A general principle for multiple testing

MOR Seminar October 14, 2025

Rianne de Heide, University of Twente and Centrum Wiskunde & Informatica Amsterdam

Bringing closure to FDR control: a general principle for multiple testing

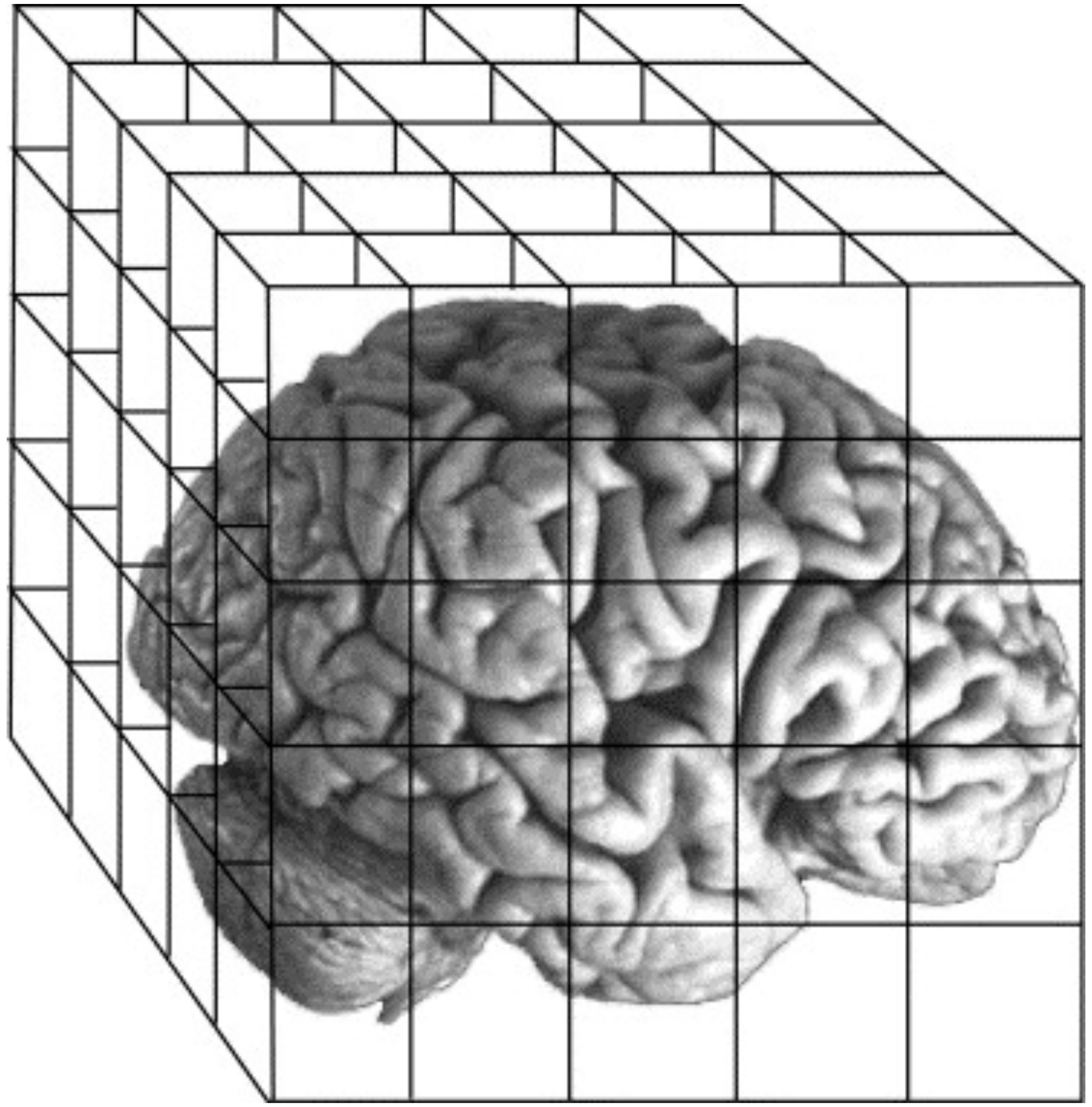
Ziyu Xu, Aldo Solari, Lasse Fischer, Rianne de Heide, Aaditya Ramdas and Jelle Goeman

