A refresher on multiple comparisons (or how to spend your α level at Christmas time)

Brice Ozenne

December 6, 2019, NRU Christmas Symposium

Let's start with an example (Ebert et al., 2019)

Aim: investigate the impact of a disease on some brain signal

```
        group    id thalamus pallidostriatum neocortex midbrain
1: concussion 125BB    2.808           3.117     2.239    3.643
2: concussion 132MH    4.292           3.893     3.158    5.050
3: concussion 133AG    9.566           7.435     5.723    9.131
4:    healthy  59HT    9.605           8.066     6.852   10.346
5:    healthy  67MF    8.543           6.742     5.419    7.944
6:    healthy  71BS    5.556           4.613     4.600    7.936
```

Strategy for handling multiple comparisons

Strategy for handling multiple comparisons

1. Avoid:
   - Focus one brain region, e.g. based on existing knowledge.

region	concussion effect (%)	p-value
cingulateGyrus	17.28	0.034

Strategy for handling multiple comparisons

1. Avoid:
   - Focus one brain region, e.g. based on existing knowledge.
     → may lead to an unacceptable loss of power

region	concussion effect (%)	p-value
cingulateGyrus	17.28	0.034

region	concussion effect (%)	p-value
thalamus	12.75	0.23

Strategy for handling multiple comparisons

1. Avoid:
   - Make a global test, i.e., absence of disease effect in all brain regions.

p-value = 0.011

Strategy for handling multiple comparisons

1. Avoid:
   - Make a global test, i.e., absence of disease effect in all brain regions.
     → loose interpretability

p-value = 0.011

Strategy for handling multiple comparisons

1. Avoid:
   - Assume the same effect in all brain regions and test it.

region	concussion effect (%)	p-value	adjusted p-value
All	10.4	0.1975	

Strategy for handling multiple comparisons

1. Avoid:
   - Assume the same effect in all brain regions and test it.
     → makes (strong but testable) assumptions

region	concussion effect (%)	p-value	adjusted p-value
All	10.4	0.1975	
thalamus	12.75		
pallidostriatum	12.03		
neocortex	4.38		
midbrain	10.4		
pons	1.56		
cingulateGyrus	17.28		
...			

Strategy for handling multiple comparisons

1. Avoid:
2. Cope with:
   - standard adjustment for multiple comparisons (Bonferroni)

region	concussion effect (%)	p-value	adjusted p-value
thalamus	12.75	0.23	1
pallidostriatum	12.03	0.177	1
neocortex	4.38	0.601	1
midbrain	10.4	0.219	1
pons	1.56	0.858	1
cingulateGyrus	17.28	0.034	0.304
...			

Strategy for handling multiple comparisons

1. Avoid:
2. Cope with:
   - standard adjustment for multiple comparisons (Bonferroni)
     → may lead to an unacceptable loss of power

region	concussion effect (%)	p-value	adjusted p-value
thalamus	12.75	0.23	1
pallidostriatum	12.03	0.177	1
neocortex	4.38	0.601	1
midbrain	10.4	0.219	1
pons	1.56	0.858	1
cingulateGyrus	17.28	0.034	0.304
...			

Strategy for handling multiple comparisons

1. Avoid:
2. Cope with:
   - Use "modern" approaches for multiple comparisons

region	concussion effect (%)	p-value	adjusted p-value
thalamus	12.75	0.23	0.395
pallidostriatum	12.03	0.177	0.358
neocortex	4.38	0.601	0.753
midbrain	10.4	0.219	0.395
pons	1.56	0.858	0.858
cingulateGyrus	17.28	0.034	0.096
...			

Strategy for handling multiple comparisons

1. Avoid:
2. Cope with:
   - Use "modern" approaches for multiple comparisons
     → more work! And choices need to be made ...

region	concussion effect (%)	p-value	adjusted p-value
thalamus	12.75	0.23	0.395
pallidostriatum	12.03	0.177	0.358
neocortex	4.38	0.601	0.753
midbrain	10.4	0.219	0.395
pons	1.56	0.858	0.858
cingulateGyrus	17.28	0.034	0.096
...			

