An introduction to graphical testing procedures for group-sequential designs

Michael Grayling
Senior Principal Statistician
Statistical Modelling and Methodology
Statistics & Decision Sciences

Yevgen Tymofyeyev
Senior Scientific Director
Statistical Modelling and Methodology
Statistics & Decision Sciences

MCP-2025
1:00-4:30pm 12th Aug 2025
Center City campus, Temple University, Philadelphia, Pennsylvania, USA

Johnson&Johnson

Short Course

Online abstract

Multiple testing problems arise regularly in the design of clinical trials due to the presence of diverse sets of research hypotheses posed by multiple endpoints, treatment arms, subgroups, and combinations of these factors. Over the past decade, there has been a great expansion in the availability of methodology for performing sequential tests of multiple hypotheses. Amongst such methods, graphical testing in a group-sequential setting (see, e.g., Maurer and Bretz, 2013) has found particular utility, having now been leveraged in numerous studies. In this course, we will provide attendees with the necessary information to evaluate, select, and implement such a design in practice. This will include discussion of nuances related to planning the timing and triggering of interim analyses and comprehensive pragmatic detailing of analysis criteria.

A brief description of graphical testing in the fixed-sample setting and group-sequential design for a single hypothesis will be provided, alongside a recap of how to implement these approaches in R, however some familiarity with these methods will be helpful. The primary focus of the course will then be on how to identify and implement the stopping rules of a group-sequential trial under a graphical testing procedure, describing the minimal information that must be specified for a design to be determined. Key options within this methodology will then be covered, including the utility of 'look back' analyses, how one can modify the alpha spending function for a hypothesis on updating the graph, and different alternatives for triggering interim analyses. We discuss both purely statistical and real-world considerations when selecting a design, and also detail how simulation can be used to estimate key marginal power quantities accounting for the correlation between all test statistics.

To help implement such approaches in practice, we also discuss a Quarto template that leverages the popular {gsDesign} and {gMCP} to dynamically and efficiently produce a design and its operating characteristics in a form directly useable with a protocol, for arbitrarily complex multiple hypothesis testing and interim analysis strategies. Throughout the course, we use several recent trials (e.g., Burtness *et al* (2019) from the KEYTRUDA program) as elucidating examples, to cover use cases with multiple arms, multiple endpoints, and multiple populations.

Brutness B, *et al*. Pembrolizumab alone or with chemotherapy versus cetuximab with chemotherapy for recurrent or metastatic squamous cell carcinoma of the head and neck (KEYNOTE-048): A randomised, open-label, phase 3 study. *Lancet* 2019;**394:**1915-28.

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

Short Course

Instructors

Dr. Michael Grayling is a Senior Principal Statistician within the Statistical Modelling and Methodology group at Johnson & Johnson, primarily supporting issues in Oncology and Immunology. Before joining J&J, Michael worked as a Research Fellow in Biostatistics at Newcastle University, where he developed a significant level of teaching experience and mentored multiple graduate students. His research interests include multi-arm multi-stage trials, crossover studies, and small sample sizes. He has published more than 50 papers in peer-reviewers journals and has authored a number of R packages and Stata modules related to trial design. Having a long record of teaching similar courses and delivering invited presentations focusing on adaptive design and multiple testing procures, including running a two-day course on this topic in five countries, attendees will in particular benefit from Michael's well-developed materials and presentation experience on the subject matter.

Dr. Yevgen Tymofyeyev is a Senior Scientific Director in the Statistical Modelling and Methodology group at Johnson & Johnson. In his current role, he serves as the statistical modelling lead for the Oncology Therapeutic area, implementing innovative designs and methods, including programs that utilize complex multi-stage designs with multiple hypothesis testing objectives, which have resulted in the successful submission of several clinical trials. He is actively involved in scientific collaborations in the field of randomization, adaptive design methodology, and software, which have led to an extensive list of publications, presentations, and implementation tools. Having 20 years of experience in pharmaceutical development, attendees will in particular benefit from Yevgen's extensive expertise in employing complex methods in practice.

Short Course

Outline

1.	**10 mins:** *Initial setup and intro to examples*	*Michael Grayling*
2.	**25 mins:** *Refresher on group-sequential design for a single hypothesis*	*Yevgen Tymofyeyev*
	○ Information fractions	
	○ Canonical joint distribution	
	○ Error spending	
3.	**25 mins:** *Refresher on graphical testing procedures in fixed sample designs*	*Michael Grayling*
	○ As a special case of closed testing	
	○ 'Epsilon' edges	
	○ Power calculation	
4.	**10 mins:** *Break*	
5.	**30 mins:** *Short practical on group-sequential design / graphical testing using R*	*Michael Grayling / Yevgen Tymofyeyev*
	○ Using {gsDesign}/{rpact} and {gMCP}/{graphicalMCP} to reproduce study design(s)	
6.	**45 mins:** *Graphical testing in group-sequential designs*	*Yevgen Tymofyeyev*
	○ Spending function updating (i.e., 'delayed' alpha recycling)	
	○ 'Look back' analyses	
	○ Analysis triggers (including in multi-arm event driven studies)	
	○ Simulation accounting for all correlations	
7.	**10 mins:** *Break*	
8.	**45 mins:** *Software demonstration and practical*	*Michael Grayling*
	○ Demonstration of the functionality of a dynamic R Markdown template	
	○ Practical on using this template to reproduce recent study design(s)	
9.	**10 mins:** *Q&A and Close*	*Michael Grayling / Yevgen Tymofyeyev*
	○ Opportunity to ask questions, from theoretical issues to implementation problems	

Software installation

Run *short_course_gtp_gsd_install_required_packages.R*

- To make the practical sessions run as smoothly as possible, we've created a short R script that installs the required packages
- Minimally, you will need **{appendMCP}** to install
- **{graphicalMCP}** and **{rpact}** may also be used in parts

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

Sample size and endpoints

- 1:1 randomization of N = 568 pts over 20 mo
- Dual primary endpoints: OS and PFS
  - One key secondary endpoint also included in the graphical testing procedure: ORR
- OS:
  - Analysis method: (Stratified) log-rank
  - Control arm: Exponential with a median of 20 mo
  - Treatment arm: HR = 0.70
  - Drop-out: 1% per year
- PFS:
  - Analysis method: (Stratified) log-rank
  - Control arm: Piece-wise exponential with a median of 6.5 mo before 6.5 mo and a median of 14.5 mo after 6.5 mo
  - Treatment arm: HR = 0.69
  - Drop-out: 13% per year
- ORR:
  - Analysis method: (Stratified) Miettinen and Nurminen
  - Control arm: 39%
  - Treatment arm: Δ = 20%

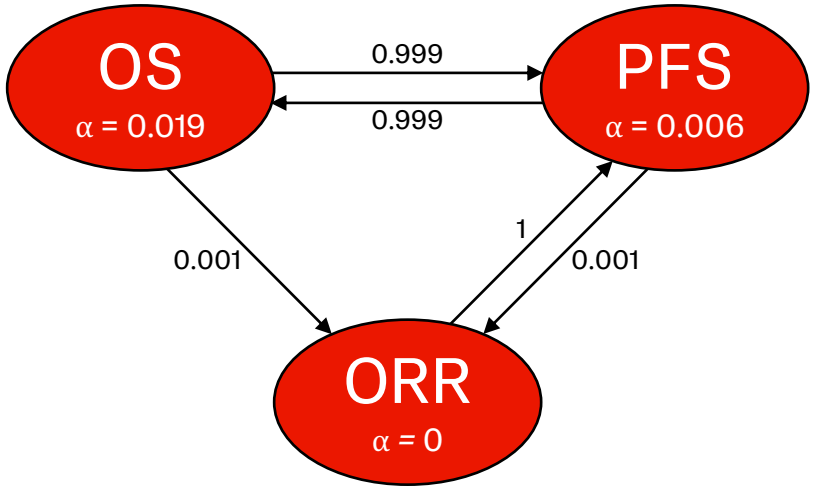
Running example 1: KEYNOTE-598

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	~361 OS 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 1: KEYNOTE-598

Efficacy boundaries and properties for OS analyses: Table 19 in Section 8.8.3

Analysis	Value	$\alpha = 0.019$	$\alpha = 0.025$
IA1: 71%[a] N: 568 Events: 255 Month: ~ 32[f]	Z	2.5592	2.4257
	p (1-sided)[b]	0.0052	0.0076
	HR at bound[c]	0.7256	0.7378
	P(Cross) if HR=1[d]	0.0052	0.0076
	P(Cross) if HR=0.7[e]	0.6135	0.6631
IA2: 85% N: 568 Events: 307 Month: ~ 39[f]	Z	2.3490	2.2316
	p (1-sided)[b]	0.0094	0.0128
	HR at bound[c]	0.7646	0.7750
	P(Cross) if HR=1[d]	0.0110	0.0151
	P(Cross) if HR=0.7[e]	0.7900	0.8228
FA[g]: N: 568 Events: 361 Month: ~ 47[f]	Z	2.1577	2.0504
	p (1-sided)[b]	0.0155	0.0202
	HR at bound[c]	0.7967	0.8058
	P(Cross) if HR=1[d]	0.0190	0.0250
	P(Cross) if HR=0.7[e]	0.8985	0.9167

[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
[g] FA will be conducted at 34 months after the enrollment of the last participant at the latest, if the event accumulation is slower than expected

Running example 1: KEYNOTE-598

Efficacy boundaries and properties for ORR analyses: Section 8.8.1

"No initial alpha is allocated to test ORR. However, the ORR p-value from IA1 (ie, no new data is added after IA1) can be compared to an α-level of eg, 0.025 if the null hypotheses for both PFS and OS are rejected at IA1 or a later time. The ***power*** *at the updated* ***α-level of 0.025 is 99.8%****, with an* ***approximate treatment difference (ΔORR)*** *required for reaching the efficacy bound being* ***8.1%****, assuming underlying* ***39% and 59%*** *ORR in the control and experimental groups, respectively."*

Running example 2

See /Examples/Running example 2/

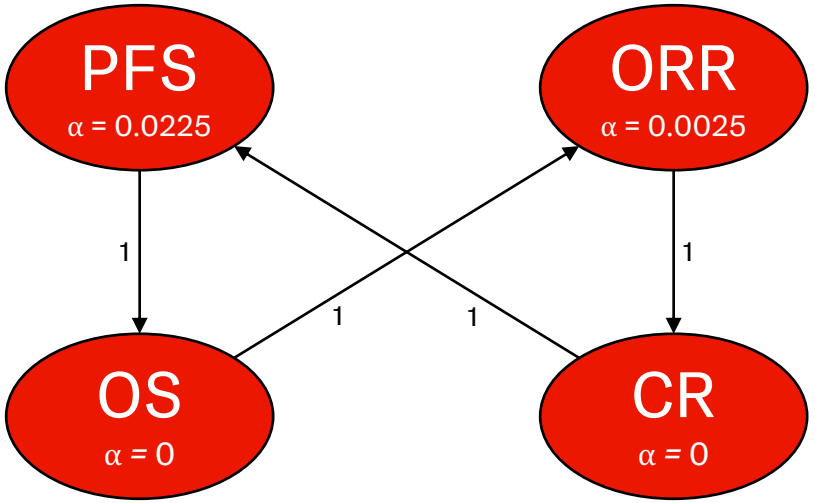
- Ph3 oncology trial, comparing CON vs TRT
- Enrollment
  - N = 450 pts, randomized 1:1
  - 10 pts/mo for months 1-2; 15 pts/mo for months 3-4; 25 pts/mo thereafter
- Four endpoints for which strong control to one-sided $\alpha = 0.025$ is ensured
  - ORR and PFS as dual primary endpoints
  - CR and 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, including nominal spending
- (Futility stopping)
- {gsDesign} and {rpact}

25 mins

Introduction to group-sequential design

What, why, and how?

- In a group-sequential design (GSD), **interim analyses** (IAs) are incorporated at which a **study may terminate early**
  - Overseen by *Independent Data Monitoring Committee*
- A type of **adaptive design** where the adaptation is about whether to reject (or not) hypotheses and potentially stop the study, **at pre-specified levels of study data**
  - I.e., timing of IAs in a GSD are generally (approximately) pre-specified
- Multiple reasons for using a GSD: can be grouped as **ethical, administrative, and economic**
  - Not just about achieving early positive result: also incentives to reach early decision if study is negative → see brief discussion later on futility monitoring

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 to control error rates to the desired level

Introduction to group-sequential design

What, why, and how?

- **Long history:** Original theory dates back to Wald in the 1940s, with early medical applications by Armitage in the 1950/60s
- Most commonly known methodology (Pocock / O'Brien-Fleming / Kim-DeMets) developed in 1970/80s
- Error spending approach by Lan and DeMets (1983)
- Now (probably) the **most commonly utilized** type of adaptive design
- Still an active area of research
  - Multiple endpoints
  - Multiple arms
- Majority of literature assumes a single endpoint of interest

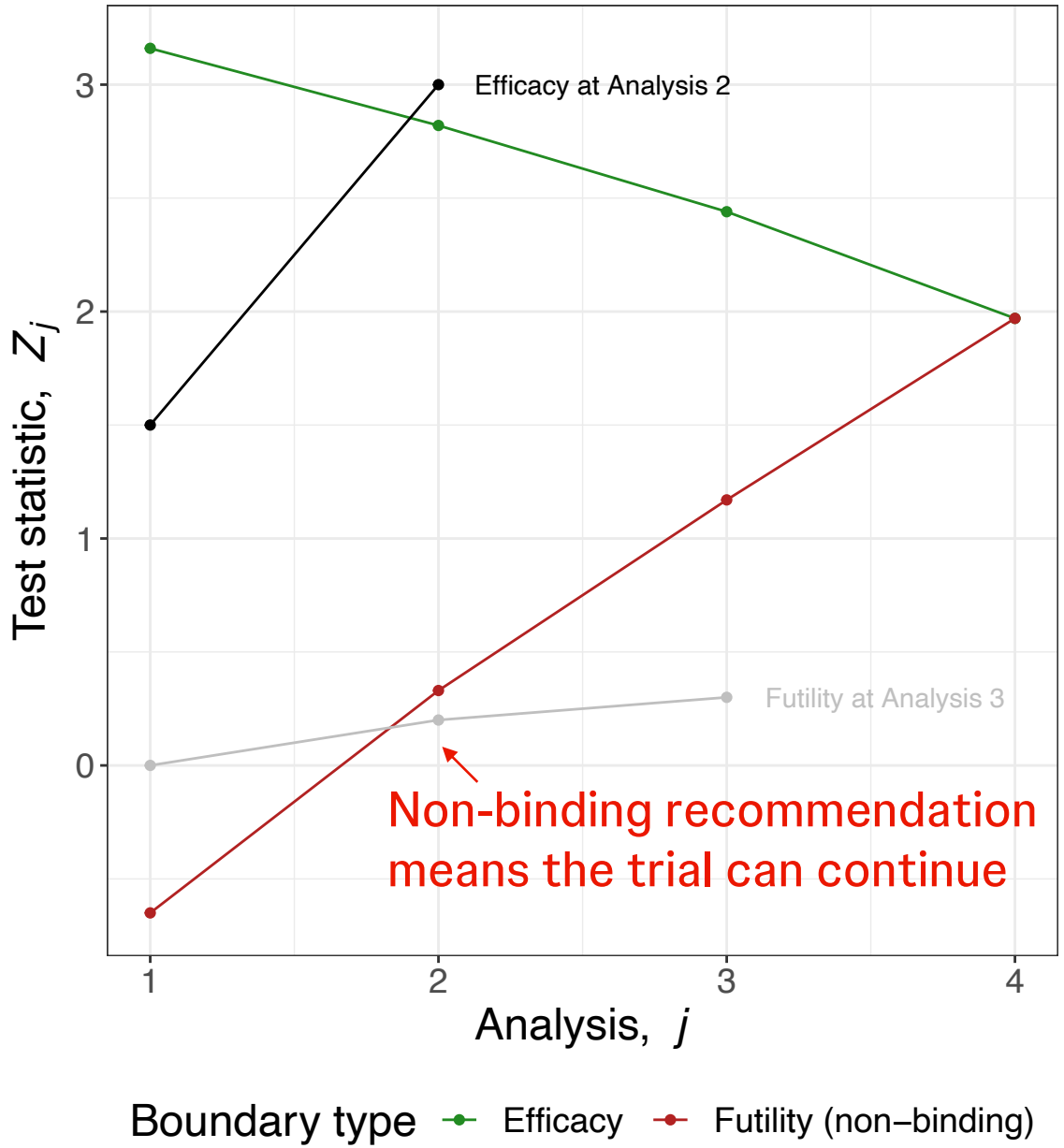
Test decisions using the Z-scale

- Suppose the **parameter of interest is θ**
  - E.g., mean difference, log(hazard ratio), log(odds ratio), absolute risk difference
- Want to test $H_0 : \theta \leq 0$ vs. $H_1 : \theta > 0$
  - Desire type I error rate of α when $\theta = 0$ and power of $1 - \beta$ when $\theta = \delta > 0$
- GSD has **at most *J* analyses** planned: analysis $j = 1, \ldots, J$ uses standardized test statistic Z_j
  - $Z_j = \hat{\theta}_j \sqrt{I_j}$, where $\hat{\theta}_j$ is the estimate of θ, and I_j the information for the estimate, at analysis j
- The standardized test statistic approach covers a wide spectrum of applications
  - Normal, binary, and survival data
  - Design and analysis types (parallel groups, linear models, and other parametric models,…)
  - See Jennison and Turnbull (2000)

Stopping rules

Example for $J = 4$

- Testing rules depend on **futility bounds $f_1, \ldots, f_J$ and efficacy bounds $e_1, \ldots, e_J$**. 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



Other forms of stopping rules

_p_-value and effect scales

- Above expressed stopping rules as $Z_j \geq e_j$
- Can equivalently write in the form $p_j \leq p_j^*$, where p_j is the unadjusted (sometimes called local) *p*-value at analysis j, and p_j^* is the corresponding threshold
- Or in terms of the observed effect for the parameter of interest, i.e., $\hat{\theta}_j \geq \theta_j^*$
  - E.g., $\widehat{HR}_j \leq HR_j^*$, where HR_j^* is an (approximate) threshold for the HR estimate at analysis j
  - This is often the most useful form for clinical teams
- We'll see in the practical(s) later that standard software often outputs stopping rules in multiple forms

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:

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

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

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

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

Futility rules ignored when calculating type I error rate

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

Computing error rates and other operating characteristics

The 'canonical joint distribution'

- 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 $j_1, j_2 = 1, \ldots, J, j_2 \geq j_1$
- So 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. E.g.

$$\mathbb{P}_\theta(Z_1 \leq e_1, Z_2 > e_2) = \int_{-\infty}^{e_1} \int_{e_2}^{\infty} \phi_2 \left\{ \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \theta \begin{pmatrix} \sqrt{I_1} \\ \sqrt{I_2} \end{pmatrix}, \begin{pmatrix} 1 & \sqrt{I_1/I_2} \\ \sqrt{I_1/I_2} & 1 \end{pmatrix} \right\} \mathrm{d}z_2 \mathrm{d}z_1$$

Computing a design

- So MVN integration can be used to compute operating characteristics
- 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

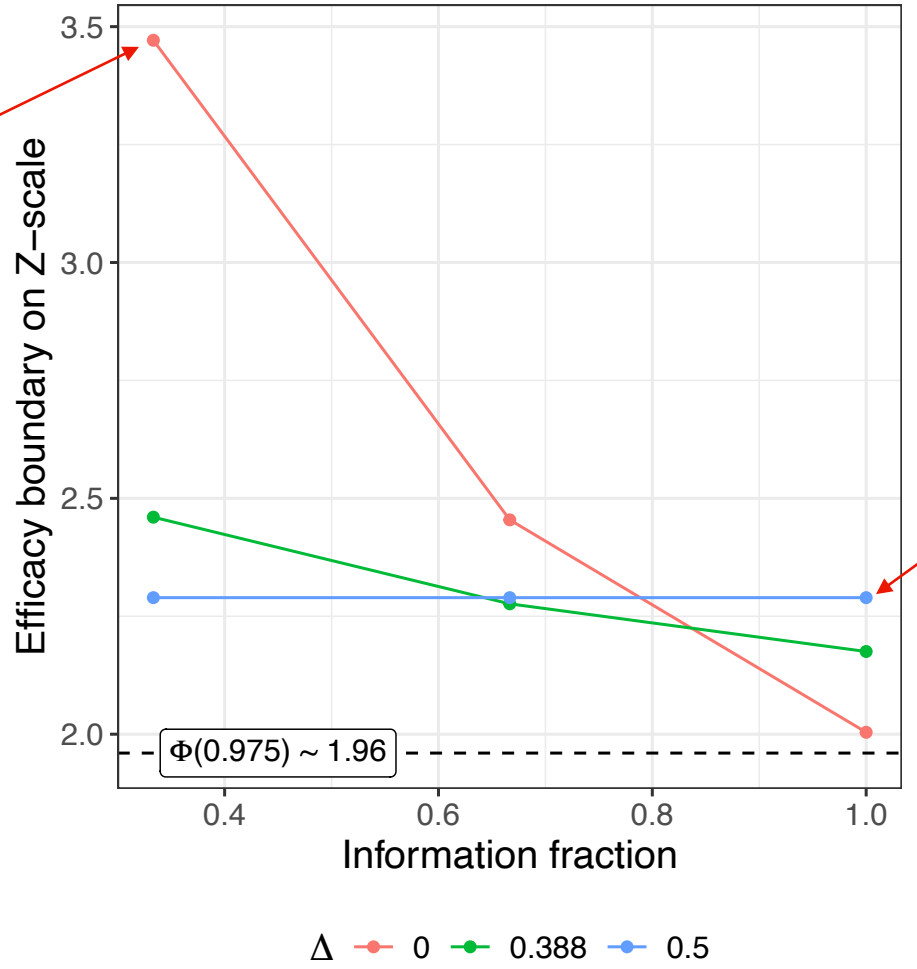
$$I_j = \frac{jI_J}{J}$$

- 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

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

Functional form efficacy bounds

I.e., the original approach

- Can easily be generalized 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 less-predictable information levels is…

