Group-level analysis

Simon Vandekar

Group-level analysis

- Estimation – in parallel across all voxels "Mass-univariate"
- Inference – uses spatial information and correlation among measurements

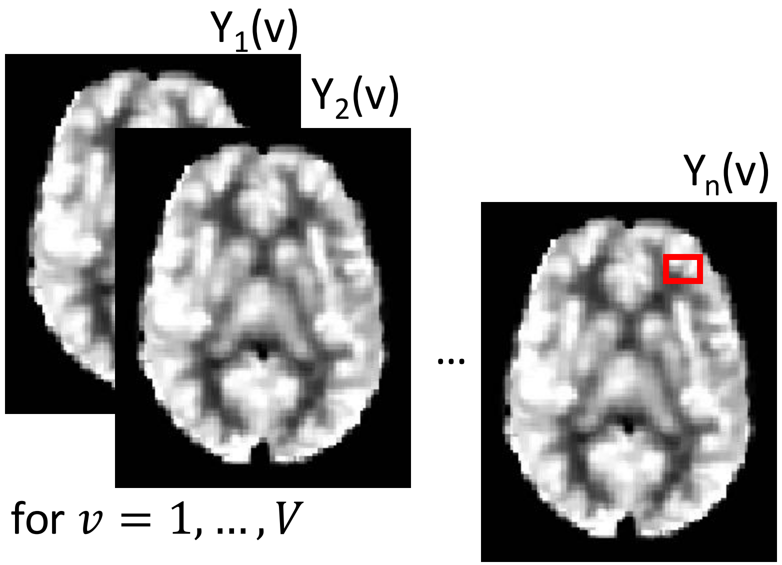
Group-level estimation

Two important features to consider in estimation:

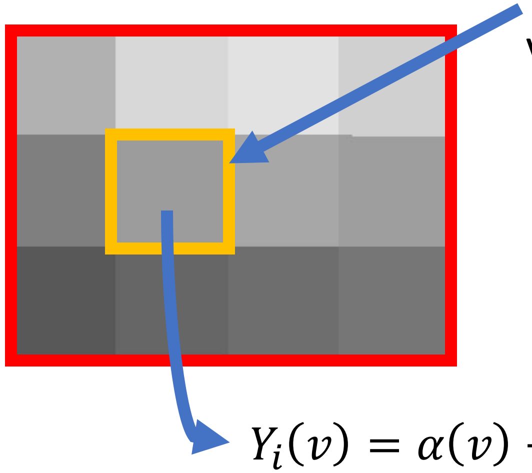
- **Heteroskedasticity** – variance differs between participants
- **Correlation** – measurements from the same individual are correlated
- Accounting for these improves estimation efficiency
  - Also needed to obtain unbiased effect size estimates and test statistics

Mass-univariate approach

- Mass-univariate estimation
  - Widely used to localize regions of association (≈18,800 studies)
  - Perform inference at each location in the image