Interpretation: p-value vs. adjusted p-value?

region	concussion effect (%)	p-value	adjusted p-value
...			
cingulateGyrus	17.28	**0.034**	**0.096**
...			

Interlude

Definition

- given a **random** variable X,
  e.g. estimator of the concussion effect
- and **a** null hypothesis H_0,
  e.g. $\mathbb{E}[X] = 0$, no concussion effect

The p-value is:

- the probability to observe a realisation of X at least as large as what we observed under H_0,
  e.g. $\mathbb{P}\left[|X| > 17.28 \middle| H_0\right]$.

⚠ P-value are relative to a fixed null hypothesis,
i.e. defined independently of the observations

Interpretation: p-value vs. adjusted p-value?

So why did you picked cingulateGyrus:

- prior knowledge → p-value
- looked at the p-values → an adjustment is necessary!

region	concussion effect (%)	p-value	adjusted p-value
...			
cingulateGyrus	17.28	**0.034**	**0.096**
...			

Handling cherry picking

Statisticians have no problem with cherry picking . . . as soon as it is correctly accounted for!

Cherry picking redefines the null hypothesis:

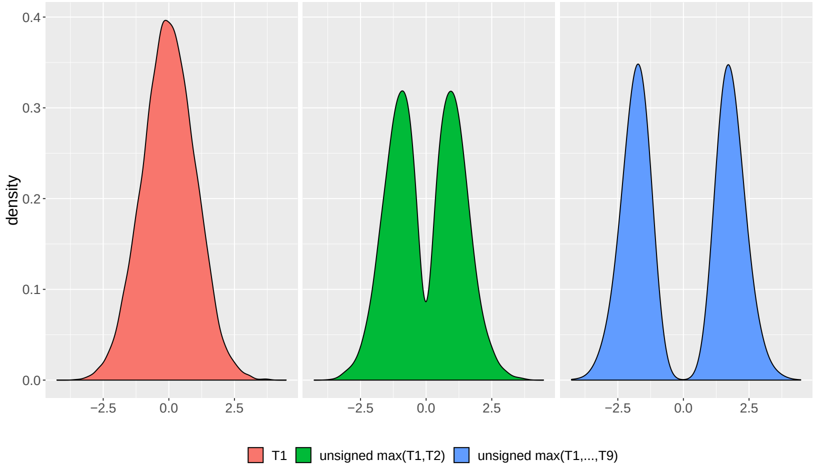
- $H_0^{\max}$: $\mathbb{E}\left[X_{\text{thalamus}}\right] = 0$
  and $\mathbb{E}\left[X_{\text{pallidostriatum}}\right] = 0$
  and . . . ,
  i.e. no effect in all regions

- i.e., denoting by $T_{.}$ the test statistics,

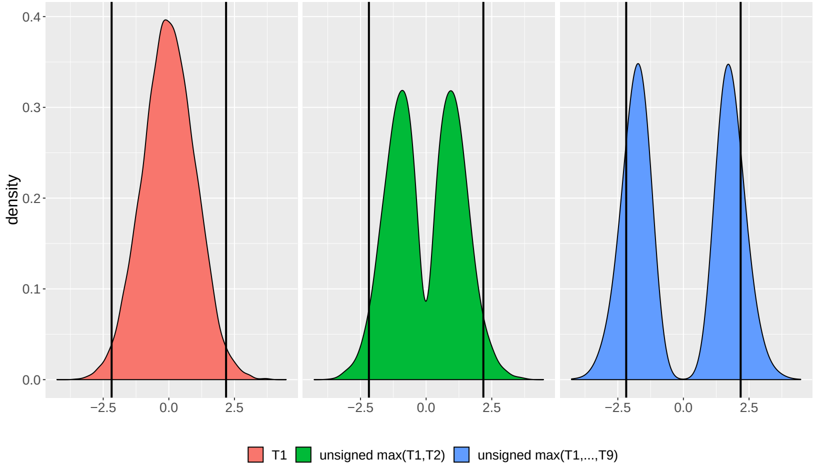
$$\mathbb{P}\left[\max\left(\left|T_{\text{thalamus}}\right|, \left|T_{\text{pallidostriatum}}\right|, \ldots\right) > 2.18 \middle| H_0^{\max}\right] = 0.096.$$

Called a max-test procedure.

