Utilising the graphical approach to multiple testing

Nick Counsell (nicholas.counsell@ucl.ac.uk)

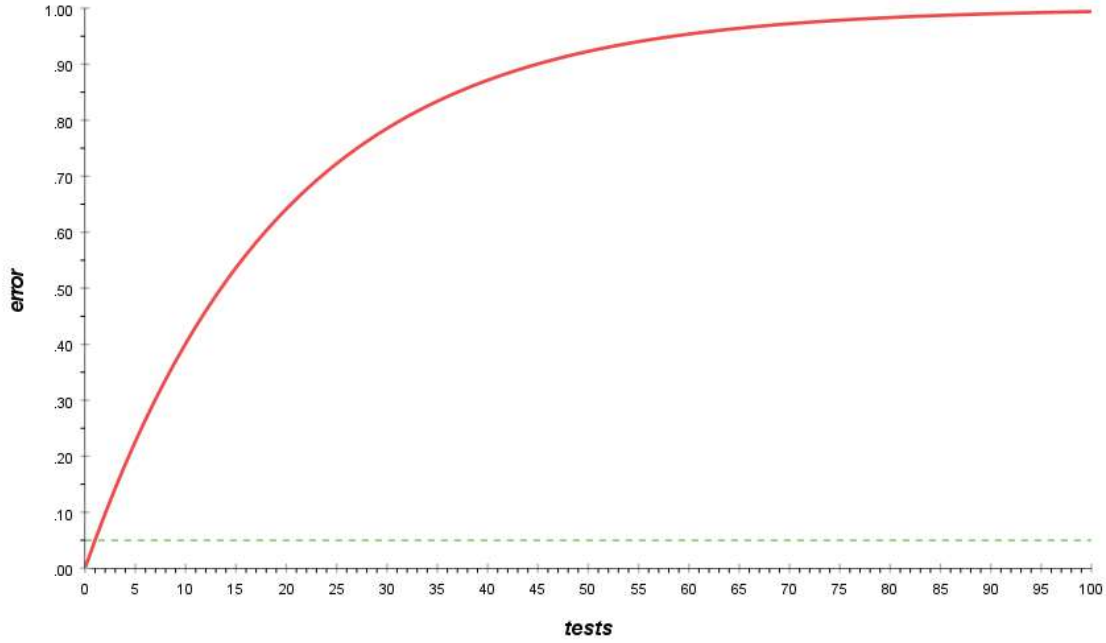
University College London

Multiple testing

We often want to address several questions of interest in an individual clinical trial – endpoints/arms/groups/times/etc.

Problem: this multiplicity inflates overall error rate, i.e. more comparisons → more likely positive chance findings

# Tests	P(Type I Error)
1	5%
2	~10%
5	>20%
7	>30%
10	>40%
14	>50%



Multiple testing

Addressing multiple objectives is a common issue in clinical trials. Regulatory guidelines are generally focused on confirmatory conclusions for proof of efficacy and decision making, also an important consideration in earlier phases

Can reduce multiplicity by prioritising objectives and then use statistical procedures to control and/or adjust as required, or justify why not required

Want to control the study-level false positive rate, the probability of rejecting at least one true null hypothesis (the global familywise error rate), through proper statistical adjustment to avoid incorrect conclusions

Controlling FWER

A test procedure has:

- weak control if the FWER ≤ α conditioning on the complete set of nulls (i.e. the intersection of all pairwise nulls)
- strong control if the FWER ≤ α under all partial nulls (i.e. regardless of which of the multiple comparisons have no true effect)

Common procedures and adjustments:

- single step (Bonferroni, PAAS, Simes, Dunnett)
- stepwise (Holm, Hochberg, Hommel, stepdown Dunnett)

Controlling FWER

These common test procedures can be conservative and data-driven…. should take clinical considerations into account as well as statistical power

Hierarchical test procedures can reflect relative importance of the multiple tests, whilst controlling FWER:

- fixed sequence procedure (order H_i and test until $p_i > \alpha$);
- fallback procedure (split α and test H_i at $\alpha_i+\alpha_{i-1}$ if $p_{i-1} < \alpha_{i-1}$);
- gatekeeping ('families' of endpoints).

Gatekeeping

Gatekeeping accounts for a hierarchical structure of our multiple hypotheses:

- Serial – requires all hypotheses in family to be rejected before proceeding to the next family (which is tested using α)
- Parallel – requires only one hypotheses in family to be rejected before proceeding to the next family (tested using the 'rejection gain factor' of α)
- Tree-structured – combinations for more complex structures

This allows us to construct powerful multiple test procedures, based on the closed testing principle to control the FWER (considering all partial nulls).

However, 2^m-1 intersections requiring many tests….

Graphical approach

Graphical approach provides a shortcut for structured hypotheses (importance, logical relationships) reducing this number to *m* through *s*tepwise gatekeeping

Flexible and powerful, combining non-hierarchical and hierarchical approaches

Intuitive and easier to specify and communicate testing strategies

Maintains strong control of FWER because of the closed test principle, and common approaches for test procedures can be represented as specific cases

Graphical approach

Directed weighted graphs represent the testing procedure with:

- nodes representing hypotheses H_i, assigned initial significance level α_i ($\sum_i \alpha_i \leq \alpha$);
- weighted directional edges g_{ij} from each node i ($\sum_j g_{ij} \leq 1$) representing, if rejected, the fraction of H_i that will be passed to the H_j;
- this graph is updated as hypotheses are rejected by removing the node, and updating significance levels and weighted edges.

Generally, maximise power if all nodes are accessible from other nodes (graph is irreducible) and $\sum_j g_{ij} = 1$ for each.

Example – ICON 9

- International phase III RCT in patients with relapsed platinum-sensitive ovarian cancer following a response to platinum-based chemotherapy

- Maintenance therapy with olaparib and cediranib or olaparib alone (randomised 1:1 stratified by BRCA, prior bev, PFI, surgery, country)

- Two primary endpoints and two groups of interest, therefore four comparisons leading to an overall type 1 error rate of ~18.5% if FWER not controlled

UCL CANCER RESEARCH UK NCRI National Cancer Research Institute ENGOT European Network of Gynaecological Oncological Trial groups GYNECOLOGIC CANCER INTERGROUP An Organization of International Cooperative Groups for Clinical Trials in Gynecologic Cancers

Example – ICON 9

Originally designed using a fixed-sequence serial gatekeeping approach:

i. PFS, all patients

ii. PFS, BRCA wild-type

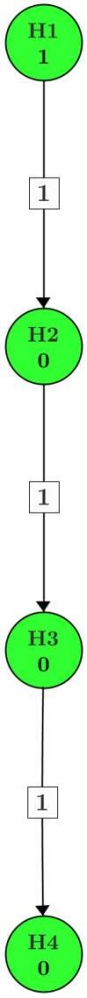
iii. OS, all patients (N=588, HR=0.75, 80% power)

iv. OS, BRCA wild-type (N=350, HR=0.70, 80% power)

Based on where most likely to detect a difference; more interested in BRCA wild-type, as less benefit from maintenance olaparib than those with BRCA-mutation.

