Optimising multiplicity adjustment in clinical trials using elicited functions of commercial value and clinical benefit

Alex Spiers, Graham Wheeler, Adrian Mander

gsk.com

Thank you to:

The PSI 2026 organising committee

Adrian Mander

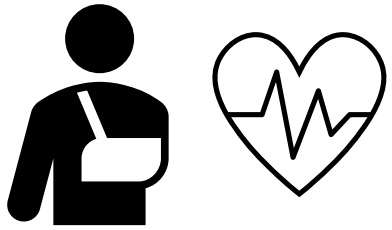
Graham Wheeler

Disclaimer:

The views expressed in these slides are my own, and do not necessarily reflect the opinions of GSK

Decision-making framework for choosing a multiplicity strategy is about what we want on the label

Clinical Benefit



Commercial

Regulatory

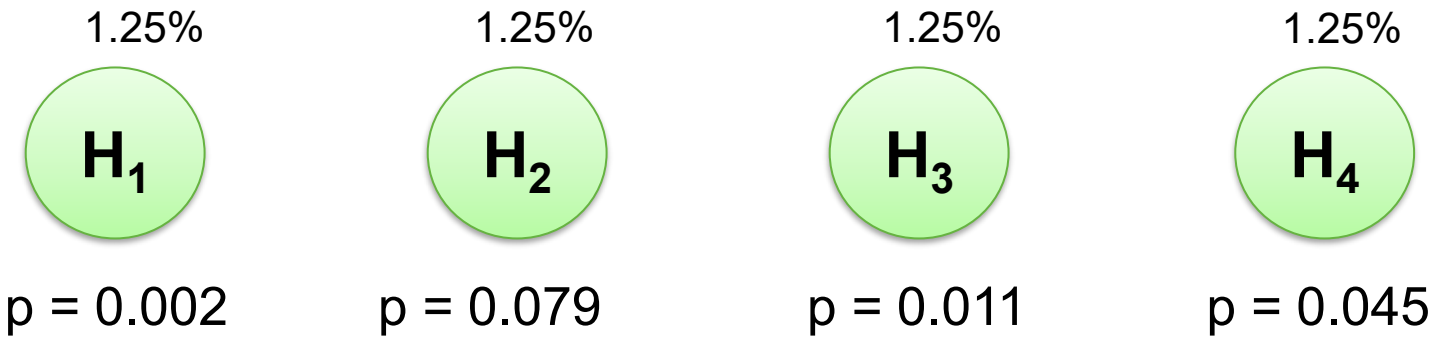
Making multiple claims on asset label requires multiple testing

When multiple claims, endpoints, or subgroups are targeted for product labelling, we want the most robust package of evidence for regulators

- Often this will mean family-wise error rate (FWER) control for primary and key secondary hypotheses
- → we need a method to strongly control FWER at e.g. 5%

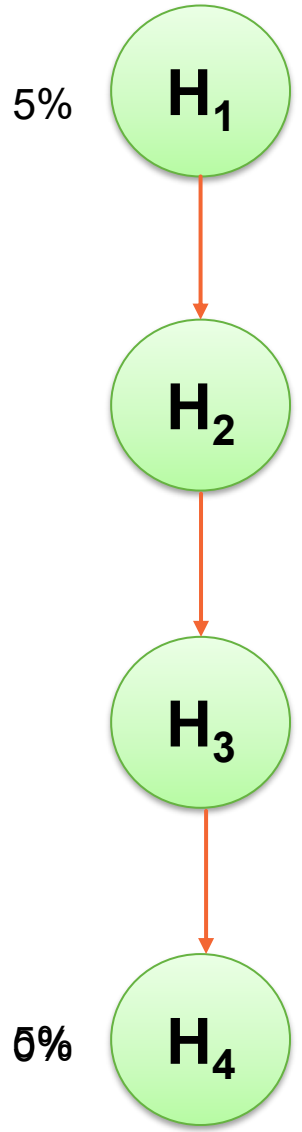
One way we can do this (**not recommended**):

- **Bonferroni** (using ***alpha-splitting***)



