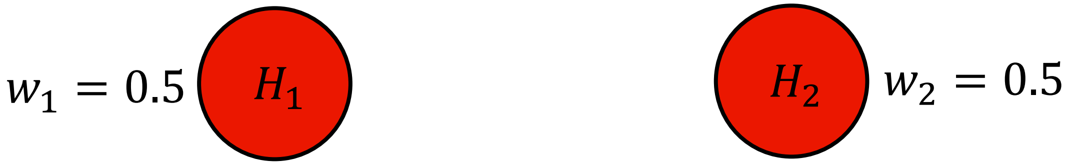
- Flexible multiple testing framework that can be **tailored to reflect the relative importance of hypotheses**
  - I.e., can deal with complex trial objectives and multiple structured hypotheses
- Built on the principle of **closed testing**
  - I.e., they can be thought of as a shortcut to specifying 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

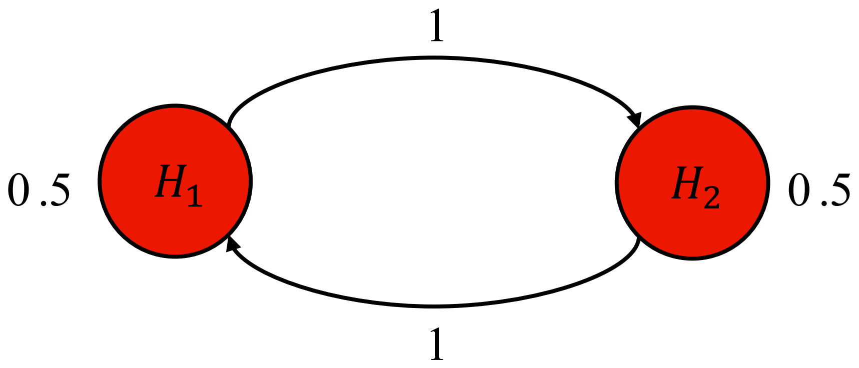
Specification

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

2. (Initial) split of significance level represented by **weights** $w_1, \dots, w_K$
   - Sometimes written in terms of $\alpha_1, \dots, \alpha_K$



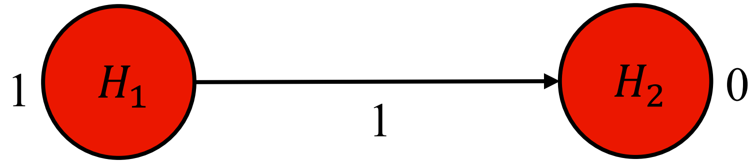
3. 'α-recycling' through **weighted directed edges**



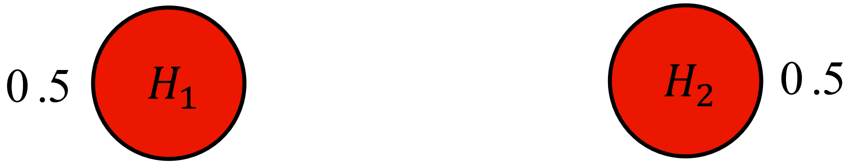
Examples

$\boldsymbol{K = 2}$

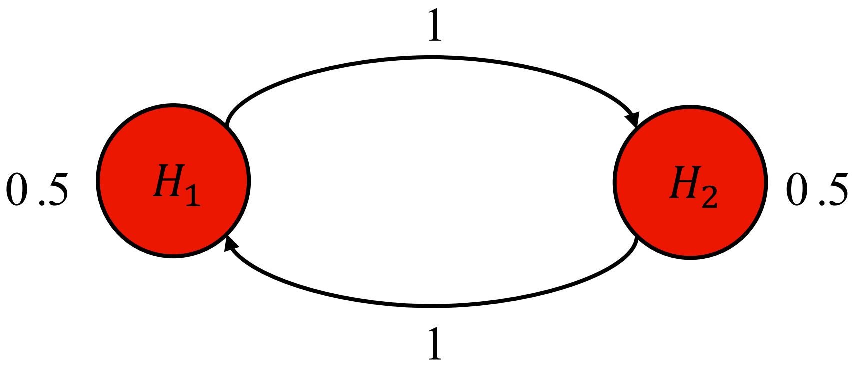
1. **Fixed sequence:** Maximises power if previous hypotheses rejected as all tests performed at level α



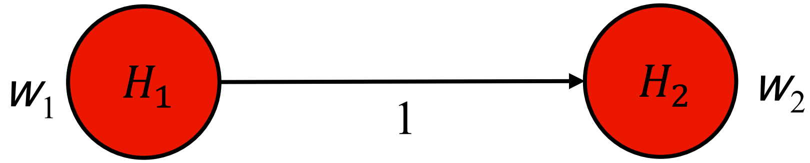
2. **Bonferroni:** No α-recycling



3. **Holm:** Everything in Bonferroni + more → more powerful



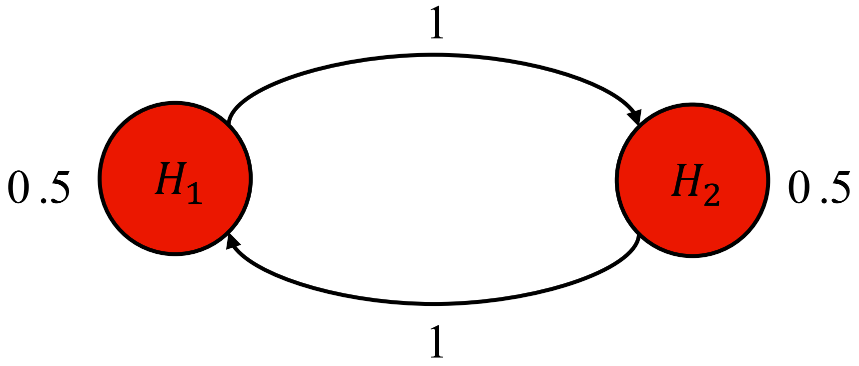
4. **Fallback**



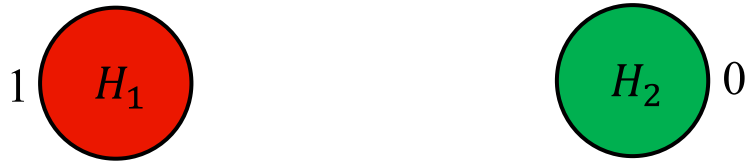
Example: Holm

$K = 2$ 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



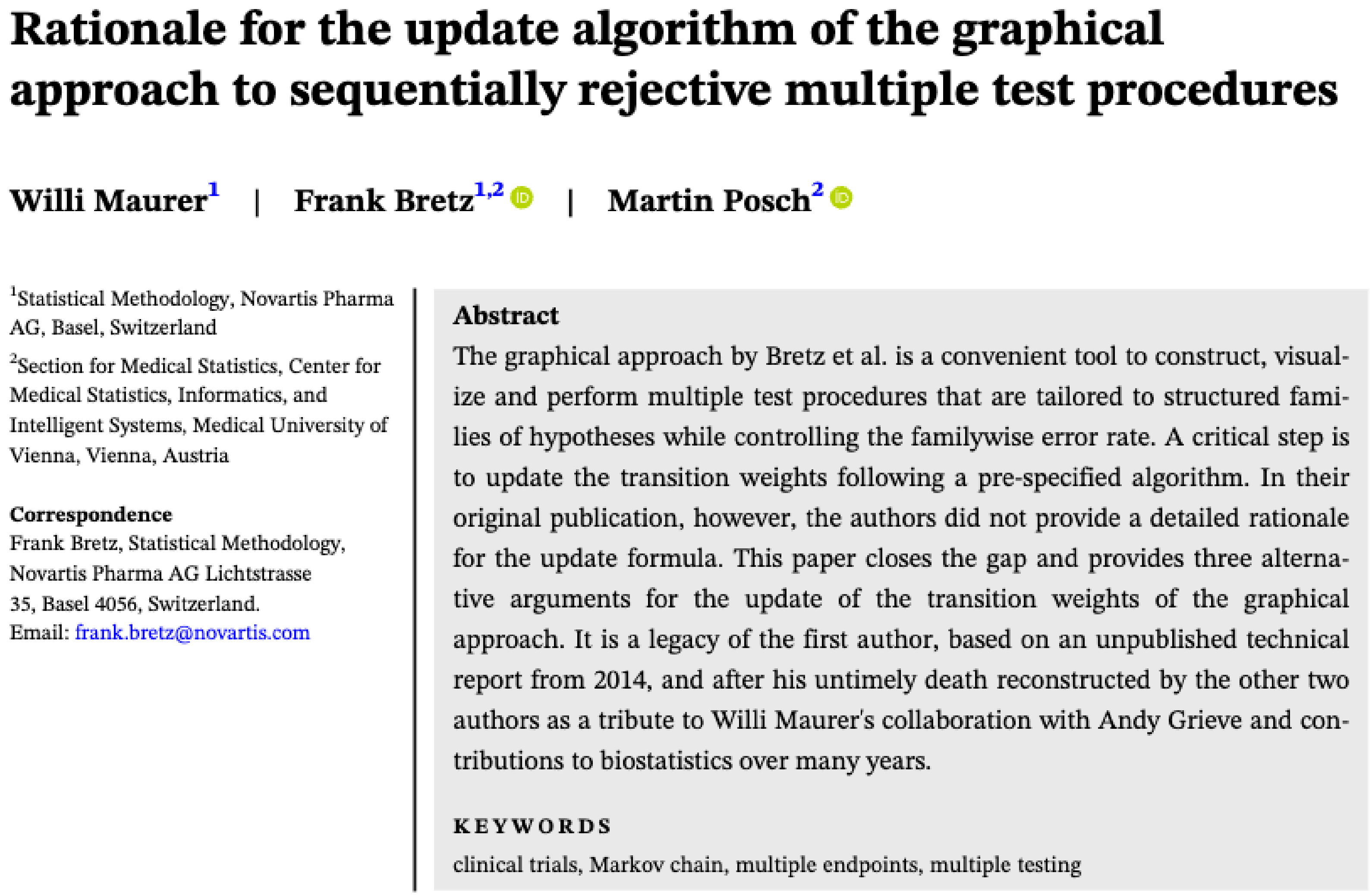
- As $p_2 = 0.01 \leq 0.0125 = 0.5(0.025) = w_2\alpha$, reject H_2 and update the graph



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

Technical basis

Graph update algorithm



Rationale for the update algorithm of the graphical approach to sequentially rejective multiple test procedures

Willi Maurer[1] | Frank Bretz[1,2] | Martin Posch[2]

[1] Statistical Methodology, Novartis Pharma AG, Basel, Switzerland

[2] Section for Medical Statistics, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria

Correspondence
Frank Bretz, Statistical Methodology,
Novartis Pharma AG Lichtstrasse
35, Basel 4056, Switzerland.
Email: frank.bretz@novartis.com

Abstract

The graphical approach by Bretz et al. is a convenient tool to construct, visualize and perform multiple test procedures that are tailored to structured families of hypotheses while controlling the familywise error rate. A critical step is to update the transition weights following a pre-specified algorithm. In their original publication, however, the authors did not provide a detailed rationale for the update formula. This paper closes the gap and provides three alternative arguments for the update of the transition weights of the graphical approach. It is a legacy of the first author, based on an unpublished technical report from 2014, and after his untimely death reconstructed by the other two authors as a tribute to Willi Maurer's collaboration with Andy Grieve and contributions to biostatistics over many years.