for $v = 1, \ldots, V$



Voxel
volumetric pixel

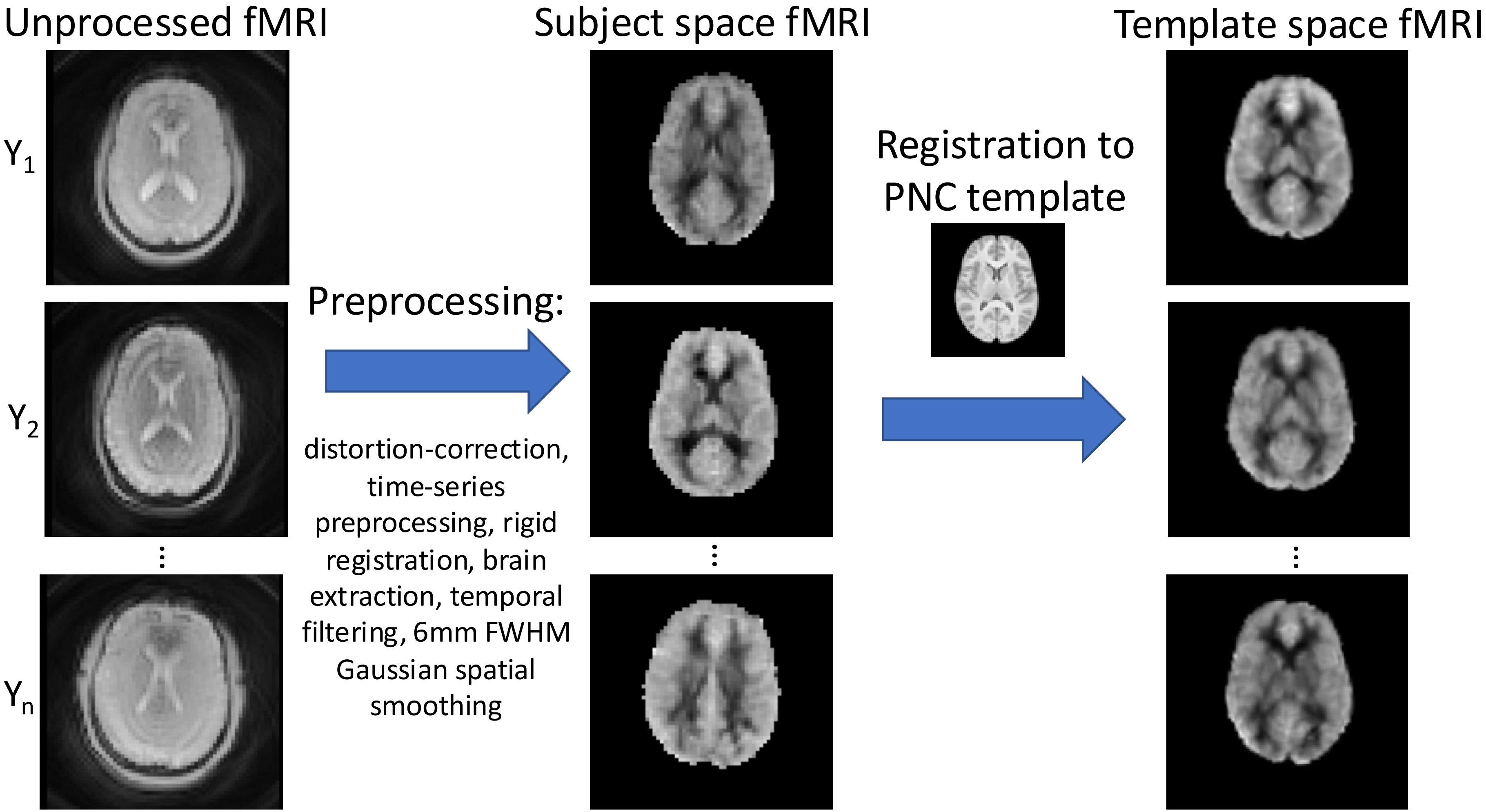
$Y_i(v) = \alpha(v) + d_i' \times \beta(v) + \epsilon_i(v),$

$\boldsymbol{H_0(v)}$**:** $\boldsymbol{\beta(v) = 0}$

Reject $\boldsymbol{H_0(v)}$ **where** $\boldsymbol{p(v) < \alpha}$

Friston et al., *Neuroimage*, 1994
Nichols, http://blogs.warwick.ac.uk/nichols/entry/bibliometrics_of_cluster/, 2016

fMRI preprocessing steps



Unprocessed fMRI

Subject space fMRI

Template space fMRI

Y_1

Y_2

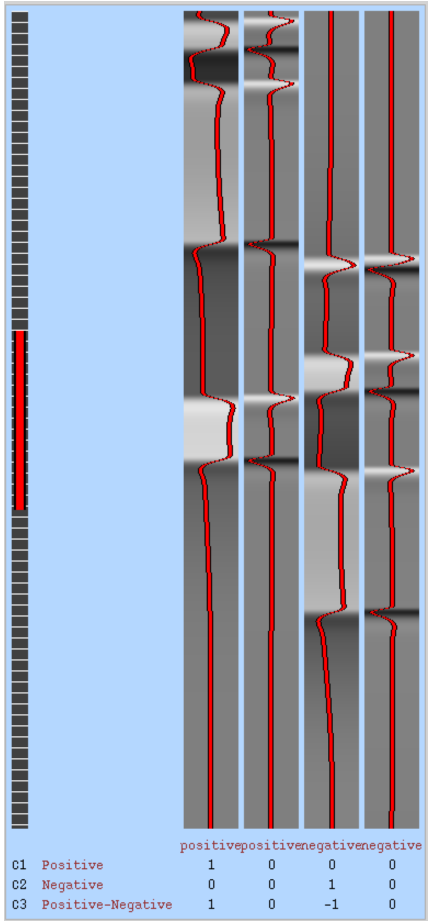
Y_n

Preprocessing:

distortion-correction, time-series preprocessing, rigid registration, brain extraction, temporal filtering, 6mm FWHM Gaussian spatial smoothing