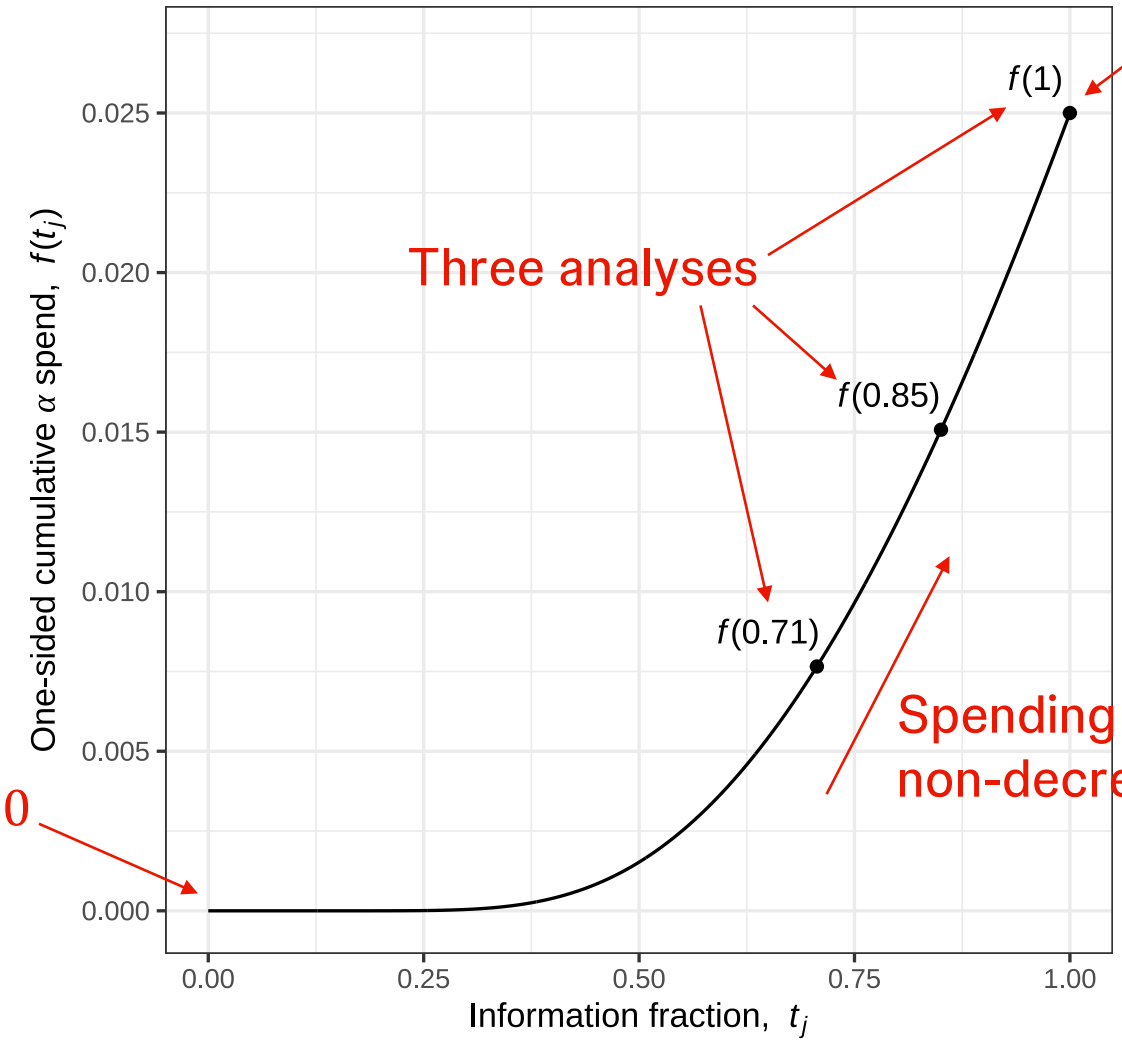
Error spending

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
- Based on **information fractions (IFs)** $t_j = \frac{I_j}{I_J}$
- And non-decreasing function $f : [0,1] \to [0, \alpha]$, that gives **cumulative α spend** at IF t_j as $f(t_j)$
- Does require information level I_j **to not depend** on $\hat{\theta}_1, \ldots, \hat{\theta}_{j-1}$
  - 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 $\alpha = 0.025$



Error spending

Technical approach

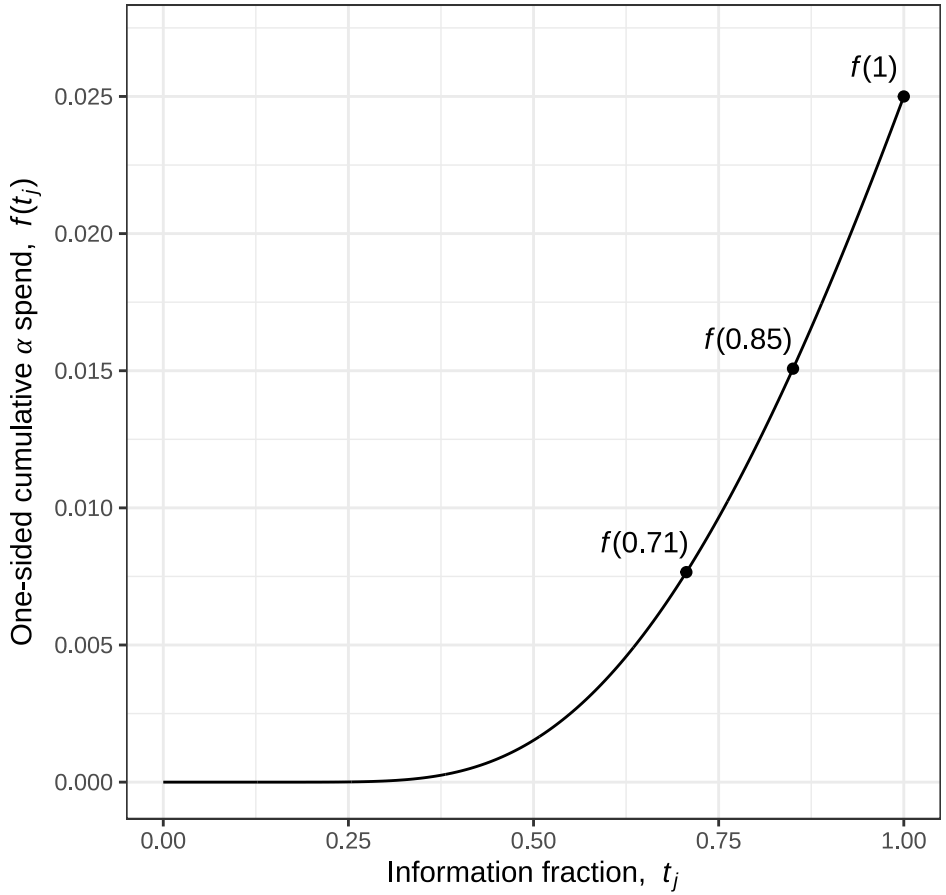
- Iterative approach used to determine e_1, then e_2, then e_3, etc.
- Analysis 1 choose e_1 such that

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

- Analysis 2 choose

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

- 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"

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

- 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

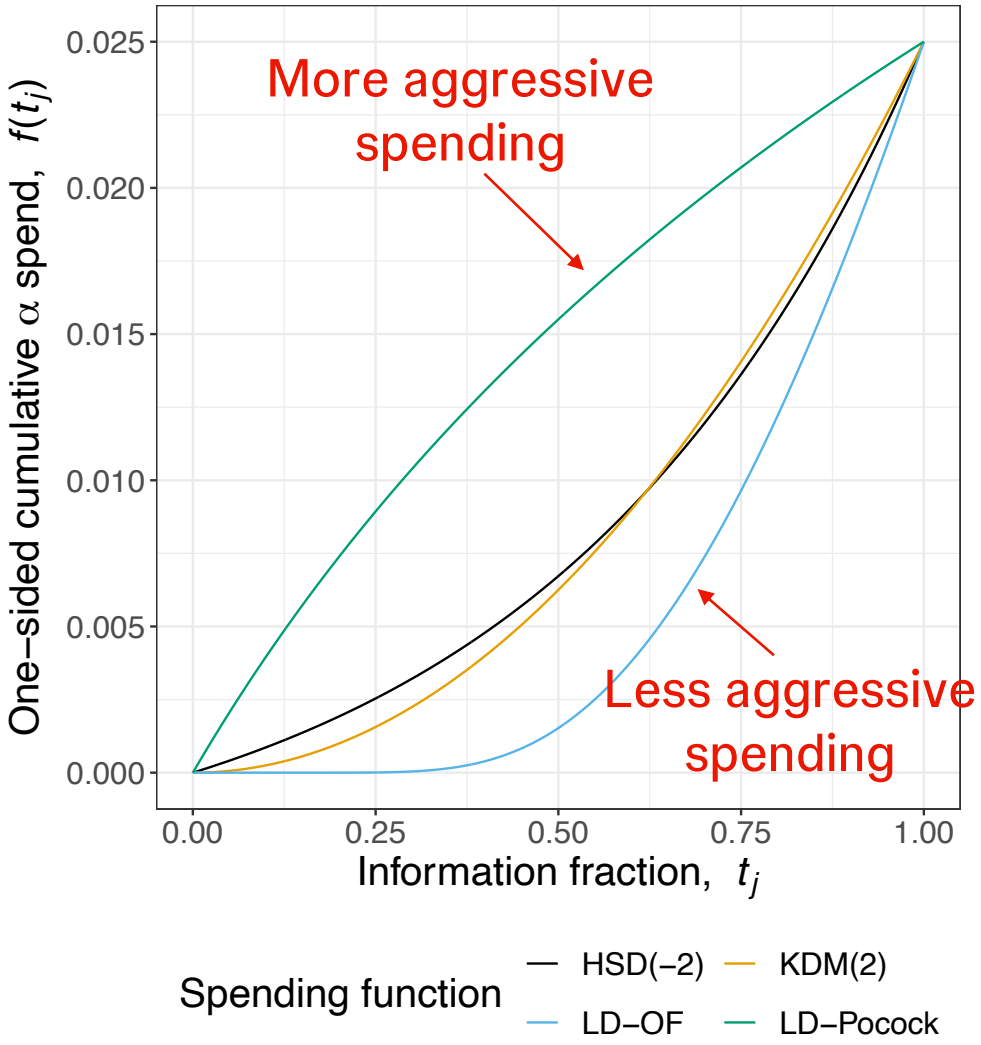
$\gamma = 1$ Similar to Pocock

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

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

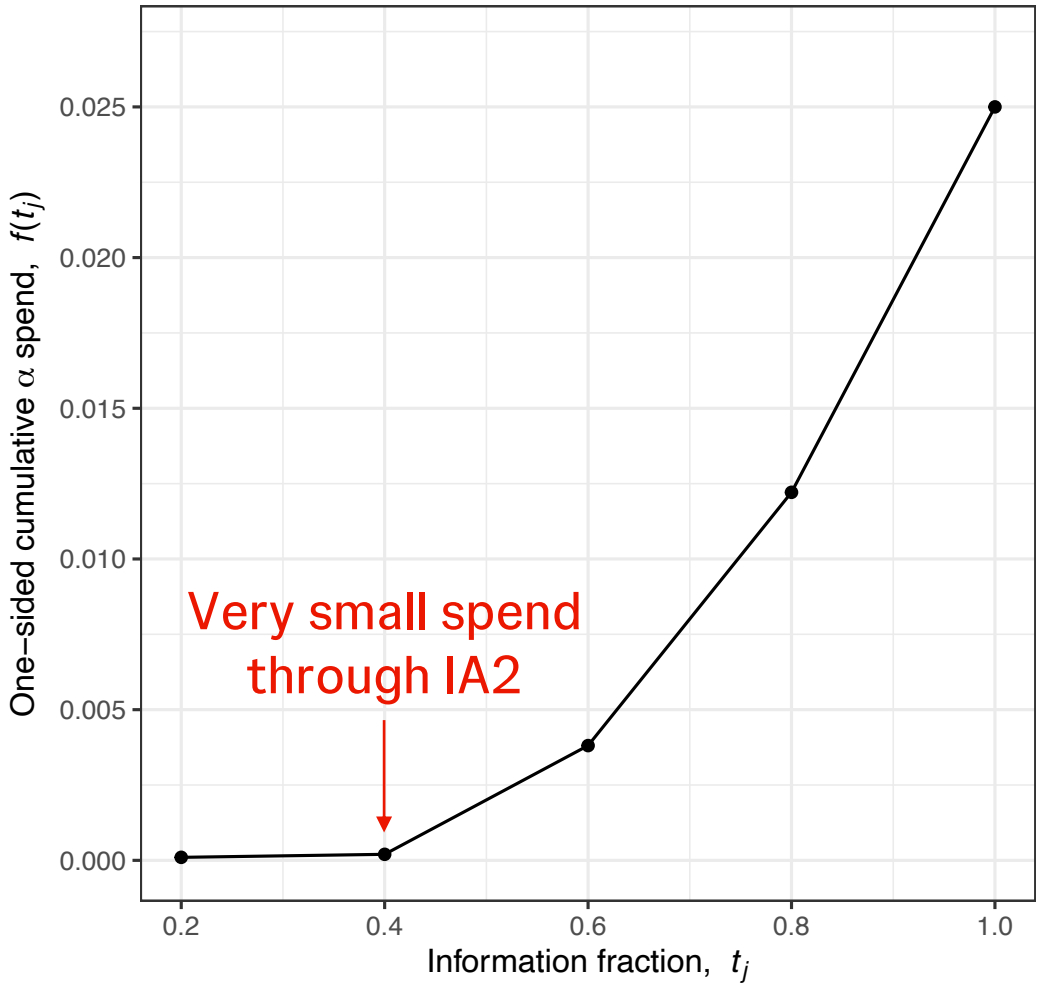
$\rho = 3$ Similar to O'Brien-Fleming

$\rho = 1$ Similar to Pocock



‘Nominal’ spending

- In the multiple endpoint case (see later), not uncommon to have an endpoint tested at very low information fraction
- E.g., IA1 timed to analyze a short-term intermediate outcome. Also possible we may analyze OS, but it will have few events
- In this case, spending (meaningful) α on OS could be argued to be a waste
- A solution is to define a bespoke spending function that spends a very small amount of alpha initially
  - At later analyses it could follow a standard spending function
- E.g., 5 equally spaced analyses. IA1-2 spends 0.0001 alpha. Then follow LDOF



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 of α 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

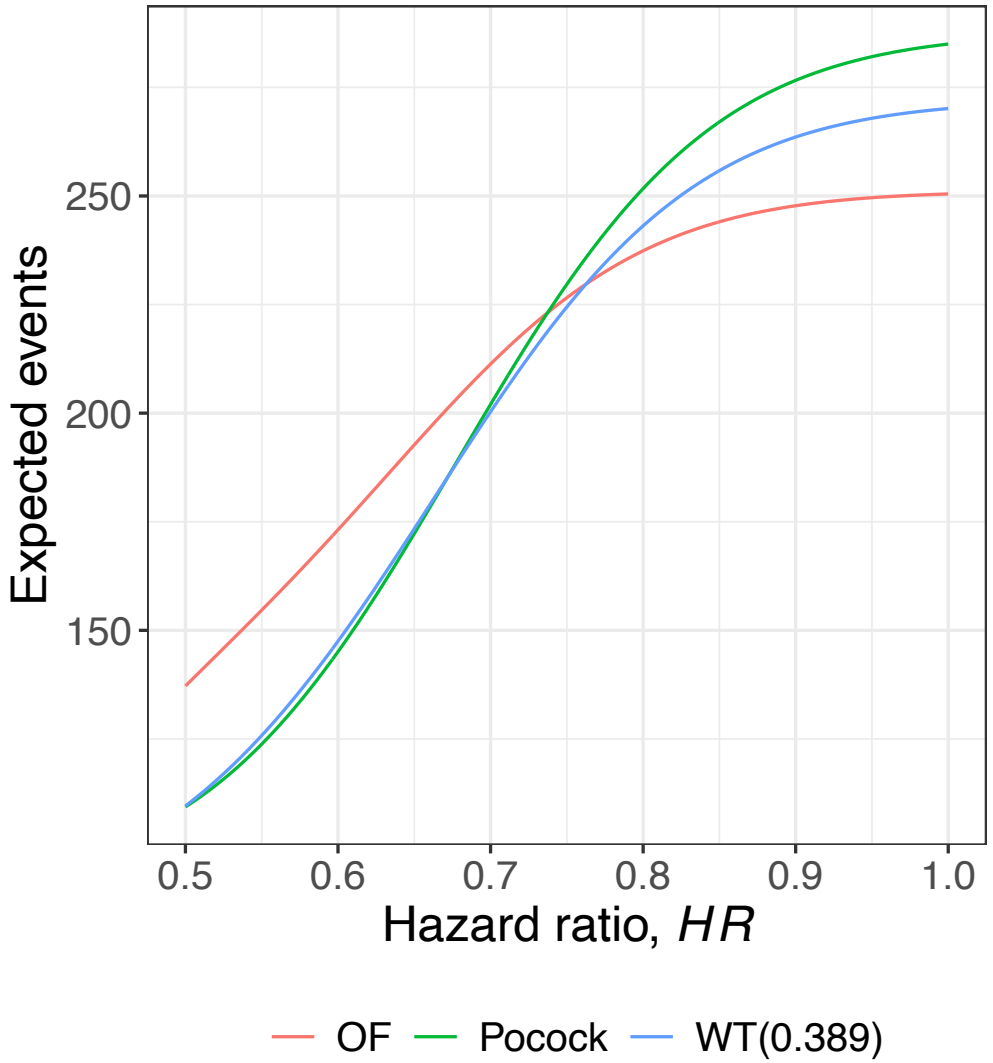
Minimizes the ESS reduction under H_1

Increasing in Δ

Spending options

Inflation factors and ESS reduction for Wang-Tsiatis bounds

- Further assume control arm mOS is 20 mo, and target HR is 0.7
  - As for OS in Running example 1: KEYNOTE-598
- WT(0.389) is optimal for the 3-stage equal stage design with alpha=2.5%, power=80%
- Power is 80%, when HR = 0.7
- WT(0.389) is only minorly better than Pocock under this setting
- Pocock and O'Brien-Fleming have opposite performance under extreme HRs



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

An aside on futility: Conventional futility monitoring

Why?

- Can be very useful to build in futility assessment(s)
  - Based on context could be for one or more endpoints
- Concession to regulatory authorities that generally costs very little
  - Set up the futility rule correctly and it is unlikely to be met unless the treatment is genuinely poor performing
- Could save money and/or time
  - Dedicated IAs to futility and/or conduct when efficacy is assessed
  - In survival case, meaningful cost savings generally could only be associated with a futility look before recruitment has completed
  - ~40%IF good benchmark for enough data for a futility look
- Can be used for ruling out harm

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

Running example 1: KEYNOTE-598

- Let's consider OS
  - Control arm: Exponential with a median of 20 mo
  - Treatment arm: HR = 0.70
  - Drop-out: 1% per year
- Assume one-sided $\alpha = 0.019$
  - More on this later though
- GSD
  - Two interim analyses at ~255 (~71%IF) and ~307 (~85%IF), with final analysis after ~361 events
  - Lan and DeMets O'Brien-Fleming (LDOF) spending function

Running example 1: KEYNOTE-598

`gsDesign::gsSurv()`

Input

```
gsDesign::gsSurv(k         = 3,
                 test.type = 1,
                 alpha     = 0.019,
                 beta      = 0.1,
                 timing    = c(0.71, 0.85, 1),
                 sfu       = gsDesign::sfLDOF,
                 lambdaC   = log(2)/20,
                 hr        = 0.7,
                 eta       = -log(1 - 0.01)/12,
                 gamma     = 568/20,
                 R         = 20,
                 minfup    = NULL,
                 T         = NULL)
```

Output

```
Time to event group sequential design with HR= 0.7
Equal randomization:          ratio=1
One-sided group sequential design with
90 % power and 1.9 % Type I Error.

  Analysis  N   Z   Nominal p  Spend
         1 258 2.55    0.0054 0.0054
         2 309 2.35    0.0094 0.0056
         3 363 2.16    0.0155 0.0080
     Total                    0.0190

++ alpha spending:
 Lan-DeMets O'Brien-Fleming approximation spending function (no
parameters).

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

Upper boundary (power or Type I Error)
          Analysis
   Theta      1      2      3 Total  E{N}
  0.0000 0.0054 0.0056 0.0080 0.019 361.5
  0.1785 0.6226 0.1696 0.1078 0.900 287.7
              T   n   Events HR efficacy
IA 1  31.41004 568 257.2980       0.728
IA 2  37.77820 568 308.0328       0.765
Final 46.33306 568 362.3915       0.797
Accrual rates:
     Stratum 1
0-20      28.4
Control event rates (H1):
      Stratum 1
0-Inf      0.03
Censoring rates:
      Stratum 1
0-Inf         0
```

Running example 1: KEYNOTE-598

`rpact::getSampleSizeSurvival()`

Input

```
rpact::getSampleSizeSurvival(
  design           = rpact::getDesignGroupSequential(
    kMax             = 3,
    alpha            = 0.019,
    beta             = 0.1,
    sided            = 1,
    informationRates = c(0.71, 0.85, 1),
    typeOfDesign     = "asOF"
  ),
  lambda2          = log(2)/20,
  hazardRatio      = 0.7,
  dropoutRate1     = 0.01,
  dropoutRate2     = 0.01,
  dropoutTime      = 12,
  accrualTime      = c(0, 20),
  accrualIntensity = 568/20
) |> summary()
```

Output

```
Sample size calculation for a survival endpoint

Sequential analysis with a maximum of 3 looks (group sequential design),
one-sided overall significance level 1.9%, power 90%.
The results were calculated for a two-sample logrank test,
H0: hazard ratio = 1, H1: hazard ratio = 0.7, control lambda(2) = 0.035,
accrual time = 20, accrual intensity = 28.4, dropout rate(1) = 0.01,
dropout rate(2) = 0.01, dropout time = 12.
```

Stage	1	2	3
Planned information rate	71%	85%	100%
Cumulative alpha spent	0.0054	0.0110	0.0190
Stage levels (one-sided)	0.0054	0.0094	0.0155
Efficacy boundary (z-value scale)	2.551	2.350	2.158
Efficacy boundary (t)	0.728	0.765	0.797
Cumulative power	0.6226	0.7922	0.9000
Number of subjects	568.0	568.0	568.0
Expected number of subjects under H1			568.0
Cumulative number of events	257.7	308.5	363.0
Expected number of events under H1			288.2
Analysis time	31.46	37.85	46.44
Expected study duration under H1			35.65
Exit probability for efficacy (under H0)	0.0054	0.0056	
Exit probability for efficacy (under H1)	0.6226	0.1696	

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

R package comparison

Cheatsheet on function arguments

	{gsDesign}	{gsDesign2}	{rpact}
Computing required events for desired power	`gsDesign::gsSurv()`	`gsDesign2::gs_design_ahr()`	`rpact::getSampleSizeSurvival()`
Computing power for specified events	`gsDesign::gsProbability()`	`gsDesign2::gs_power_ahr()`	`rpact::getPowerSurival()`
Number of analyses	`k`	`info_frac`, `event`, and/or `analysis_time` Inferred by the package through one or more of these arguments	`kMax`, inside `design` `design` formatted through `rpact::getDesignGroupSequential()`
One-sided type I error-rate	`alpha`, with `sided = 1`	`alpha`	`alpha`, with `sided = 1`, inside `design` `design` formatted through `rpact::getDesignGroupSequential()`
Desired power / type II error-rate	`beta`	`beta`	`beta`, inside `design` `design` formatted through `rpact::getDesignGroupSequential()`
Interim analysis timings (information fractions or events)	`timing`	`info_frac`, `event`, and/or `analysis_time`	`informationRates`, inside `design`, with `maxNumberOfEvents` for power `design` formatted through `rpact::getDesignGroupSequential()`
Alpha spending function and any associated parameters	`sfu` and `sfupar`	`upar`	`typeOfDesign` and `gammaA`, inside `design` `design` formatted through `rpact::getDesignGroupSequential()`
Control arm survival assumptions	`lambdaC`	`fail_rate` `fail_rate` formatted through `gsDesign2::define_fail_rate()`	`median2` or `lambda2`
Treatment arm survival assumptions	`hr`		`median1`, `lambda1`, or `hazardRatio`
Dropout rate(s)	`eta` and `etaE`		`dropoutRate1`, `dropoutRate2`, and `dropoutTime`
Enrollment rate(s) over time	`gamma`, `R`, and `ratio`	`enroll_rate` and `ratio` `enroll_rate` formatted through `gsDesign2::define_enroll_rate()`	`accrualTime`, `accrualIntensity`, and `accrualIntensityType`

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. See Practical 1 later

3. Refresher on graphical testing procedures in fixed sample designs

- Basic multiplicity corrections
- Introduction to graphical testing procedures
- ‘Epsilon’ edges
- Partial correlation

25 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
- We've already seen one reason this can arise: the presence of interim analyses
- Also routinely 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
  - (Whether FWER control is needed across multiple treatment arms can be a little more complex)
- **Multiple testing procedures** are methods for achieving such FWER control

Multiple testing procedures

How?