α is available to be carried forward to the next objective in the sequence, as soon as there is a non-statistically significant result, α is used up and none is available for further tests. Conclusions can only be made if all previous analyses $p<0.05$.

Fixed sequence procedure



PFS, all patients

PFS, BRCA wild-type

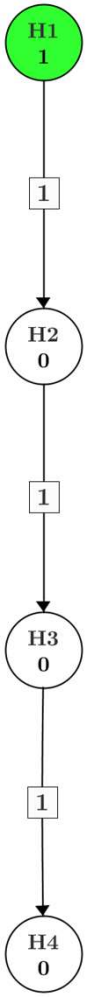
OS, all patients

OS, BRCA wild-type

$$\boldsymbol{\alpha} = (1\ 0\ 0\ 0)$$

$$\boldsymbol{G} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Fixed sequence procedure



PFS, all patients

Test H_1 at level $\alpha_1 = \alpha*1 = \alpha$

PFS, BRCA wild-type

Test H_2 at level $\alpha_2 = \alpha*0 = 0$

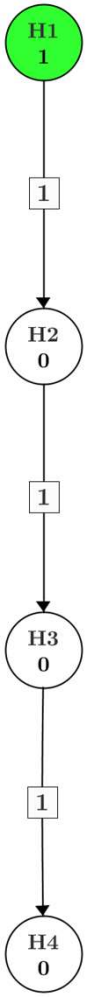
OS, all patients

Test H_3 at level $\alpha_3 = \alpha*0 = 0$

OS, BRCA wild-type

Test H_4 at level $\alpha_4 = \alpha*0 = 0$

Fixed sequence procedure

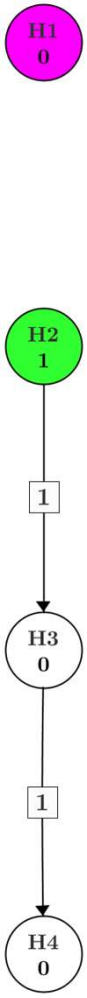


PFS, all patients

***H*$_1$ rejected at level α_1 if p_1<0.05**

node removed and α_1 passed to *H*$_2$ with weight 1

Fixed sequence procedure



PFS, all patients

H_1 rejected

PFS, BRCA wild-type

Test H_2 at level $\alpha_2 = \alpha^*1 = \alpha$

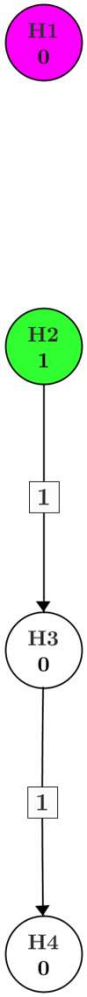
OS, all patients

Test H_3 at level $\alpha_3 = \alpha^*0 = 0$

OS, BRCA wild-type

Test H_4 at level $\alpha_4 = \alpha^*0 = 0$

Fixed sequence procedure



PFS, all patients

***H*$_1$ rejected**

PFS, BRCA wild-type

***H*$_2$ rejected at level α_2 if p_2<0.05**

node removed and α_2 passed to *H*$_3$ with weight 1

Fixed sequence procedure

PFS, all patients

H_1 rejected

PFS, BRCA wild-type

H_2 rejected

OS, all patients

Test H_3 at level $\alpha_3 = \alpha^*1 = \alpha$

OS, BRCA wild-type

Test H_4 at level $\alpha_4 = \alpha^*0 = 0$

Fixed sequence procedure

PFS, all patients

***H*$_1$ rejected**

PFS, BRCA wild-type

***H*$_2$ rejected**

OS, all patients

***H*$_3$ rejected at level α_3 if p_3<0.05**

node removed and α_3 passed to *H*$_4$ with weight 1

Fixed sequence procedure

PFS, all patients

***H*$_1$ rejected**

PFS, BRCA wild-type

***H*$_2$ rejected**

OS, all patients

***H*$_3$ rejected**

OS, BRCA wild-type

Test *H*$_4$ at level $\alpha_4 = \alpha * 1 = \alpha$

Fixed sequence procedure

PFS, all patients

***H*$_1$ rejected**

PFS, BRCA wild-type

***H*$_2$ rejected**

OS, all patients

***H*$_3$ rejected**

OS, BRCA wild-type

***H*$_4$ rejected at level α_4 if p_4<0.05**

Fixed sequence procedure

PFS, all patients

***H*$_1$ rejected**

PFS, BRCA wild-type

***H*$_2$ rejected**

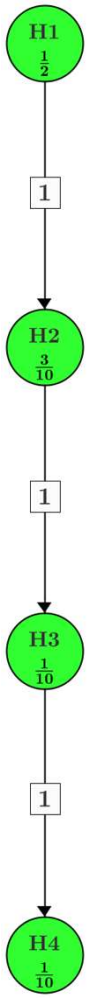
OS, all patients

***H*$_3$ rejected**

OS, BRCA wild-type

***H*$_4$ rejected**

Fallback procedure



PFS, all patients

PFS, BRCA wild-type

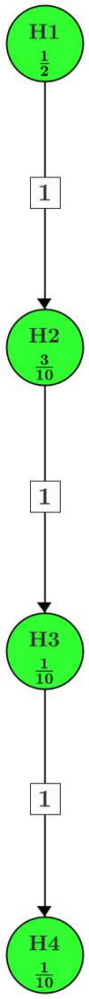
OS, all patients

OS, BRCA wild-type

$$\boldsymbol{\alpha} = (0.5\ 0.3\ 0.1\ 0.1)$$

$$\boldsymbol{G} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Fallback procedure



PFS, all patients

Test H_1 at level $\alpha_1 = 0.5\alpha$

PFS, BRCA wild-type

Test H_2 at level $\alpha_2 = 0.3\alpha$

OS, all patients

Test H_3 at level $\alpha_3 = 0.1\alpha$

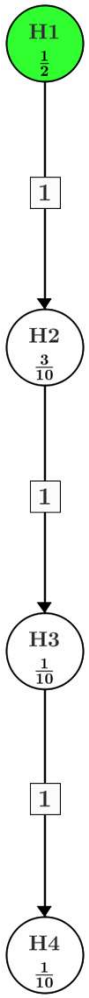
OS, BRCA wild-type

Test H_4 at level $\alpha_4 = 0.1\alpha$

$$\boldsymbol{\alpha} = (0.5\ 0.3\ 0.1\ 0.1)$$

$$\boldsymbol{G} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Fallback procedure