KEYWORDS

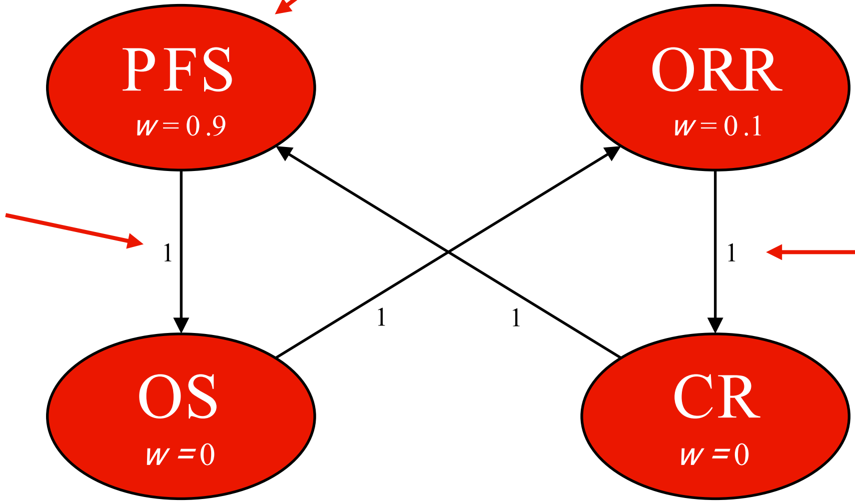
clinical trials, Markov chain, multiple endpoints, multiple testing

Running example 2

Initial graph

α split initially between the dual primary outcomes

PFS recycles all of its α to OS:
1. to maximise minimal alpha assigned to OS
2. Because less value to short term outcomes after success on PFS



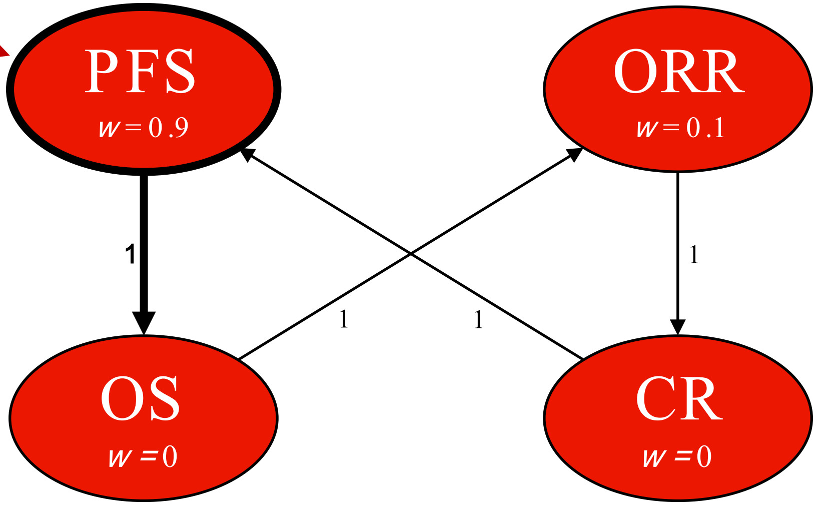
ORR recycles all its α to CR+, because of their similar maturation time

Include all edges allowed: all four hypotheses have edges totalling 1 leaving them

Running example 2

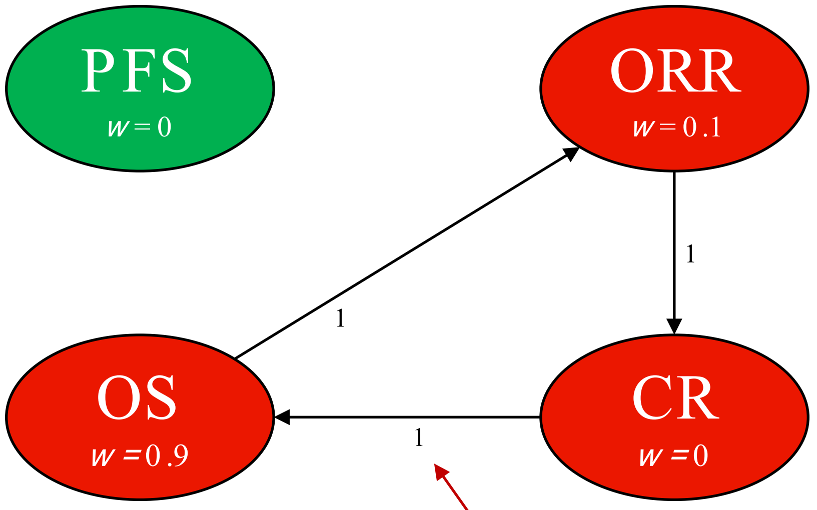
Sequential updating

Suppose we achieve significance for PFS



Running example 2

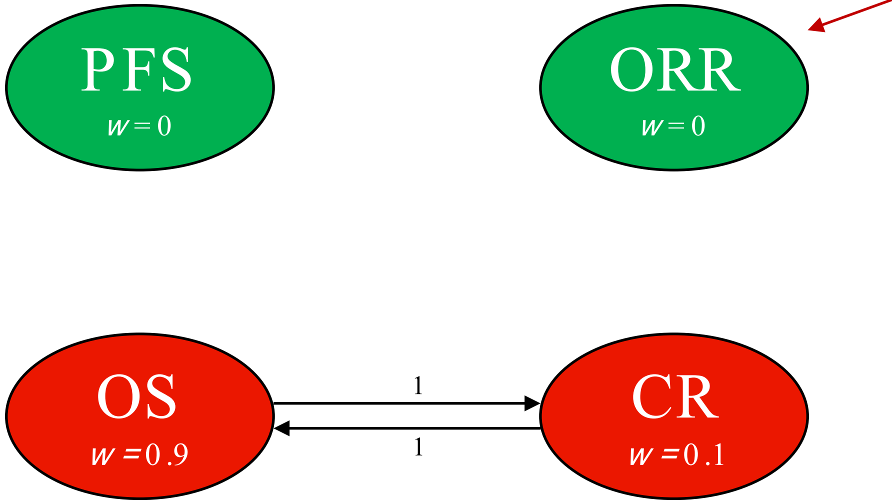
Sequential updating



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

Running example 2

Sequential updating

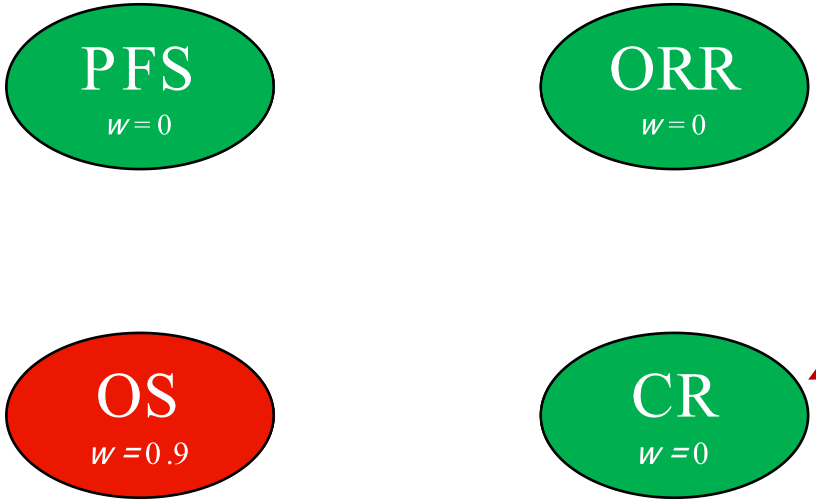


Next suppose we achieve significance for ORR

The graph would look like this regardless of whether PFS was rejected and then ORR, or ORR was rejected and then PFS

Running example 2

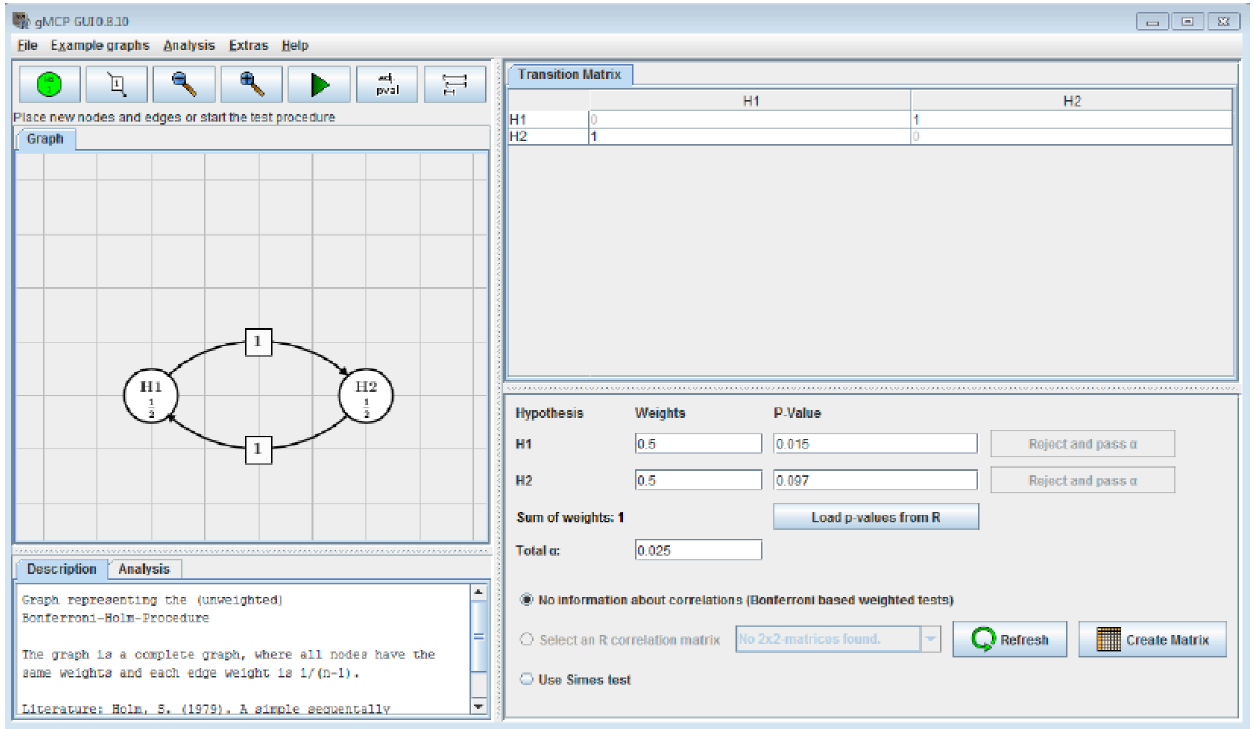
Sequential updating



Now suppose we achieve significance for CR+

Software

- R
  - {gsDesign}: Helps draw, but not evaluate graphs
  - {gMCP}: Can now be quite challenging to install. Has a GUI
  - {gMCPLite}: Will install, but no GUI
  - {graphicalMCP}: New option
- Web (R Shiny): GraphApp
- https://mrc-bsu.shinyapps.io/20MRC_BSU_GraphApp/