- Error-spending is a type of multiplicity correction that controls across multiple analyses
- Here focus on corrections for the case of multiple hypotheses in a fixed-sample setting
  - Simple corrections (Bonferroni, Dunnett, …)
    - Many others we don't discuss (Hochberg, Hommel, Šidák, …)
  - Closed testing procedures
  - **Graphical testing procedures** (GTPs)

Bonferroni

- Assume there are K null hypotheses, H_k for $k = 1, \dots, K$
- Uses an adjusted significance level for each hypotheses test of α/K
- Comes from:

$$\mathbb{P}(\text{Reject at least one of } H_1, \dots, H_K) \leq \sum_{k=1}^{K} \mathbb{P}(\text{Reject } H_k) = K\left(\frac{\alpha}{K}\right) = \alpha$$

- 'Weighted Bonferroni' would instead use a different level for each hypothesis, with the levels adding up to α
  - H_k tested at level α_k, with $\alpha_1 + \cdots + \alpha_K = \alpha$

Holm

- Suppose the p-values for the hypotheses are $p_1, \ldots, p_K$
- Holm is a stepwise procedure that first orders the p-values from lowest to highest

$$p_{(1)}, p_{(2)}, \ldots, p_{(K)}$$

- Compare each ordered p-value $p_{(i)}$ against $\frac{\alpha}{K+1-i}$
- Find the smallest i for which this comparison is false
- Reject all nulls up to $H_{(i)}$ and do not reject the others

Dunnett

- Developed to correct for testing multiple treatment arms against a shared control arm (many-to-one comparison)
- Same logic easily extends to any situation in which the joint distribution of the test statistics is multivariate normal
- I.e., suppose $\boldsymbol{Z} = (Z_1, \dots, Z_K)^\top \sim MVN(\mathbf{0}, \Sigma)$ when all nulls are true, where Z_k is the test statistics for hypothesis H_k
- Then if H_k is rejected if $Z_k > z_{1-\alpha'}$ we have

$$\begin{aligned}\mathbb{P}(\text{Reject at least one of } H_1, \dots, H_K) &= 1 - \mathbb{P}\{\max(Z_1, \dots, Z_K) \leq z_{1-\alpha'}\} \\ &= 1 - \int_{-\infty}^{z_{1-\alpha'}} \cdots \int_{-\infty}^{z_{1-\alpha'}} f(\boldsymbol{z}, \mathbf{0}, \Sigma)\, \mathrm{d}\boldsymbol{z}\end{aligned}$$

- So choose α' so that this probability is α

Hierarchical testing procedures

- Hypotheses can be ordered following a *pre-specified* hierarchy before the data are observed
  - Use clinical/commercial importance, power considerations, maturity time, …
- Hierarchical testing procedures test the hypotheses in this order
  - Fixed sequence procedure
  - Fallback procedure

Fixed sequence

- Each hypothesis is tested in the pre-specified sequence at level α until the first non-rejection
- Rejection rule:
  - If $p_1 \leq \alpha$, reject H_1 and continue; else stop
  - If $p_2 \leq \alpha$, reject H_2 and continue; else stop
  - …
  - $p_K \leq \alpha$, reject H_K and continue; else stop
- I.e., as soon as a hypothesis H_k cannot be rejected, because $p_k > \alpha$, the procedure stops and all remaining hypotheses $H_{k+1}, \ldots, H_k$ are not rejected
- Advantages:
  - Simple procedure that controls the FWER
  - Optimal (maximizes power) if previous hypotheses rejected
- Disadvantages:
  - Order of testing sequence is critical (and we may not have great data to choose it)
  - Minimizes power is a previous hypothesis is not rejected
  - Once a hypothesis is not rejected, no further testing is allowed

Fallback

- As for fixed sequence, test each hypothesis in the pre-specified sequence, but split the α between the hypotheses
- Assign α_k to hypothesis H_k, with $\alpha_1 + \cdots + \alpha_K = \alpha$
- H_1 is tested at level $\alpha_1' = \alpha_1$. Then for $k \geq 2$, H_k is tested at level α_k'

$$\alpha_1' = \begin{cases} \alpha_k & : \text{if } H_{k-1} \text{ is not rejected} \\ \alpha_k + \alpha_{k-1}' & : \text{if } H_{k-1} \text{ is rejected} \end{cases}$$

- Allows all hypotheses to be tested, even if initial hypotheses are not rejected

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

Closed testing procedures

- Very general and powerful methodology for constructing multiple testing procedures that strongly control the FWER
- Include many well-known procedures as special cases
- Relies on construction of tests for **intersection hypotheses**: see Backup
- Main **disadvantage**: Can be a very large number of intersection hypotheses to test
- We focus on a particular *shortcut* to specifying all intersection hypotheses

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, …

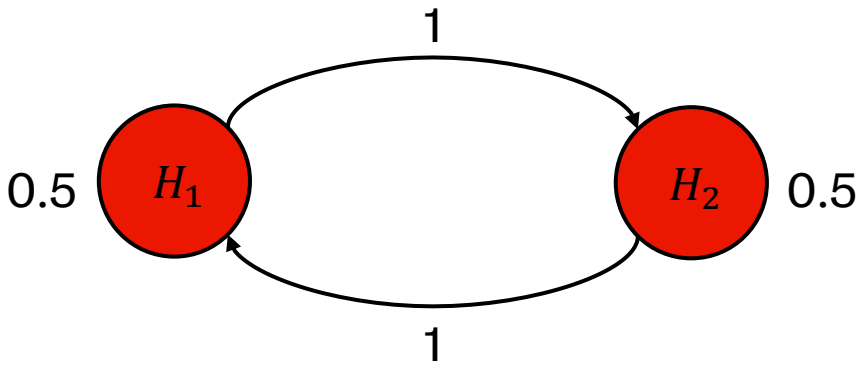
The graph

Specification

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

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

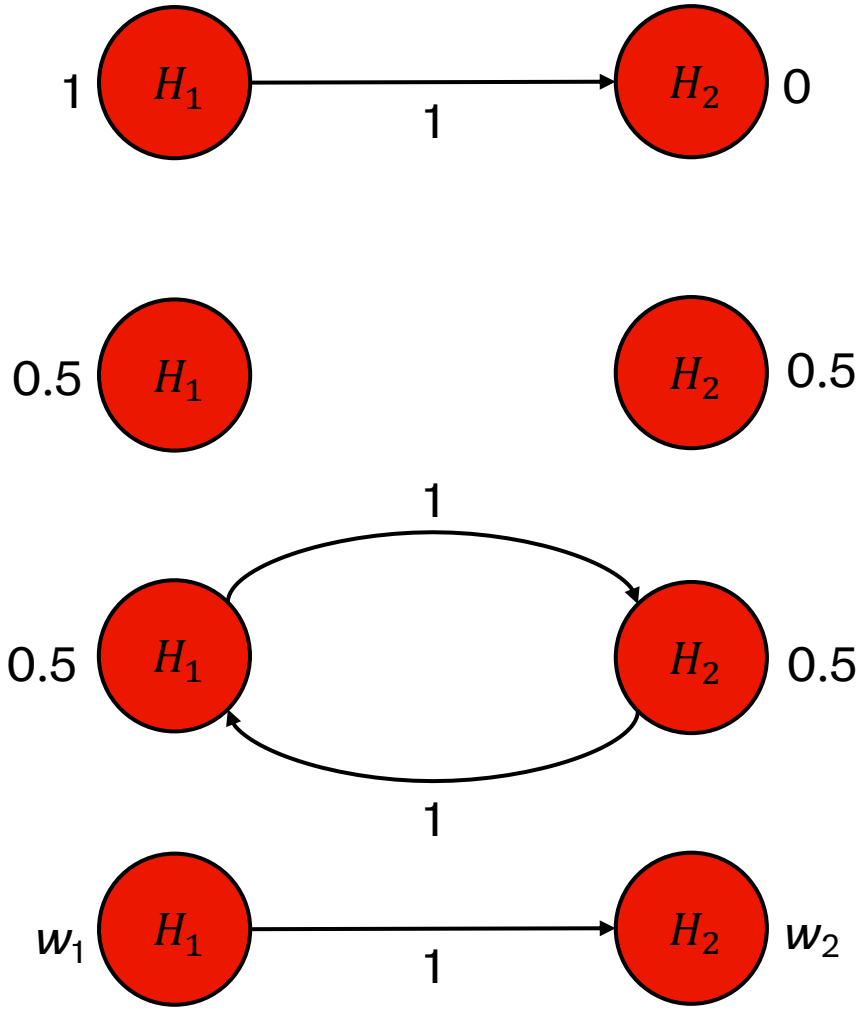
3. 'α-recycling' through **weighted directed edges**



Examples

$\boldsymbol{K = 2}$

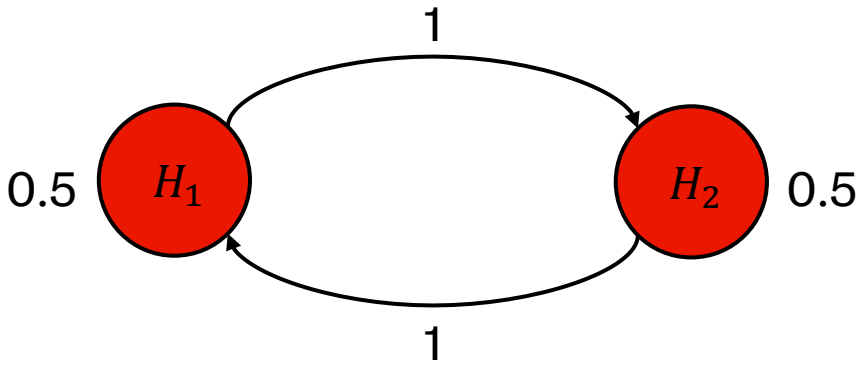
1. **Fixed sequence:** Maximises power if previous hypotheses rejected as all tests performed at level α
2. **Bonferroni:** No α-recycling
3. **Holm:** Everything in Bonferroni + more → more powerful
4. **Fallback**



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

- The graph defines a closed testing procedure with **weighted Bonferroni tests** 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 $0 \leq g_{ij} \leq 1$, $g_{ii} = 0$ and $\sum_{k=1}^{K} g_{ik} = 1$ for $i, j = 1, \ldots, K$

0. Set $\mathcal{K} = \{1, \ldots, K\}$
1. Select a $k \in \mathcal{K}$ such that $p_k \leq w_k \alpha$ 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 $|\mathcal{K}| \geq 1$, go to Step 1; otherwise stop

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

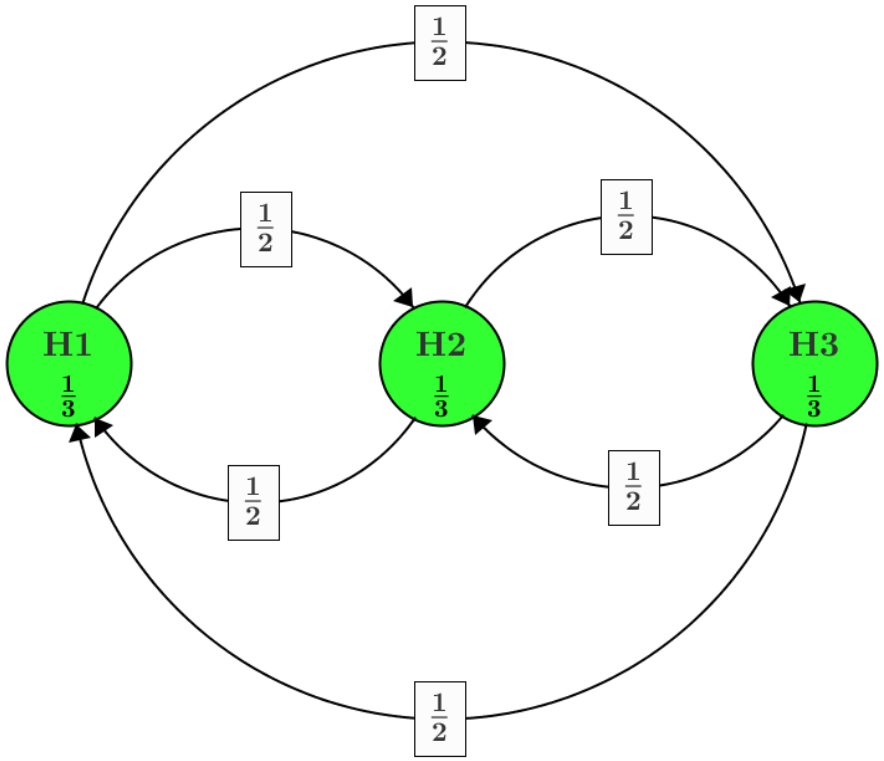
The graphical approach by Bretz et al. is a convenient tool to construct, visualize and perform multiple test procedures that are tailored to structured families of hypotheses while controlling the familywise error rate. A critical step is to update the transition weights following a pre-specified algorithm. 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

Graphical testing procedure defines intersection hypothesis tests in a closed test

Example



J	$w_1(J)$	$w_2(J)$	$w_3(J)$
$H_1 \cap H_2 \cap H_3$	1/3	1/3	1/3
$H_1 \cap H_2$	0.5	0.5	-
$H_1 \cap H_3$	0.5	-	0.5
$H_2 \cap H_3$	0	0.5	0.5
H_1	1	-	-
H_2	-	1	-
H_3	-	-	1

Consonance

Graphical testing procedure guarantees consonance

Let p_k denote an unadjusted observed p-value for H_k

J	$w_1(J)$	$w_2(J)$	$w_3(J)$	$p_3 < w_3(J)\alpha$
$H_1 \cap H_2 \cap H_3$	1/3	1/3	1/3	**Yes**
$H_1 \cap H_2$	1/2	1/2	-	
$H_1 \cap H_3$	1/2	-	1/2	Yes
$H_2 \cap H_3$	0	1/2	1/2	Yes
H_1	1	-	-	
H_2	-	1	-	
H_3	-	-	1	Yes

$p_3 < w_3(J)\alpha$ implies

So can immediately reject H_3

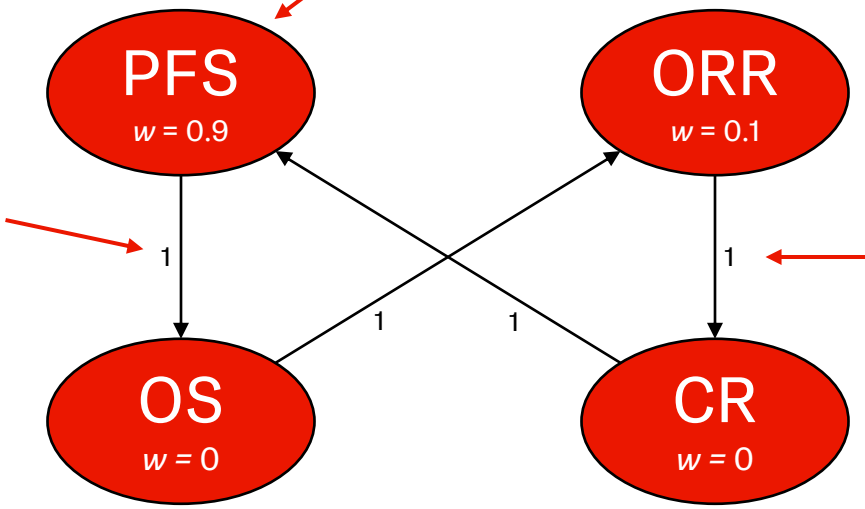
Consonance is ensured because $w(J_1) \leq w(J_2)$ for $J_2 \subseteq J_1$

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



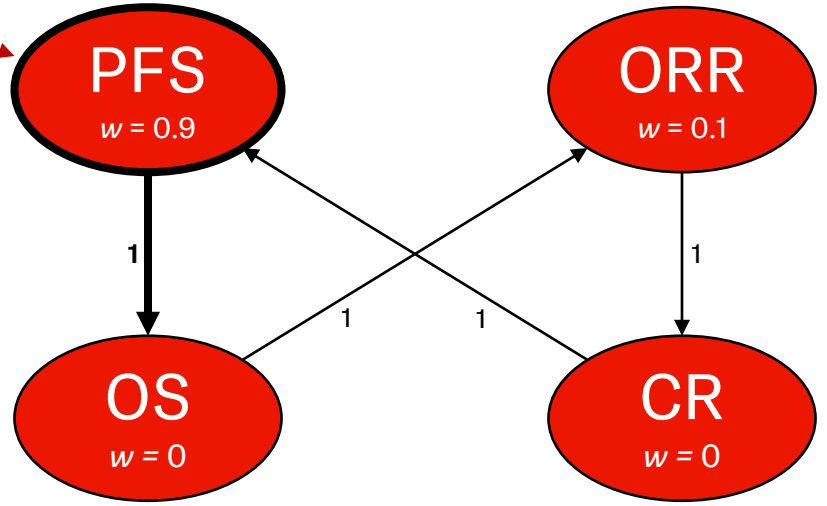
ORR recycles all its α to CR+, because of their similar maturation time

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

Running example 2

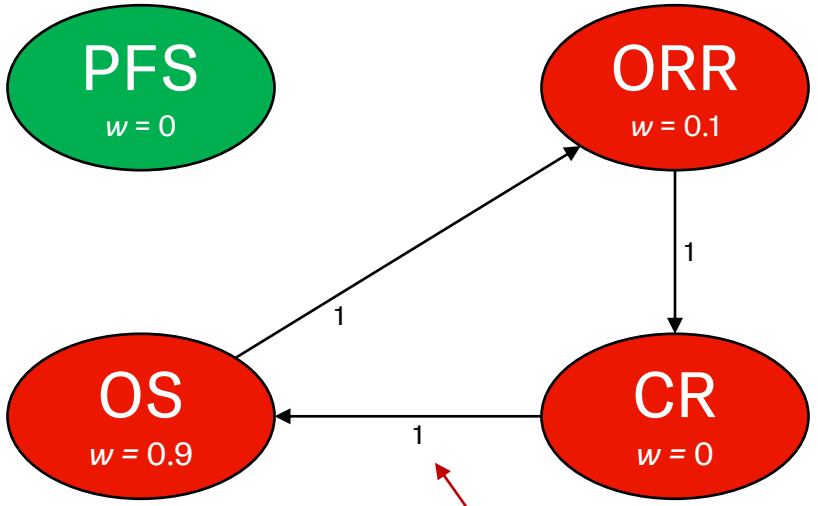
Sequential updating

Suppose we achieve significance for PFS



Running example 2

Sequential updating

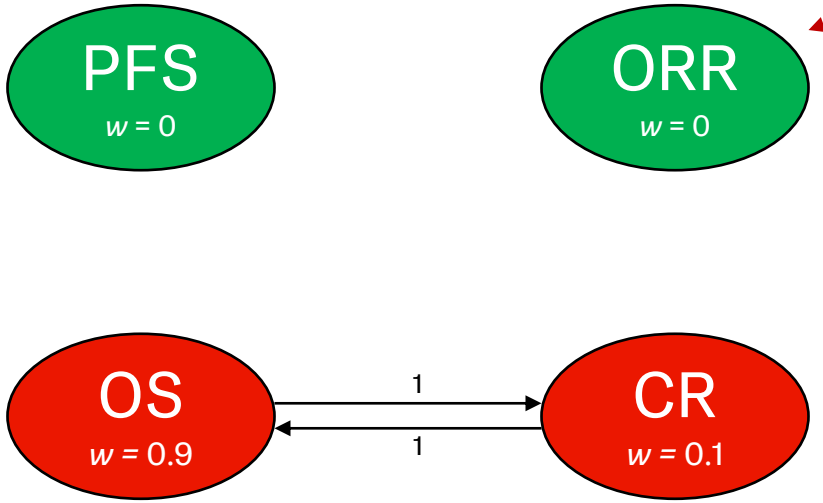


There are now edges that weren't previously in the graph

Running example 2

Sequential updating

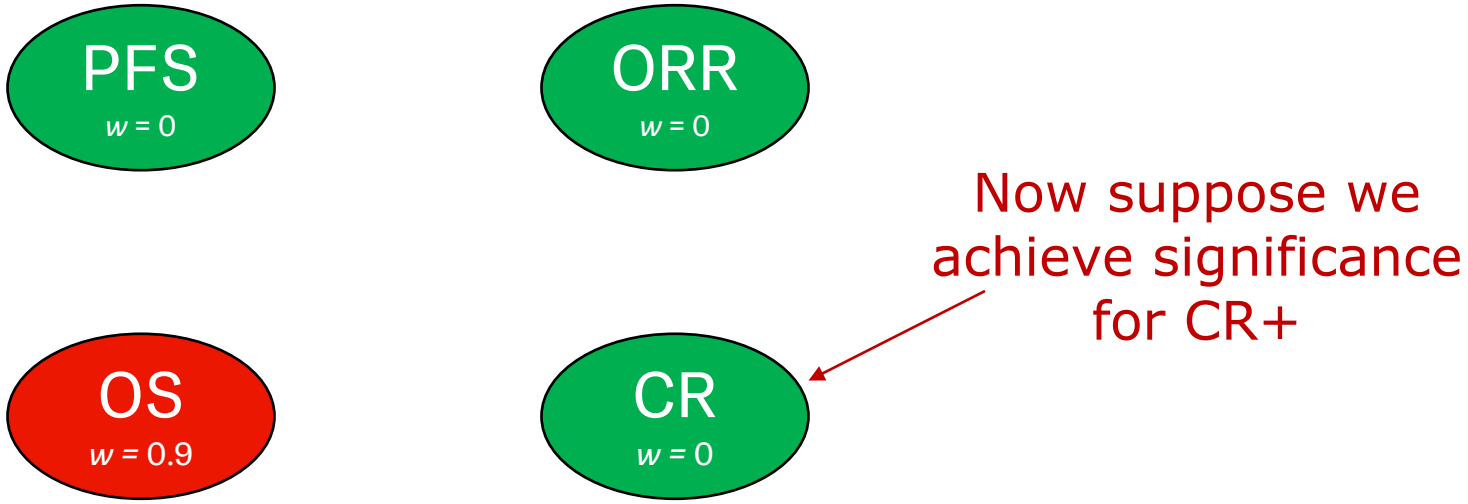
Next suppose we achieve significance for ORR