Impact of the cherry picking on the distribution of the test statistic



Impact of the cherry picking on the distribution of the test statistic



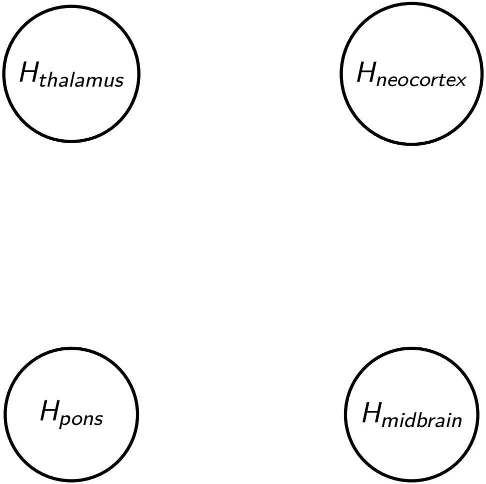
Modern multiplicity adjustment methods

How can we improve[1] the Bonferroni adjustment?

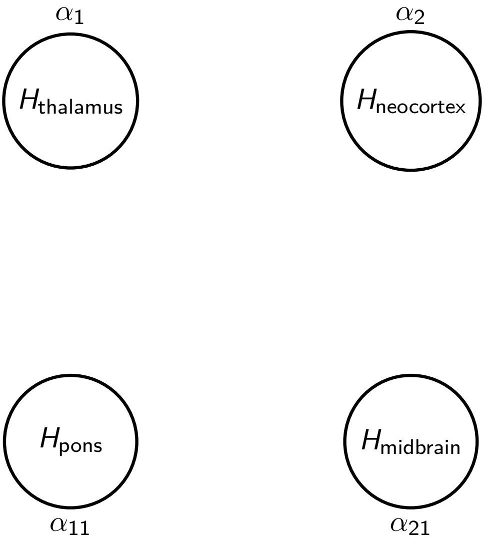
- account for the **correlation** between the test statistics,
  if we do twice the same test, only correct for one
- account for **logical** restrictions,
  when testing $\mu_1 = \mu_2 = \mu_3$, if $\mu_1 \neq \mu_3$ and $\mu_2 \neq \mu_3$ then $\mu_1 \neq \mu_2$.
- account for the **ordering** between the hypothesis
  graphical approach proposed by Bretz et al. (2009)

[1] Higher power while controling the FWER, Alosh et al. (2014)

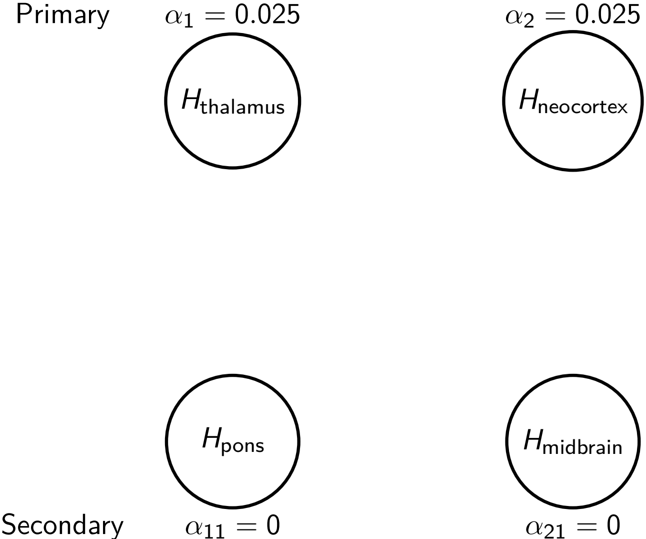
Step 1: write down null hypotheses



Step 2: Spread the α level ($\alpha_1 + \alpha_2 + \alpha_{11} + \alpha_{21} = 0.05$)