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 sequentially
```

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

4. Graphical testing in group-sequential designs

- Combining the graphical and group-sequential methodologies
- Analysis triggers
- ‘Look-back’ analyses
- Delayed vs immediate α-recycling

20 mins

History

- Long history of methods/applications of GSDs in clinical trials
- The same is true for multiplicity corrections such as GTPs
- However, the development of methods for correction across multiple hypotheses in a group-sequential setting has primarily occurred over the last 10-15 years
- Much of this development 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)?
- 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** $\boldsymbol{\alpha}$ (regardless of IA or FA)

Hierarchical testing of a primary and one secondary endpoint

- It was demonstrated that the naive approach does not control the FWER
- Depending on the correlation between the endpoints, FWER could be as much as 4.1%
- Therefore, specialized methodology is required for FWER control

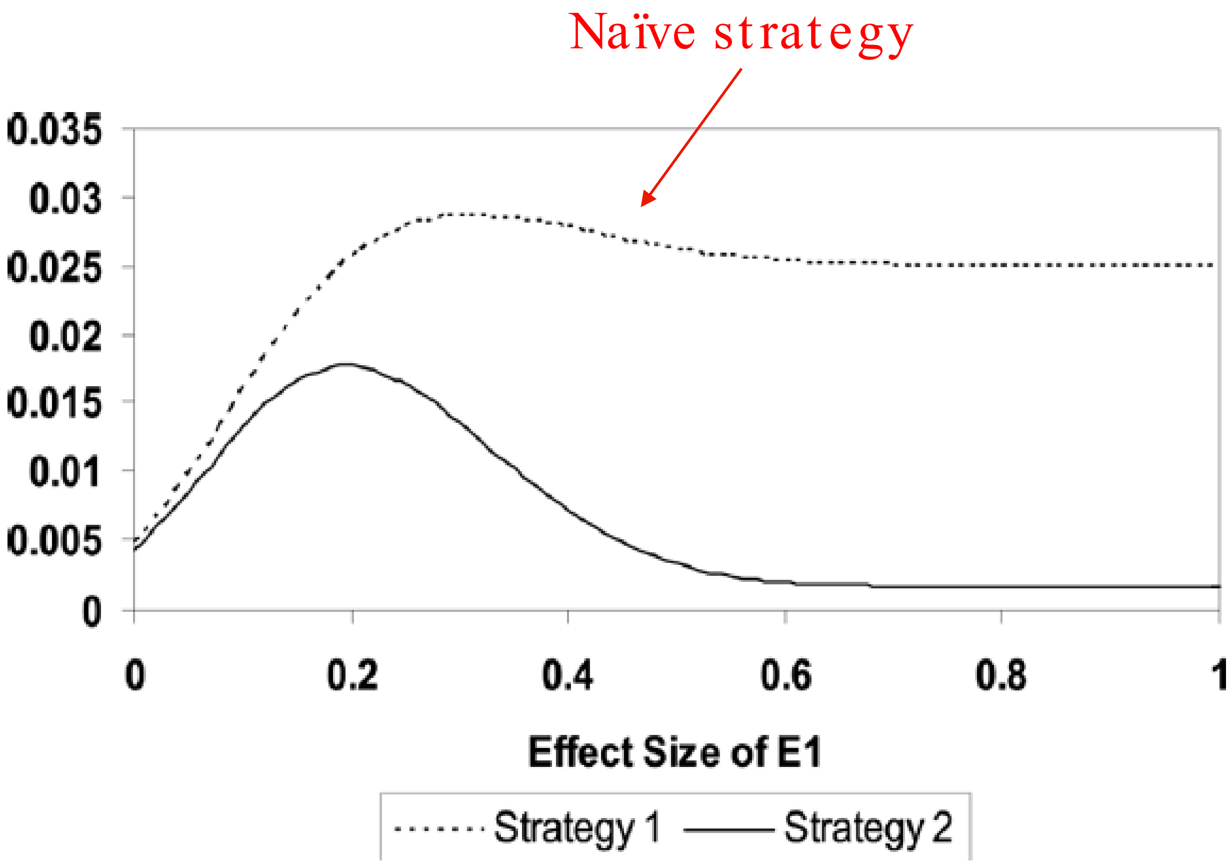


Figure 1 Type I error rate of E2 ($\rho = 0.5$).

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
  - There are some restrictions assumed in the paper that aren't necessary; with these relaxed the methodology covers vast majority of trial use cases
- **Take home message: Essentially, all you have to do is specify your initial GTP and the 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 most of the way there

“Well ordered” rejection boundaries

- Suppose we have a single hypotheses and analyses $j = 1, \dots, J$
  - Information fraction at analysis j is t_j, with $t_1 \leq t_2 \leq \cdots \leq t_J = 1$
  - **Suppose that the allowed significance level for the hypothesis is** $\boldsymbol{\gamma}$
  - Let
    - $\boldsymbol{f(\gamma, t_j)}$ **denote the spending function, with** $\boldsymbol{f(\gamma, 1) = \gamma}$
    - $p_j^*(\gamma)$ is the corresponding nominal p-value
- Need a special condition called a ‘**well ordered** ’boundary:

$$p_j^*(\gamma) \leq p_j^*(\gamma') \text{ if } \gamma \leq \gamma' \text{ for } j = 1, \dots, J$$

Defining spending function for each hypothesis

- Now, for each hypothesis $k = 1, \ldots, K$ consider:
  - Let $f_k(\gamma, t)$ denote the level-α spending function, t is information fraction
  - $f_k(\gamma, t)$ must be 'well-ordered'
  - Current hypothesis weight in GTP determines α spend for hypothesis k, $w_k\alpha$
  - Hence, the current allowed "local α", $\gamma = w_k\alpha$

Testing algorithm

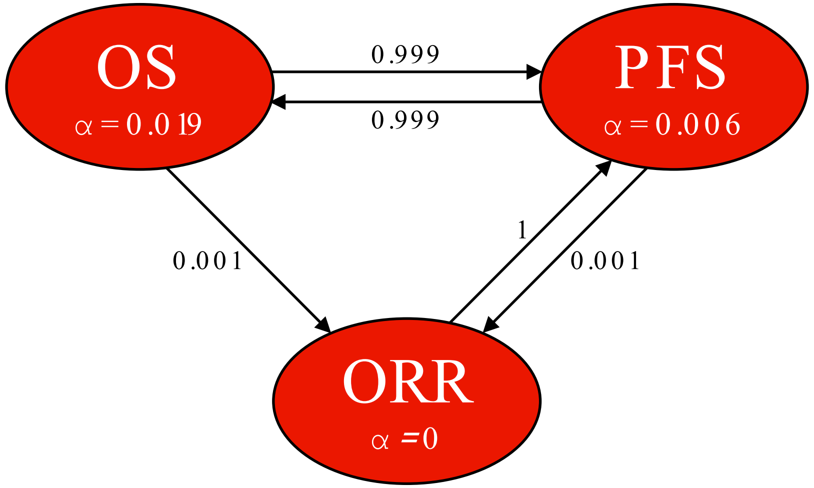
- Start with $j = 1$

1. Test each hypothesis k (not previously rejected at or before) analysis j:
   - Compute the nominal p-value threshold $p^*_{lk}(w_k\alpha)$, $l = 1, \dots, j$, based on $f_k\left(w_k\alpha, t_{jk}\right)$
   - Note: $p^*_{lk}(w_k\alpha)$ may change for some $l < j$, compared to a threshold calculated at previous IAs
   - If the nominal observed p-value $p_{lk} \leq p^*_{lk}(w_k\alpha)$ for any $l = 1, \dots, j$, reject hypothesis k
2. If any hypothesis was rejected, relocate w_k per GTP and go back to 1, otherwise move to the next analysis