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

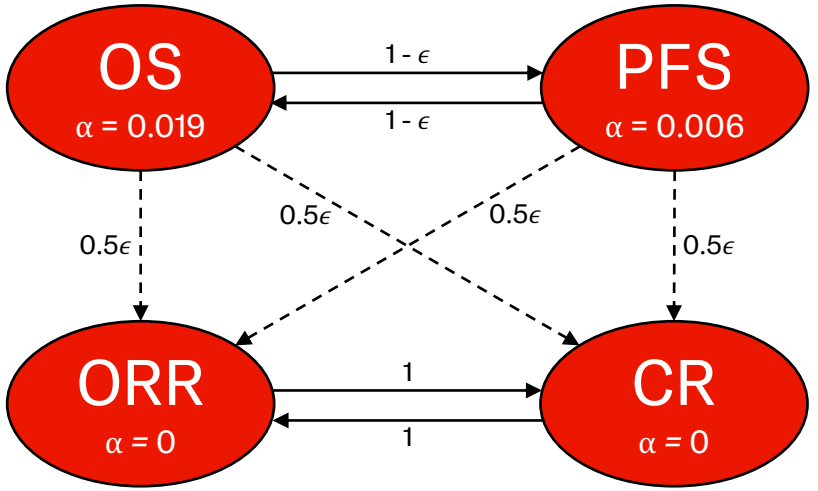


‘Epsilon’ edges

A form of sequential gatekeeping

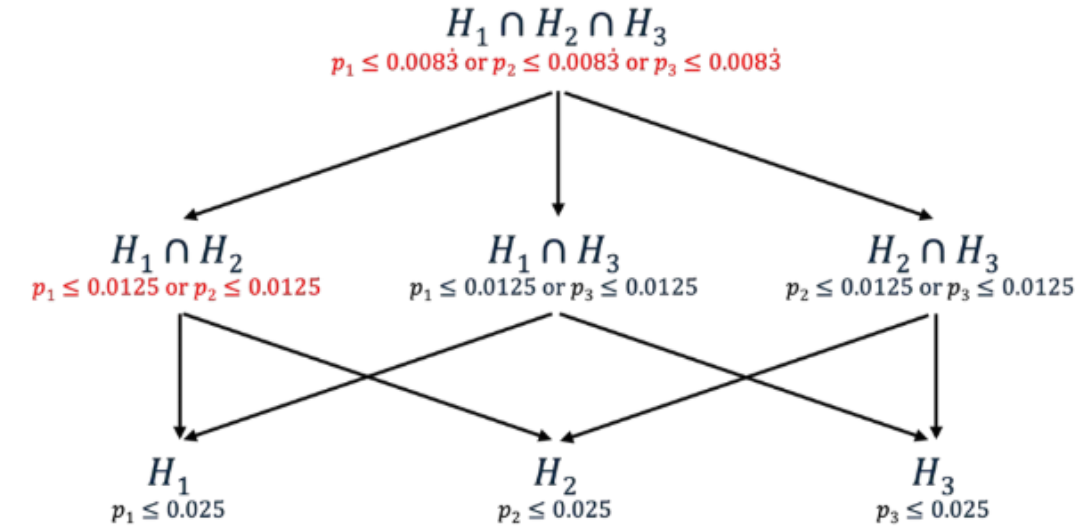
- Infinitesimally small values on recycling edges can be used to indicate that several hypotheses must be rejected before any alpha is recycled on to some additional set of hypotheses

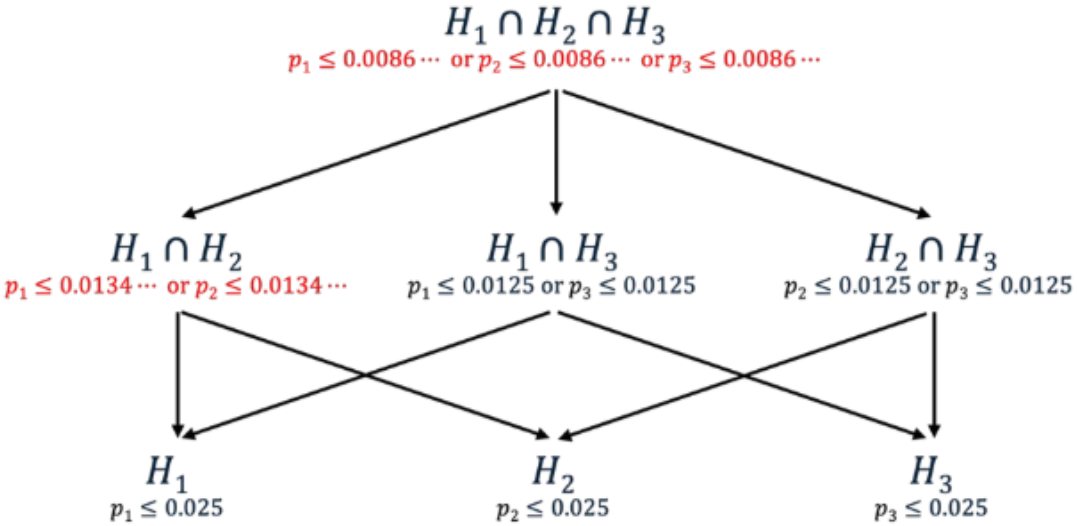
- E.g., modifying the KEYNOTE-598 graph
  - Suppose CR is also to be tested
  - ORR and CR only tested after both OS and PFS significance



Partial correlation

- Standard GTP makes no assumption of knowledge about the correlation between tests performed for the various hypotheses
- If we do wish to use such information, it is relatively straight-forward to do so
  - E.g., if we know the correlation by design (sub- and overall population tested for the same endpoint)
  - See Bretz *et al.* (2011) for further details
- Suppose that $K = 3$ and we know that $Corr(Z_1, Z_2) = 0.5$. Then Holm's procedure with and without use of partial correlation is as shown for $\alpha = 0.025$
  - Nothing assumed about other correlations
- In practice, power gain from using partial correlation is often small: α-recycling is the more important component of the multiplicity strategy





Operating characteristics

- Typically, we are most interested in the power for each individual hypothesis at specific levels of alpha availability
  - E.g., what is the power for PFS when tested at the one-sided 0.02 level?
- This is a type of conditional power that doesn't require multiplicity-specific software for calculation
- If we wish to calculate more complex powers, we will need to set an assumption regarding the joint distribution of the test statistics for the various hypotheses
  - E.g., disjunctive/conjunctive powers across some set of hypotheses
  - E.g., multiplicity-adjusted power for a specific hypothesis
- Assuming they are uncorrelated is often not an awful approach
- Simulation is typically the best approach to then calculating these powers
  - See `gMCP::calcPower()`

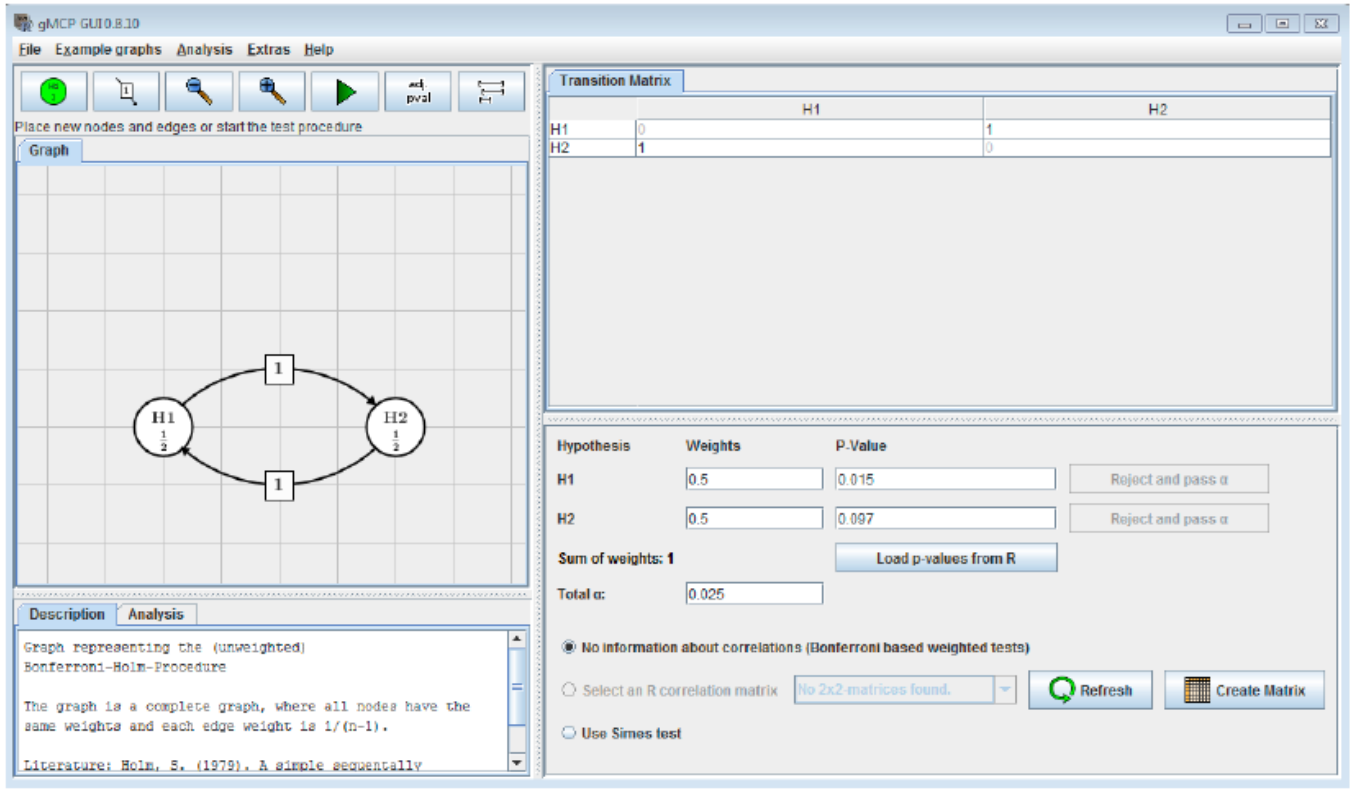
Selecting a GTP

And more generally how to choose a multiplicity correction

- In a purely statistical sense, splitting and recycling alpha in more complex way will generally provide a more robust testing procedure
  - Will see this if you search for an optimal GTP
- But can provide little benefit. In practice we typically
  - Factor in rate at which endpoints mature (e.g., OS last; short-term outcomes first)
  - Test short-term outcomes (that are often highly correlated) in a hierarchical manner. E.g., ORR and then CR+
  - Give small amount of alpha to short-term dual primary outcome, so minimal risk to/impact on conventional primary outcome
  - Recycle some alpha between dual primary outcomes, to increase chance we win on both
  - Select edges based on trade-offs. I.e., how much does power go up for one hypothesis compared to how much it goes down for another
- A lot of publications and regulatory guidance available
  - E.g., Dmitrienko and D'Agostino (2013); Dmitrienko and D'Agostino (2018); Li *et al.* (2017); Wang *et al.* (2011)
  - EMA: Guideline on multiplicity issues in clinical trials
  - FDA: Multiple endpoints in clinical trials guidance for industry

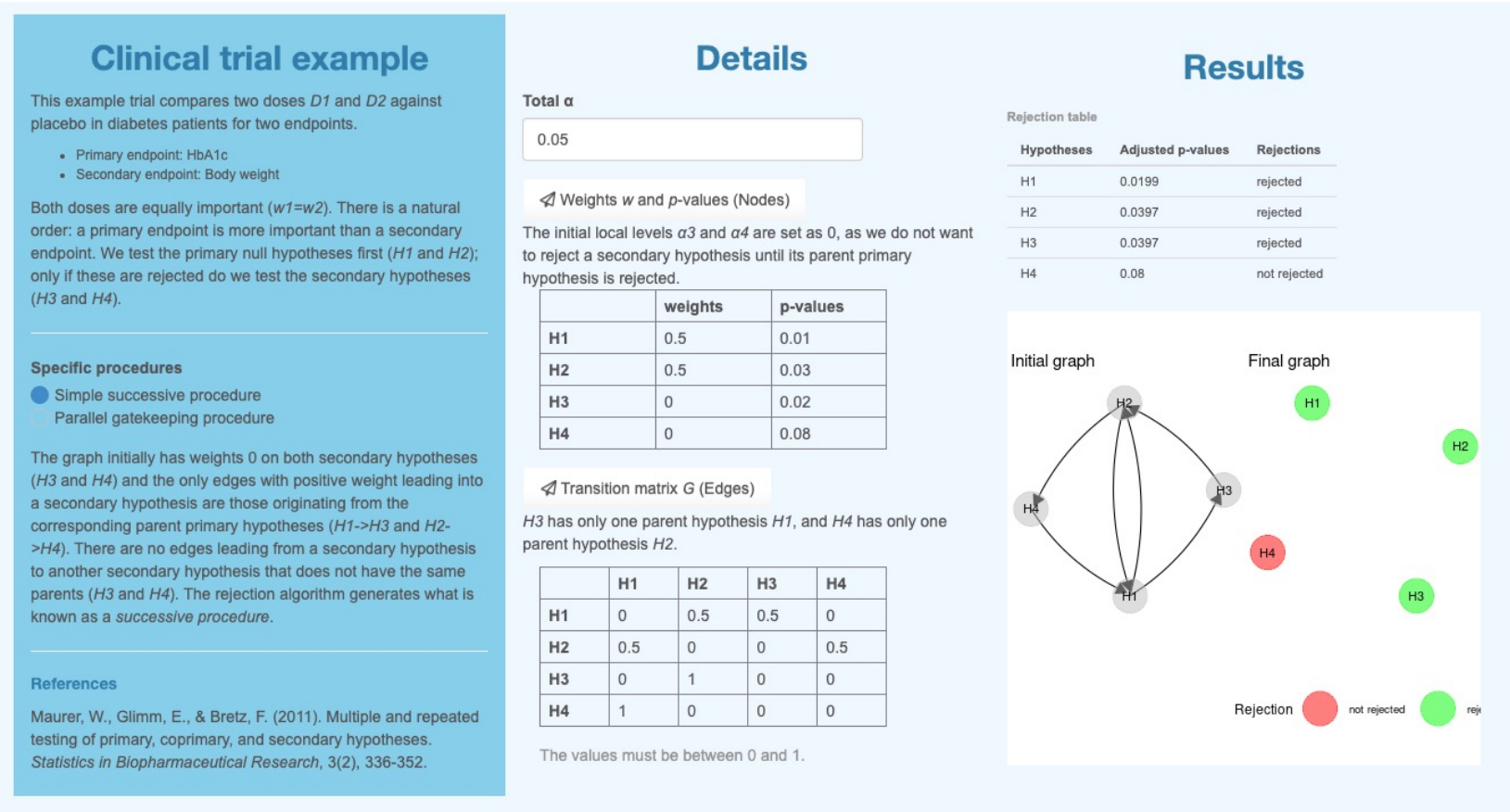
Software

- 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



GraphApp

https://mrc-bsu.shinyapps.io/20MRC_BSU_GraphApp/



Clinical trial example

This example trial compares two doses *D1* and *D2* against placebo in diabetes patients for two endpoints.

- Primary endpoint: HbA1c
- Secondary endpoint: Body weight

Both doses are equally important (*w1=w2*). There is a natural order: a primary endpoint is more important than a secondary endpoint. We test the primary null hypotheses first (*H1* and *H2*); only if these are rejected do we test the secondary hypotheses (*H3* and *H4*).

Specific procedures

- (●) Simple successive procedure
- ( ) Parallel gatekeeping procedure

The graph initially has weights 0 on both secondary hypotheses (*H3* and *H4*) and the only edges with positive weight leading into a secondary hypothesis are those originating from the corresponding parent primary hypotheses (*H1->H3* and *H2->H4*). There are no edges leading from a secondary hypothesis to another secondary hypothesis that does not have the same parents (*H3* and *H4*). The rejection algorithm generates what is known as a *successive procedure*.

References

Maurer, W., Glimm, E., & Bretz, F. (2011). Multiple and repeated testing of primary, coprimary, and secondary hypotheses. *Statistics in Biopharmaceutical Research*, 3(2), 336-352.

Details

Total α

0.05

Weights *w* and *p*-values (Nodes)

The initial local levels *α3* and *α4* are set as 0, as we do not want to reject a secondary hypothesis until its parent primary hypothesis is rejected.

	weights	p-values
H1	0.5	0.01
H2	0.5	0.03
H3	0	0.02
H4	0	0.08

Transition matrix *G* (Edges)

H3 has only one parent hypothesis *H1*, and *H4* has only one parent hypothesis *H2*.

	H1	H2	H3	H4
H1	0	0.5	0.5	0
H2	0.5	0	0	0.5
H3	0	1	0	0
H4	1	0	0	0

The values must be between 0 and 1.

Results

Rejection table

Hypotheses	Adjusted p-values	Rejections
H1	0.0199	rejected
H2	0.0397	rejected
H3	0.0397	rejected
H4	0.08	not rejected

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. Break

10 mins

5. Short practical on group-sequential design / graphical testing using R

- {gsDesign} and {rpact}
- {gMCP}/{gMCPLite} and {graphicalMCP}

30 mins

Practical 1

Recreating the PFS, OS, ORR, and CR designs in Running example 2

- See practical1.pdf
  - As well as practical1_solutions.R and practical1_solutions.pdf
- Primary goal is to recreate each of the (group-sequential or fixed sample) designs for the four endpoints included in the multiplicity plan
- Then, ignoring the presence of interim analyses, to compute multiplicity-adjusted power
- Feel free to use {gsDesign} or {rpact} and {gMCP}/{gMCPLite} or {graphicalMCP} as you prefer
  - And also the GUI interface to {gsDesign} or {rpact} if preferred
- Other exercises also available (e.g., on optimal group-sequential design)
- Do ask questions 🙂

6. Graphical testing in group-sequential designs

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

45 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** α (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

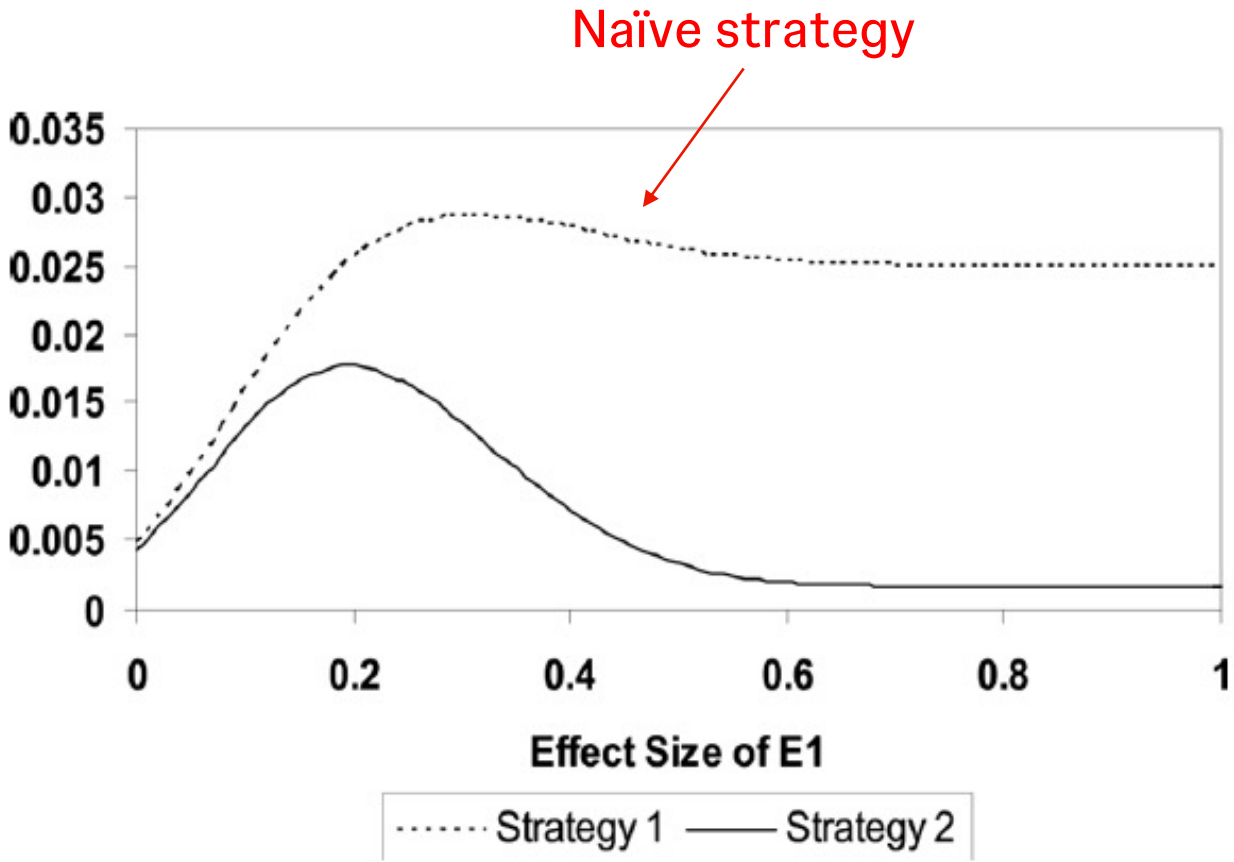


Figure 1 Type I error rate of E2 ($\rho = 0.5$).