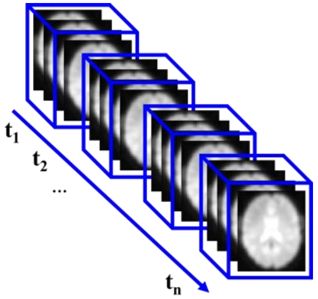
Registration to PNC template

Despicable Me first-level analysis

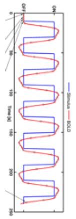


		positive	positive	negative	negative
C1	Positive	1	0	0	0
C2	Negative	0	0	1	0
C3	Positive-Negative	1	0	-1	0

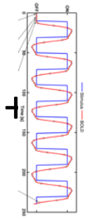
Subject-level time series model:



$=$



$\times \; \beta_{i,positive}(v) +$



$\times \; \beta_{i,negative}(v)$

$+ \;\; \epsilon_i(v,t)$

Group-level analysis:

$$Y_i(v) := \; \hat{\beta}_{i,positive}(v) \; - \hat{\beta}_{i,negative}(v)$$

Outcome is increased activation in 2-back condition

Heteroskedasticity in group-level analysis

- *Despicable Me* data
  - $Y_i(v) := \hat{\beta}_{i,positive}(v) - \hat{\beta}_{i,negative}(v)$
  - It’s possible that $\mathrm{Var}\{Y_i(v) | X_i\} = g(X_i)$, which implies unequal variances, also called **heteroskedasticity/nonexchangeability**
    - E.g. motion causes increased noise
  - Even worse: $\mathrm{Cov}\{Y_i(v), Y_i(w) | X_i\} = g(X_i)$ implies nonexchangeability
- Other examples:
  - Cortical thickness, functional connectivity
  - Any subject-level estimates

Multilevel models in neuroimaging

- A group-level analysis might use a multivariate model

$$Y_i(v) \coloneqq [\hat{\beta}_{i,positive}(v), \hat{\beta}_{i,negative}(v)]$$

- Multiple measurements per participant – requires accounting for **correlation within participant**
- There could be many conditions included in the analysis
  - Some tasks have multiple runs (e.g. they scan over three 8 min sessions). These are included as repeated measurements
- Random effects analysis

Potential data structures

- And appropriate methods

		Variance Type	
		Homosked.	**Heterosked.**
Number of Measurements	**One**	Classical statistical methods	Variance weighted GLM; Robust standard errors (sPBJ)
	Multiple	Random effects models	FLAME; GEE (sPBJ, SwE)

Options for group-level analysis

- Classical approach – using linear model at group-level
- Random effects analysis – with multiple measurements on the same participants
- FLAME – random effects **with** heteroskedasticity
- Semiparametric bootstrap joint inference (sPBJ) – research from my group. Correlated data **with** robust standard errors
    - Also implemented in SwE toolbox in SPM
- Permutation testing

Classic approach – ignores first-level model

- Assume

$$Y_i = X_i\beta + \epsilon_i$$

- $\epsilon_i \sim N(0, \sigma^2)$
- If σ_i^2 does not depend on X_i, this model is still valid – why?
- Normality assumption is not strictly necessary – why?
- What is the estimator for β and what is the variance of $\hat{\beta}$?

Robust (heteroskedasticity-consistent) standard errors approach

- Assume

$$Y_i = X_i\beta + \epsilon_i$$

- $\epsilon_i \sim N\left(0, \sigma_i^2\right)$, participants have different variances
- If σ_i^2 depends on X_i, possibly in an unknown way
- Normality assumption is not strictly necessary – why?
- What is an estimator for β and what is the variance of the estimator?
  - The least squares estimator is ok

Multi-level (random effects) model

- A random effects model is used to account for correlation among repeated measurements
  - First-level time-series model, matrix equation:
$$Y_k = X_k \beta_k + \epsilon_k$$
  - $\mathrm{Cov}(\epsilon_k) = V$
  - Group-level (second-level) model:
$$\beta_k^T = X_{Gk} \beta_G + \eta_k$$
  - β_k - can think of this as a column vector or a scalar parameter
  - X_{Gk} is a matrix where each row is the same set of participant-level covariates (e.g. participant *k*'s age, sex, and task performance.
  - $\mathrm{Cov}([\eta_1, \dots, \eta_n]) = V_G$ - for n independent subjects, η_k are assumed IID