Running example 1: KEYNOTE-598

Initial GTP and GSD for each hypothesis

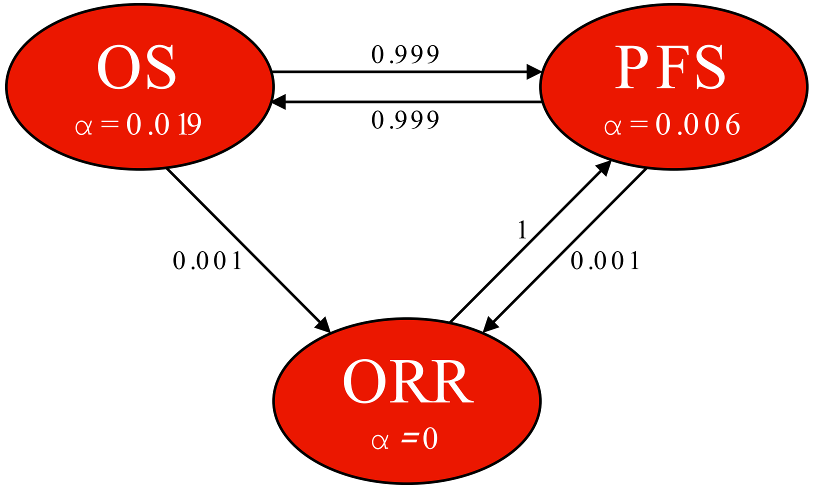
- OS
  - Two IAs at ~71%IF and ~85%IF
  - LDOF spending function
  - Initially it has alpha of 0.019 (weight of 0.76)
- PFS
  - One IA at ~92%IF
  - LDOF spending function
  - Initially it has alpha of 0.006 (weight of 0.24)
- ORR
  - No IAs
  - Initially it has weight of 0
- Overall one-sided $\alpha = 0.025$



Running example

Focus on PFS

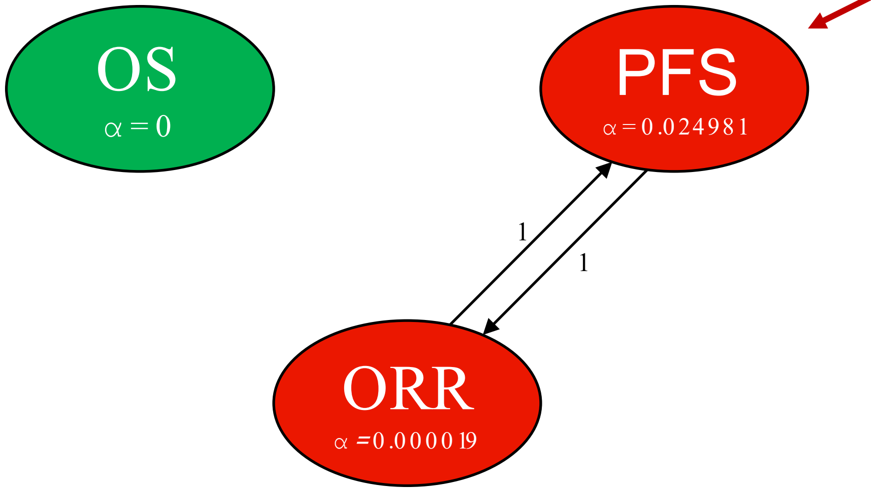
To begin with, can only spend $\alpha = 0.006$ in total



Running example

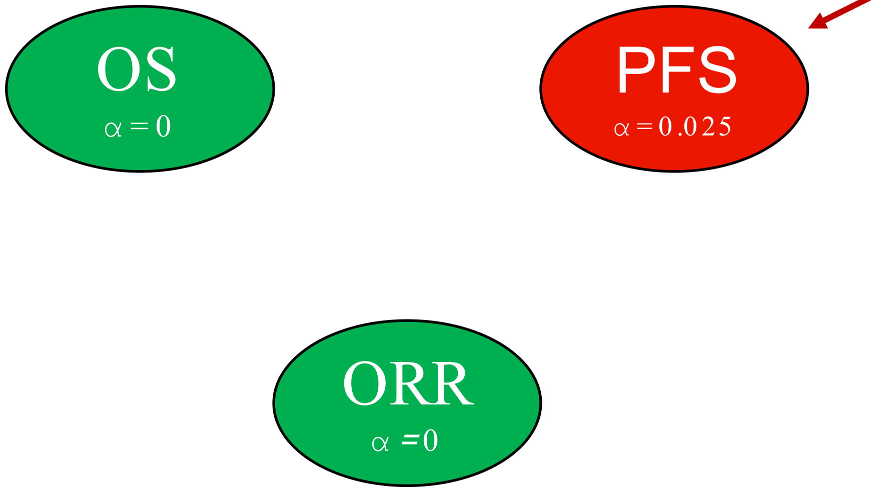
Focus on PFS

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



Running example

Focus on PFS

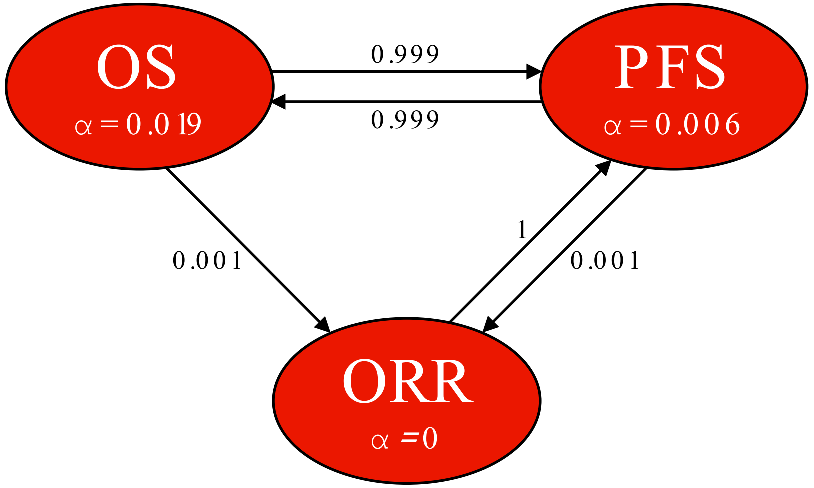


If the graph updates the allowed total spend, the whole spending function updates

Running example

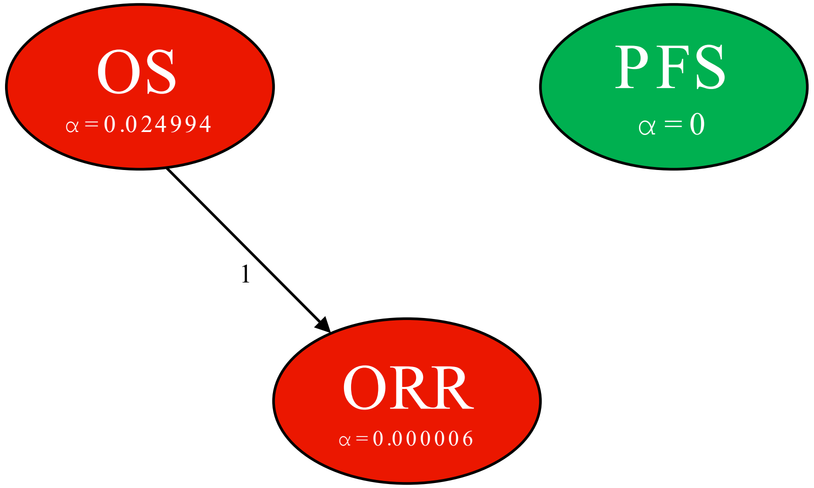
Focus on ORR

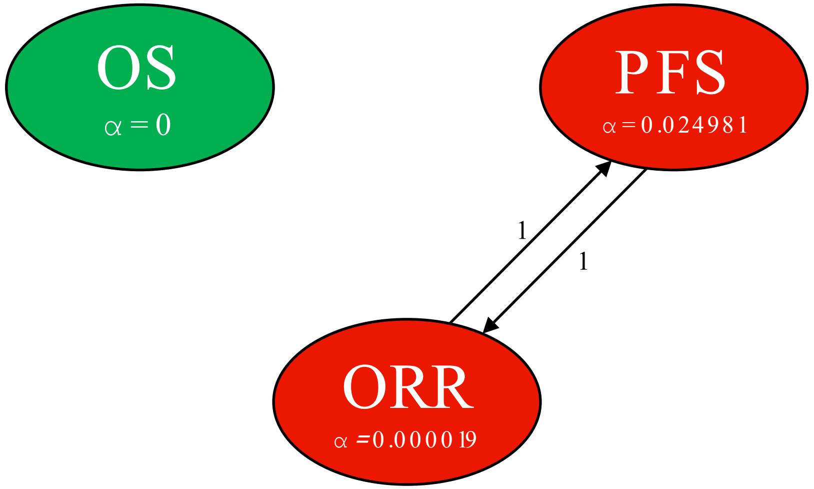
To begin with, cannot spend any α



Running example

Focus on ORR

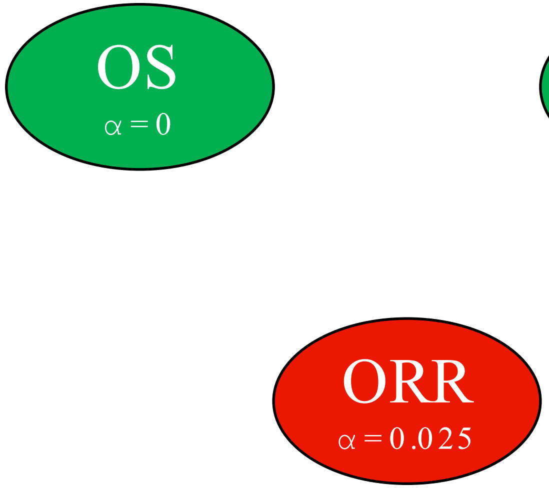




Significance on either OS or PFS would require very low *p*-value for significance

Running example

Focus on ORR



Significance on both OS or PFS allows for higher likelihood of ORR success

‘Look back’ analyses

Significant if $w_{\text{PFS}} = 1$,
but not if $w_{\text{PFS}} = 0.24$

- The algorithm allows for what has been termed ‘look back’ analyses
- E.g., consider PFS in the KEYNOTE-598 example
- Suppose that at IA1 we have to stay at $w_{\text{PFS}} = 0.24$ (because OS wasn’t rejected). Then we aren’t able to reject H_{PFS} based on the black dot in the plot

‘Look back’ analyses

Significant if $w_{\text{PFS}} = 1$,
but not if $w_{\text{PFS}} = 0.24$

- If we reach $w_{\text{PFS}} = 1$ at PFS’s FA, we are technically allowed to ‘look back’ and claim significance for this hypothesis based on the IA1 result
- 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
- It usually shouldn’t matter, provided there isn’t a strong trend in the treatment effect
  - It would only lead to a gain in power if the PFS test statistic is below the orange dot at its FA

‘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
- This is the case for ORR in Running example 1: KEYNOTE-598

Immediate recycling

The approach for PFS in Running example 1: KEYNOTE-598

- 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

J&J

Delayed recycling

- A way around this α wasting is to prospectively say that additional α will only be used at the FA if more weight becomes available
- Can think of this 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?

- 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
- Delayed recycling does the opposite: by pushing spend later in the trial it increases power, at the cost of expecting significance to occur later
- 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 define 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

Changing the spending function

What and why?

- Could alternatively think of delayed recycling as a particular case of changing the spending function after recycling
  - Changing it to delay recycling as much as possible