Step 2: Spread the α level $(\alpha_1 + \alpha_2 + \alpha_{11} + \alpha_{21} = 0.05)$



Primary

$\alpha_1 = 0.025$

$\alpha_2 = 0.025$

H_{thalamus}

$H_{\text{neocortex}}$

H_{pons}

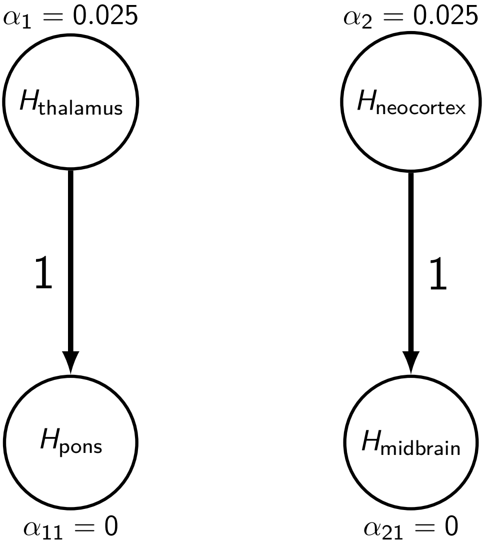
H_{midbrain}

Secondary

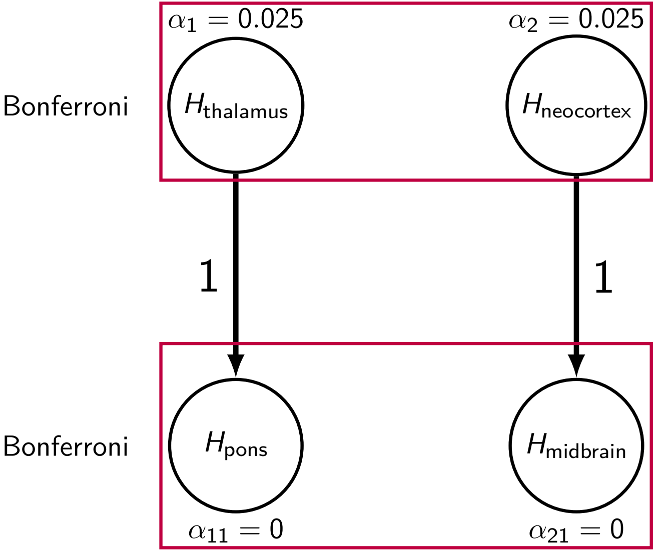
$\alpha_{11} = 0$

$\alpha_{21} = 0$

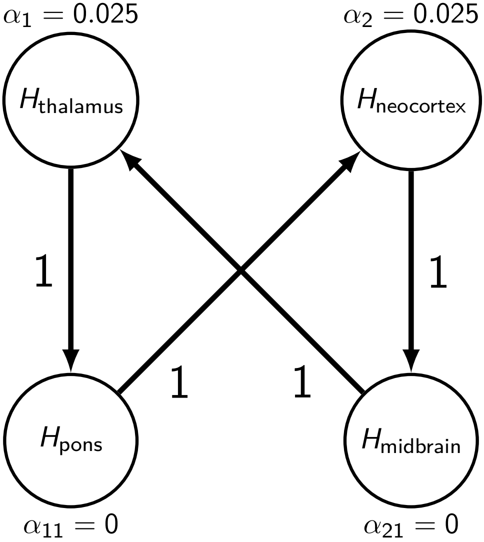
Step 3: Define the α propagation



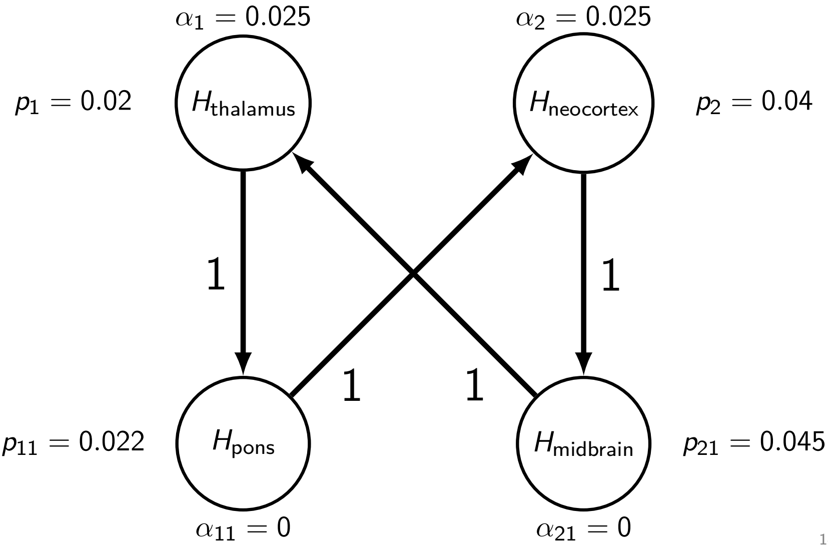
Step 3: Define the α propagation



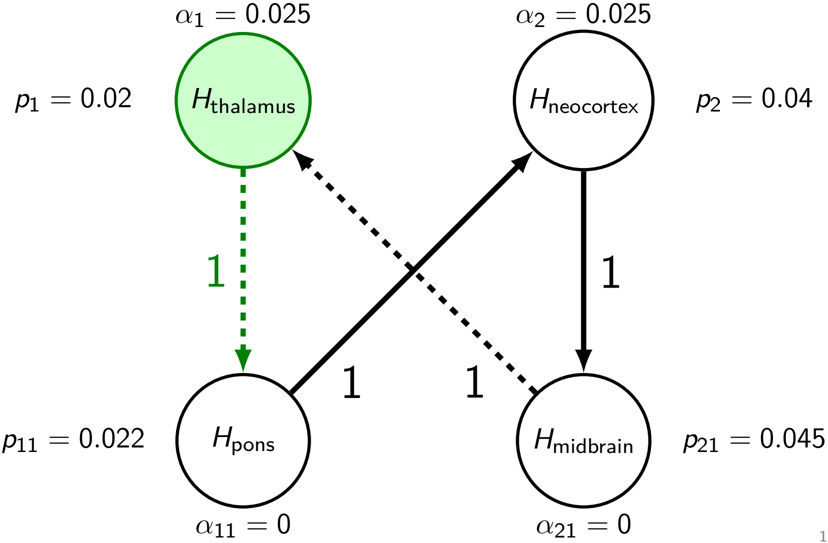
Step 3: A more powerful α propagation



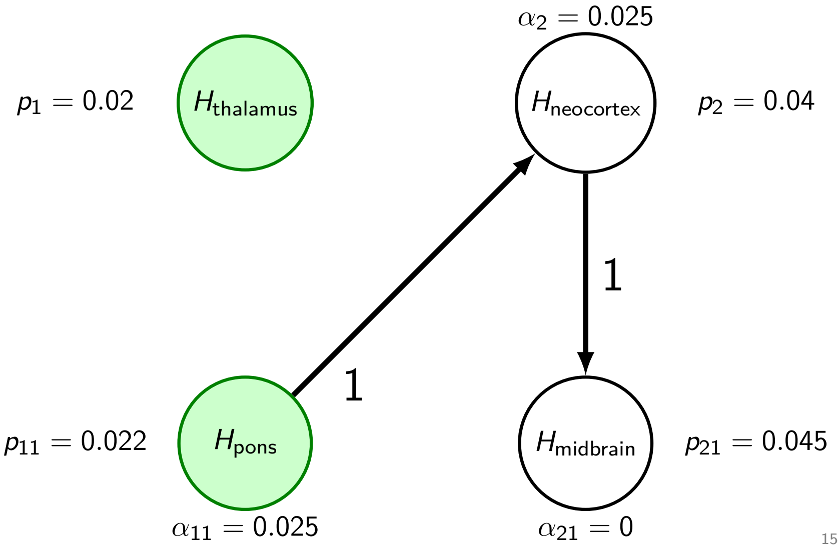
Step 4: Add the (uncorrected) p-values



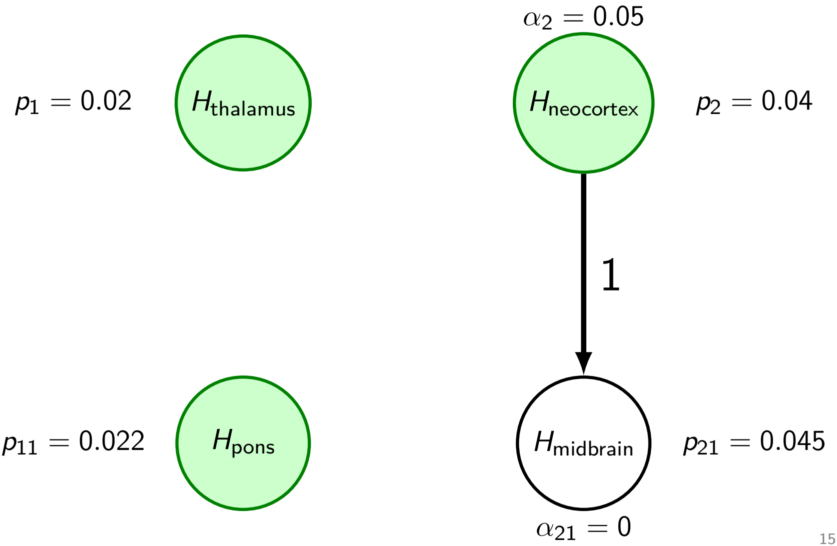
Step 5: Run the algorithm



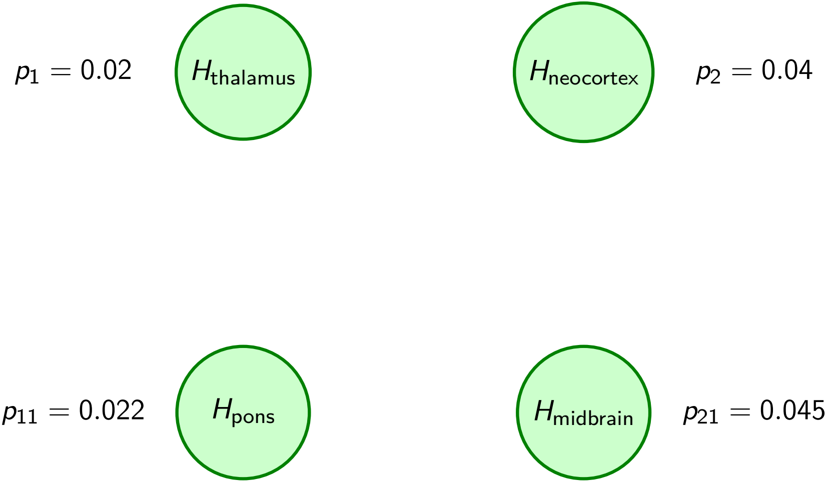
Step 5: Run the algorithm