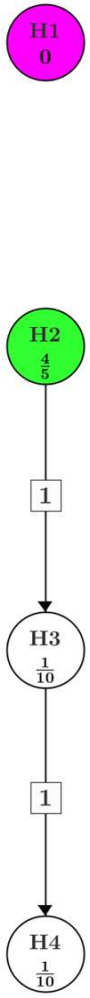


PFS, all patients

***H*$_1$ rejected at level α_1 if p_1<0.025**

node removed and α_1 passed to *H*$_2$ with weight 1

Fallback procedure



PFS, all patients

H_1 rejected

PFS, BRCA wild-type

Test H_2 at level $\alpha_2 = 0.3\alpha + 0.5\alpha = 0.8\alpha$

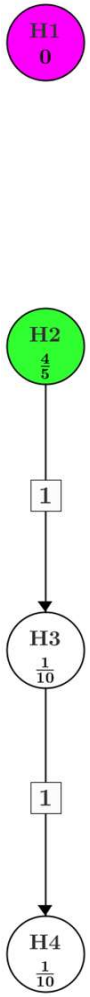
OS, all patients

Test H_3 at level $\alpha_3 = 0.1\alpha$

OS, BRCA wild-type

Test H_4 at level $\alpha_4 = 0.1\alpha$

Fallback procedure



PFS, all patients

***H*$_1$ rejected**

PFS, BRCA wild-type

***H*$_2$ rejected at level α_2 if p_2<0.04**

node removed and α_2 passed to *H*$_3$ with weight 1

Fallback procedure

PFS, all patients

***H*$_1$ rejected**

PFS, BRCA wild-type

***H*$_2$ rejected**

OS, all patients

Test *H*$_3$ at level $\alpha_3 = 0.1\alpha + 0.8\alpha = 0.9\alpha$

OS, BRCA wild-type

Test *H*$_4$ at level $\alpha_4 = 0.1\alpha$

Fallback procedure

PFS, all patients

***H*$_1$ rejected**

PFS, BRCA wild-type

***H*$_2$ rejected**

OS, all patients

***H*$_3$ rejected at level α_3 if p_3<0.045**

node removed and α_3 passed to *H*$_4$ with weight 1

Fallback procedure

PFS, all patients

***H*$_1$ rejected**

PFS, BRCA wild-type

***H*$_2$ rejected**

OS, all patients

***H*$_3$ rejected**

OS, BRCA wild-type

Test H_4 at level $\alpha_4 = 0.1\alpha + 0.9\alpha = \alpha$

Fallback procedure

PFS, all patients

***H*$_1$ rejected**

PFS, BRCA wild-type

***H*$_2$ rejected**

OS, all patients

***H*$_3$ rejected**

OS, BRCA wild-type

***H*$_4$ rejected at level α_4 if p_4<0.05**

Fallback procedure

PFS, all patients

***H*$_1$ rejected**

PFS, BRCA wild-type

***H*$_2$ rejected**

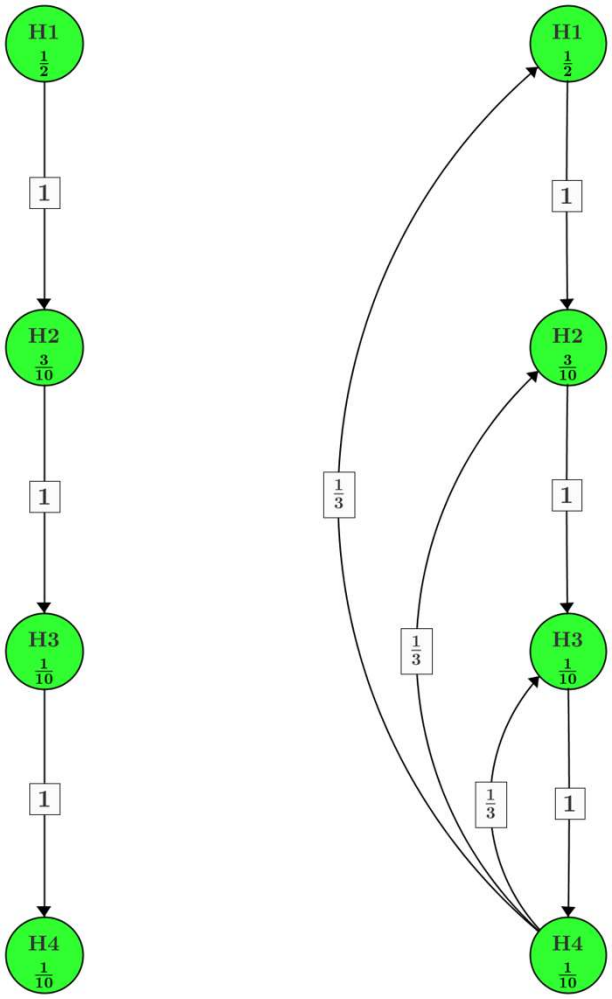
OS, all patients

***H*$_3$ rejected**

OS, BRCA wild-type

***H*$_4$ rejected**

Modified fallback procedures

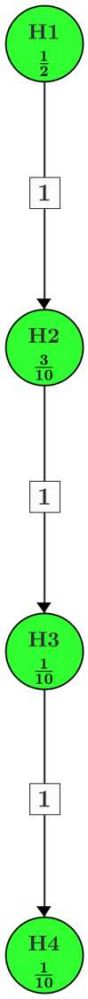


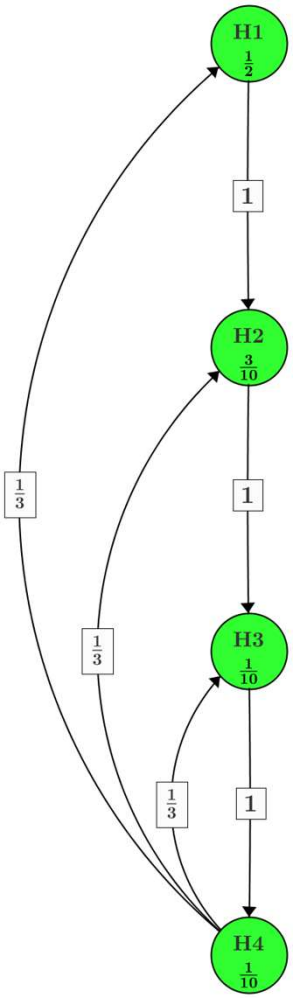
No ‘wasted’ alpha

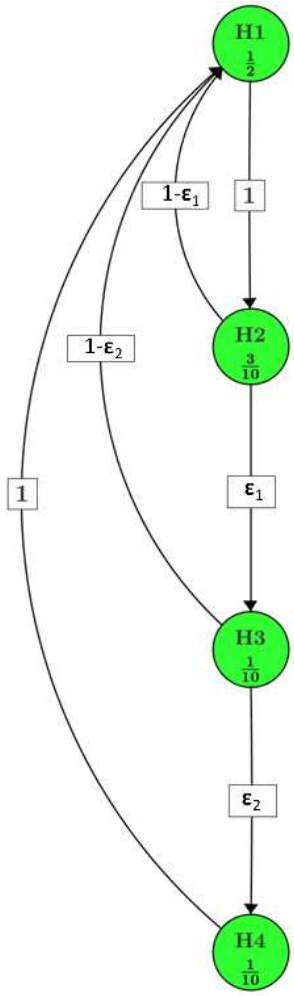