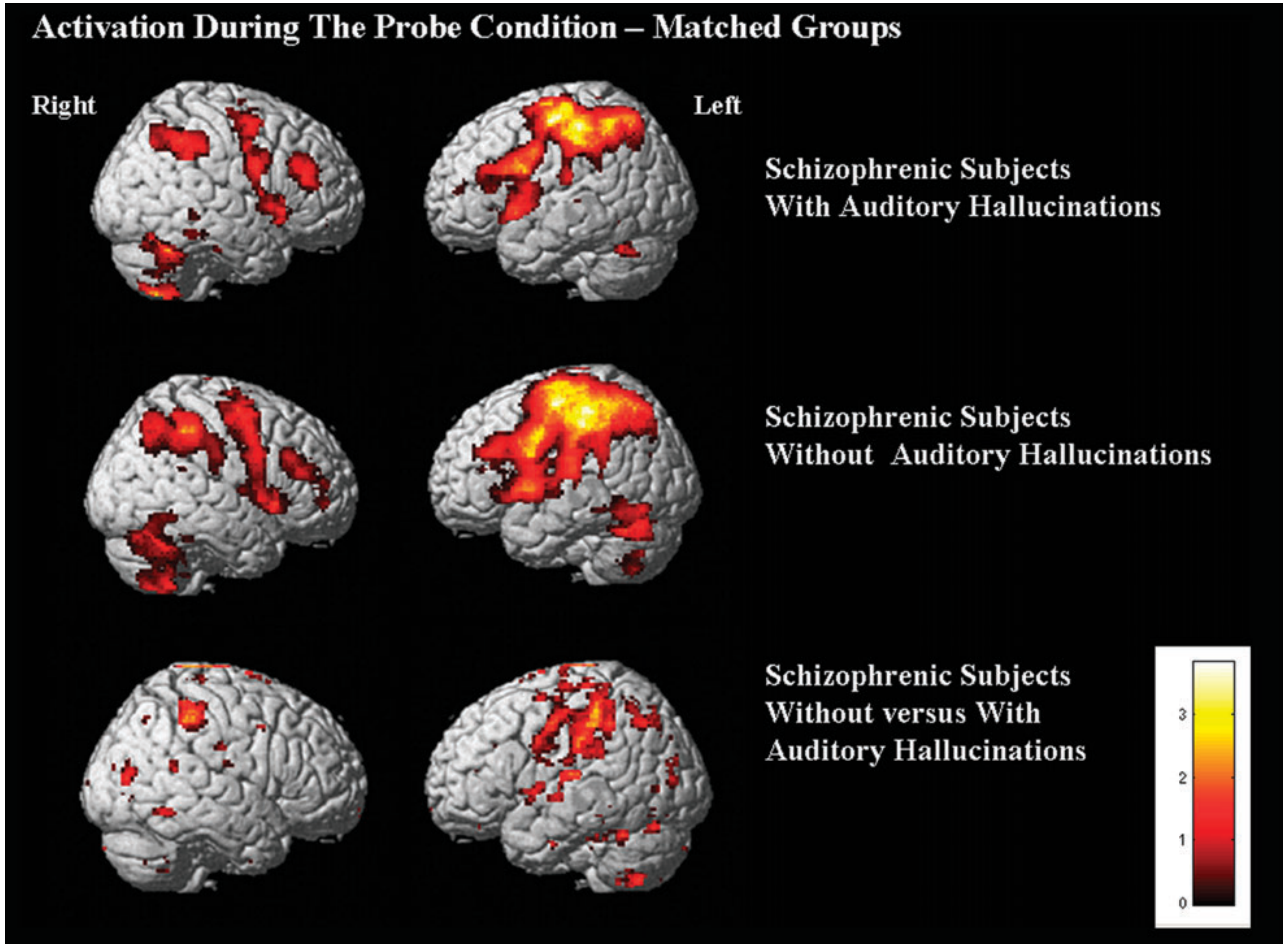
https://arxiv.org/pdf/2509.02517

What is multiple testing?