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 $j = 1, \ldots, J$
  - 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 γ
  - Let
    - $f(\gamma, t_j)$ denote the spending function, with $f(\gamma, 1) = \gamma$
    - $\pi_j(\gamma) = f(\gamma, t_j) - f(\gamma, t_{j-1})$
    - $p_j^*(\gamma)$ is the corresponding threshold for the nominal p-value at analysis j
- Need a special condition called a ‘**well ordered**’ boundary:

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

  - For $\gamma \leq \gamma'$, $\pi_j(\gamma) \leq \pi_j(\gamma')$ for all $j \Longrightarrow p_j^*(\gamma) \leq p_j^*(\gamma')$
    - Another formulation: $f(\gamma, t_j) \leq f(\gamma', t_j)$ if $\gamma \leq \gamma'$ for all j and the second mixed partial derivative of $f(.,.) \geq 0$

Defining spending function for each hypothesis

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

Testing algorithm

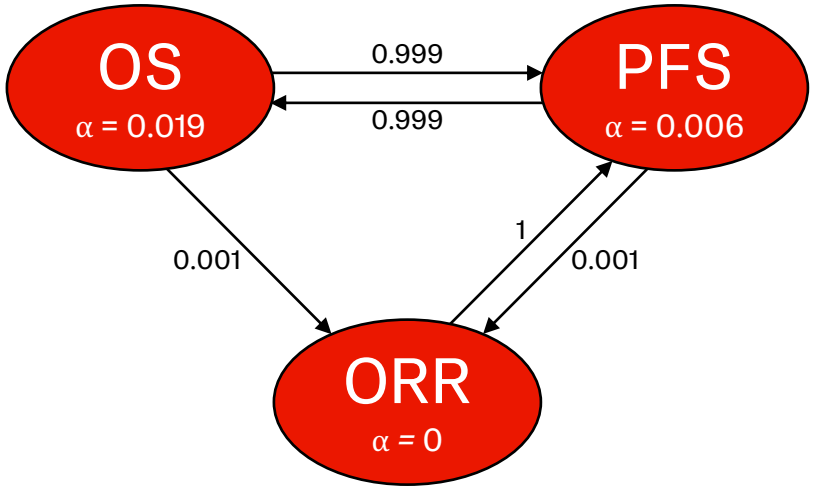
- Start with $j = 1$

1. Test each hypothesis k (not previously rejected at or before) analysis j:
   - Compute the nominal p-value threshold $p^*_{lk}(w_k\alpha)$, $l = 1, \ldots, j$, 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 $l = 1, \ldots, j$, 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

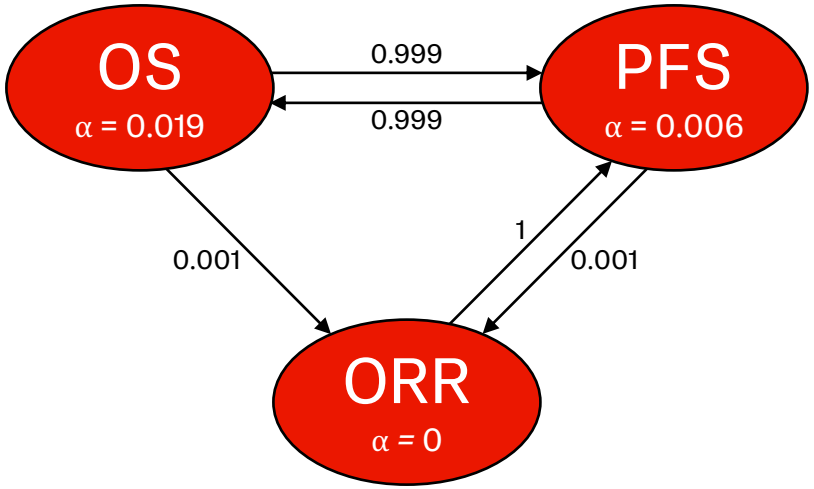
Initial GTP and GSD for each hypothesis

- 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 $\alpha = 0.025$

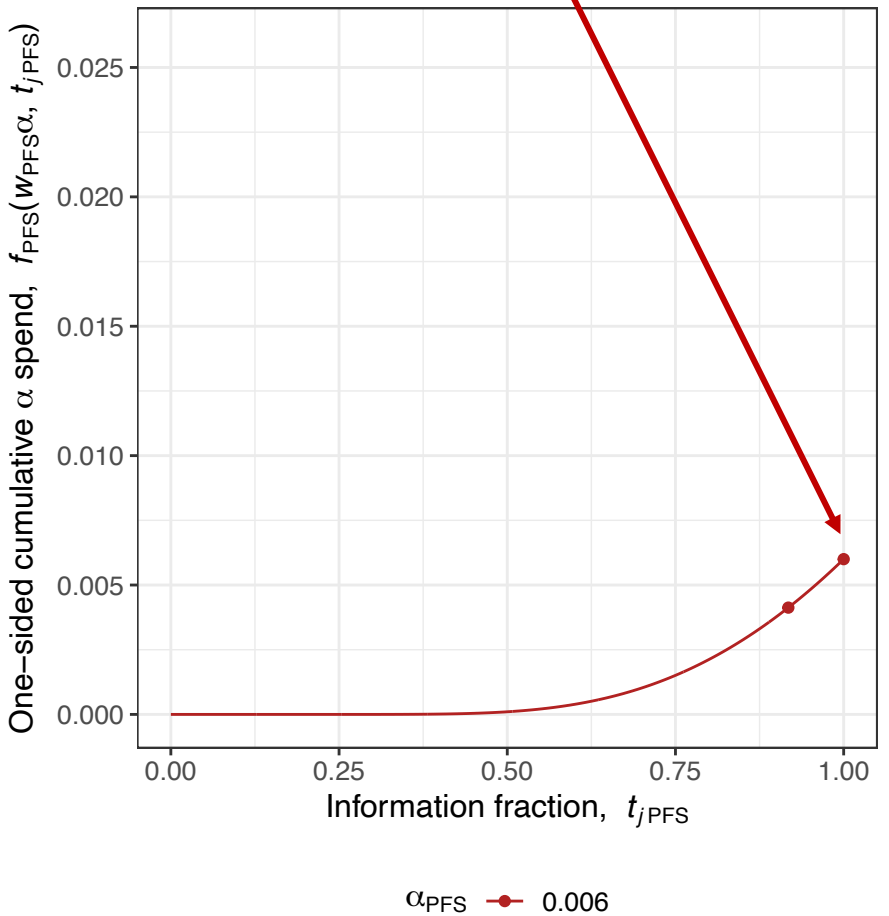


Running example

Focus on PFS

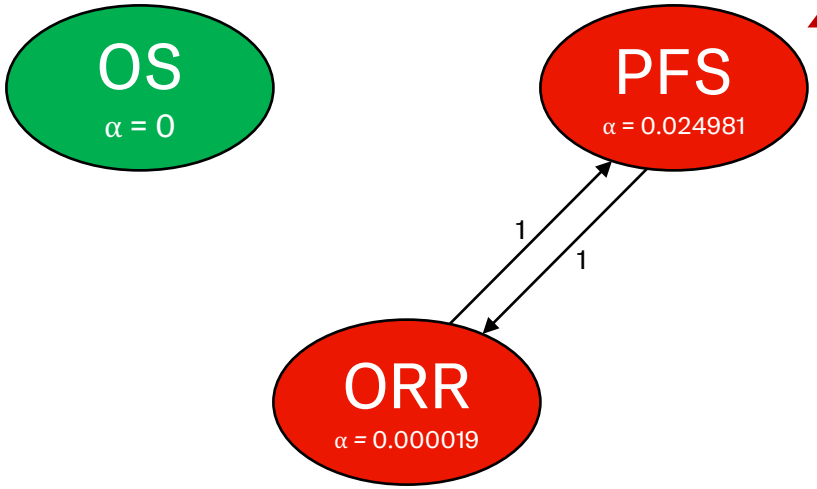


To begin with, can only spend
$\alpha = 0.006$ in total

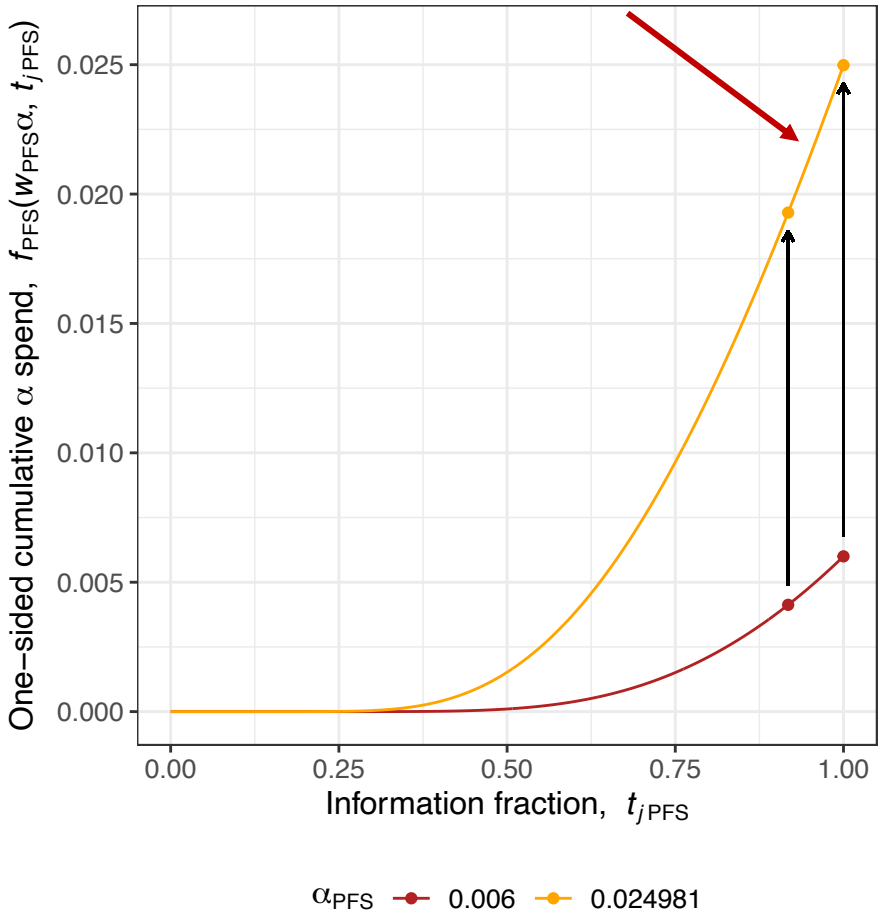


Running example

Focus on PFS

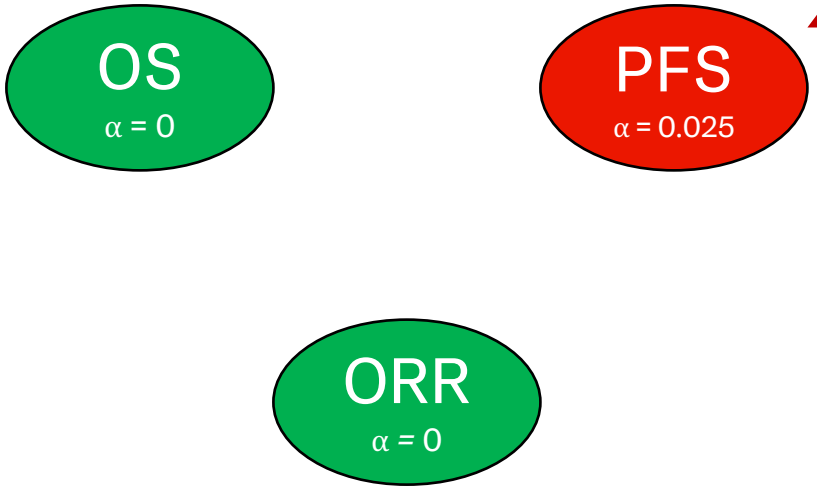


If the graph updates the allowed total spend, the whole spending function updates

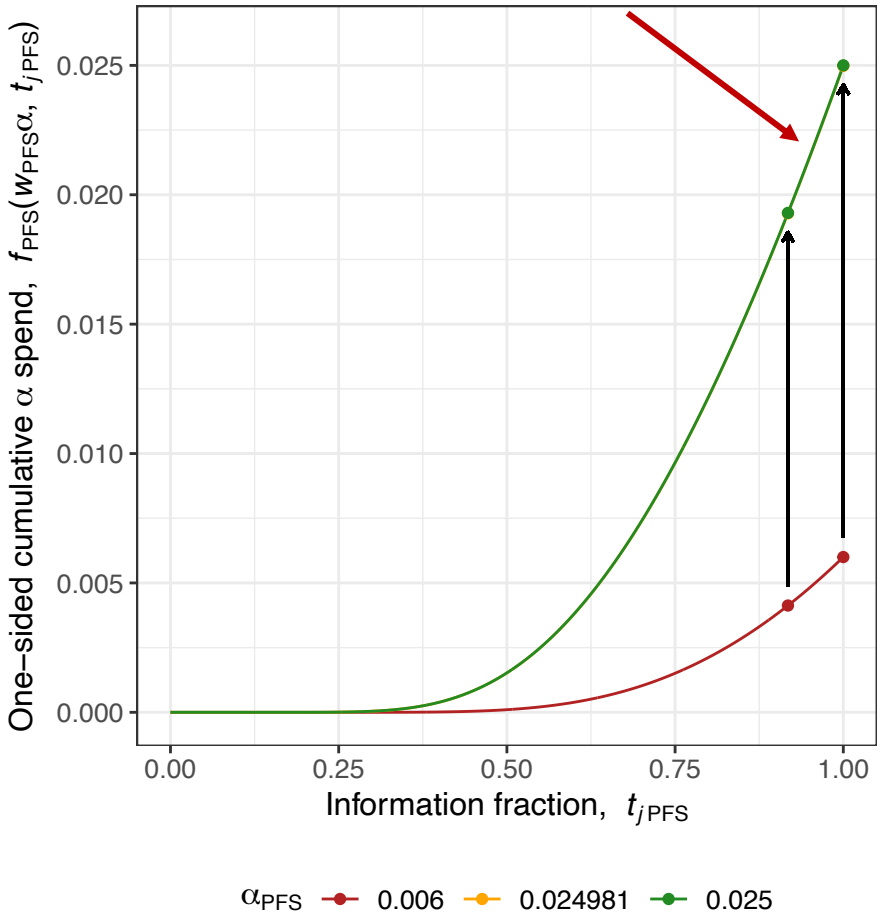


Running example

Focus on PFS

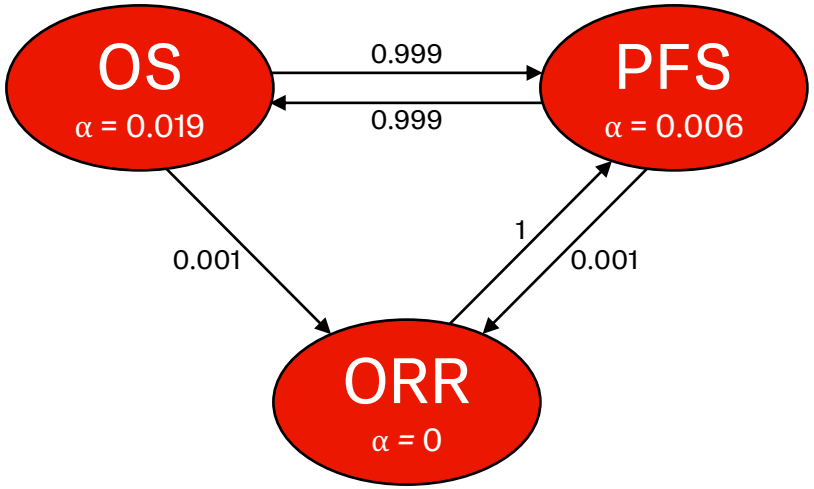


If the graph updates the allowed total spend, the whole spending function updates

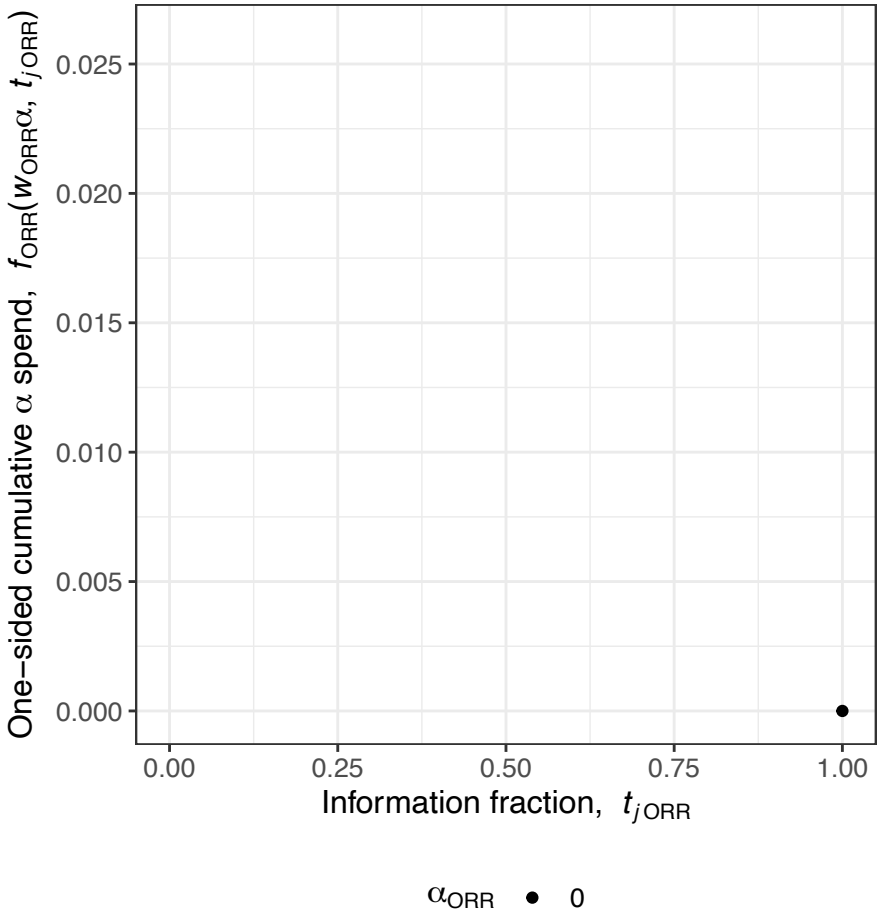


Running example

Focus on ORR

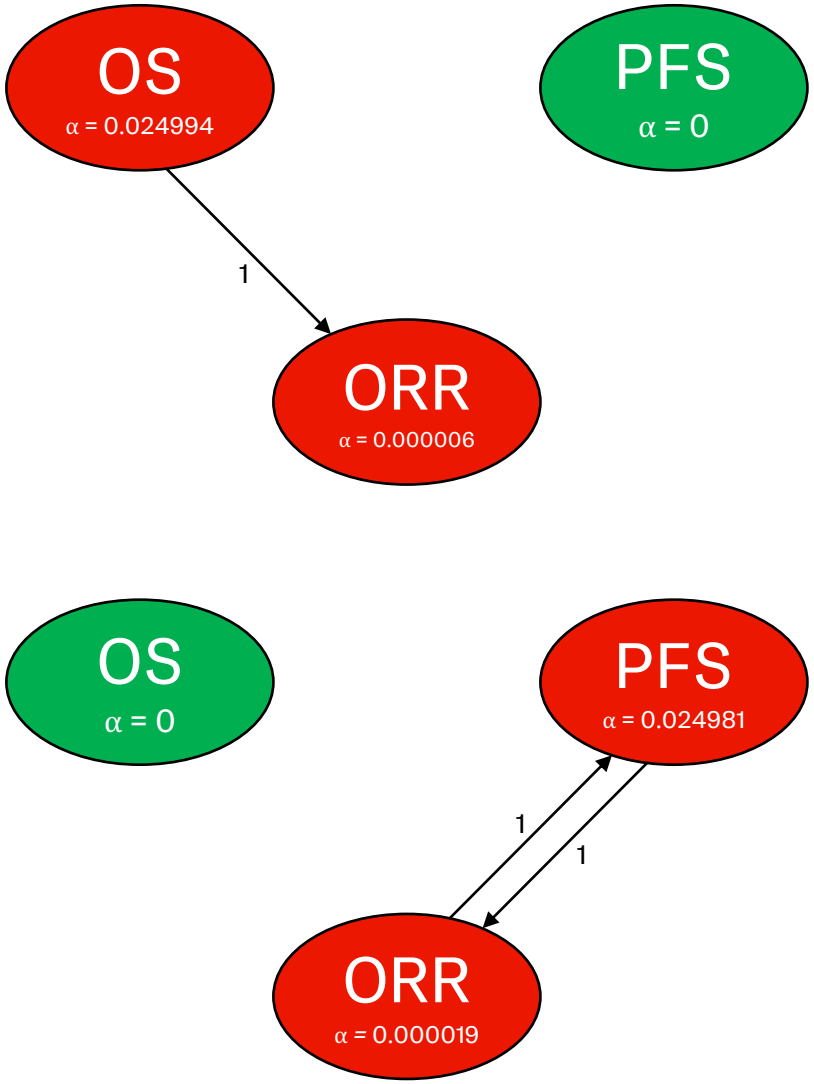


To begin with, cannot spend any α

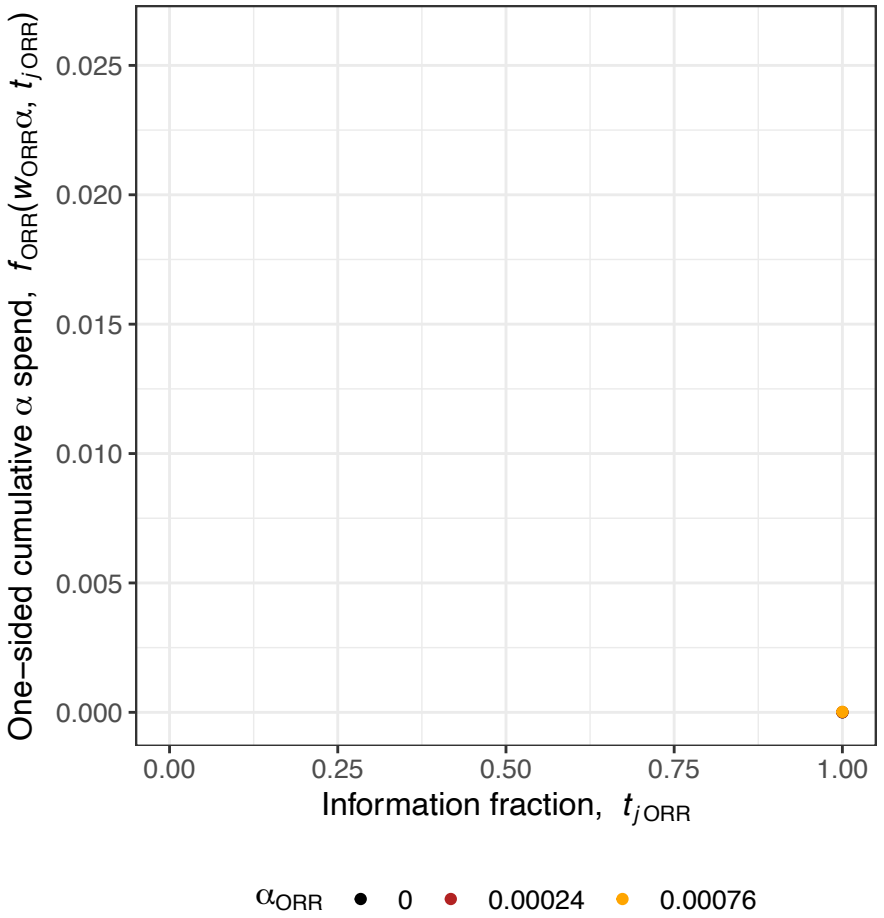


Running example

Focus on ORR

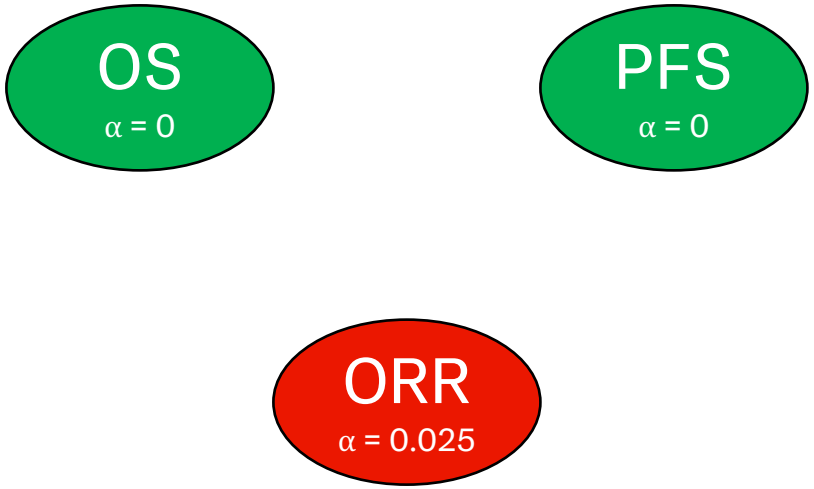


Significance on either OS or PFS would require very low *p*-value for significance