- May change the spending function to recycle more alpha earlier
  - Recall what we said earlier about the 'well ordered' boundary requirement: **this needs to be checked!**
- Makes sense when after success on one hypothesis, the value of another diminishes over time
  - E.g. 1, three-arm design where need significance on both experimental arms at some time
  - E.g. 2, short term outcome value reduced after success on conventional endpoint
- E.g., PFS switching from LDOF to KDM(1) in Running example 1: KEYNOTE-598

J&J

Running example 1: KEYNOTE-598

Example implementation in practice: IA1

Accrued events / sample size for each hypothesis

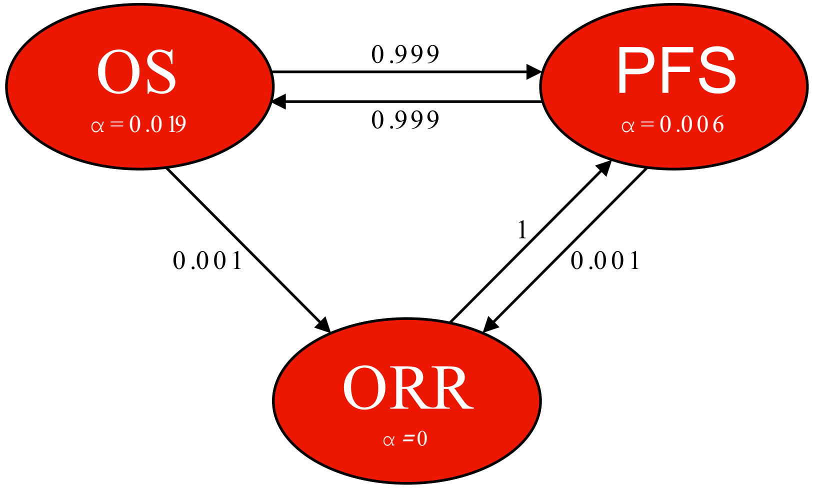
Analysis	OS	PFS	ORR
IA1	255	356	568
IA2			
FA			

Test statistic for each hypothesis

Analysis	OS	PFS	ORR
IA1	0.016	0.006	0.009
IA2			
FA			

Current efficacy boundaries (p -value scale)

Current graphical testing procedure



Running example 1: KEYNOTE-598

Example implementation in practice: IA2

Accrued events / sample size for each hypothesis

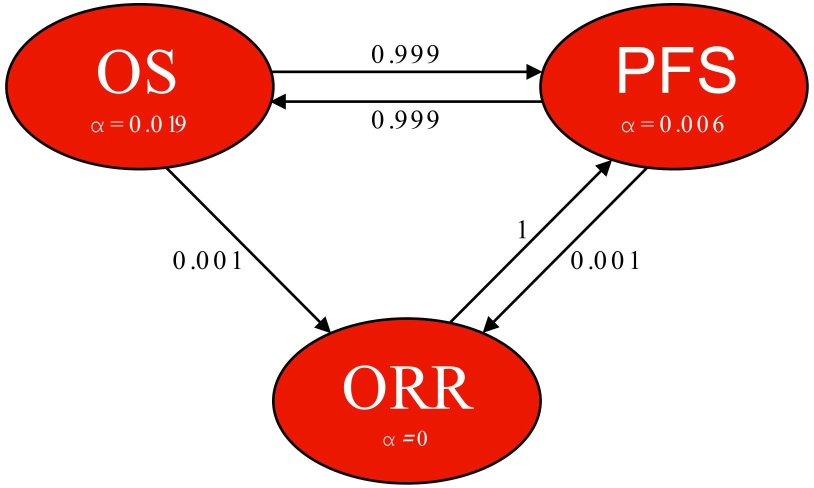
Analysis	OS	PFS	ORR
IA1	255	356	568
IA2	307	388	-
FA			

Test statistic for each hypothesis

Analysis	OS	PFS	ORR
IA1	0.016	0.006	0.009
IA2	0.014	0.003	-
FA			

Current efficacy boundaries (p -value scale)

Current graphical testing procedure



Running example 1: KEYNOTE-598

Example implementation in practice: PFS is rejected and the graph updates, which triggers updating the efficacy boundaries

Accrued events / sample size for each hypothesis

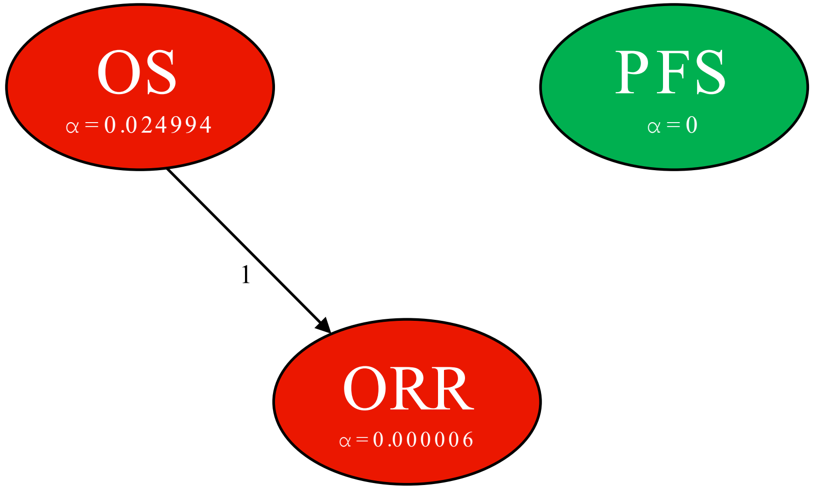
Analysis	OS	PFS	ORR
IA1	255	356	568
IA2	307	388	-
FA			

Test statistic for each hypothesis

Analysis	OS	PFS	ORR
IA1	0.016	0.006	0.009
IA2	0.014	0.003	-
FA			

Current efficacy boundaries (p-value scale)

Current graphical testing procedure



J&J

Running example 1: KEYNOTE-598

Example implementation in practice: FA

Accrued events / sample size for each hypothesis

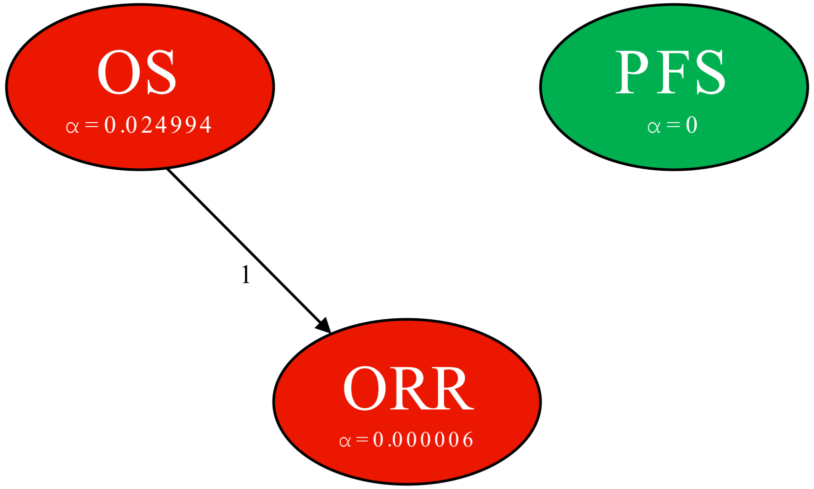
Analysis	OS	PFS	ORR
IA1	255	356	568
IA2	307	388	-
FA	361	-	-

Test statistic for each hypothesis

Analysis	OS	PFS	ORR
IA1	0.016	0.006	0.009
IA2	0.014	0.003	-
FA	0.011	-	-

Current efficacy boundaries (p-value scale)

Current graphical testing procedure



J&J

Running example 1: KEYNOTE-598

Example implementation in practice: OS is rejected and the graph update , which triggers updating the efficacy boundaries

Accrued events / sample size for each hypothesis

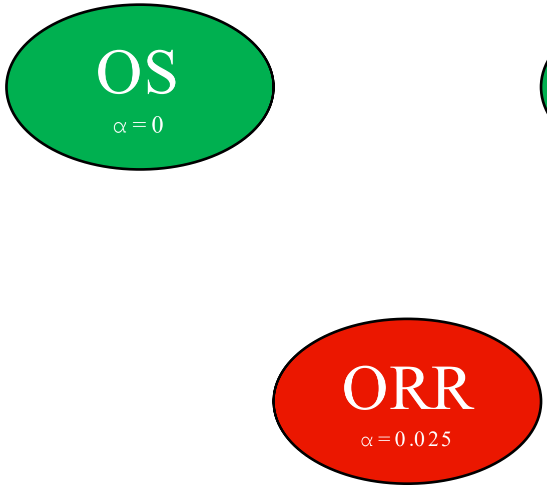
Analysis	OS	PFS	ORR
IA1	255	356	568
IA2	307	388	-
FA	361	-	-

Test statistic for each hypothesis

Analysis	OS	PFS	ORR
IA1	0.016	0.006	0.009
IA2	0.014	0.003	-
FA	0.011	-	-

Current efficacy boundaries (p -value scale)

Current graphical testing procedure



Running example 1: KEYNOTE-598

Example implementation in practice: ORR is rejected

Accrued events / sample size for each hypothesis

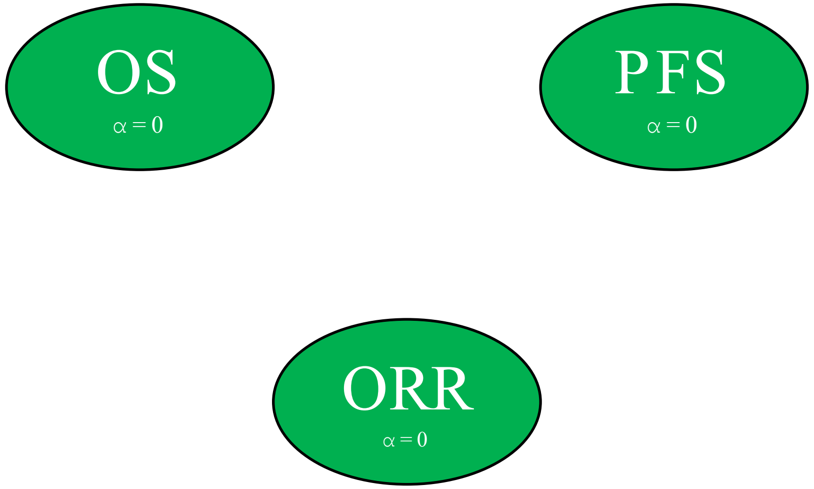
Analysis	OS	PFS	ORR
IA1	255	356	568
IA2	307	388	-
FA	361	-	-

Test statistic for each hypothesis

Analysis	OS	PFS	ORR
IA1	0.016	0.006	0.009
IA2	0.014	0.003	-
FA	0.011	-	-

Current efficacy boundaries (p-value scale)

Current graphical testing procedure



OS
α = 0

PFS
α = 0

ORR
α = 0