Combining the two equations

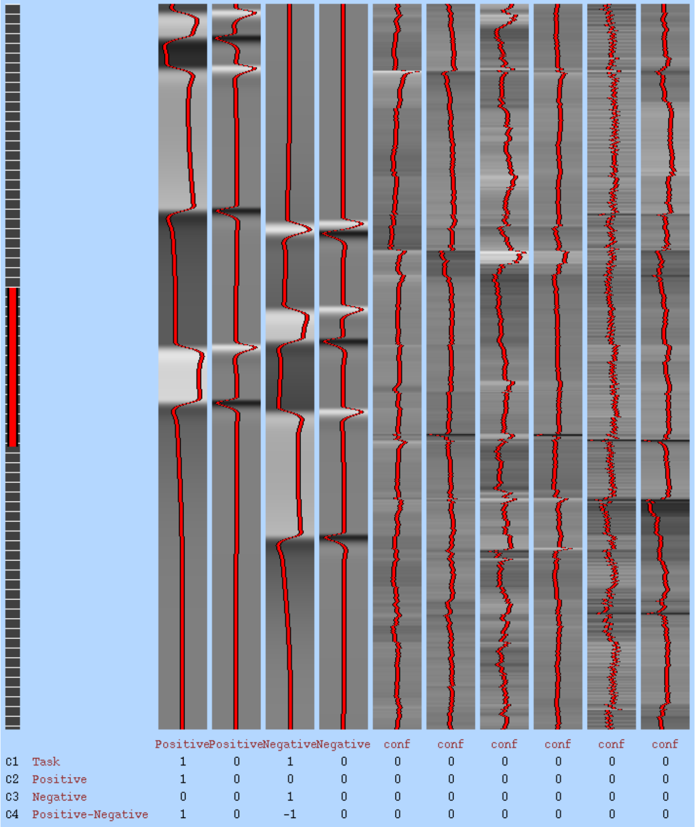
- Combining the two equations gives a typical formula for a random effects model

$$Y_k = X_k X_{Gk}^T \beta_G + X_k \eta_k + \epsilon_k$$

- $\epsilon_k \sim N(0, V)$ – these are the error variances (time series dependence)
- $\eta_k \sim N(0, V_\eta)$ – these are the random effect term with variance
- ϵ_k and η_k are independent

Examples of design matrices in my notation

$X_k =$



		Positive	Positive	Negative	Negative	conf	conf	conf	conf	conf	conf
C1	Task	1	0	1	0	0	0	0	0	0	0
C2	Positive	1	0	0	0	0	0	0	0	0	0
C3	Negative	0	0	1	0	0	0	0	0	0	0
C4	Positive-Negative	1	0	-1	0	0	0	0	0	0	0

Dims: T x 10

$$\beta_k = \begin{bmatrix} \beta_{1k} \\ \beta_{2k} \\ \cdots \\ \beta_{10,k} \end{bmatrix}$$

Dims: 1 x 10

$$X_{Gk} = \begin{bmatrix} \text{age}_k & \cdots & \text{sex}_k \\ \vdots & \ddots & \vdots \\ \text{age}_k & \cdots & \text{sex}_k \end{bmatrix}$$

Dims: 1 x p

$$\beta_G = [\beta_{G1} \quad \cdots \quad \beta_{G,10}]$$

Dims: p x 10

Notation

- I found the notation in the FLAME paper a little confusing
- My notation differs by using subject level notation and treating β_k as a vector
- In their paper they seem to collapse β_k as a single vector across all subjects (instead of a matrix), and so X_{Gk} would need to have a Kronecker product kind of shape

FLAME

- FLAME is FSL's approach that uses Bayesian estimation
  - It is a multilevel random effects model with unequal variances
  - First-level time-series model, matrix equation:

$$Y_k = X_k \beta_k + \epsilon_k$$

  - $\mathrm{Cov}(\epsilon_k) = V_k$. Covariance being indexed by k implies what?
  - Group-level (second-level) model:

$$\beta_k^T = X_G \beta_G + \eta_k$$

  - β_k - can think of this as a column vector or a scalar parameter
  - $\mathrm{Cov}([\eta_1, \ldots, \eta_n]) = V_G$ - for n independent subjects, η_k are assumed IID.
- Notation from the reading: Beckmann paper ("flame.pdf")
- What is the difference from the random effects model?

FLAME estimation

- Estimation is complex –occurs in multiple levels, which is convenient (Section 2.C of Beckmann paper)
- More complex due to estimation of the variance terms

Semiparametric bootstrap joint (sPBJ) inference

- This is what I called the “robust approach” on a previous slide and is essentially an imaging version of GEE
- Can accommodate heteroskedasticity and longitudinal models
- Uses “sandwich” covariance estimates for asymptotically unbiased spatial covariance function estimator
- Guillaume et al., 2014 and Vandekar et al., 2019 (see homework/reading)