Example: Multiple testing in neuroimaging

130.000 voxels





Wible, Cynthia G. et al. "fMRI activity correlated with auditory hallucinations during performance of a working memory task: data from the FBIRN consortium study." Schizophrenia bulletin 35 1 (2009): 47-57 .

Multiple testing: the problem

- If we test n true null hypotheses at level α, then on average we will (falsely) reject αn of them.
- Examples:
  - testing whether some of 20.000 genes are linked to a disease
  - fMRI: 100.000 voxels
  - DNA methylation: 500.000 sites
- We need other measures of acceptance/rejection errors.
- We need statistical procedures to control these measures of errors.

Error rates

$N \subseteq [m]$ hypotheses are true null; the rest are potential discoveries

Famous error rates:

- Familywise error rate (FWER): $P(|R \cap N| > 0)$
- Per-family error rate: $\mathbb{E}(|R \cap N|)$
- False Discovery rate (FDR): $\mathbb{E}\left(\frac{|R \cap N|}{R}\right)$

General form

Control some expected loss: $\mathbb{E}(f_N(R))$

FWER history

- Closure principle: necessary and sufficient for the construction of valid methods (Sonnemann, 1982, 2008)

FWER history

- Closure principle: necessary and sufficient for the construction of valid methods (Sonnemann, 1982, 2008)
- Any method that controls FWER is a special case of a closed testing procedure

FWER history

- Closure principle: necessary and sufficient for the construction of valid methods (Sonnemann, 1982, 2008)
- Any method that controls FWER is a special case of a closed testing procedure
- Challenged by the Partitioning Principle (Finner and Strassburger, 2002)

FWER history

- Closure principle: necessary and sufficient for the construction of valid methods (Sonnemann, 1982, 2008)
- Any method that controls FWER is a special case of a closed testing procedure
- Challenged by the Partitioning Principle (Finner and Strassburger, 2002)
- They are equivalent: Goeman et al. (2021)

False Discovery Proportion (FDP) history

- Genovese and Wasserman (2006) and Goeman and Solari (2011) have extended closed testing to control of false discovery proportions (FDPs)

False Discovery Proportion (FDP) history

- Genovese and Wasserman (2006) and Goeman and Solari (2011) have extended closed testing to control of false discovery proportions (FDPs)

- Goeman et al. (2021) showed that all methods controlling a quantile of the distribution of FDP are either equivalent to a closed testing procedure or are dominated by one, extending the Closure Principle to all methods controlling FDP.

Why is the closure principle nice?

- reduces the complex task of constructing a multiple testing method to the simpler task of choosing hypothesis tests for intersection hypotheses

Why is the closure principle nice?

- reduces the complex task of constructing a multiple testing method to the simpler task of choosing hypothesis tests for intersection hypotheses
- helps to handle complex situations such as restricted combinations

Why is the closure principle nice?

- reduces the complex task of constructing a multiple testing method to the simpler task of choosing hypothesis tests for intersection hypotheses
- helps to handle complex situations such as restricted combinations
- methods constructed using closed testing often allow for some user flexibility, permitting researchers to modify some aspects of the multiple testing procedure post hoc without compromising error control

FDR history

- Blanchard and Roquain (2008) formulated two quite general sufficient conditions, self-consistency and dependence control, under which, if both hold, FDR control is guaranteed.