Historical default: fixed sequence

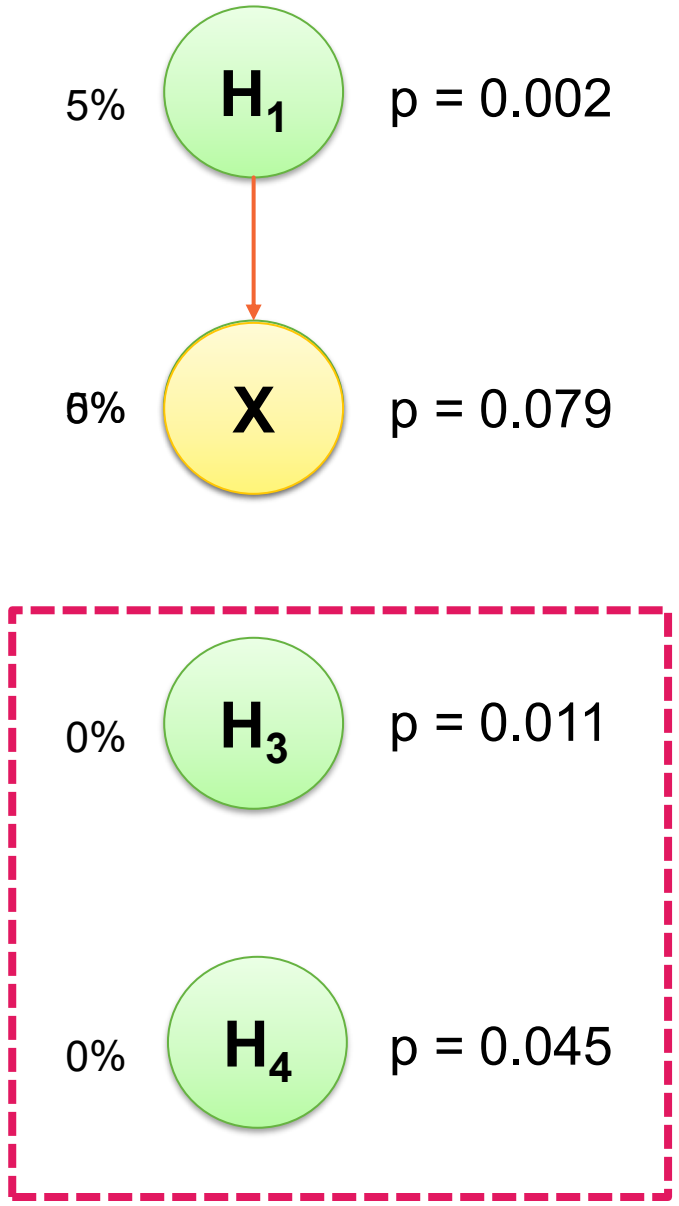
A better way of controlling FWER by is by ***alpha-recycling*** using a **fixed sequence test** (hierarchy)



Risk of fixed sequence

But this imposes **risk**

In reality, we do not know which hypotheses will get the lowest p-values!



Graphical tests – Bretz et al. (2009)

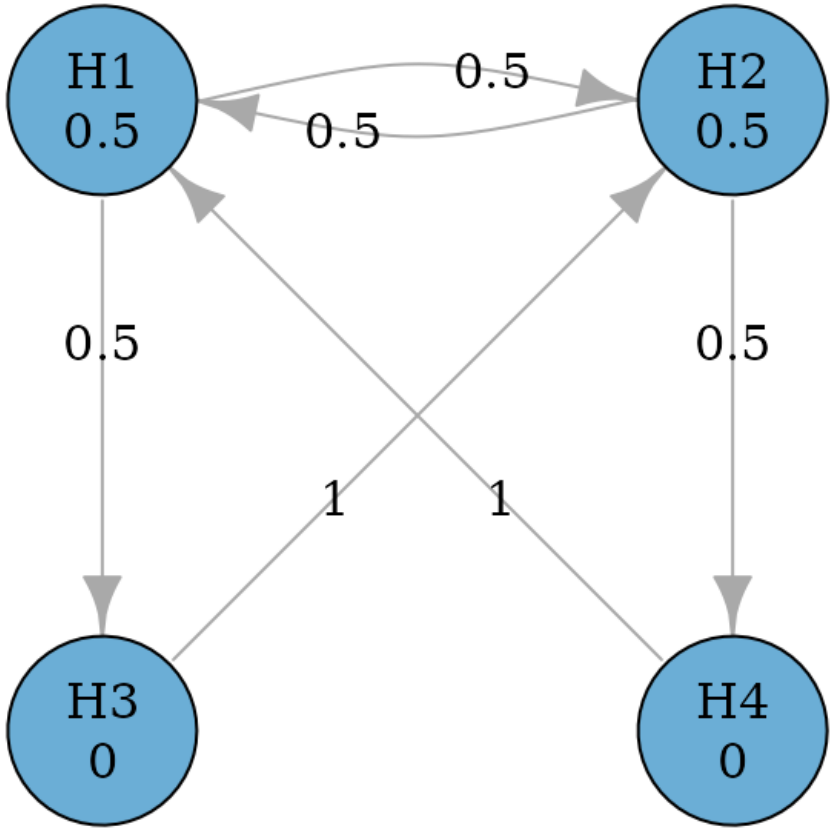
Why use graphical tests?