$$\boldsymbol{\alpha} = (0.5\ 0.3\ 0.1\ 0.1)$$

$$\boldsymbol{G} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 & 0 \end{pmatrix}$$

Modified fallback procedures







α is propagated to the most important hypothesis that has not been rejected

$$\boldsymbol{\alpha} = (0.5\ 0.3\ 0.1\ 0.1)$$

$$\boldsymbol{G} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1-\varepsilon_1 & 0 & \varepsilon_1 & 0 \\ 1-\varepsilon_2 & 0 & 0 & \varepsilon_2 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Bonferroni procedure

PFS, all patients

PFS, BRCA wild-type

OS, all patients

OS, BRCA wild-type

$$\boldsymbol{\alpha} = (¼\ ¼\ ¼\ ¼)$$

$$\boldsymbol{G} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Bonferroni procedure

PFS, all patients

H_1 rejected at level α_1 if p_1<0.0125

PFS, BRCA wild-type

H_2 rejected at level α_2 if p_2<0.0125

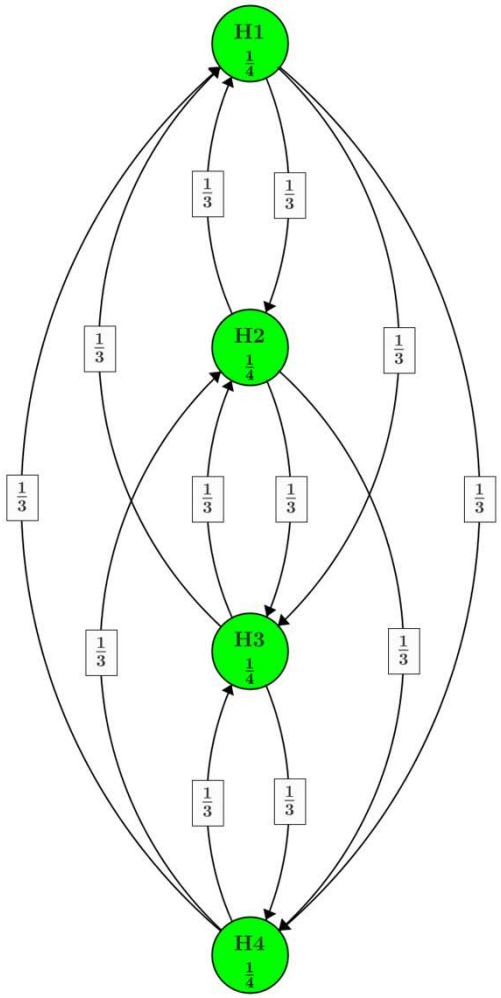
OS, all patients

H_3 rejected at level α_3 if p_3<0.0125

OS, BRCA wild-type

H_4 rejected at level α_4 if p_4<0.0125

Holm procedure



PFS, all patients

PFS, BRCA wild-type

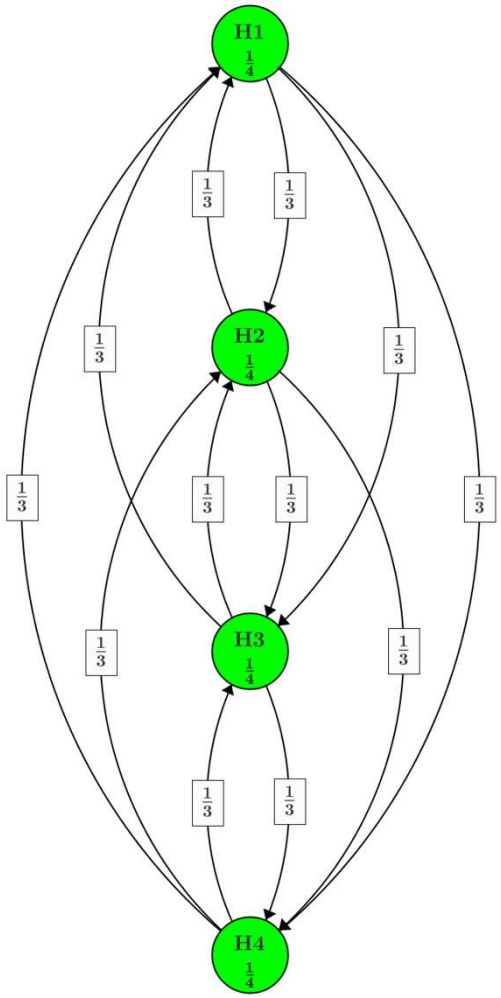
OS, all patients

OS, BRCA wild-type

$$\boldsymbol{\alpha} = \left(\tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}\right)$$

$$\boldsymbol{G} = \begin{pmatrix} 0 & \tfrac{1}{3} & \tfrac{1}{3} & \tfrac{1}{3} \\ \tfrac{1}{3} & 0 & \tfrac{1}{3} & \tfrac{1}{3} \\ \tfrac{1}{3} & \tfrac{1}{3} & 0 & \tfrac{1}{3} \\ \tfrac{1}{3} & \tfrac{1}{3} & \tfrac{1}{3} & 0 \end{pmatrix}$$

Holm procedure



PFS, all patients

H_1 rejected at level α_1 if p_1<0.0125

PFS, BRCA wild-type

H_2 rejected at level α_2 if p_2<0.0125

OS, all patients

H_3 rejected at level α_3 if p_3<0.0125

OS, BRCA wild-type

H_4 rejected at level α_4 if p_4<0.0125

Generalised graphical approach

- **α**=(α_1,…., α_m) the local significance levels such that $\sum_i \alpha_i \leq \alpha$
- **G**=(g_{ij}) the m by m transition matrix such that $0 \leq g_{ij} \leq 1$, $g_{ii} = 0$, and $\sum_j g_{ij} \leq 1$
- Algorithm:

0. Set $I = M$.
1. Let $j = \arg\min_{i \in I} p_i/\alpha_i$
2. If $p_j \leqslant \alpha_j$, reject H_j; otherwise stop.
3. Update the graph:

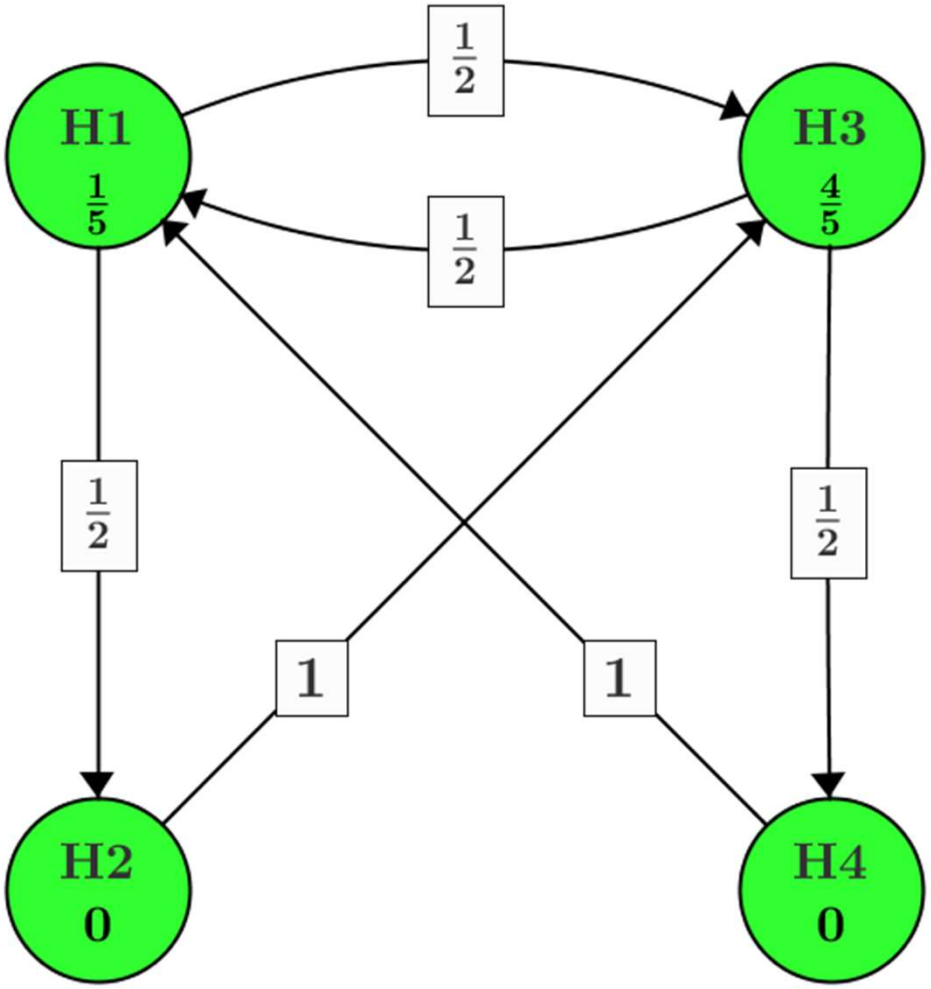
$$I \to I \setminus \{j\}$$

$$\alpha_\ell \to \begin{cases} \alpha_\ell + \alpha_j g_{j\ell}, & \ell \in I \\ 0 & \text{otherwise} \end{cases}$$

$$g_{\ell k} \to \begin{cases} \dfrac{g_{\ell k} + g_{\ell j} g_{jk}}{1 - g_{\ell j} g_{j\ell}}, & \ell, k \in I, \ \ell \neq k \\ 0 & \text{otherwise} \end{cases}$$

4. If $|I| \geqslant 1$, go to step 1; otherwise stop.

Alternative test procedure



H1: PFS, all patients

H2: PFS, BRCA wild-type

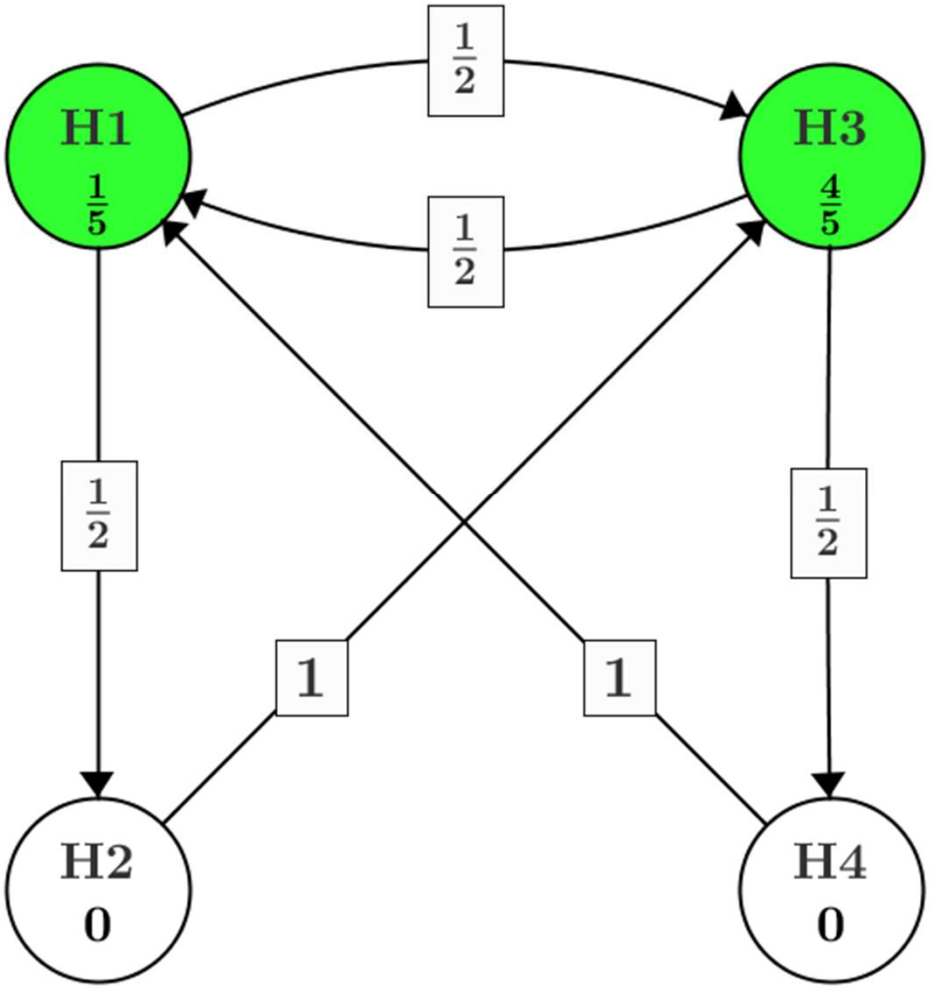
H3: OS, all patients

H4: OS, BRCA wild-type

$$\boldsymbol{\alpha} = \left(\frac{1}{5}\ 0\ \frac{4}{5}\ 0\right)$$

$$\boldsymbol{G} = \begin{pmatrix} 0 & ½ & ½ & 0 \\ 0 & 0 & 1 & 0 \\ ½ & 0 & 0 & ½ \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Alternative test procedure