FDR history

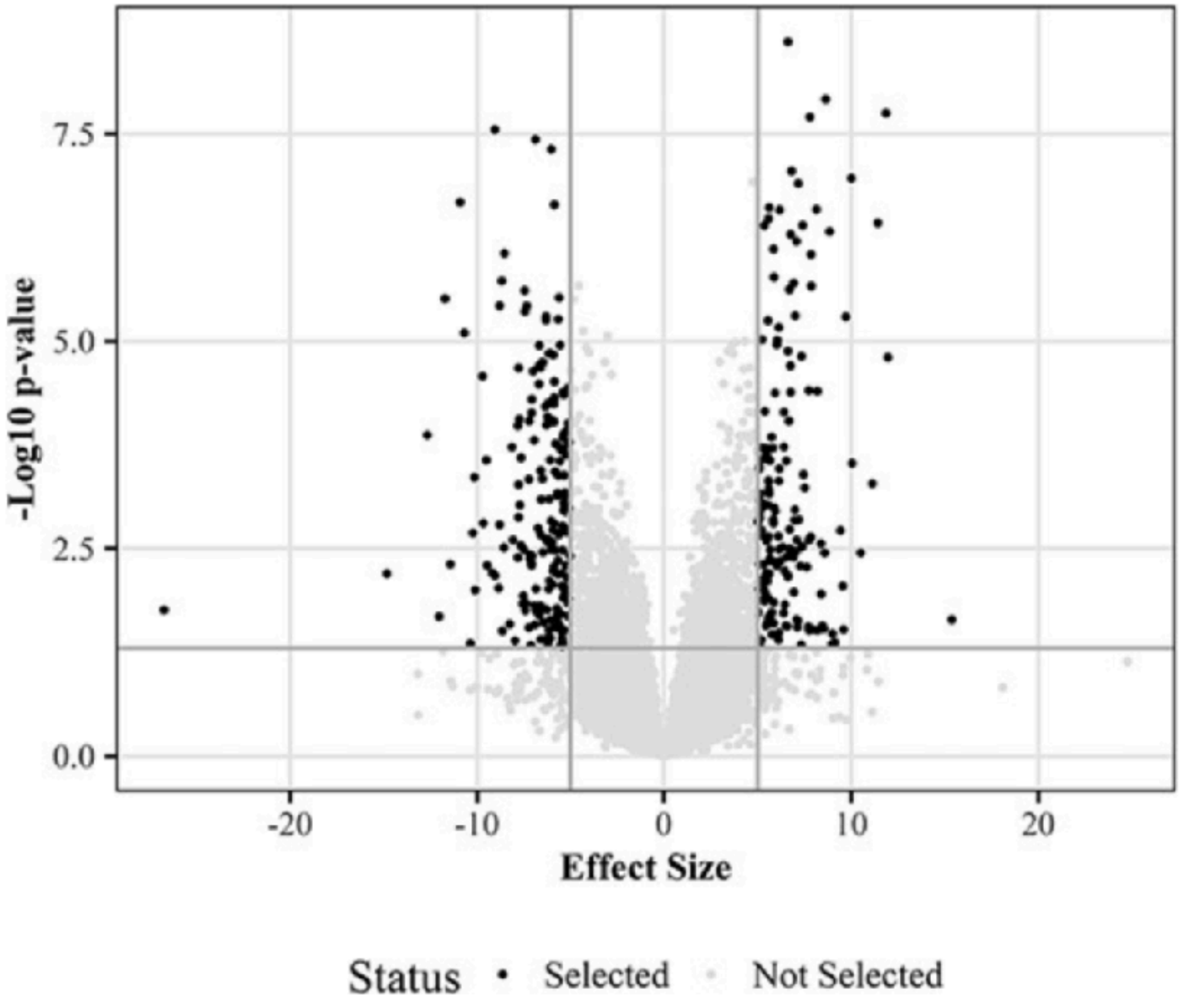
- Blanchard and Roquain (2008) formulated two quite general sufficient conditions, self-consistency and dependence control, under which, if both hold, FDR control is guaranteed.
- However, self-consistency is sufficient but not necessary for FDR control: Solari and Goeman (2017) show uniform improvements of self-consistent methods by a non-self-consistent method.

FDR history

- Blanchard and Roquain (2008) formulated two quite general sufficient conditions, self-consistency and dependence control, under which, if both hold, FDR control is guaranteed.
- However, self-consistency is sufficient but not necessary for FDR control: Solari and Goeman (2017) show uniform improvements of self-consistent methods by a non-self-consistent method.
- No post-hoc user flexibility. Why is that problematic? Example on the next slide.

FDR control and volcano plots: no guarantees!

Ebrahimpoor & Goeman (2021)



The e-closure principle

How to design a multiple testing procedure

- e-Closure
  A general recipe for making multiple testing methods

- Building blocks
  Intersection hypotheses and e-values

- Contributions

  - Recovers the Closure Principle for FWER
  - Extends to FDR
  - Uniformly improves a.o. eBH and BY
  - Introduces unprecedented flexibility in multiple testing

The e-variable

- **Definition:** e-variable

  An e-variable E for $\mathscr{P}$ is a non-negative random variable satisfying $\mathbb{E}_P[E] \leq 1$ for all $P \in \mathscr{P}$.