- Strongly control FWER

- **Combines the best** of Bonferroni and Hierarchical testing:
  - Alpha-splitting (test in any order)
  - Alpha-recycling (use all alpha)

- Established method accepted by regulators

- **Can tailor multiplicity adjustment procedures** to the importance of hypotheses



Optimal graphical testing procedures

How to find “optimal” design parameters for a graphical test

The optimal parameters for a graphical test can be found by maximizing a utility function over the search space of possible design parameters for our graph (hypothesis weights $\boldsymbol{w}$, and transition matrix $\mathbf{G}$)

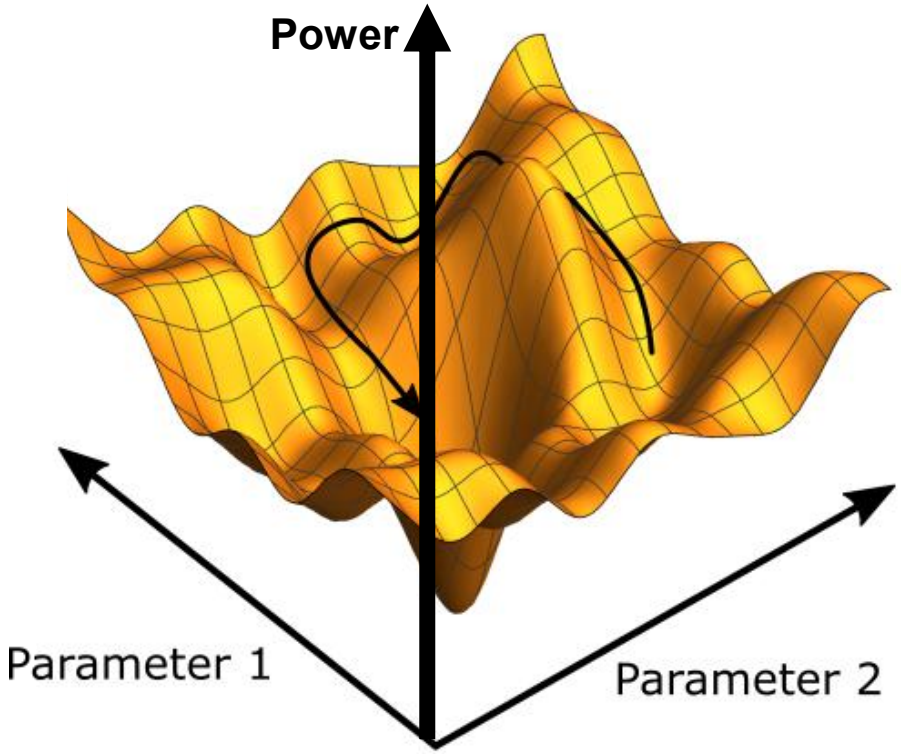
$$\underset{\boldsymbol{w},\,\mathbf{G}}{\text{maximize}}$$

Objective function

$$\text{subject to}$$

Graph regularity conditions

Conditional on: Expected distribution of test statistics



Optimisation workflow

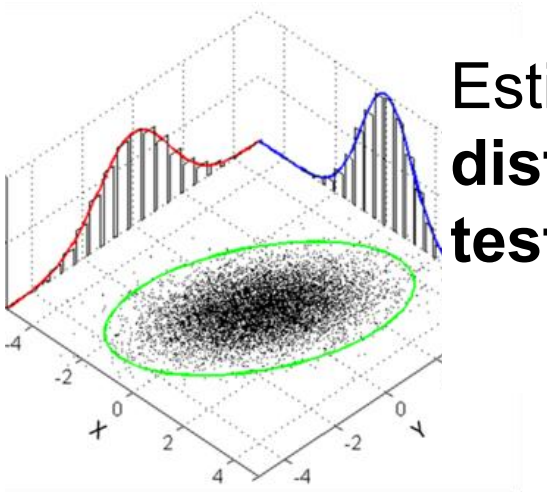
Demonstrating patient benefit

Regulatory approval likelihood

Commercial viability and market access

Definition of "**trial success**"

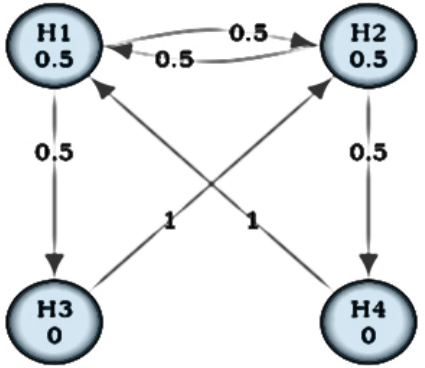
- Relative value of each rejected hypothesis on product label



Estimate of **joint distribution of test statistics**:

$$z \sim \mathrm{MVN}$$

multigrain



GSK

Defining a multiple-endpoint definition of trial success

Utility as a function of null hypothesis rejections

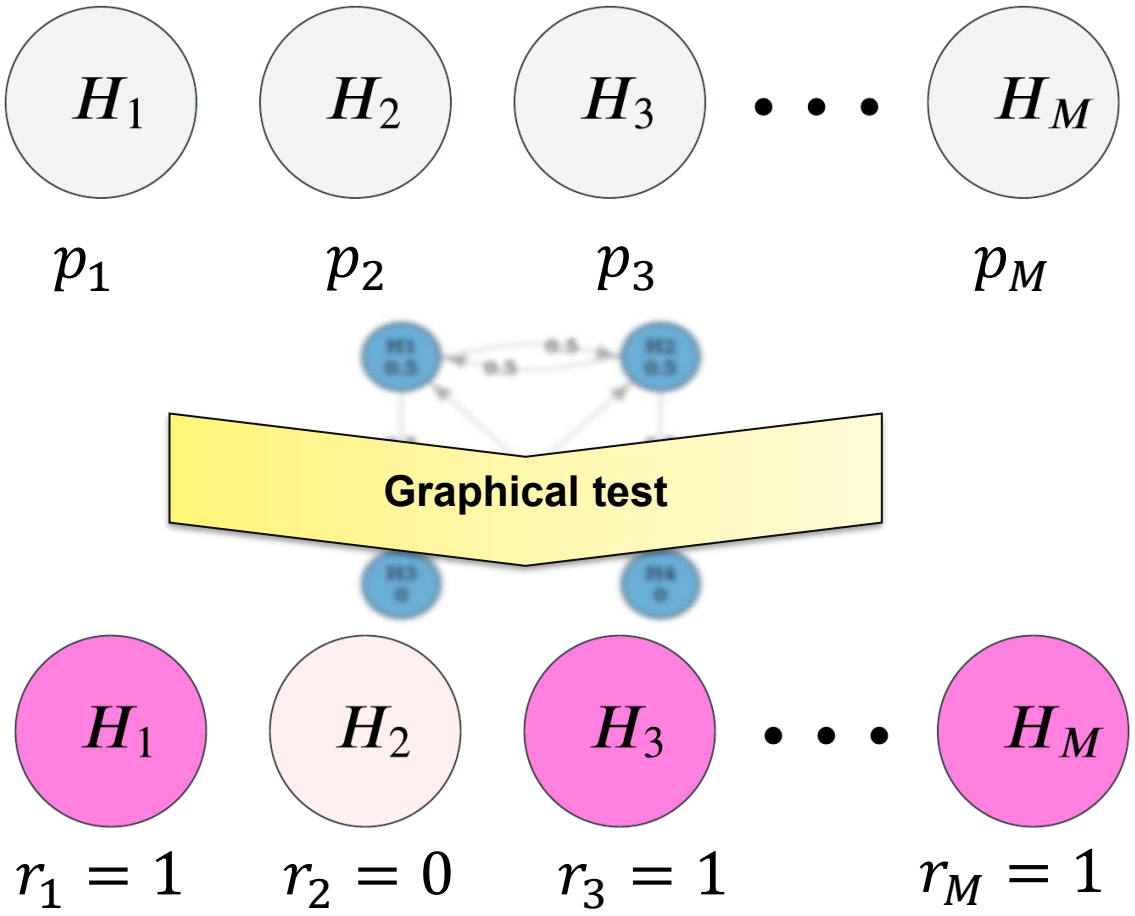
Graphical tests map M p-values to $\{0,1\}^M$

We can denote this as:

$$f(p_1, p_2, \ldots, p_M) \rightarrow \boldsymbol{r} = (r_1, r_2, \ldots, r_M)$$

where $\boldsymbol{r}$ is the **rejection vector**:

- $r_i = 1$ when H_i is rejected,
- $r_i = 0$ otherwise



Defining a utility function for trial success

We define a ***rejection utility function***:

$$\psi: \{0,1\}^M \rightarrow \mathbb{R}^+$$

which assigns a value (or “reward”) to a given rejection pattern. Commonly used functions are:

$$\psi(\mathbf{r}) = \begin{cases} 1, & \text{if } r_1 \text{ OR } r_2 \text{ OR } \cdots \text{ OR } r_M = 1, \\ 0, & \text{otherwise} \end{cases}$$

(*disjunctive power*)

Defining a utility function for trial success

We define a ***rejection utility function***:

$$\psi: \{0,1\}^M \rightarrow \mathbb{R}^+$$

which assigns a value (or “reward”) to a given rejection pattern. Commonly used functions are:

$$\psi(\mathbf{r}) = \begin{cases} 1, & \text{if } r_1 \text{ AND } r_2 \text{ AND } \cdots \text{ AND } r_M = 1, \\ 0, & \text{otherwise} \end{cases}$$

(*conjunctive power*)

Defining a utility function for trial success

We define a ***rejection utility function***:

$$\psi: \{0,1\}^M \rightarrow \mathbb{R}^+$$

which assigns a value (or “reward”) to a given rejection pattern. Commonly used functions are:

$$\psi(\mathbf{r}) = \frac{1}{M} \sum_{i=1}^{M} r_i$$

(*proportion of rejections*: also called *average power*)

GSK

Utility function for trial success

By using a combination of the above operations **conjunctive**, **disjunctive** or **additive operators,** we can construct composite functions for $\psi(\boldsymbol{r})$ that *accurately describe our criteria for trial success*.