Selecting an interim analysis and multiplicity plan

General considerations

- Advantageous to trigger early analyses based on maturation of data for short-term outcome(s)
  - Typically subject to less uncertainty around timing; characterizing this uncertainty can be helpful
- Subsequent analysis often then the earliest point HAs will accept for the conventional primary outcome
- Select carefully analysis triggers reflecting on expectation in calendar time
  - Regulatory authorities may ask for number of events specifications rather than calendar times
  - Wrong assumptions might substantially deviate calendar timing of IAs
  - Wrong assumptions might substantially deviate from your alpha-spending plan
  - It is always easier to remove analyses, rather than add them.
- Keep in mind the required delta for significance on the short-term outcome when selecting its initial alpha
- Regulatory authorities may ask for OS to be powered for the lowest amount of alpha it can be tested with

Summary

- GTPs can easily be incorporated in a GSD framework
- Specify at a minimum
    - Initial graph
    - Spending function and IFs for each hypothesis
- Preferably specify the five components: Hypotheses, analyses, enrollment information, distributional information, initial graph
- 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., the process involves specifying what you would for a GTP in a fixed -sample trial and what you would for each hypothesis in a GSD**

5. Implementation

- Describing the method in a Protocol/SAP
- R Markdown for automated interim analysis and multiplicity strategy appendix generation

45 mins

J&J

Simple graphs

- Easy to determine all α levels each hypothesis may be tested at
- Feed each one into your favourite GSD software to determine all possible stopping rules for that hypothesis
- Assuming the test statistics for the hypotheses are uncorrelated, can even compute power fairly easily
- Describing in the protocol/SAP is also relatively easy as there's not much to describe

→ Business as usual

More complex graphs

- If you determine all possible α levels a given hypothesis can be tested at, can do as for a simple graph
  - But becomes much more labor intensive / more challenging as graph complexity increases
- Tools for automation become more helpful…
- Becomes logical to have a dedicated protocol/SAP Appendix on the multiplicity strategy
- {gMCPLite} article discusses how to produce tables like the ones shown earlier for Running example 1: KEYNOTE-598
  - https://merck.github.io/gMCPLite/articles/GraphicalMultiplicity.html

- We will use some R Markdown wrapped code in **{appendMCP}**

What to include?

- We’ve taken a systematic approach and searched for available examples of where a GTP has been used in a GSD framework
- ~45 or so examples with published protocols/SAPs available
    - Often redacted in parts, but still useful

ADAURA	CEPHEUS	ENDEAVOR	IMpower132	KEYNOTE-048	KEYNOTE-355	KEYNOTE-598	KEYNOTE-689	PERSEUS
ANDROMEDA	CHANGE AFIB	ESSENCE	IMpower133	KEYNOTE-091	KEYNOTE-361	KEYNOTE-604	KEYNOTE-716	PROpel
ANNOUNCE	CLARION	EVOKE-01	IND227	KEYNOTE-183	KEYNOTE-394	KEYNOTE-641	KEYNOTE-826	TROPiCS-02