- The value taken by the e-variable after observing the data is called the e-value. However, often, as also happens with the infamous p-value (p-variable), the random variable E itself is also often called e-value.

Tests and the type I error guarantee

- **Definition:** binary test

  A binary test ϕ is a $\{0,1\}$-valued random variable. The type-I error of a test ϕ for P is $\mathbb{E}_P[\phi]$. A test has level $\alpha \in [0,1]$ for $\mathscr{P}$ if its type-I error is at most α for every $P \in \mathscr{P}$.

- **Markov's inequality for e-variables**

  Let E be an e-variable for $\mathscr{P}$. We have $P(E \geq 1/\alpha) \leq \alpha$ for all $P \in \mathscr{P}$ and $\alpha \in (0,1]$. Hence, $\mathbf{1}_{\{E \geq 1/\alpha\}}$ is a binary test of level α.

Intersection hypotheses

Intersection hypothesis

For $S \subseteq [m], H_S = \bigcap_{i \in S} H_i$, which is true iff all $H_i, i \in S$ true

The e-collection

$E = (e_S)_{S \subseteq [m]}$: local e-values such that E$(e_N) \leq 1$

Sufficient

Each e_S is an e-value for $H_S, S \subseteq [m]$

The e-Closure Principle

- The e-Closed Procedure

$$\mathscr{R}_\alpha(E) = \left\{R \subseteq [m] : \alpha e_S \geq \mathrm{f}_S(R) \quad \forall S \subseteq [m]\right\}$$

- The e-Closure Principle

R controls $\mathrm{E}(\mathrm{f}_N(R)) \leq \alpha$ iff $R \in \mathscr{R}_\alpha(E)$ for e-collection E

- Simultaneous control

$\mathrm{E}(\mathrm{f}_N(R)) \leq \alpha$ simultaneously over $R \in \mathscr{R}_\alpha(E)$: $\quad \mathrm{E}\left(\max_{R \in \mathscr{R}_\alpha(E)} \mathrm{f}_N(R)\right) \leq \alpha$