These criteria should be based on:

- **Regulatory, clinical** or **commercial criteria** for **trial success**
- **Clinical and market relevance**, where greater weight is assigned to rejection of hypotheses that offer broader clinical adoption or reimbursement potential

Trial success utility function = $\mathbb{E}_{Z\sim \mathrm{MVN}}[\psi(\boldsymbol{r})]$

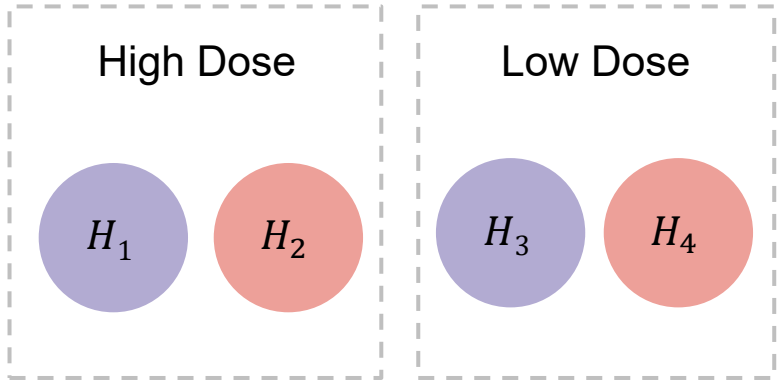
GSK

Trial success function with a clear "win criterion"

LIBERTY QUEST trial: Dupilumab to treat uncontrolled asthma – Castro et al. (2018)

A three-arm parallel confirmatory clinical trial to compare a high and low dose of dupilumab in patients with asthma versus placebo

- Regulatory success defined as:
  - Meeting **BOTH** incidence rate of exacerbations endpoint **AND** FEV1 for **at least one dose**



H_1: High dose **Rate exacs.**
H_2: High dose **FEV1**
H_3: Low dose **Rate exacs.**
H_4: Low dose **FEV1**

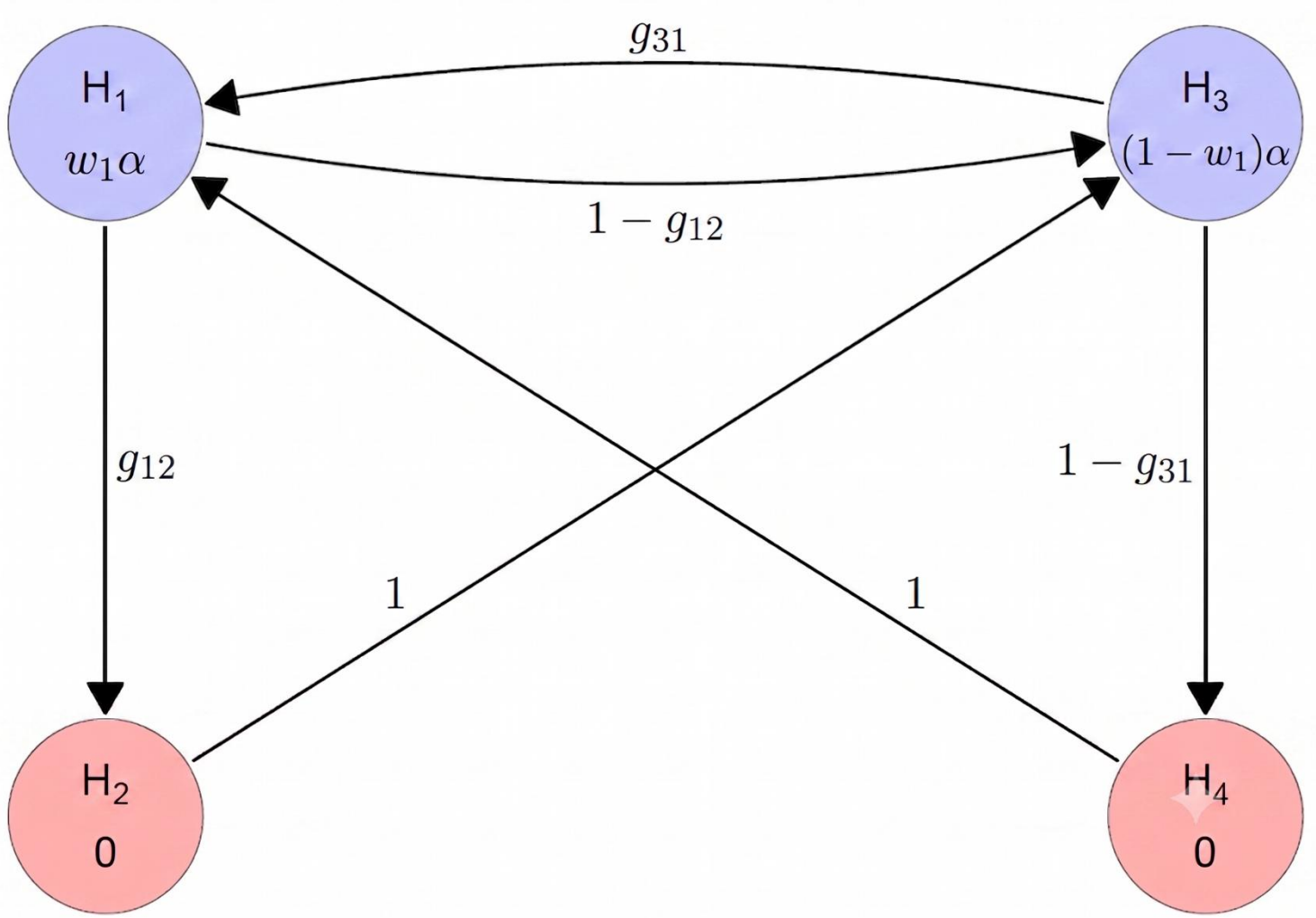
A sensible rejection utility might be: $\psi(\boldsymbol{r}) = \mathbb{I}\{(r_1 \text{ AND } r_2) \text{ OR } (r_3 \text{ AND } r_4)\}$