H1: *p*=0.001

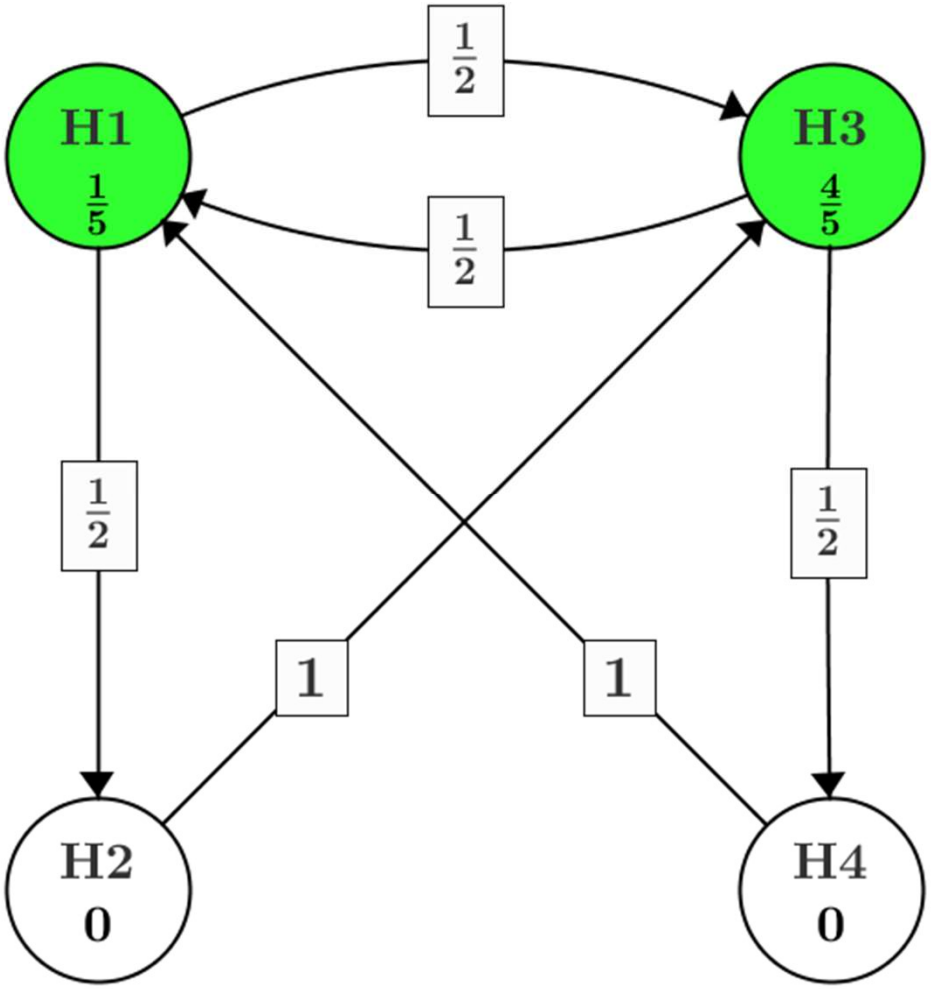
H2: *p*=0.001

H3: *p*=0.04

H4: *p*=0.06

(hypothetical p-values)

Alternative test procedure



H1: p=0.001 rejected

node H_1 removed and α_1 passed using

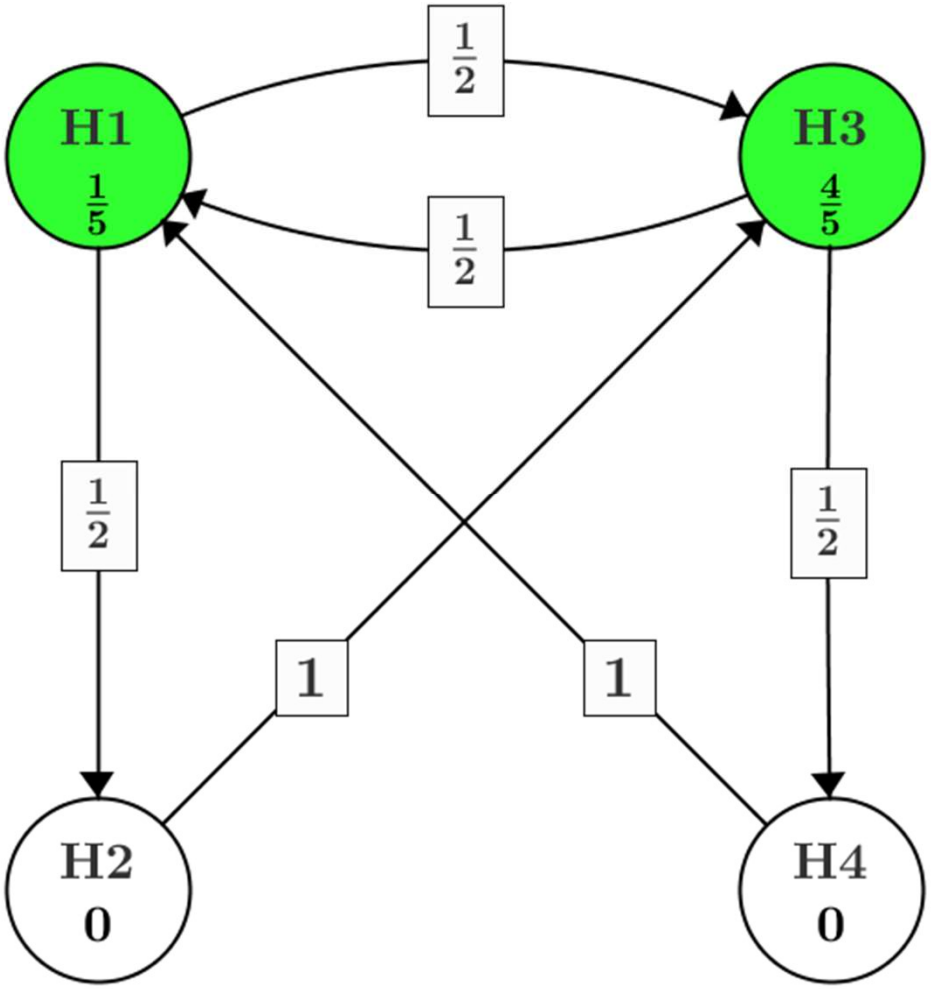
$$\alpha_\ell \rightarrow \begin{cases} \alpha_\ell + \alpha_j g_{j\ell}, & \ell \in I \\ 0 & \text{otherwise} \end{cases}$$