Post hoc error rate

- All error rates

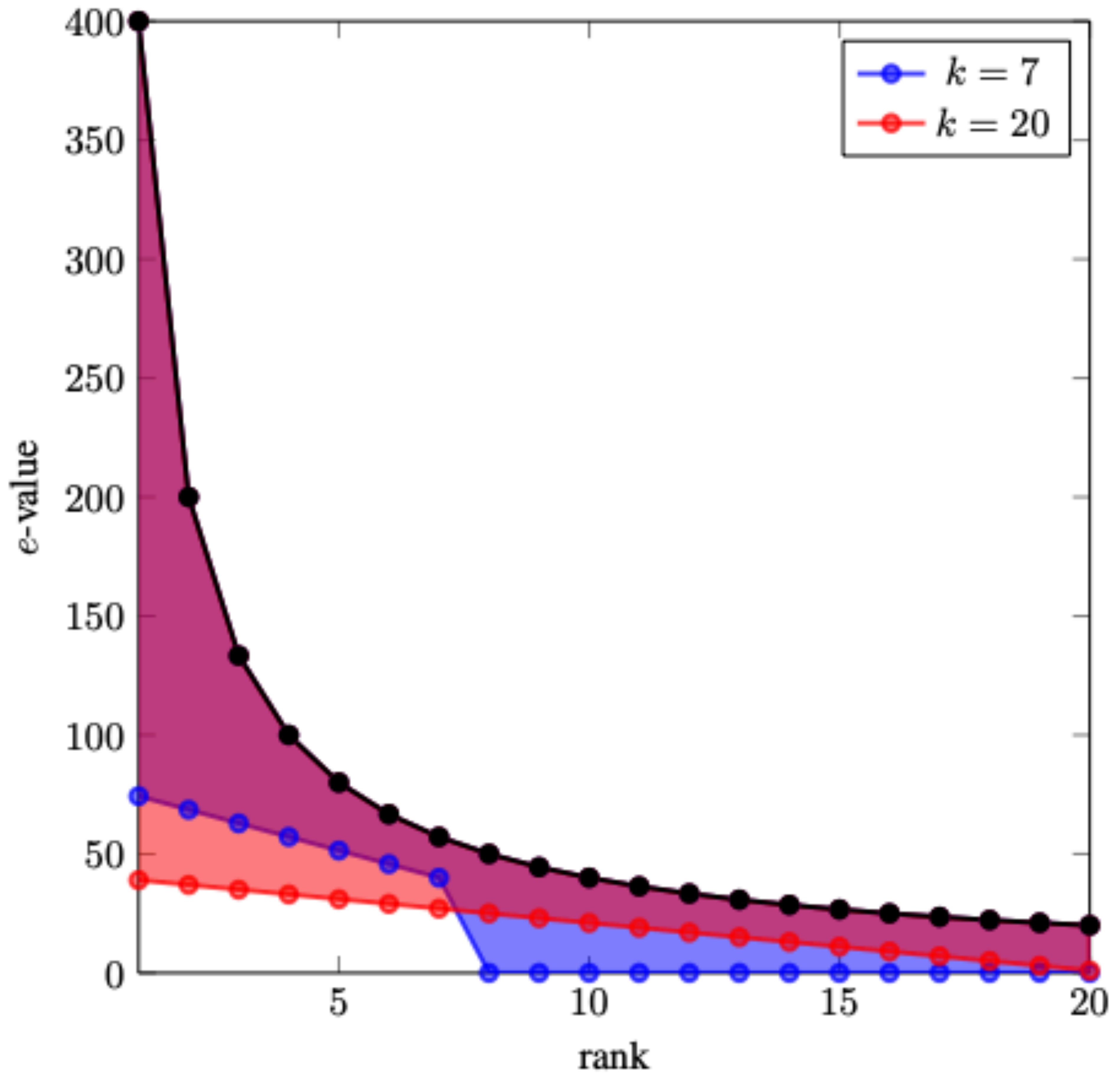
$$\mathscr{F} = \{\text{all functions } \mathrm{f}_N(R)\}$$