Disjunctive OR operator for "at least one dose" condition

Conjunctive AND condition for regulatory requirements

GSK

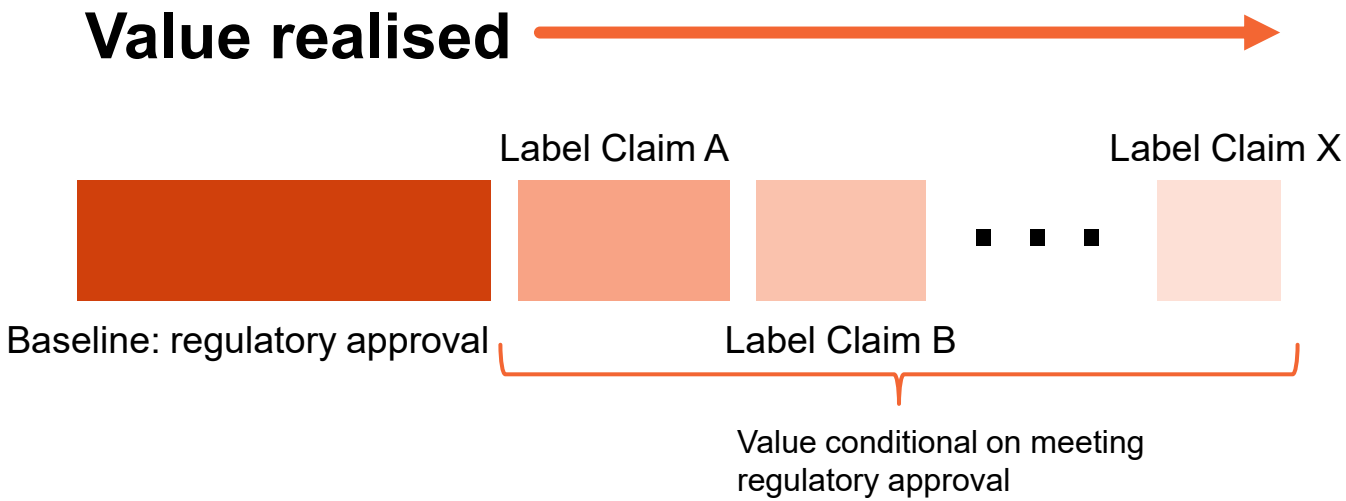
Optimal graph has the following structure:



- Alpha split between doses
- Secondary hypothesis can only be tested after primary is rejected for the same dose
- Magnitude of weights depend on the test statistic distribution (treatment effects and correlations between endpoints)

Most trials have "nice-to-haves" beyond the regulatory hurdle

- Most trials also test key secondary endpoints that support additional label claims that have commercial value.



- Trial success function should encode the **conditional value structure**:

$$\psi(\boldsymbol{r}) = \mathbb{I}_{\text{reg}}(\boldsymbol{r}) \cdot (v_{\text{base}} + v_1 r_1 + v_2 r_2 + \cdots)$$

Zero unless regulatory hurdle cleared

Value of approval alone

Value of label claim 1

Value of label claim 2

- To find $v_{\text{base}}, v_1, v_2, \ldots$, we need **elicitation**

The elicitation process

Individual stakeholders (ideally at least one representing each of **clinical**, **commercial**, and **regulatory** perspectives) score each secondary endpoints (without conferring). This should be based on their expert opinion.

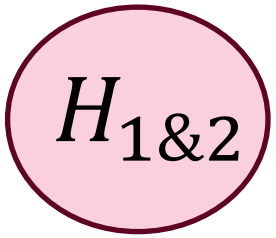
Scores collated, and median / range of scores are shared in a group. Proceed to facilitated discussion of discrepancies and factors.

After discussion: final opportunity to rescore → Trial success definition **finalised**

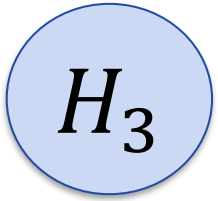
Multigrain optimises the multiplicity strategy using the shared consensus definition of trial success.