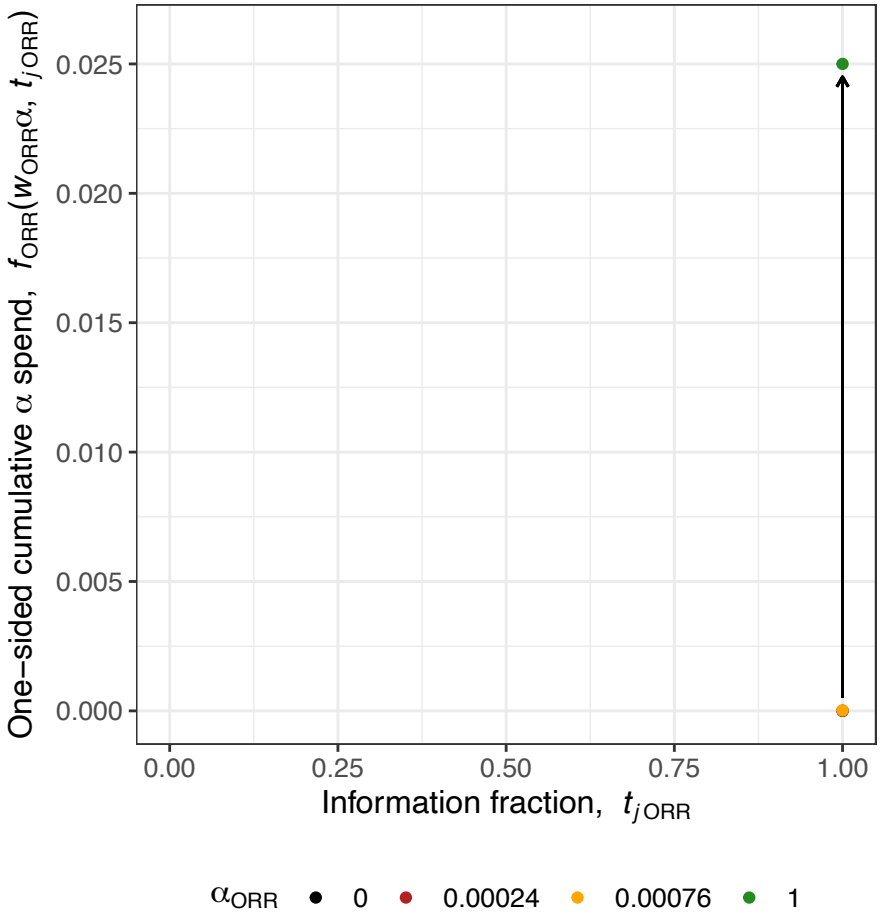


Running example

Focus on ORR

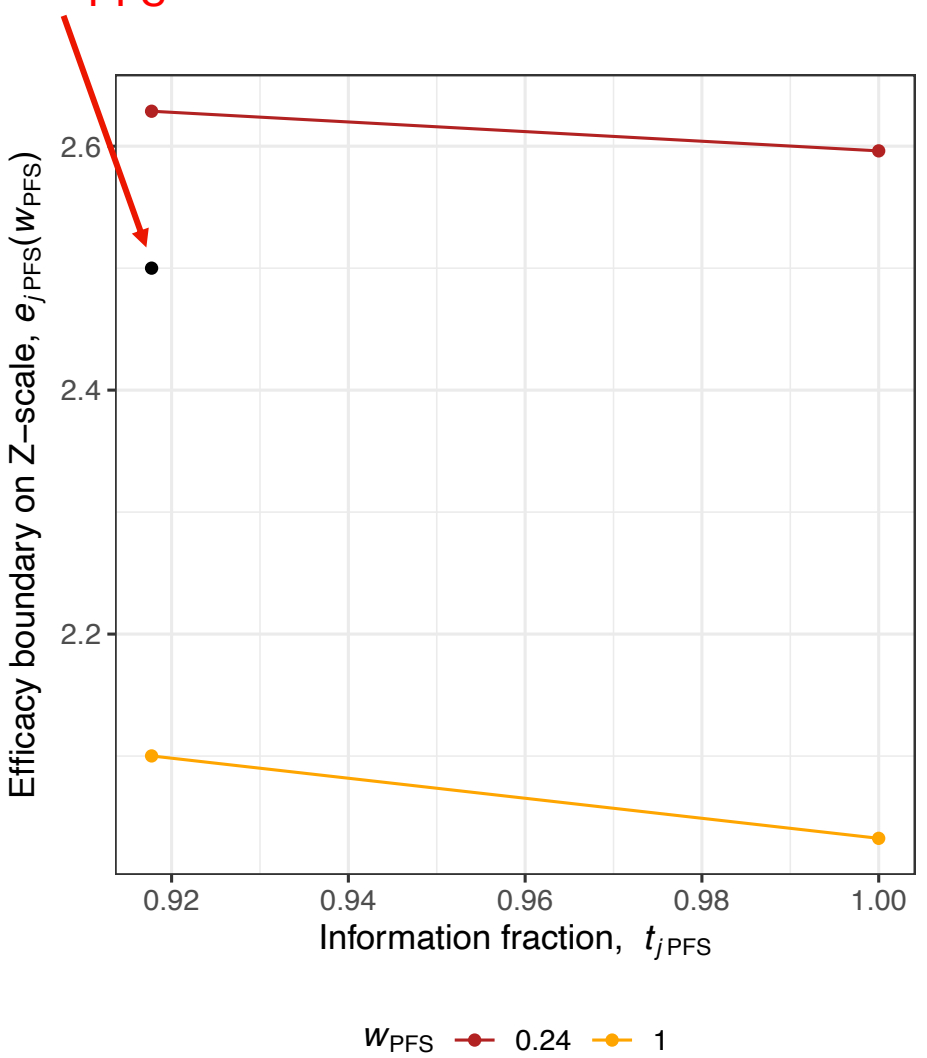


Significance on both OS or PFS allows for higher likelihood of ORR success



‘Look back’ analyses

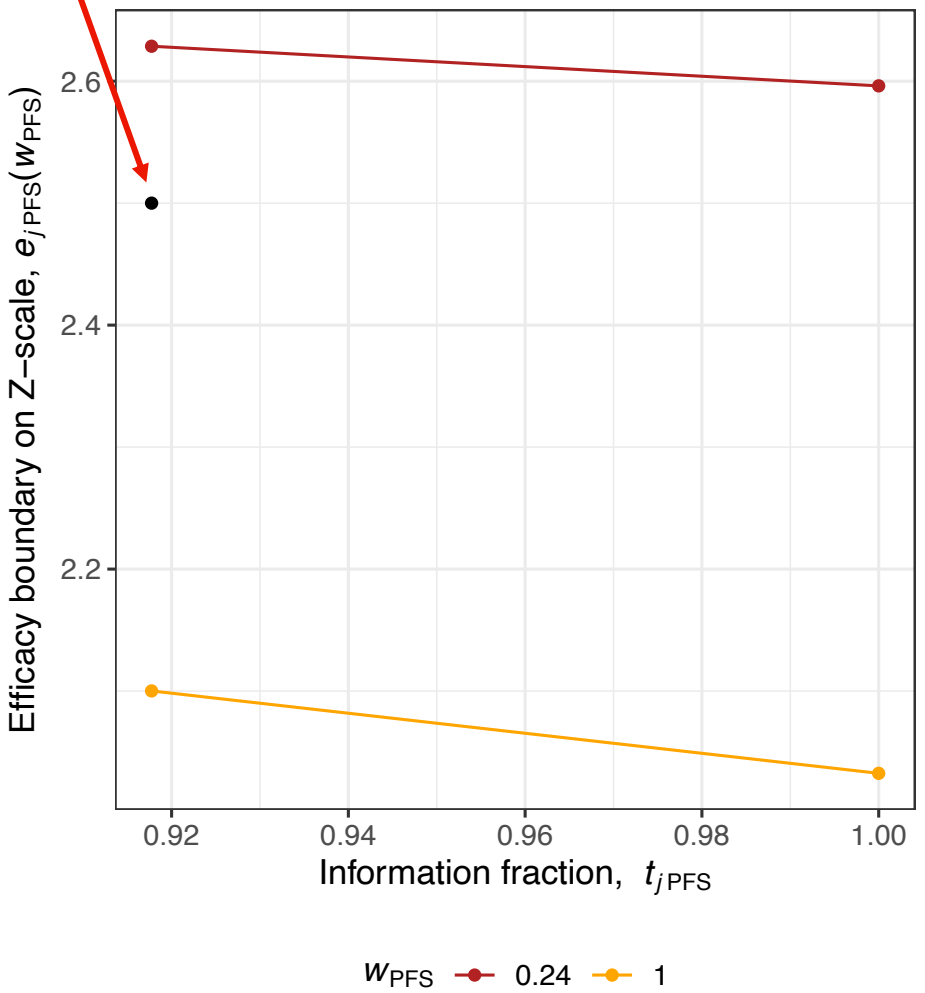
- The algorithm allows for what has been termed ‘look back’ analyses
- E.g., consider PFS in the KEYNOTE-598 example
- Suppose that at IA1 we have to stay at $w_{\mathrm{PFS}} = 0.24$ (because OS wasn’t rejected). Then we aren’t able to reject H_{PFS} based on the black dot in the plot



‘Look back’ analyses

- If we reach $w_{\mathrm{PFS}} = 1$ 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

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



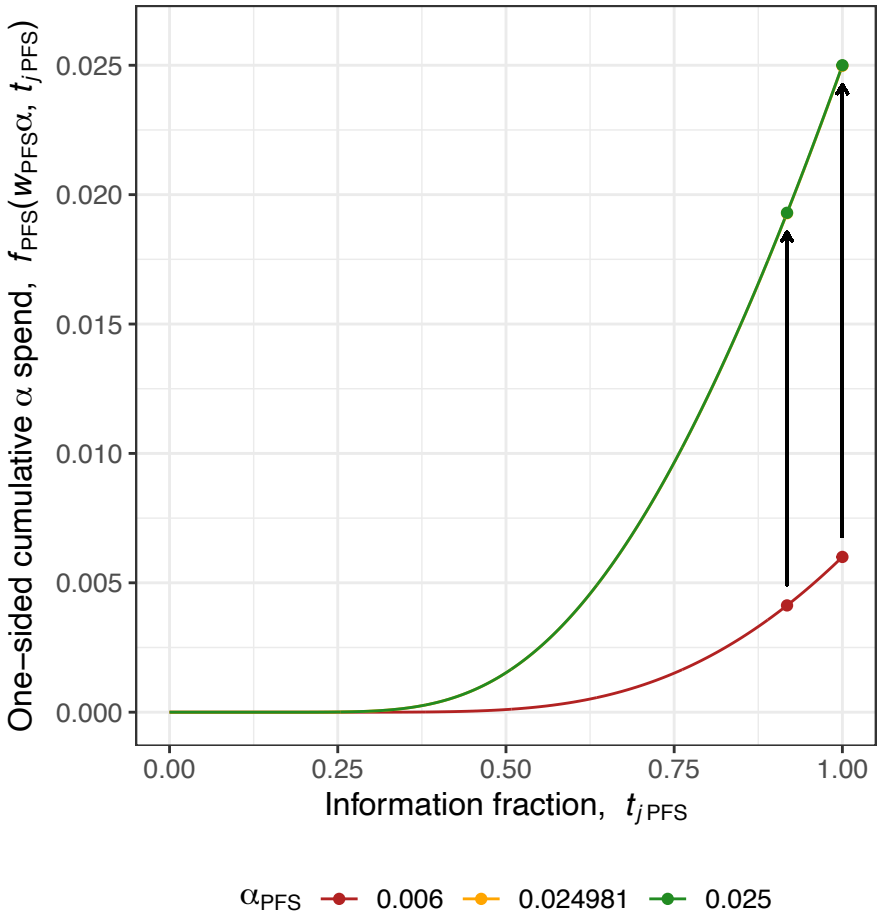
‘Look back’ analyses

- 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

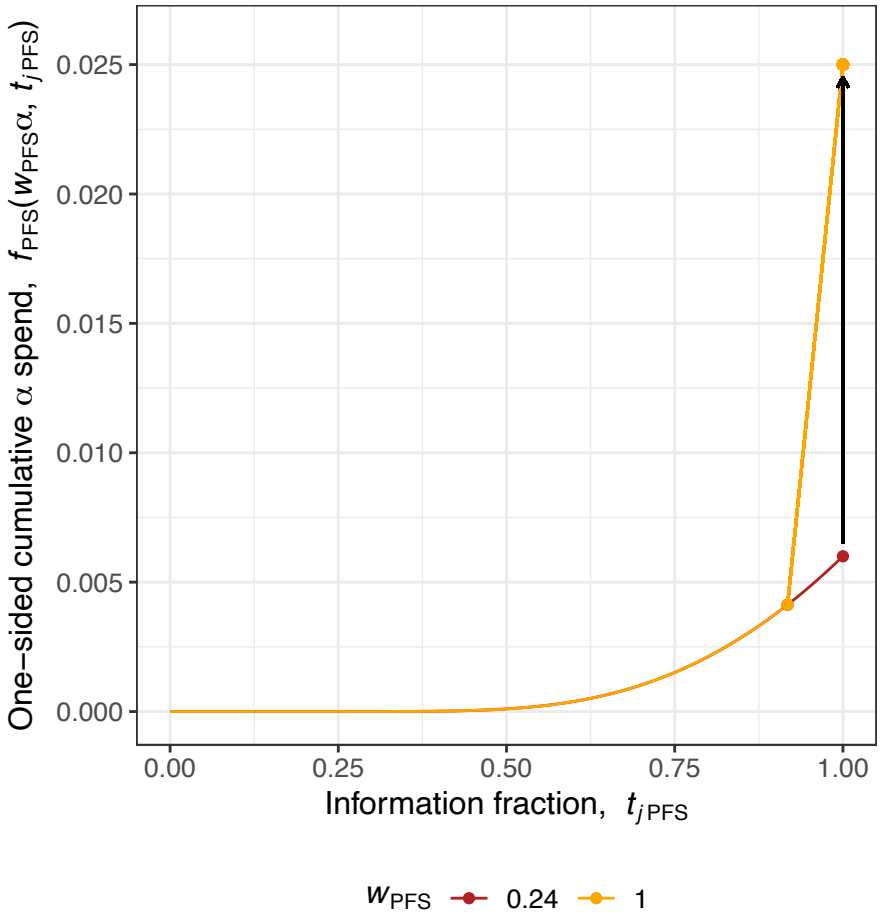
The approach for PFS in Running example 1: KEYNOTE-598

- 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



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

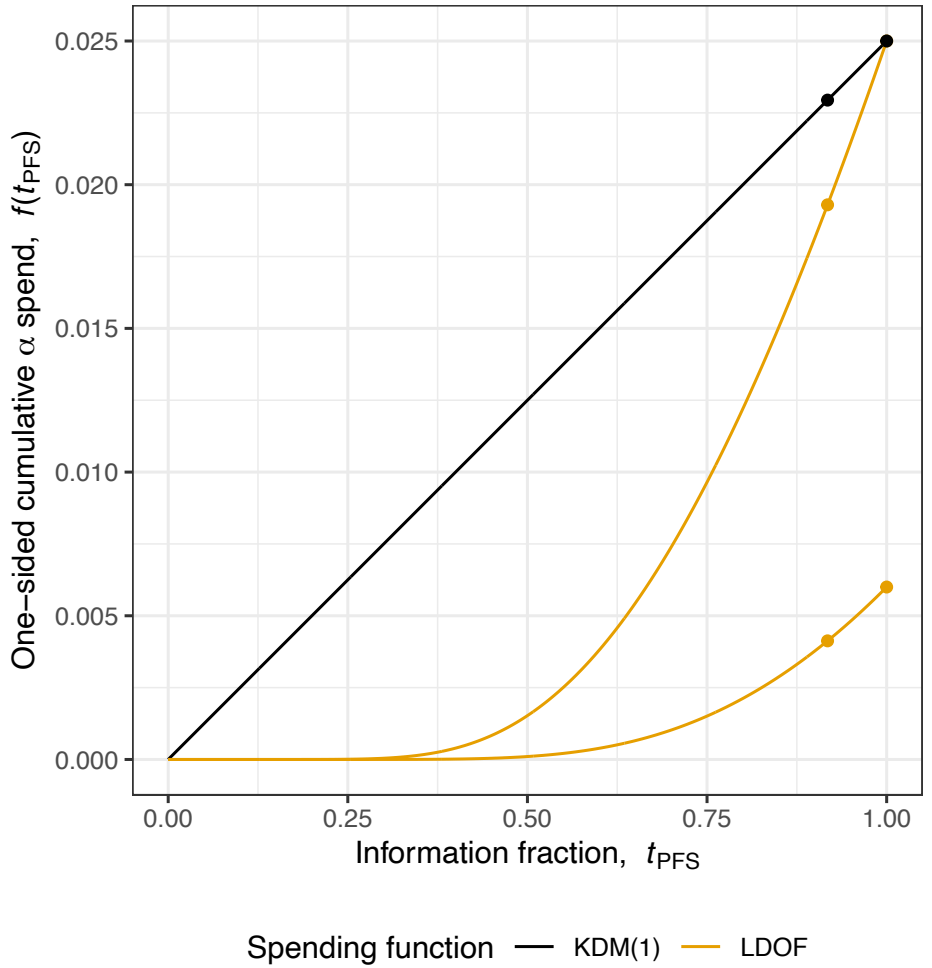
Which is best?