AtTEnd	CLEAR	HER2CLIMB	innovaTV 301	KEYNOTE-204	KEYNOTE-522	KEYNOTE-671	KEYNOTE-A18	VERTIS CV
ATTRACTION-4	EMBER-3	IMforte	KEYLYNK-010	KEYNOTE-240	KEYNOTE-564	KEYNOTE-671	NRG-GY018	VITALITY-HFpEF

- Determined what has been included in these examples, and built the output from our code around this

Recipe book for fully specifying an interim analysis and multiplicity plan

Five components: Hypotheses, analyses, enrollment information, distributional assumptions, GTP

1. **Enrollment information:** Speed and duration of enrollment over time to each of the treatment arms, by sub-population if needed

2. **Hypotheses:** Define each hypothesis included in the GTP precisely
   a) What treatments are compared?
   b) In what sub-populations?
   c) For which endpoint?
   d) At what analyses?
   e) Using what spending function(s)?

3. **Analyses:** Specify what triggers each of the analyses
   a) Is an endpoint used (e.g., PFS) or is it calendar-based?
   b) How many events / what sample size is required? With what follow up? In what sub-population(s) and for what treatment arms?

1. **Distributional information:** For all hypotheses, need to assume effect sizes to evaluate power

2. **Graphical testing procedure:** The initial graph uniquely defines the plan for sharing alpha across hypotheses

Running example 1: KEYNOTE-598

Output: High level summary of hypotheses, their assumptions, and testing strategy

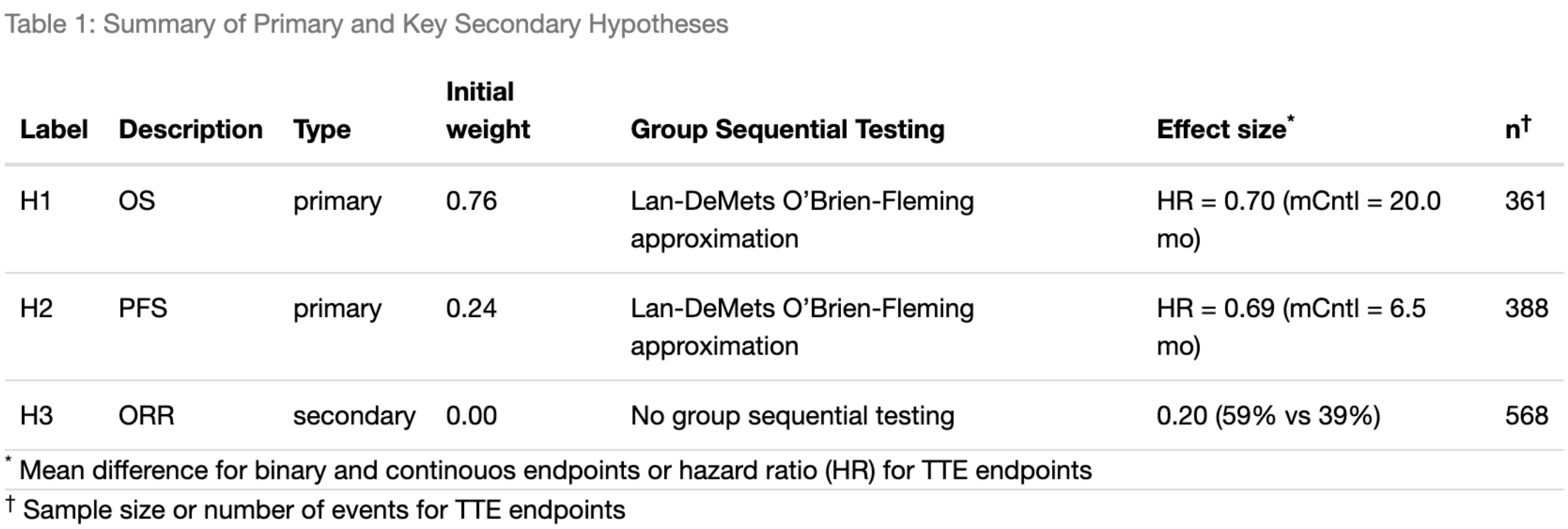


Table 1: Summary of Primary and Key Secondary Hypotheses

Label	Description	Type	Initial weight	Group Sequential Testing	Effect size[*]	n[†]
H1	OS	primary	0.76	Lan-DeMets O'Brien-Fleming approximation	HR = 0.70 (mCntl = 20.0 mo)	361
H2	PFS	primary	0.24	Lan-DeMets O'Brien-Fleming approximation	HR = 0.69 (mCntl = 6.5 mo)	388
H3	ORR	secondary	0.00	No group sequential testing	0.20 (59% vs 39%)	568

[*] Mean difference for binary and continouos endpoints or hazard ratio (HR) for TTE endpoints

[†] Sample size or number of events for TTE endpoints

Running example 1: KEYNOTE-598

Output: Data accrual

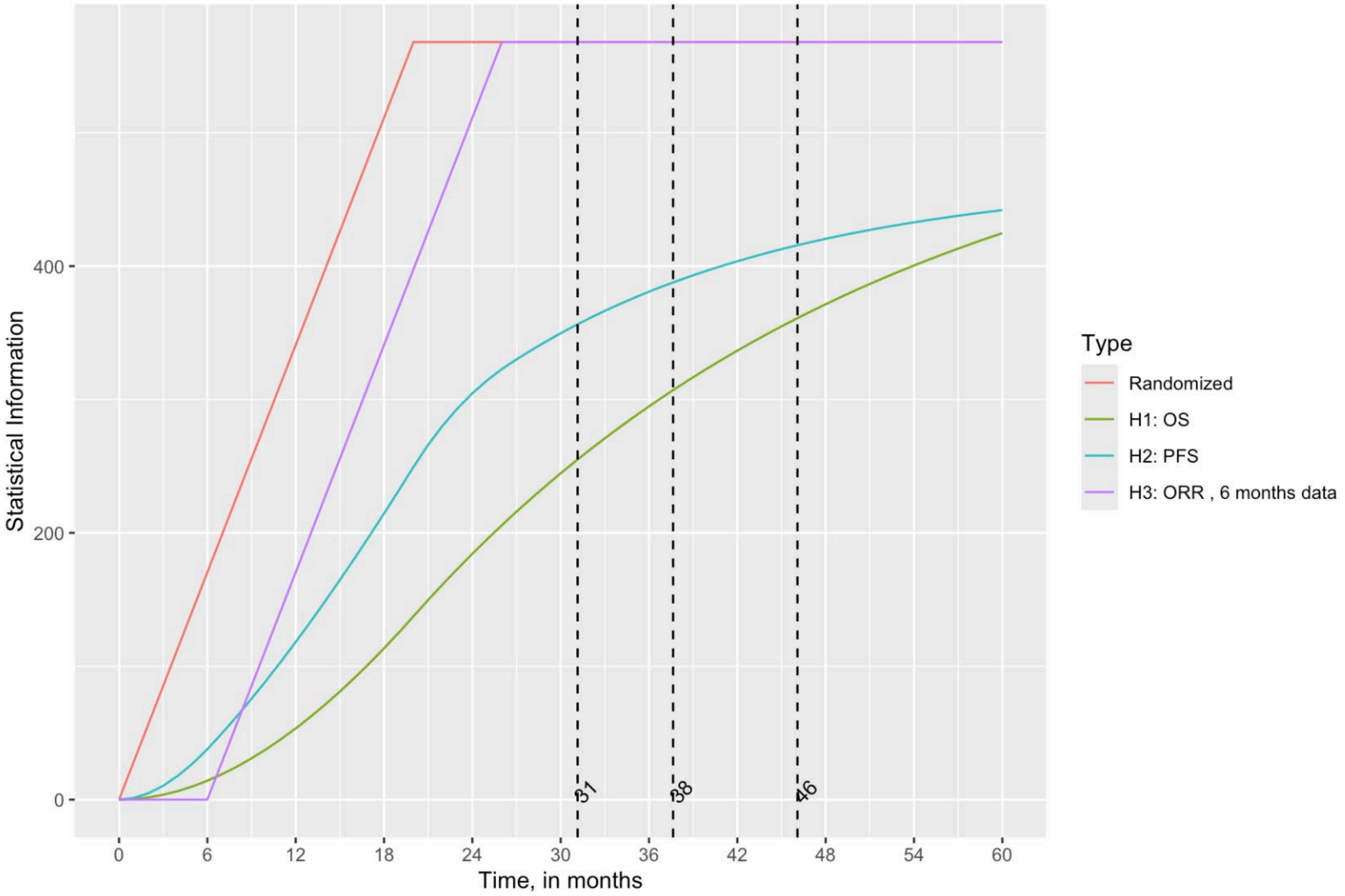


Figure 2: Timelines.

J&J

Running example 1: KEYNOTE-598

Output: When the analyses are expected by hypothesis

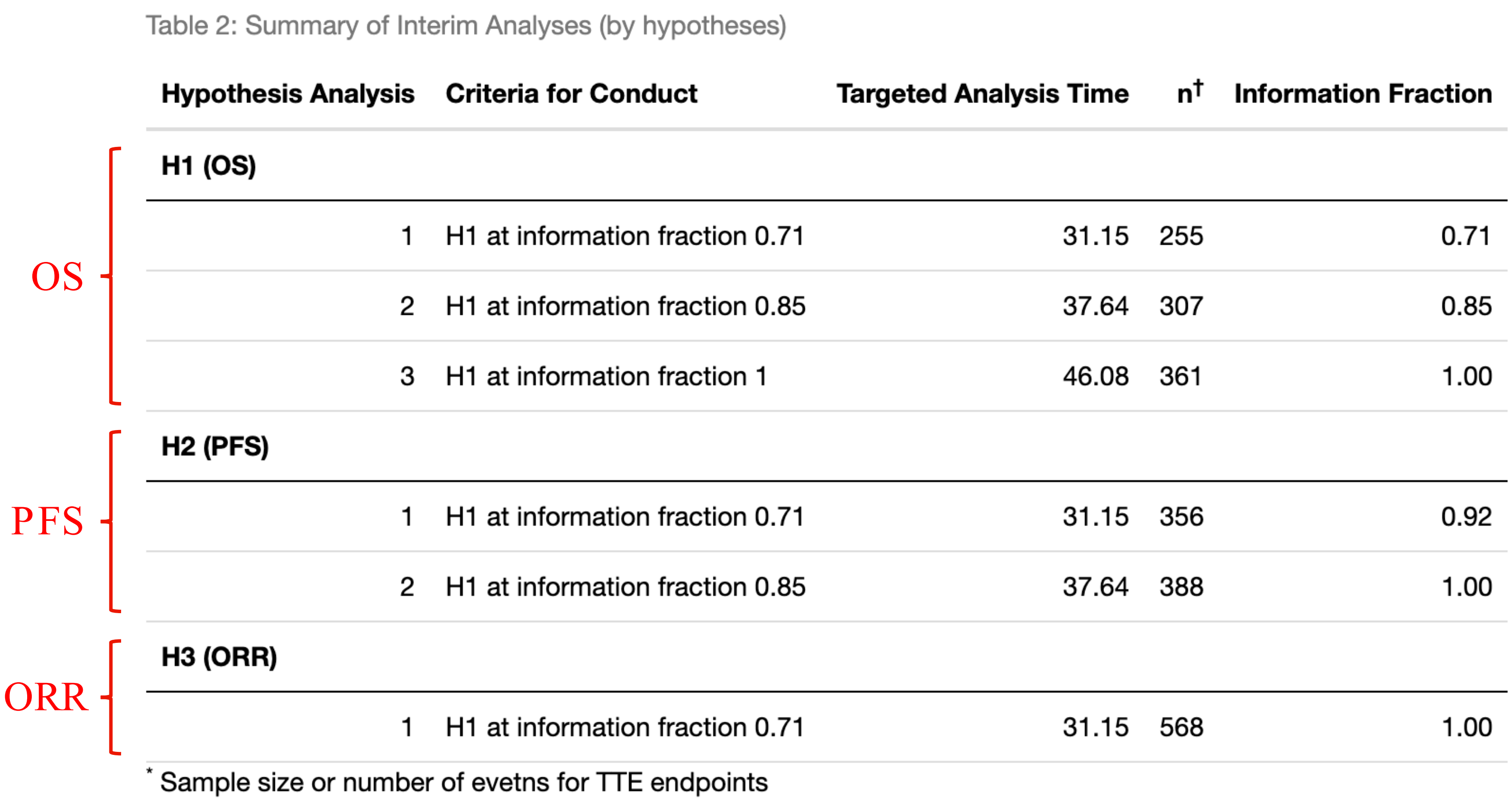


Table 2: Summary of Interim Analyses (by hypotheses)

	Hypothesis Analysis	Criteria for Conduct	Targeted Analysis Time	n†	Information Fraction
	H1 (OS)				
OS	1	H1 at information fraction 0.71	31.15	255	0.71
	2	H1 at information fraction 0.85	37.64	307	0.85
	3	H1 at information fraction 1	46.08	361	1.00
	H2 (PFS)				
PFS	1	H1 at information fraction 0.71	31.15	356	0.92
	2	H1 at information fraction 0.85	37.64	388	1.00
	H3 (ORR)				
ORR	1	H1 at information fraction 0.71	31.15	568	1.00

* Sample size or number of evetns for TTE endpoints

Running example 1: KEYNOTE-598

Output: What hypotheses are analysed by each analysis

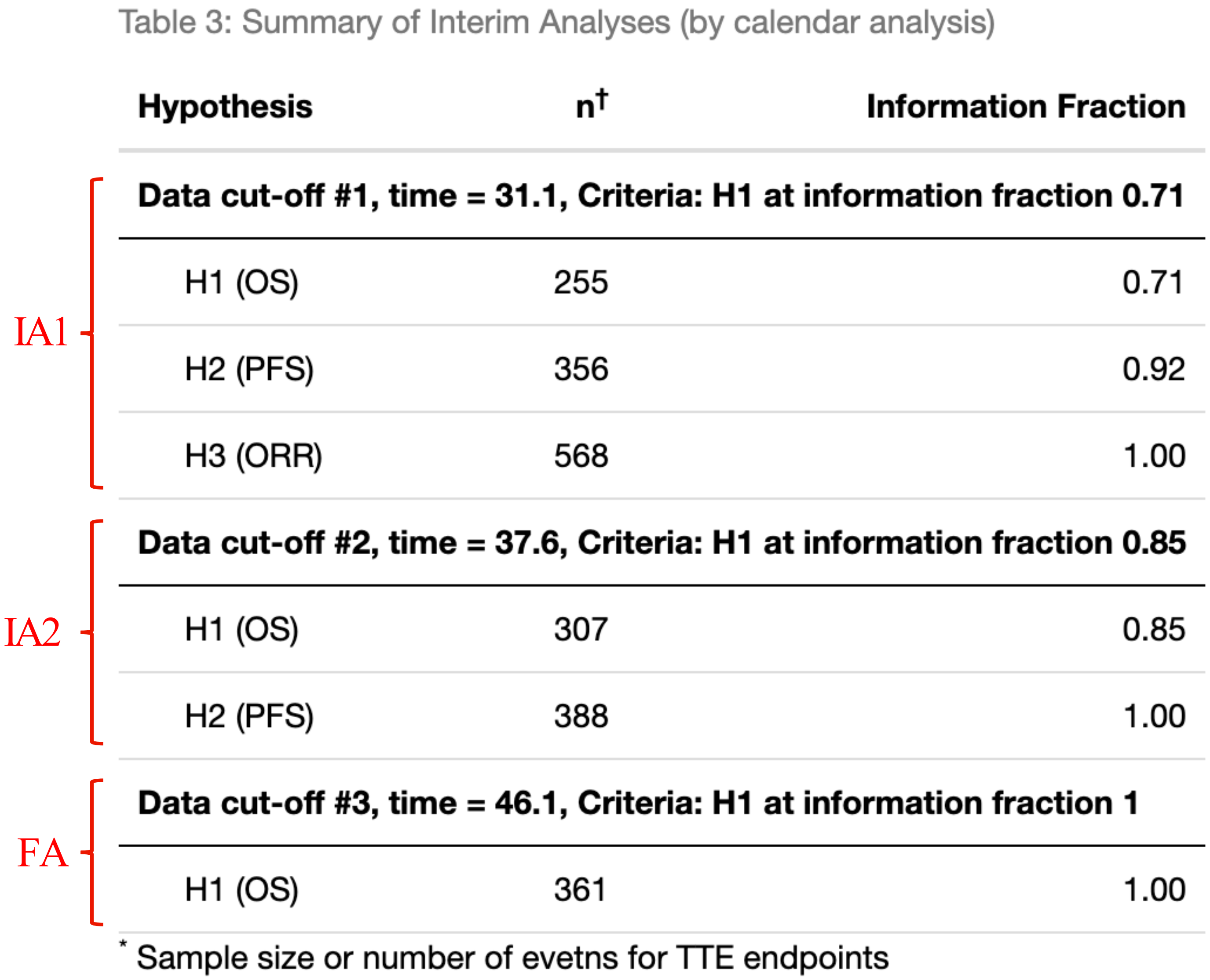


Table 3: Summary of Interim Analyses (by calendar analysis)

Hypothesis	$n^{\dagger}$	Information Fraction
Data cut-off #1, time = 31.1, Criteria: H1 at information fraction 0.71		
H1 (OS)	255	0.71
H2 (PFS)	356	0.92
H3 (ORR)	568	1.00
Data cut-off #2, time = 37.6, Criteria: H1 at information fraction 0.85		