Case Study: ECZTRA trial - Silverberg et al. (2021)

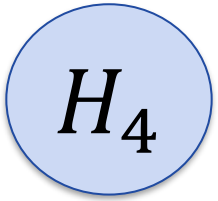
Tralokinumab corticosteroid combo vs placebo control to treat atopic dermatitis



IGA0/1 & EASI75



Pruritus



DLQI

SCORAD

- The trial objective requires the co-primary endpoints (H_1, H_2) to be met for the three pivotal secondary endpoints (H_3, H_4, H_5) to add value.

- ⇒ The trial success utility has the following structure:

$$\psi(\boldsymbol{r}) = r_1 \cdot r_2 \cdot (v_{\text{base}} + v_3 r_3 + v_4 r_4 + v_5 r_5)$$

Conditional on meeting co-primary endpoints

Relative value added from each endpoint if successful

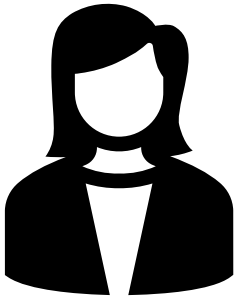
GSK

The elicitation process: stakeholder scoring

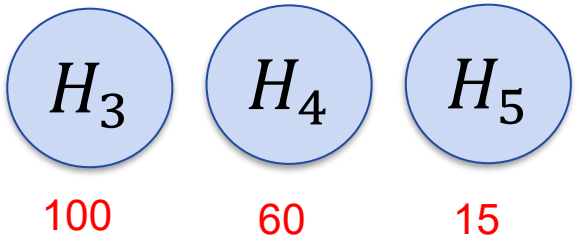
Assume the co-primary endpoints are met. Now: how valuable are the secondary claims?

Itch is what patients ask about first. If the pruritus claim = 100, quality of life feels like 60, SCORAD is maybe 15

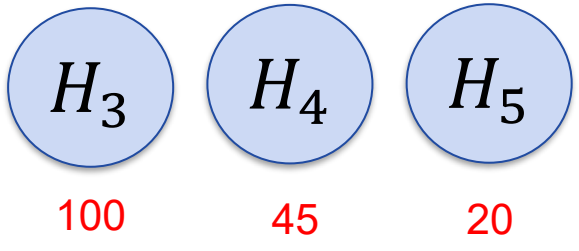
Baseline approval revenue ~$200M/year. Secondaries add ~$160M: mostly from the pruritus claim ($100M), because it's what differentiates us with payers.



Individual rating:



Individual rating:



Aggregate results (anonymised)

Hypothesis / Endpoint	H_3	H_4	H_5
Relative Value	100%	53%	16%
(Range)	(100 - 100)	(45 - 60)	(15 - 20)

GSK

From scores to utility function for trial success

What the panel said:

Secondary Endpoint	Median Score
Pruritus (H_3)	100
DLQI (H_4)	53
SCORAD (H_5)	16

- Co-primary success alone = **$200M** revenue
- Co-primary success + all three secondaries = **$360M**

Allocate the $160M of additional value in proportion to elicitation scores:

Utility function for trial success:

$$\psi(\boldsymbol{r}) = r_1 \cdot r_2 \cdot (200 + 95r_3 + 50r_4 + 15r_5)$$

Value approval alone → $200M

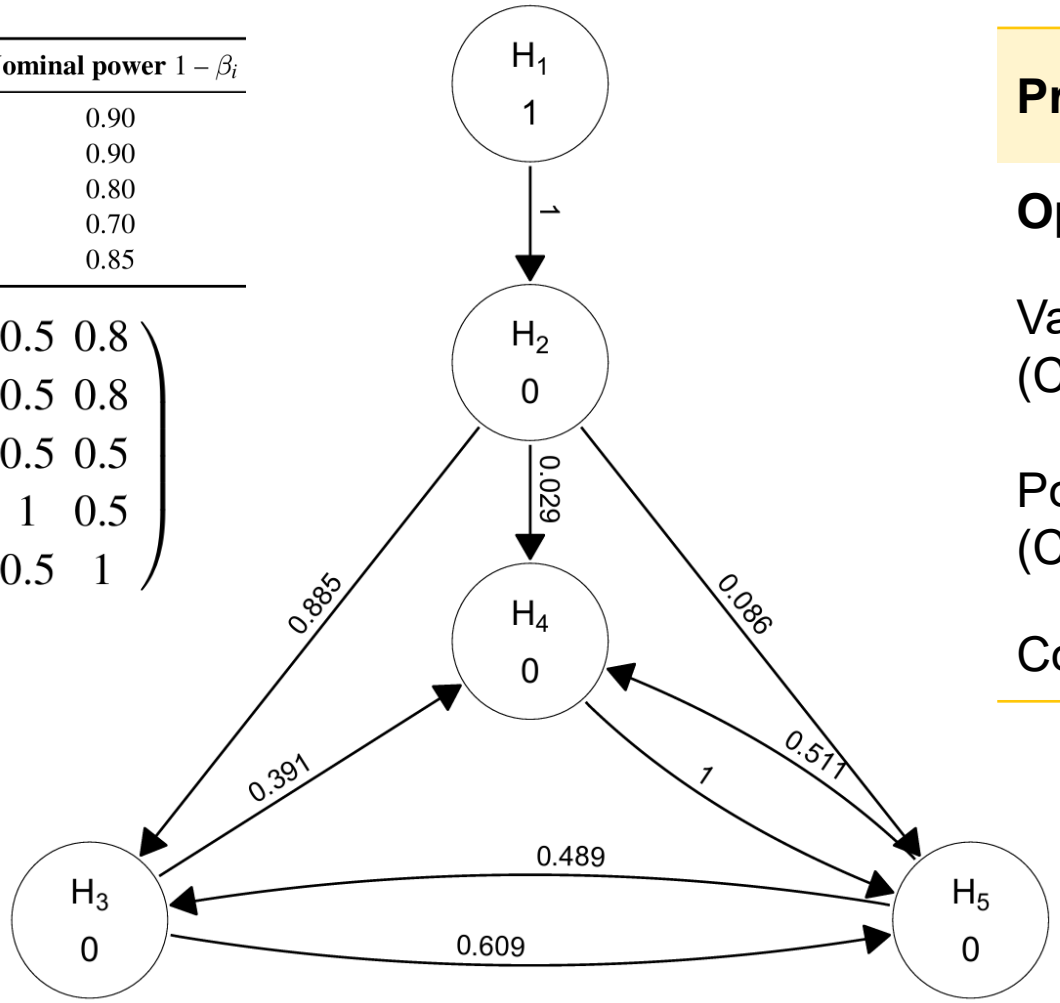
Value conditional on meeting regulatory approval → total = $160M