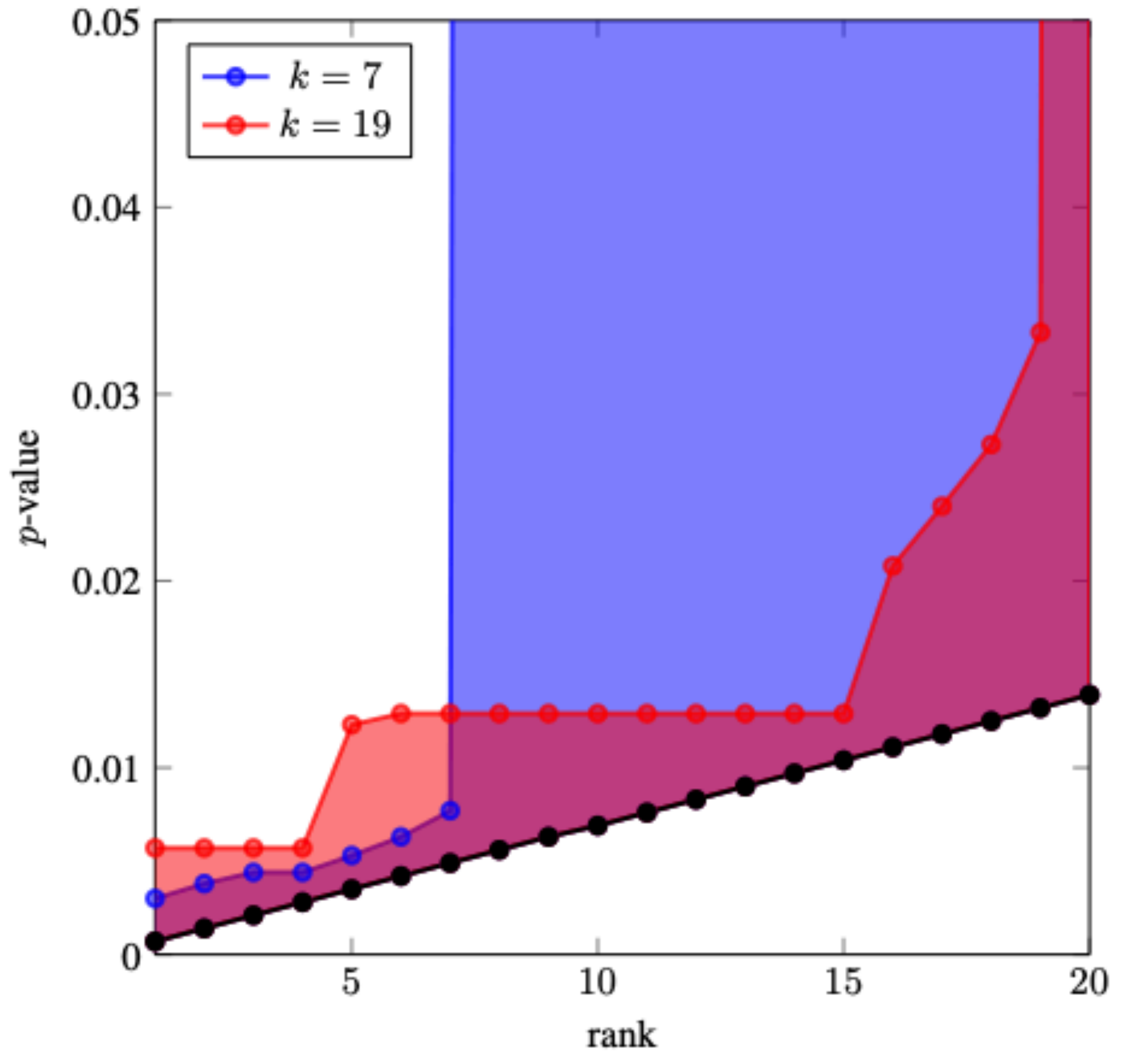
- Simultaneous (= post hoc) choice of error

$$\mathrm{E}\left(\sup_{\mathrm{f} \in \mathscr{F}} \max_{R \in \mathscr{R}_\alpha^{\mathrm{f}}(E)} \mathrm{f}_N(R) \right) \leq \alpha$$

- So: possible to switch from FWER to FDR if not much signal present

Improving existing procedures: eBH and BY





BY vs closed BY in standard data sets

Dataset	m	BY / $\overline{BY}$ rejections		source
		$\alpha = 5\%$	$\alpha = 10\%$	
APSAC	15	3 / 3	3 / 5	BH '95
NAEP	34	6 / 8	8 / 11	BH '00
PADJUST	50	12 / 15	17 / 20	`p.adjust`
PVALUES	4289	129 / 145	225 / 275	`fdrtool`
VANDEVIJVER	4919	614 / 677	779 / 866	Goeman Solari '14
GOLUB	7128	617 / 648	743 / 799	Efron Hastie '16

More properties: post hoc α

- Choose rejected set post hoc
- Choose error loss post hoc
- One step further: choose α post hoc (Koning 2023)
- Requires: e-value does not depend on α. Then we have:

$$\mathrm{E}\left(\sup_{\alpha \in (0,1)} \sup_{\mathrm{f} \in \mathscr{F}} \max_{R \in \mathscr{R}_\alpha^{\mathrm{f}}(E)} \frac{\mathrm{f}_N(R)}{\alpha} \right) \leq 1$$

More properties: restricted combinations

- Logically related hypotheses, for example pairwise combinations

$$H_{1=2} : \mu_1 = \mu_2; \quad H_{1=3} : \mu_1 = \mu_3; \quad H_{2=3} : \mu_2 = \mu_3$$

- Logical relationships = gain in power

  Up to now only known for FWER, unknown for FDR

Summary: e-Closure

- General Necessary and Sufficient Principle: unites all multiple testing methods
- Simplifies multiple testing: Choose how to summarise evidence against H_S; rest is computation
- Flexibility: Simultaneous over rejected sets, error rates, α
- Power: Uniformly improves known methods

A general recipe for making multiple testing methods