Step 5: Run the algorithm



Step 5: Run the algorithm



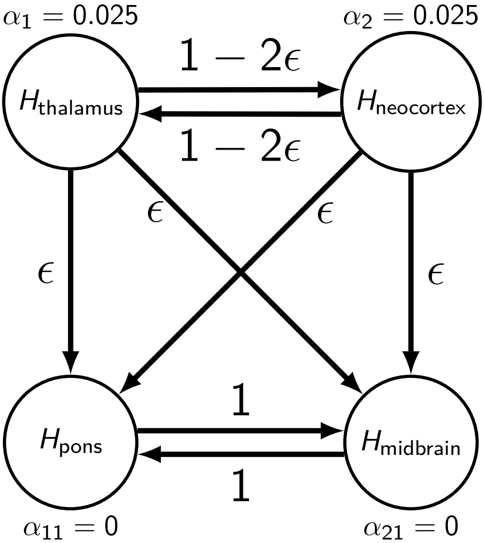
$p_1 = 0.02$ H_{thalamus}

$H_{\text{neocortex}}$ $p_2 = 0.04$

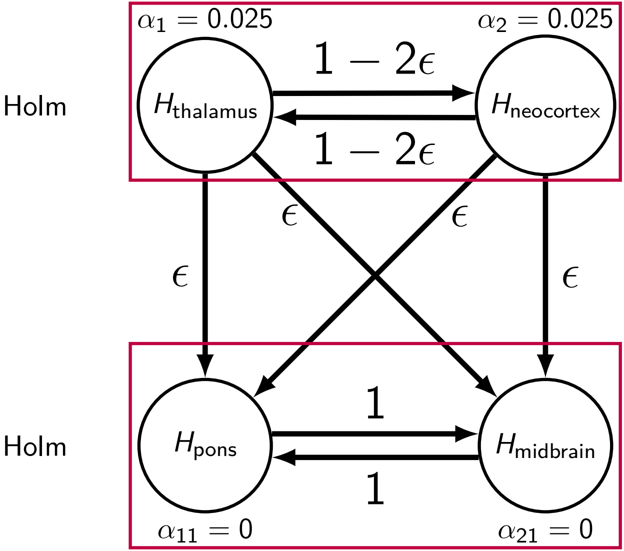
$p_{11} = 0.022$ H_{pons}

H_{midbrain} $p_{21} = 0.045$

Many other possible options



Many other possible options



Alosh, M., Bretz, F., and Huque, M. (2014). Advanced multiplicity adjustment methods in clinical trials. *Statistics in medicine*, 33(4):693–713.

Bretz, F., Maurer, W., Brannath, W., and Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in medicine*, 28(4):586–604.

Ebert, S. E., Jensen, P., Ozenne, B., Armand, S., Svarer, C., Stenbaek, D. S., Moeller, K., Dyssegaard, A., Thomsen, G., Steinmetz, J., et al. (2019). Molecular imaging of neuroinflammation in