From stakeholder scores to trial success utility function

Test statistic distribution assumptions:

Hypothesis	Endpoint	Nominal power $1-\beta_i$
H_1	IGA 0/1	0.90
H_2	EASI-75	0.90
H_3	Pruritus	0.80
H_4	DLQI	0.70
H_5	SCORAD	0.85

$$\Sigma = \begin{pmatrix} 1 & 0.8 & 0.5 & 0.5 & 0.8 \\ 0.8 & 1 & 0.5 & 0.5 & 0.8 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.8 & 0.8 & 0.5 & 0.5 & 1 \end{pmatrix}$$

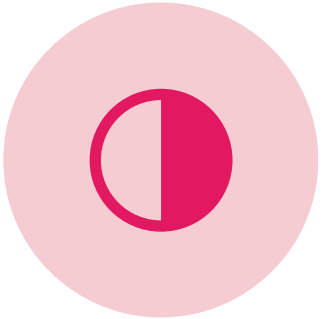


Optimised graph and performance:

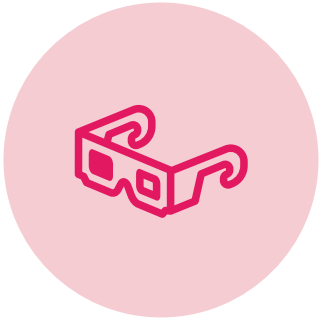
Procedure	Expected utility*
Optimised graph (left)	**0.774**
Value-ordered sequence (Co-primary → H_3 → H_4 → H_5)	0.767
Power-ordered sequence (Co-primary → H_5 → H_3 → H_4)	0.769
Co-primary → Holm{H_3, H_4, H_5}	0.770

*scaled such that the range is $[0,1]$

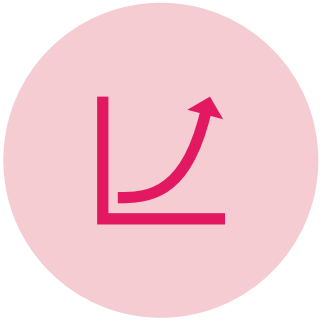
Lessons from practice at GSK



1. Separate “value” from likelihood of rejection



2. Elicitation forces transparency



3. Endpoint value is not always additive – interactions are possible

4. The process is part of the value – the yielded strategy is explicitly justified

GSK

Summary

- **Trial success utility functions** capture how trials release value: approval when regulatory criteria are met, then additional label claims yield conditional commercial value
- **Structured elicitation** turns cross-functional judgements into the objective function
- **multigrain** R package for optimisation: open-source on GitHub: https://github.com/GSK-Biostatistics/multigrain

References

- Bretz, F., et al. (2009). A graphical approach to sequentially rejective multiple test procedures. *Stat Med, 28(4), 586-60*

- Castro, M., et al. (2018). Dupilumab Efficacy and Safety in Moderate-to-Severe Uncontrolled Asthma. *N Engl J Med, 378(26), 2486-2496.*

- Silverberg, I., et al. (2021). Tralokinumab plus topical corticosteroids for the treatment of moderate-to-severe atopic dermatitis: results from the double-blind, randomized, multicentre, placebo-controlled phase III ECZTRA 3 trial*. *British Journal of Dermatology, 184(3), 450-463.*

- Spiers, A., et al. (2026). *Gain-function optimisation of graphical multiple testing procedures for confirmatory clinical trials* [Unpublished manuscript].

gsk.com