- 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



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

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

1. **Enrollment information:** 568 pts, randomized 1:1 over 20 mo
2. **Hypotheses:**
   - i. OS: Analysed at all 3 analyses, using LDOF spending function with immediate recycling
   - ii. PFS: Analyses at the first 2 analyses, using LDOF spending function with immediate recycling
   - iii. ORR: Matures at the first analysis
3. **Analyses:**
   - IA1. 255 OS events
   - IA2. 307 OS events
   - FA. 361 OS events
4. **Distributional information:**
   - i. OS: Exponential with a median of 20 mo. HR = 0.7. 1% yearly drop-out
   - ii. PFS: Piece-wise exponential with a median of 6.5 mo before 6.5 mo and a median of 14.5 mo after 6.5 mo. HR = 0.69. 13% yearly drop-out
   - iii. ORR: 59% vs 39%
5. **Graphical testing procedure:** See earlier

Running example 1: KEYNOTE-598

Example implementation in practice: IA1

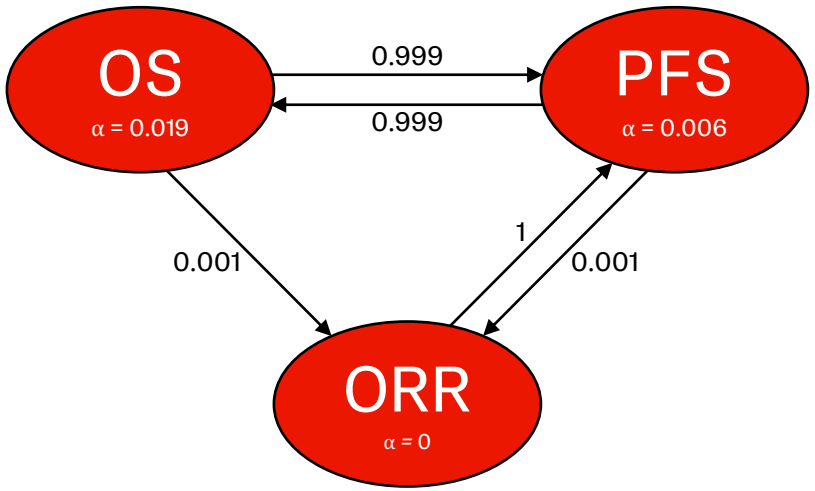
Accrued events / sample size for each hypothesis

Analysis	OS	PFS	ORR
IA1	255	356	568
IA2			
FA			

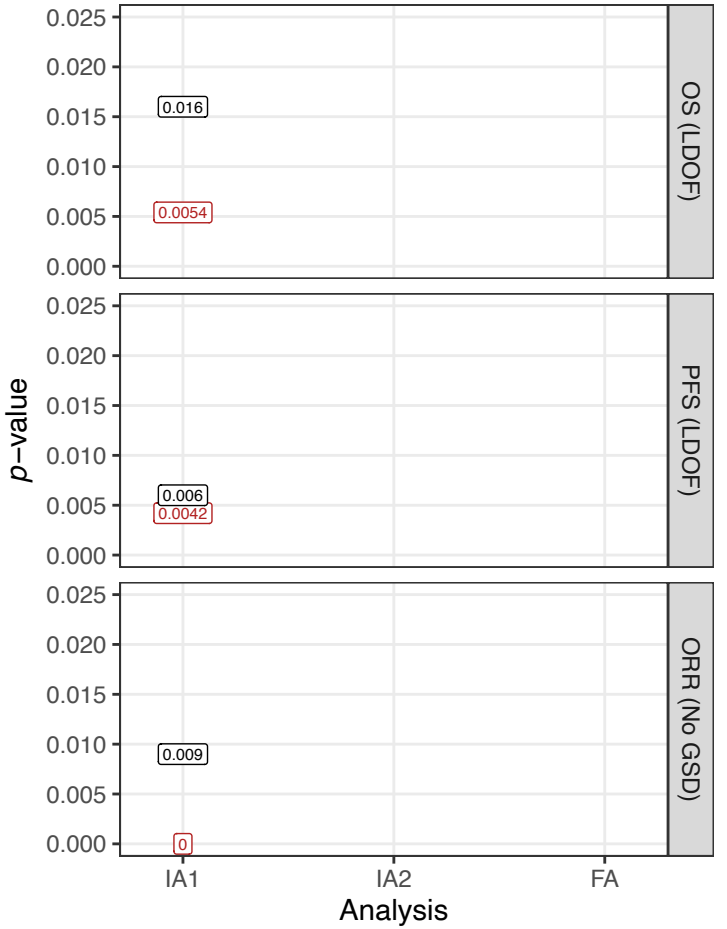
Test statistic for each hypothesis

Analysis	OS	PFS	ORR
IA1	0.016	0.006	0.009
IA2			
FA			

Current graphical testing procedure



Current efficacy boundaries (p-value scale)



Running example 1: KEYNOTE-598

Example implementation in practice: IA2

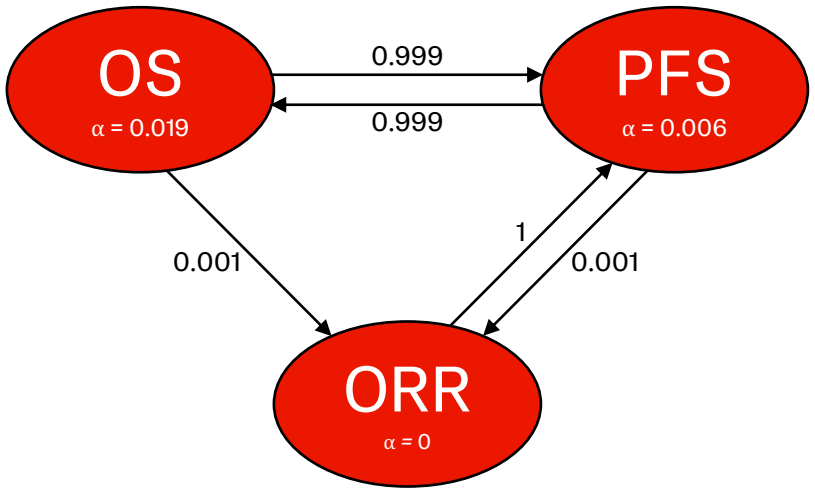
Accrued events / sample size for each hypothesis

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

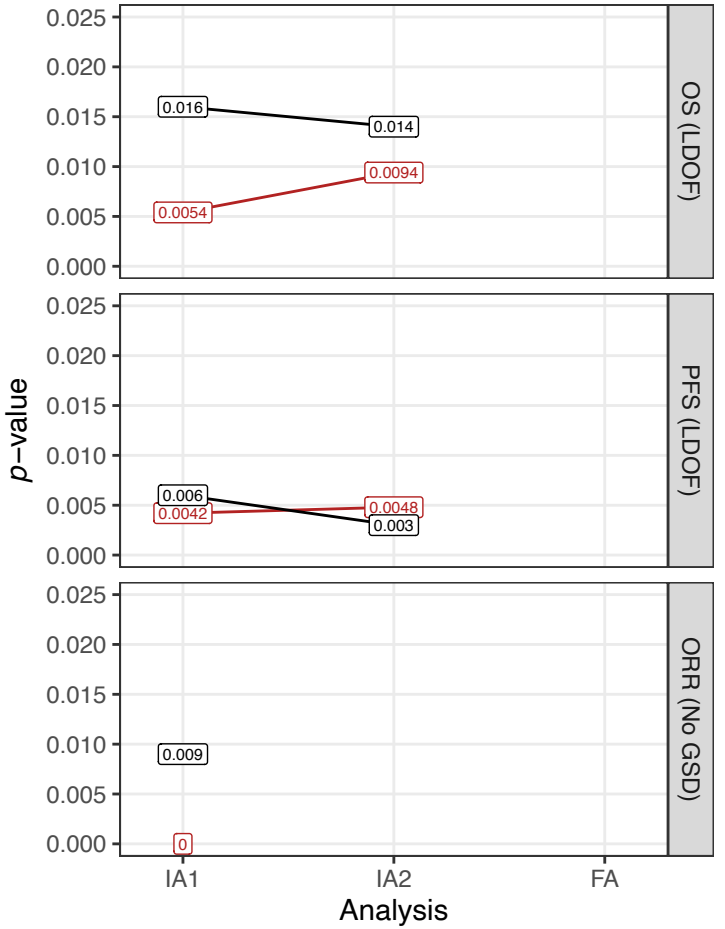
Test statistic for each hypothesis

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

Current graphical testing procedure



Current efficacy boundaries (p-value scale)



Running example 1: KEYNOTE-598

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

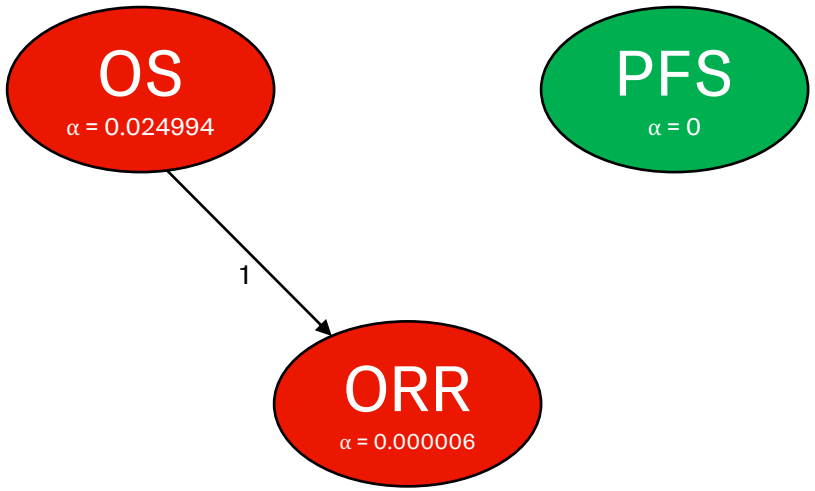
Accrued events / sample size for each hypothesis

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

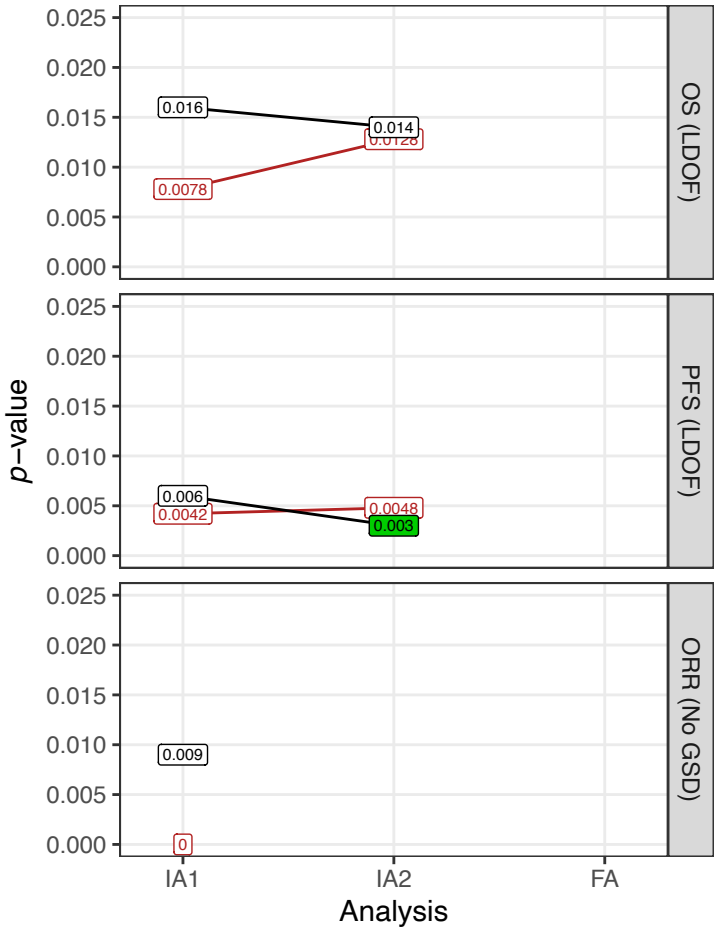
Test statistic for each hypothesis

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

Current graphical testing procedure



Current efficacy boundaries (p-value scale)



Running example 1: KEYNOTE-598

Example implementation in practice: FA

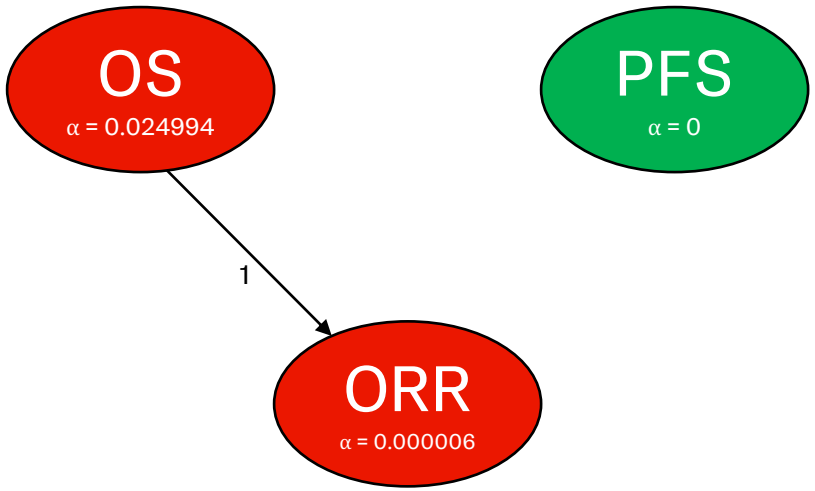
Accrued events / sample size for each hypothesis

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

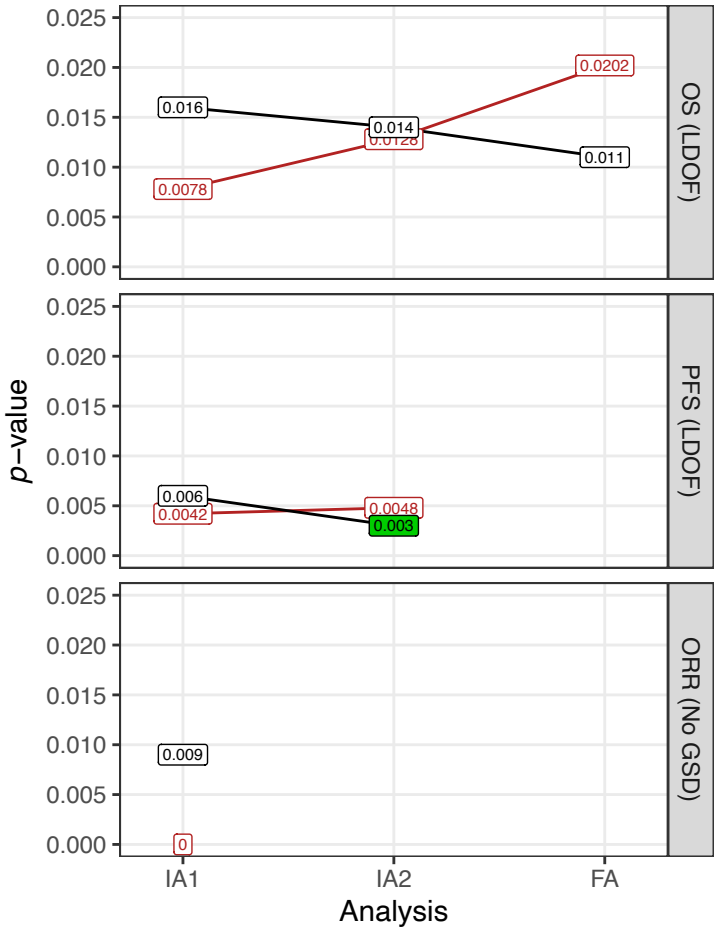
Test statistic for each hypothesis

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

Current graphical testing procedure



Current efficacy boundaries (p-value scale)



Running example 1: KEYNOTE-598

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

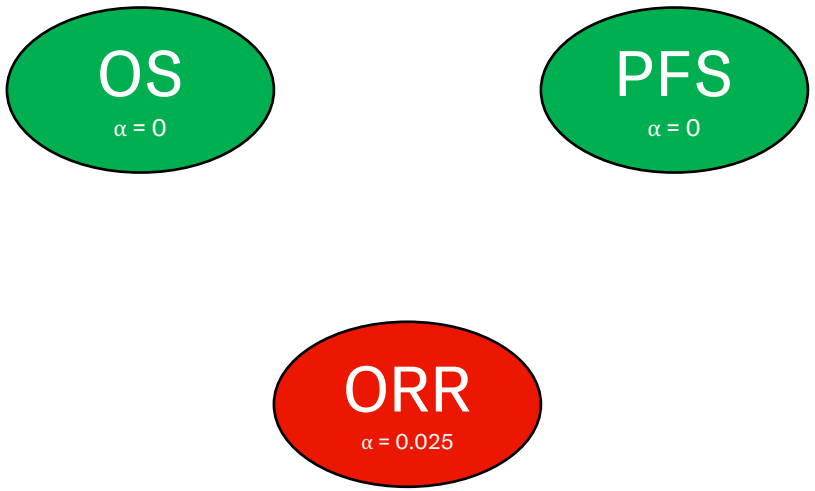
Accrued events / sample size for each hypothesis

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

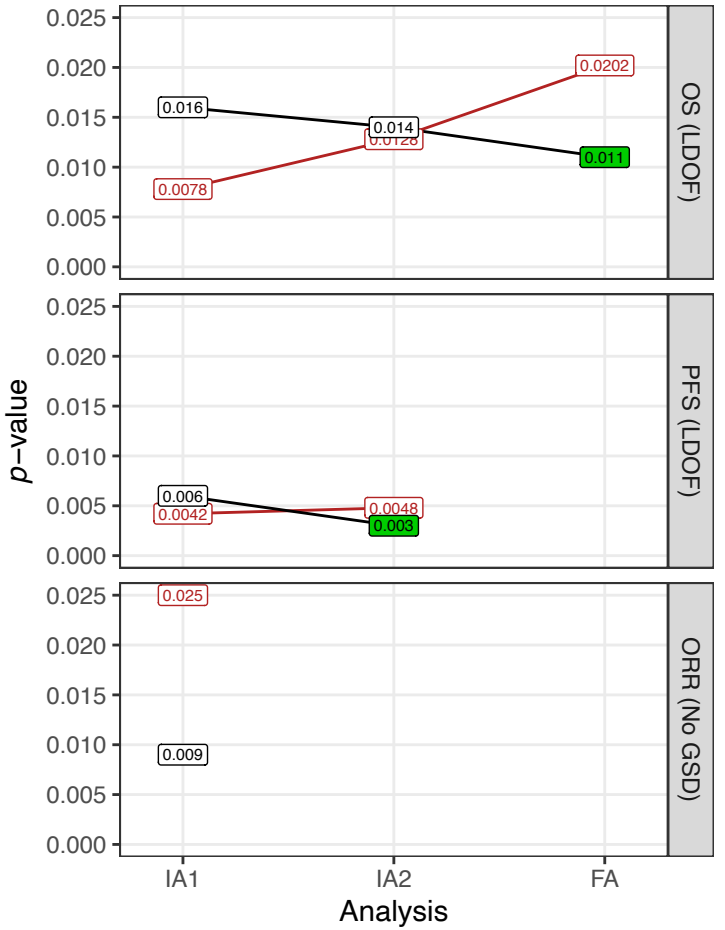
Test statistic for each hypothesis

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

Current graphical testing procedure



Current efficacy boundaries (p-value scale)



Running example 1: KEYNOTE-598

Example implementation in practice: ORR is rejected

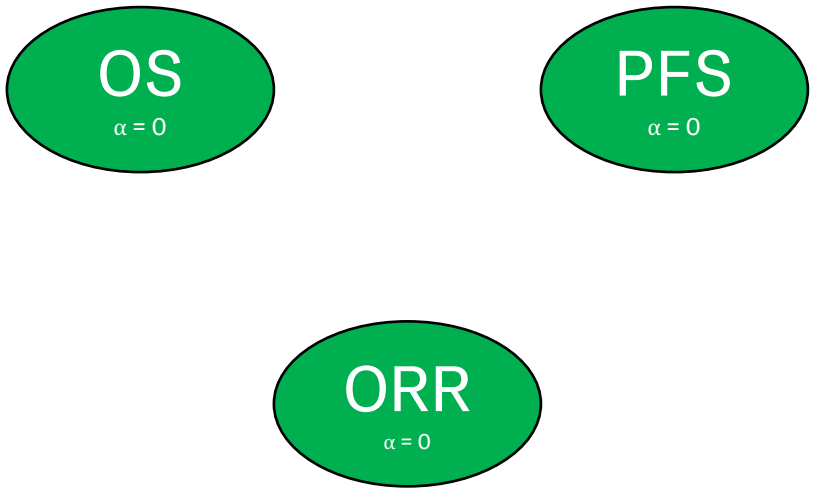
Accrued events / sample size for each hypothesis

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

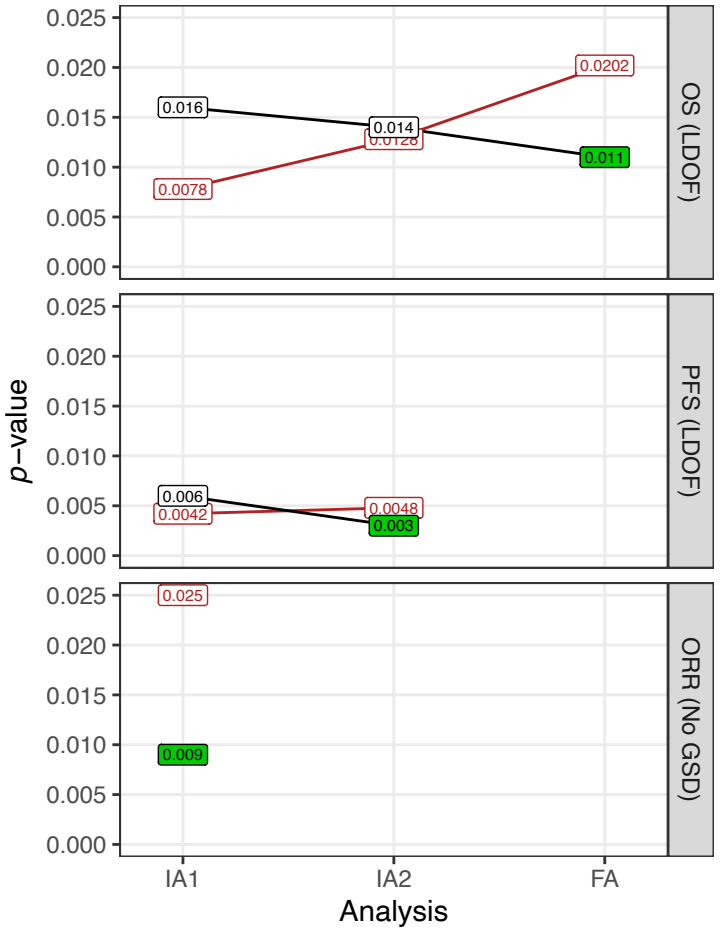
Test statistic for each hypothesis

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

Current graphical testing procedure



Current efficacy boundaries (p-value scale)



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
  - 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**

7. Break

10 mins

8. Software demonstration and practical

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

45 mins

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 dicusses how to produce tables like the ones shown earlier for Running example 1: KEYNOTE-598
  - https://merck.github.io/gMCPLite/articles/GraphicalMultiplicity.html
  - See Keaven Anderson's talk in Session PSW8 on Wednesday afternoon
- 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

Running example 1: KEYNOTE-598

Input: Stay in the input code block

Should only ever need to modify code here

```
```{r input, fig.height=5, fig.width=5, fig.align="center"}
######### << START INPUT >> ##################################################
                                    ⋮
                                    ⋮
                                    ⋮
######### << END INPUT >> ####################################################
```
```

Running example 1: KEYNOTE-598

Input: Options to tweak outputting

```
##### 0. Global parameters
# Number of hypotheses included in the testing strategy with FWER control
numHyp         ← 3
# Allowed one-sided FWER
alphaTotal     ← 0.025
# Number of digits to report for p-value boundaries
pdigits        ← 5
# Number of rounding digits for information fractions
idigits        ← 2
# Data availability at IA in percent relative to max
plotInPercent ← FALSE
# Time limit in plot
Tmax_atPlot    ← 60
# Study specific parameter for multiple testing procedure
mtParam        ← 0.999
```

Helpful approach if the multiplicity plan may be subject to revision

Running example 1: KEYNOTE-598

Input: Enrollment information

```
##### 1. Enrollment
# Overall study enrollment rate object
enrollmentAll ← tibble::tibble(stratum  = "All",
                               duration = 20,
                               rate     = 568/20)
```

Running example 1: KEYNOTE-598

Input: Basic hypothesis data, GTP weights, and spending functions

Basic information about the hypotheses

Spending function for each hypothesis

```
##### 2-4. Main input tibble: Hypotheses, analyses, and distributional
##### information
# One row per hypothesis
hypotheses    ← tibble::tibble(
  # Hypothesis IDs
  id            = paste0("H", 1:numHyp),
  # Hypothesis 'tags', used in graph and output tables
  tag           = c("OS", "PFS", "ORR"),
  # The fields 'regimen', 'ep', and 'suffix' can be pasted together internally
  # into a description field for the output table defining hypotheses
  regimen       = rep("Pembro", numHyp),
  ep            = c("OS", "PFS", "ORR"),
  suffix        = rep("all", numHyp), # E.g., for subgroups
  # Type of hypothesis (primary or secondary)
  type          = c("primary", "primary", "secondary"),
  # initial weights in graphical multiple testing procedure
  w             = c(0.019, 0.006, 0)/0.025,
  # Spending functions; use NULL if no group sequential test for Hi
  grSeqTesting  = list(
    H1 = list(sfu = gsDesign::sfLDOF),
    H2 = list(sfu = gsDesign::sfLDOF),
    H3 = NULL
  ),
```