patients after mild traumatic brain injury: a longitudinal 123i-clinde single photon emission computed tomography study. *European journal of neurology*.

Conditions

Let $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_m)$ denote the local significance levels, such that $\sum_{i=1}^{m}\alpha_i\leqslant\alpha$.
Let $\mathbf{G}=(g_{ij})$ denote an $m\times m$ transition matrix with freely chosen entries g_{ij}
that are subject to the regularity conditions

$$0\leqslant g_{ij}\leqslant 1,\quad g_{ii}=0 \text{ and } \sum_{k=1}^{m} g_{ik}\leqslant 1 \quad \text{for all } i,j=1,\ldots,m \tag{1}$$

The weight g_{ij} determines the fraction of the local level α_i that is allocated to H_j
in case H_i was rejected

and the transition matrix $\mathbf{G}$ thus fully determines the directed edges.

Algorithm

Based on the observed p-values $p_i \;\; i \in M = \{1, \ldots, m\}$,
we define a sequentially rejective test procedure through the following algorithm:

Algorithm 1

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.

In the Appendix we show that a graph $\mathscr{G} = (\boldsymbol{\alpha}, \mathbf{G})$ together with the updating rules from Algorithm 1 defines a short cut for a consonant closed test procedure where each intersection hypothesis is tested with a weighted Bonferroni test. Together with Algorithm 1, a graph $\mathscr{G} = (\boldsymbol{\alpha}, \mathbf{G})$ thus defines a sequentially rejective multiple test procedure that strongly controls the FWER at level α, where $\boldsymbol{\alpha}$ and $\mathbf{G}$ are subject to the constraints above.