$\alpha_2 = 0 + (\frac{1}{5} * \frac{1}{2}) = \frac{1}{10}$

$\alpha_3 = \frac{4}{5} + (\frac{1}{5} * \frac{1}{2}) = \frac{9}{10}$

$\alpha_4 = 0 + (\frac{1}{5} * 0) = 0$

Alternative test procedure



H1: p=0.001 rejected

node H_1 removed and weights updated

$$g_{\ell k} \rightarrow \begin{cases} \dfrac{g_{\ell k} + g_{\ell j} g_{jk}}{1 - g_{\ell j} g_{j\ell}}, & \ell, k \in I, \ \ell \neq k \\ 0 & \text{otherwise} \end{cases}$$

g_{23} = (1+(0*½))/(1-(0*½)) = 1/1 = 1

g_{24} = (0+(0*0))/(1-(0*½)) = 0/1 = 0

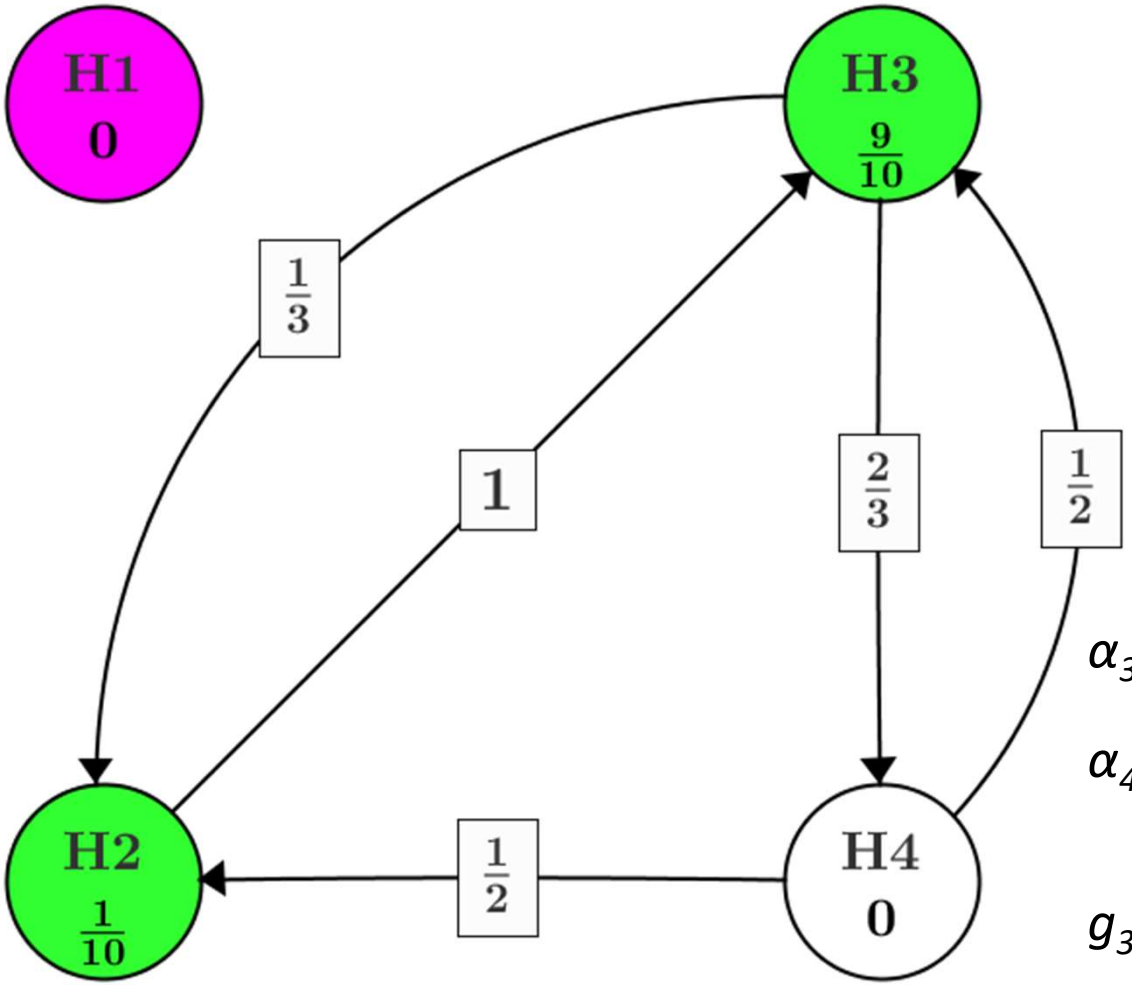
g_{32} = (0+(½*½))/(1-(½*½)) = ¼/¾ = 1/3