Weights in GTP

Running example 1: KEYNOTE-598

Input: When are the hypotheses analysed and when do the analysed end

```
# 'iaSpec' and 'hypN' are used to derive the information fractions and timing
# (calendar time since study start) of the analyses.
# For each hypothesis, set criteria that trigger analyses through
# list(A1_list, A2_list, ... , Aj_list), where
# Aj_list = list(H = 1, atIF = 0.5) means that hypothesis analysis j takes
# place when H1 is at 0.5 information fraction
iaSpec         = list(
  list(A1 = list(H = 1, atIF = 255/361),
       A2 = list(H = 1, atIF = 307/361),
       A3 = list(H = 1, atIF = 361/361)),
  list(A1 = list(H = 1, atIF = 255/361),
       A2 = list(H = 1, atIF = 307/361)),
  list(A1 = list(H = 1, atIF = 255/361))
),
# Set total N for each hypothesis (sample size or events); leave NA if
# 'enrollment' and 'iaSpec' would define N
hypN           = c(361, NA, NA),
```

OS is analyzed three times, at specified IFs

PFS is analysed twice, and ORR once, based on the OS maturity

OS has a final information level; the others can be inferred

Running example 1: KEYNOTE-598

Input: Distributional information

The class defines the type of endpoint

```
# To define hypothesis test statistics Zi ~ N(., 1), use 'endpointParam'
# Class of 'endpointParam' is used to derive effect delta and standardized
# effect. Standardized effect size is used for power calculations. Several
# options are available to set the test for binary endpoints
endpointParam = list(
  structure(
    list(
      p1              = 0.70*log(2)/20,
      p2              = log(2)/20,
      dropoutHazard = -log(1 - 0.01)/12
    ),
    class = "tte_exp"
  ),
  structure(
    list(
      p1              = 0.69*log(2)/c(6.5, 14.5),
      p2              = log(2)/c(6.5, 14.5),
      durations       = 6.5,
      dropoutHazard = -log(1 - 0.13)/12
    ),
    class = "tte_pwe"
  ),
  structure(
    list(
      p1              = 0.59,
      p2              = 0.39,
      maturityTime    = 6
      ),
    class = "binomial_pooled")
),
```

Exponential distributions for the two arms and the dropout hazard

PWE distributions for the two arms and the dropout hazard

39% vs 59%, assuming it takes 6 mo for the ORR data to ‘mature’

Running example 1: KEYNOTE-598

Input: Further enrollment information

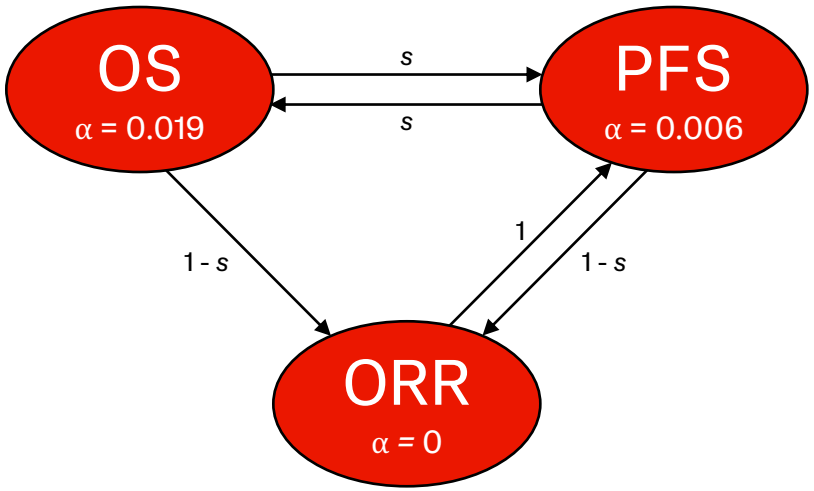
```
# Allocation ratio; trt/control
allocRatio    = 1,
# Prevalence of the hypotheses
prevalence    = c(1, 1, 1),
# Compute enrollment to each hypothesis using its prevalence and the
# previously set enrollment information
enrollment    = lapply(prevalence, function(a) {
  purrr::modify_at(enrollmentAll, "rate", ~{a*.x})
})
```

If the hypothesis is in a sub-population, specify its prevalence

Running example 1: KEYNOTE-598

Input: Parameterized GTP

```r
##### 5. Define graphical testing procedure
graphProc     ← function(s, hypNames = NULL) {
  # s - split parameter
  m             ← matrix(0, numHyp, numHyp)
  m[1, 2]       ← s
  m[1, 3]       ← 1 - s
  m[2, 1]       ← s
  m[2, 3]       ← 1 - s
  m[3, 2]       ← 1
  if (!is.null(hypNames)) {
    colnames(m) ← rownames(m) ← hypNames
  }
  new("graphMCP", m = m, weights = inputD$w)
}
G             ← graphProc(mtParam,
                          hypNames = paste(inputD$id, inputD$tag, sep = ": "))
```



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: 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
OS	**H1 (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
PFS	**H2 (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
ORR	**H3 (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†	Information Fraction
IA1	**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
IA2	**Data cut-off #2, time = 37.6, Criteria: H1 at information fraction 0.85**		
	H1 (OS)	307	0.85
	H2 (PFS)	388	1.00
FA	**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

Running example 1: KEYNOTE-598

Output: Data accrual

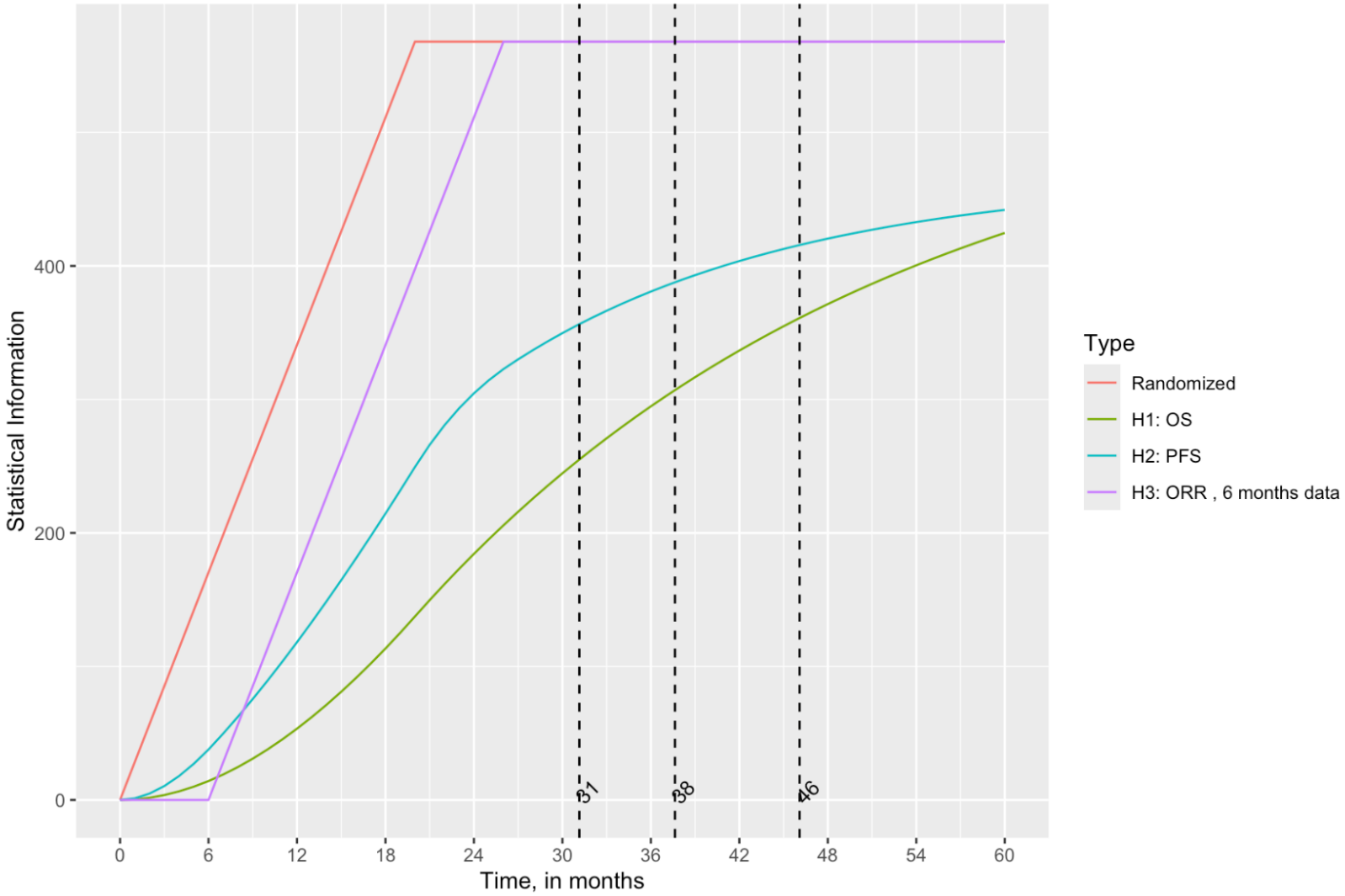


Figure 2: Timelines.

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**

Practical 2

Recreating the complete interim analysis and multiplicity plan for Running example 2

- See practical2.pdf
  - As well as practical2_solutions.R and practical2_solutions.pdf
- The objective is to recreate the full interim analysis and multiplicity plan for Running example 2
  - Start with the template for Running example 1 and convert as needed
- As before, do ask questions 🙂

9. Q&A and Close

10 mins

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)

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
  - Tang and Geller

Feedback and contact information

- Any and all feedback on how to improve the course in the future would be very gratefully received!
- Any issues / questions / etc. please do reach out
  - mgraylin@its.jnj.com
  - ytymofye@its.jnj.com

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

10. Backup

- Closed testing procedures

Closed testing procedures

- General methodology to construct multiple testing procedures that strongly control the FWER (Marcus *et al*, 1976)
- Closed testing procedures consider all intersection hypotheses

$$H_{\mathcal{K}} = \bigcap_{k \in \mathcal{K}} H_k\,, \qquad \mathcal{K} \subseteq \{1, \ldots, K\}$$

- **Closure principle:** An individual hypothesis H_k is rejected at familywise level α only if every intersection hypothesis $H_{\mathcal{K}}$ with $k \in \mathcal{K}$ is rejected at local level α

$$\text{Reject } H_i \Leftrightarrow \text{all } H_J : \; i \in J \text{ are rejected}$$

- **Example ($K = 3$):** To reject H_1 (while controlling FWER at level α) need to reject all intersection hypotheses $H_1, H_1 \cap H_2, H_1 \cap H_3, H_1 \cap H_2 \cap H_3$ (by α-level tests)

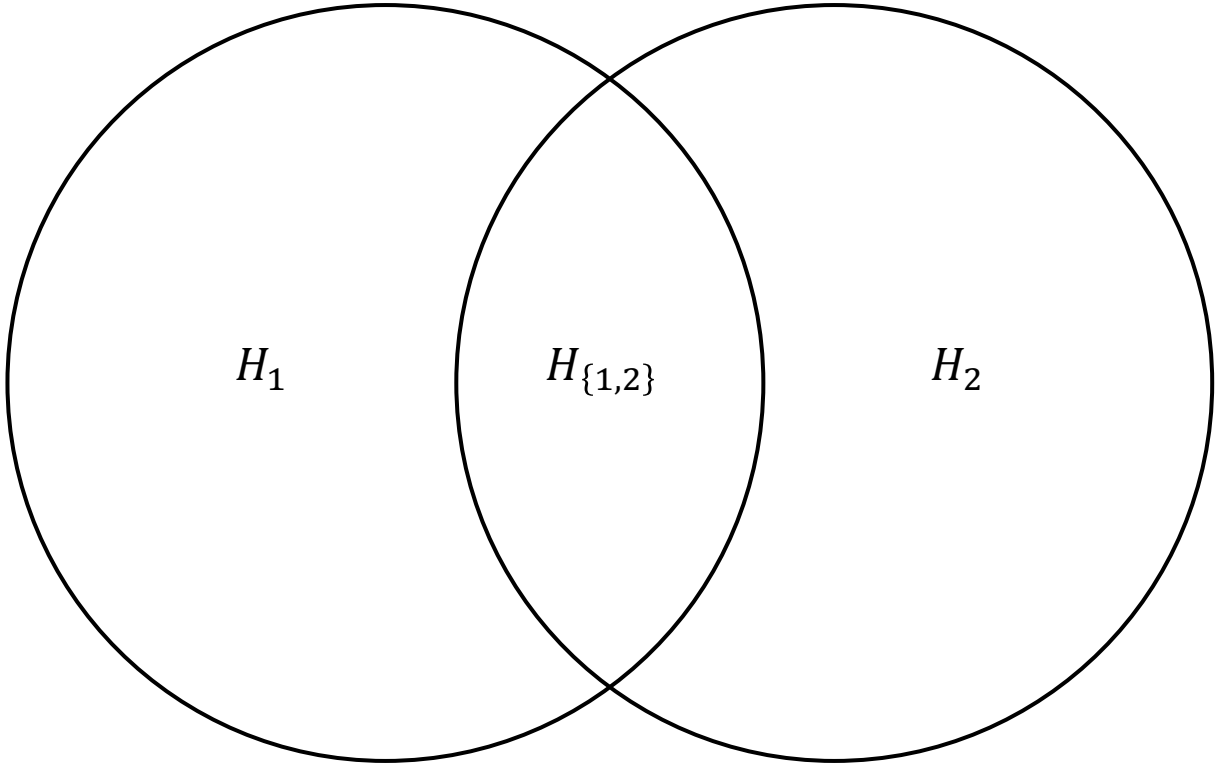
Closed testing procedures

- **Advantages:**
  - Includes many well-known procedures as special cases
  - Closed test procedures are more powerful than the procedures they are derived from
  - By construction, they are coherent: if null H_I is rejected, all subsets $H_J \subseteq H_I$ are rejected as well
  - Any non-coherent multiple testing procedure can be replaced by a coherent one that is at least as powerful
  - Any coherent multiple test controlling FWER is a closed test
- **Disadvantages:**
  - No natural point estimates or confidence intervals
  - Can be a very large number of intersection hypotheses to test as K increases: worst-case is $2^K - 1$

Closure principle

Venn diagram for $K = 2$

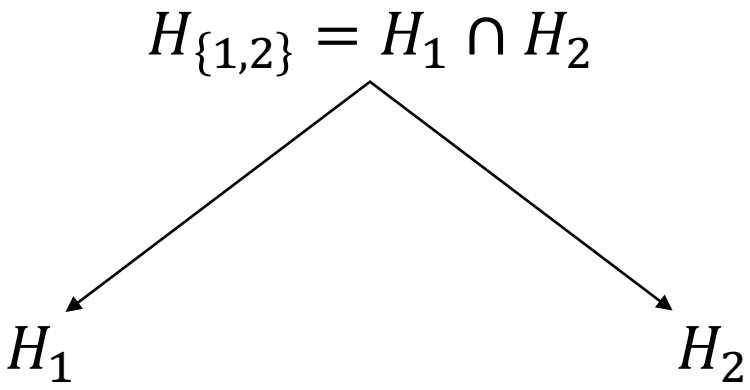
- Test $H_{\{1,2\}}$ using Bonferroni or Dunnett etc. at level α
- Test H_1 and H_2 using a level α test



Closure principle

$\boldsymbol{K = 2}$

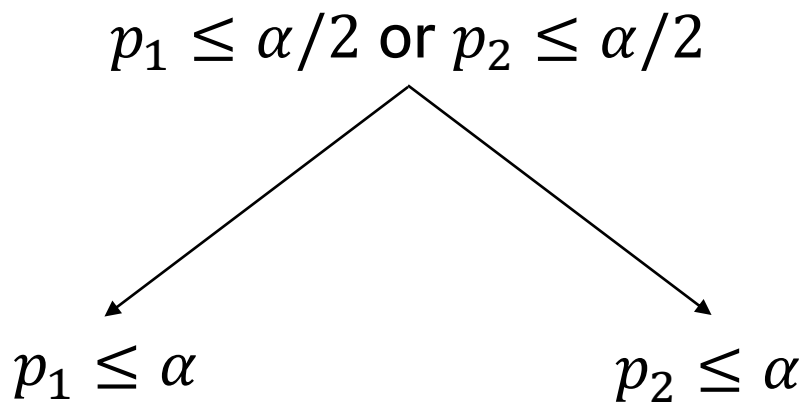
- Reject H_1 overall if $H_{\{1,2\}}$ and H_1 are rejected locally at level α
- If $K > 2$, several intersection hypotheses have to be tested
- Different tests can be chosen for each intersection hypothesis



Closure principle

Holm for $K = 2$

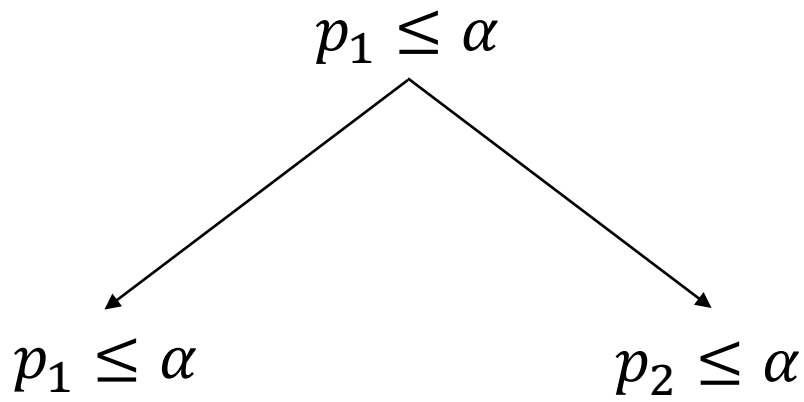
- Holm's procedure is the closure principle applied to Bonferroni
- $H_{\{1,2\}}$ is rejected if either $p_1 \leq \alpha/2$ or $p_2 \leq \alpha/2$



Closure principle

Fixed sequence for $K = 2$

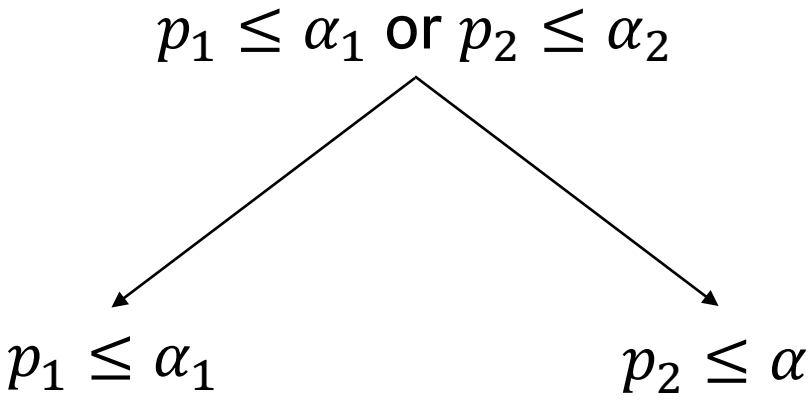
- With *a priori* fixed order $H_1 \rightarrow H_2$



Closure principle

Fallback procedure for $K = 2$

- With *a priori* fixed order $H_1 \rightarrow H_2$ and alpha α_k assigned to H_k



Closure principle

Holm for $K = 3$

$$H_{\{1,2,3\}} = H_1 \cap H_2 \cap H_3$$
$$p_1 \leq \alpha/3 \text{ or } p_2 \leq \alpha/3 \text{ or } p_3 \leq \alpha/3$$

$$H_{\{1,2\}} = H_1 \cap H_2$$
$$p_1 \leq \alpha/2 \text{ or } p_2 \leq \alpha/2$$

$$H_{\{1,3\}} = H_1 \cap H_3$$
$$p_1 \leq \alpha/2 \text{ or } p_3 \leq \alpha/2$$

$$H_{\{2,3\}} = H_2 \cap H_3$$
$$p_2 \leq \alpha/2 \text{ or } p_3 \leq \alpha/2$$

$$H_1$$
$$p_1 \leq \alpha$$

$$H_2$$
$$p_2 \leq \alpha$$

$$H_3$$
$$p_3 \leq \alpha$$

Closure principle

Fixed sequence for $K = 3$

- With *a priori* fixed order $H_1 \rightarrow H_2 \rightarrow H_3$

$$H_{\{1,2,3\}} = H_1 \cap H_2 \cap H_3$$
$$p_1 \leq \alpha$$

$$H_{\{1,2\}} = H_1 \cap H_2$$
$$p_1 \leq \alpha$$

$$H_{\{1,3\}} = H_1 \cap H_3$$
$$p_1 \leq \alpha$$

$$H_{\{2,3\}} = H_2 \cap H_3$$
$$p_2 \leq \alpha$$

$$H_1$$
$$p_1 \leq \alpha$$

$$H_2$$
$$p_2 \leq \alpha$$

$$H_3$$
$$p_3 \leq \alpha$$

Graphical testing procedures

As a short-cut to a closed testing procedure

- In general, a CTP requires on the order of 2^K tests to be done
- "Shortcut" (no need to check all tests) if using Bonferroni test for $H_J = \cap_{j \in J} H_j$

$$\text{Reject } H_J \text{ if for at least one } k \in J \; p_k < w_k(J)\alpha$$

- Monotonicity condition holds:

$$w_k(J) < w_k(U) \; \forall k \text{ when } U \subseteq J, k \in J$$

- If p-value for H_k for small enough to reject global intersection test, the corresponding individual H_k is guaranteed rejection (**consonance**)
- CTP reduced to the order K tests
- Each successful rejection removes one hypothesis from the set

Non-consonance

Example

- The Simes test (Simes, 1986)
  - Consider ordered unadjusted p-values $p_{(1)} \leq \cdots \leq p_{(K)}$
  - Reject H_I if there is k such that $p_{(k)} \leq \frac{k}{m}\alpha$
- **Example ($K = 3$):** As in Hommel (2007)
  - Equal weights, i.e., $w_{(i)}(I) = \frac{1}{|I|}$. Using $\alpha = 0.05$
  - $p_1 = p_2 = 0.03, p_3 = 0.07$
  - Can reject the global null $H_1 \cap H_2 \cap H_3$ using the Simes test because
    - $p_{(2)} \leq \frac{2}{3}0.05$
  - Can not reject neither $H_1 \cap H_3$ nor $H_2 \cap H_3$, hence no rejection of individual H_k
  - Using weights, e.g., defined by graphical approach, applying the Simes test for interception hypotheses will result in more powerful procedure, but no “shortcut”. See Bretz *et al.* (2011)

Johnson&Johnson

Johnson&Johnson