H1 (OS)	307	0.85
H2 (PFS)	388	1.00
Data cut-off #3, time = 46.1, Criteria: H1 at information fraction 1		
H1 (OS)	361	1.00

[*] Sample size or number of evetns for TTE endpoints

J&J

Running example 1: KEYNOTE-598

Output: Requirements for specific α levels for each hypothesis

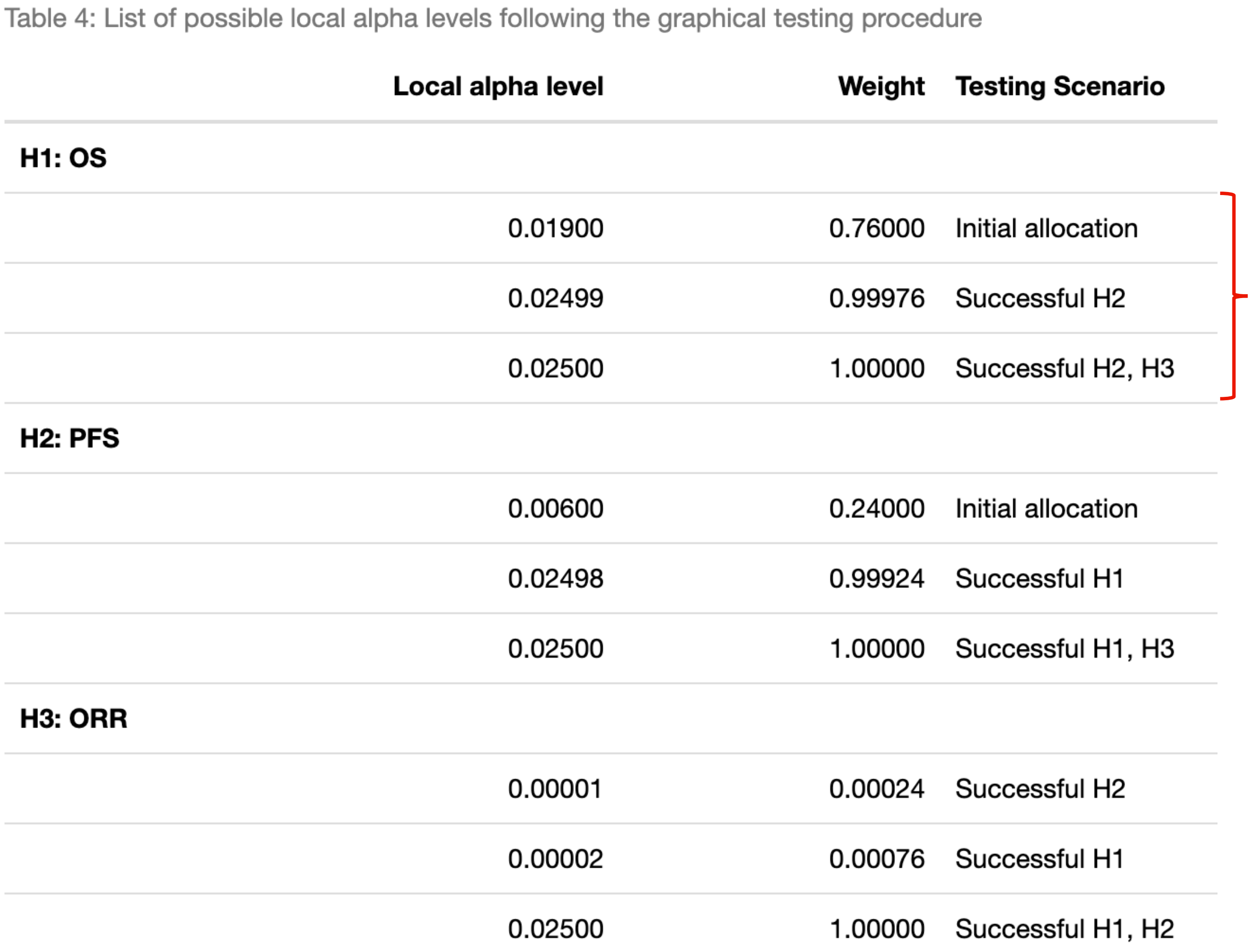


Table 4: List of possible local alpha levels following the graphical testing procedure

	Local alpha level	Weight	Testing Scenario
H1: OS			
	0.01900	0.76000	Initial allocation
	0.02499	0.99976	Successful H2
	0.02500	1.00000	Successful H2, H3
H2: PFS			
	0.00600	0.24000	Initial allocation
	0.02498	0.99924	Successful H1
	0.02500	1.00000	Successful H1, H3
H3: ORR			
	0.00001	0.00024	Successful H2
	0.00002	0.00076	Successful H1
	0.02500	1.00000	Successful H1, H2

As hypotheses are rejected, the α for OS increased

Running example 1: KEYNOTE-598

Output: Cumulative powers and significance thresholds for each hypothesis for each α level

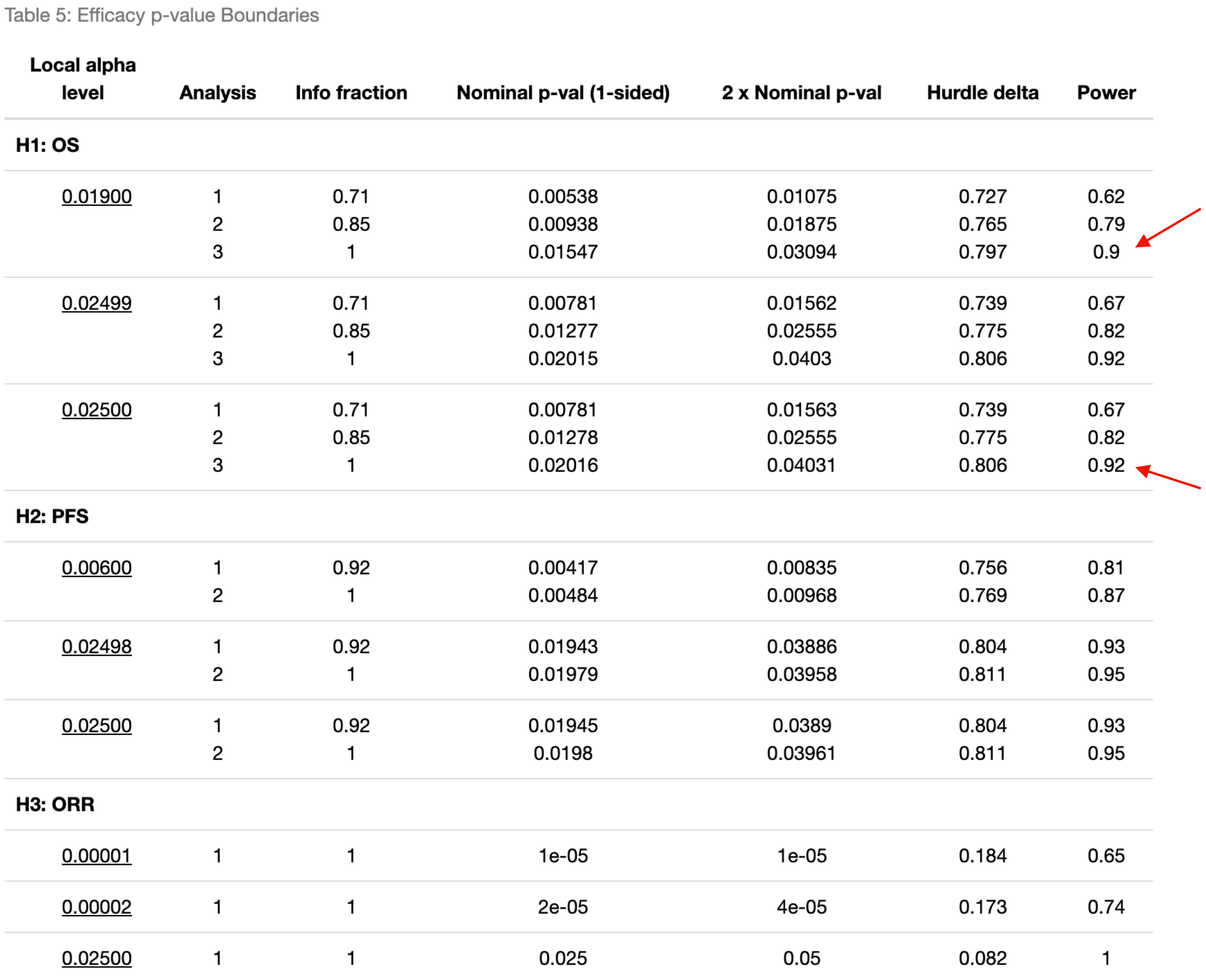


Table 5: Efficacy p-value Boundaries

Local alpha level	Analysis	Info fraction	Nominal p-val (1-sided)	2 x Nominal p-val	Hurdle delta	Power
H1: OS						
0.01900	1	0.71	0.00538	0.01075	0.727	0.62
	2	0.85	0.00938	0.01875	0.765	0.79
	3	1	0.01547	0.03094	0.797	0.9
0.02499	1	0.71	0.00781	0.01562	0.739	0.67
	2	0.85	0.01277	0.02555	0.775	0.82
	3	1	0.02015	0.0403	0.806	0.92
0.02500	1	0.71	0.00781	0.01563	0.739	0.67
	2	0.85	0.01278	0.02555	0.775	0.82
	3	1	0.02016	0.04031	0.806	0.92
H2: PFS						
0.00600	1	0.92	0.00417	0.00835	0.756	0.81
	2	1	0.00484	0.00968	0.769	0.87
0.02498	1	0.92	0.01943	0.03886	0.804	0.93
	2	1	0.01979	0.03958	0.811	0.95
0.02500	1	0.92	0.01945	0.0389	0.804	0.93
	2	1	0.0198	0.03961	0.811	0.95
H3: ORR						
0.00001	1	1	1e-05	1e-05	0.184	0.65
0.00002	1	1	2e-05	4e-05	0.173	0.74
0.02500	1	1	0.025	0.05	0.082	1

OS

Initial α

After PFS significance

After PFS and ORR significance

90% power initially for OS

α splitting does not cost OS much

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 automation can expedite things substantially
- Support in {appendMCP} extensive and growing
  - E.g., allows 'nominal' spends at early IAs
- Don't reinvent the wheel: multiplicity appendix provides a way to clearly explain the plan to regulatory authorities
- For all graphs, certain 'conditional powers' are easy to get: if you need **unconditional powers** , you likely need **simulation**

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)

Thank you for listening!

Any questions?

References

Closed testing procedures / Graphical testing procedures in fixed -sample designs

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

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

Group-sequential design

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

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

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

Misc.

Kunzmann K, Pilz M, Herrmann C, Rauch G, Kieser M (2021) The adoptr package: Adaptive optimal designs for clinical trials in R. *J Stat Soft* **98:** 1-21

Pilz M, Kunzmann K, Herrmann C, Rauch G, Kieser M (2021) Optimal planning of adaptive two-stage designs. *Stat Med* **40 :**3196-213