g_{34} = (½+(½*0))/(1-(½*½)) = ½/¾ = 2/3

g_{42} = (0+(1*½))/(1-(1*0)) = ½/1 = ½

g_{43} = (0+(1*½))/(1-(1*0)) = ½/1 = ½

Alternative test procedure



H1: p=0.001 rejected

H2: p=0.001 rejected

H3: p=0.04

H4: p=0.06

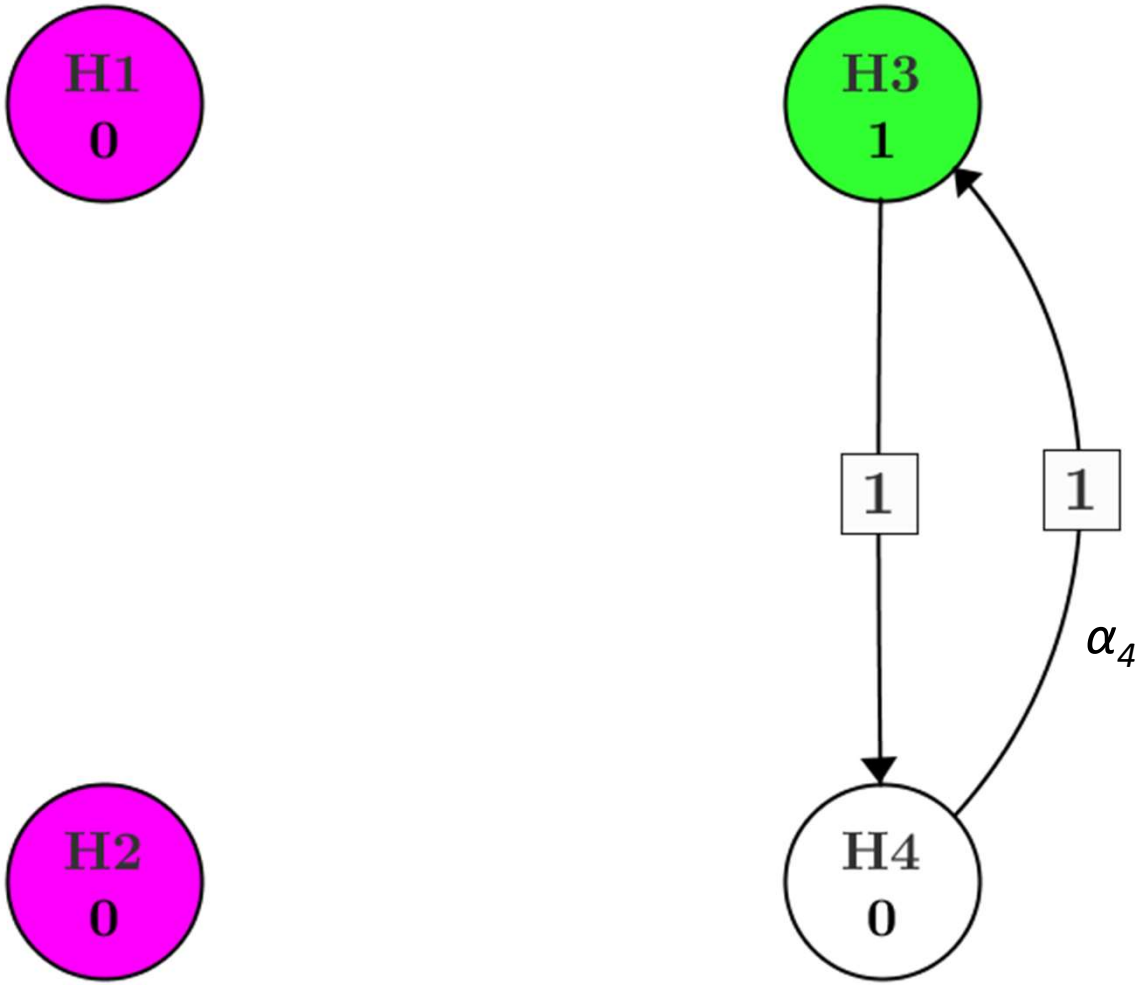
$\alpha_3 = \frac{9}{10} + (\frac{1}{10}*1) = 1$

$\alpha_4 = 0 + (\frac{1}{10}*0) = 0$

$g_{34} = (\frac{2}{3}+(\frac{1}{3}*0))/(1-(\frac{1}{3}*1)) = \frac{2}{3}/\frac{2}{3} = 1$

$g_{43} = (\frac{1}{2}+(\frac{1}{2}*1))/(1-(\frac{1}{2}*0)) = 1/1 = 1$

Alternative test procedure



H1: *p*=0.001 rejected

H2: *p*=0.001 rejected

H3: *p*=0.04 rejected

H4: *p*=0.06

$\alpha_4 = 0 + (1*1) = 1$

Alternative test procedure

H1: *p*=0.001 rejected

H2: *p*=0.001 rejected

H3: *p*=0.04 rejected

H4: *p*=0.06 not rejected

Adjusted p-values

- To use adjusted p-values p_1^{adj}, …., p_m^{adj}, define $\mathbf{w}=(w_1, \ldots, w_m)=(\alpha_1, \ldots, \alpha_m)/\alpha$
- Algorithm:

0. Set $I = M$ and $p_{\max} = 0$.
1. Let $j = \arg\min_{i \in I} p_i / w_i$.
2. Calculate $p_j^{\text{adj}} = \max\{p_j / w_j, p_{\max}\}$ and set $p_{\max} = p_j^{\text{adj}}$
3. Update the graph:

$$I \to I \setminus \{j\}$$

$$w_\ell \to \begin{cases} w_\ell + w_j g_{j\ell}, & \ell \in I \\ 0 & \text{otherwise} \end{cases}$$

$$g_{\ell k} \to \begin{cases} \dfrac{g_{\ell k} + g_{\ell j} g_{jk}}{1 - g_{\ell j} g_{j\ell}}, & \ell, k \in I, \ \ell \neq k \\ 0 & \text{otherwise} \end{cases}$$

4. If $|I| \geqslant 1$, go to step 1; otherwise stop.
5. Reject all hypotheses H_j with $p_j^{\text{adj}} \leqslant \alpha$.

Conclusions

The graphical approach is a powerful and efficient way to control the study-level false positive rate for multiple testing in clinical trials

Intuitive method and easy to communicate with a hierarchical structure reflecting the relative importance of the multiple objectives

Choosing the testing procedure at the design stage has to balance statistical power (α-splitting) with allowing the testing of lower-order hypotheses

References

R-packages:

- http://cran.r-project.org/web/packages/gMCP/
  (gMCP: Graph Based Multiple Comparison Procedures)
- http://cran.r-project.org/web/packages/multxpert/
  (multxpert: Common Multiple Testing Procedures and Gatekeeping Procedures)

Regulatory guidelines:

- EMA “Guideline on multiplicity issues in clinical trials”
- FDA “Multiple endpoint analyses”

References

- Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. Stat Med. 2009 Feb 15;28(4):586-604.
- Bretz F, Maurer W, Hommel G. Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. Stat Med. 2011 Jun 15;30(13):1489-1501.
- Burman CF, Sonesson C, Guilbaud O. A recycling framework for the construction of Bonferroni-based multiple tests. Stat Med. 2009 Feb 28;28(5):739-761.
- Dmitrienko A, Offen WW, Westfall PH. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. Stat Med. 2003 Aug 15;22(15):2387-2400.
- Elyashiv O, Ledermann J, Parmar G, Farrelly L, Counsell N, Feeney A, El-Khouly F, Macdonald I, Neto A, Arthur-Darkwa E, Burnett E, Jayson GC, Mileshkin L, Gourley C, Nicum S. ICON 9-an international phase III randomized study to evaluate the efficacy of maintenance therapy with olaparib and cediranib or olaparib alone in patients with relapsed platinum-sensitive ovarian cancer following a response to platinum-based chemotherapy. Int J Gynecol Cancer. 2021 Jan;31(1):134-138.
- Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. Biometrika. 1976 Dec;63(3):655–660.