Estimating equations approach

- Target parameters can be written as solutions to estimating equations

$$\hat{\beta}(v) = \mathrm{argmax}_{\beta(v)} - \sum_i^n W_i(v)\left(Y_i(v) - X_i^T\beta(v)\right)^2 = \Psi(Y(v), \beta(v))$$

- Differentiating:

$$\frac{\partial\Psi}{\partial\beta(v)} = \sum_i^n W_i(v)\{Y_i(v) - X_i^T\beta(v)\}X_i$$

$$\frac{\partial^2\Psi}{\partial\beta^2(v)} = -\sum_i^n W_i(v)X_iX_i^T$$

Taylor expansion of estimating equation

- $\frac{\partial \Psi}{\partial \beta(v)}\left\{\hat{\beta}(v)\right\} = 0$
- Taylor expansion:

$$0 = \sum_{i}^{n} W_i(v)\left\{Y_i(v) - X_i^T \beta(v)\right\} X_i^T - \sum_{i}^{n} W_i(v) X_i X_i^T \left\{\hat{\beta}(v) - \beta(v)\right\}$$

- Which implies

$$\sqrt{n}\left\{\hat{\beta}(v) - \beta(v)\right\} = \left\{n^{-1} \sum_{i}^{n} W_i(v) X_i X_i^T\right\}^{-1} n^{-1/2} \sum_{i}^{n} W_i(v)\left\{Y_i(v) - X_i^T \beta(v)\right\} X_i^T$$

Asymptotic joint distribution

- $n^{-1} \sum_i^n W_i(v) X_i X_i^T \to_P A(v)$
- $n^{-1/2} \sum_i^n W_i(v)\{Y_i(v) - X_i^T \beta(v)\} X_i \to_D$

$$N(0, B(v,v)),$$

where $B(v,v) = \mathbb{E}\{W_i(v)^2 \{Y_i(v) - X_i \beta(v)\}^2 X_i^T X_i\}$

- For imaging we are particularly interested in the covariance terms

$$Cov\left\{\sqrt{n}\left(\hat{\beta}(v) - \beta(v)\right), \sqrt{n}\left(\hat{\beta}(w) - \beta(w)\right)\right\} = A(v)^{-1} B(v,w) A(w)^{-1}$$

where

$$B(v,w) = \mathbb{E}\{W_i(v) W_i(w)\{Y_i(v) - X_i^T \beta(v)\}\{Y_i(w) - X_i^T \beta(w)\} X_i X_i^T\}$$

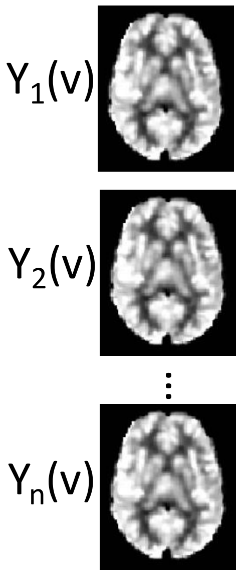
Joint distribution of test statistic image $T_{m_1}(v)$

- $T_{m_1}(v) = n \times \left(\hat{\beta}(v) - \beta(v)\right)^T A(v) B(v,v)^{-1} A(v) \left(\hat{\beta}(v) - \beta(v)\right)$
- So $T_{m_1}(v) = Z(v)^T Z(v)$, where

$$Z(v) = \sqrt{n} \times B(v,v)^{-1/2} A(v) \left(\hat{\beta}(v) - \beta(v)\right) \sim N(0, I_{m_1})$$

Choosing subject weights for the sPBJ

- Optimal weights are the inverse of the observation level variance
- Optimal weight: $W_i(v) \approx \Sigma_{\mathrm{i}}(v, v)^{-1}$
- Good weight: W_i is inverse of in-scanner motion
- No weights: still good asymptotically



$\mathrm{Cov}\{Y_1(v), Y_1(w) | \, X_1\} = \Sigma_1(v, w)$

$\mathrm{Cov}\{Y_2(v), Y_2(w) | \, X_2\} = \Sigma_2(v, w)$

$\mathrm{Cov}\{Y_n(v), Y_n(w) | \, X_n\} = \Sigma_{\mathrm{n}}(v, w)$

Some computational practicalities

- Estimation in neuroimaging often occurs in parallel across all locations
- Statistical Inference accounts for correlation among voxels
- sPBJ uses bootstrapping to estimate the joint distribution. More on this in Guillaume et al., 2014 and Vandekar et al., 2019 (readings)

Statistical Inference

This is where the spatial aspect of the data in incorporated

Options for statistical inference

- Types of inference
  - Voxel-wise inference
  - Cluster extent inference (spatial extent inference)
  - Threshold-free cluster enhancement
- Methods of inference
  - Gaussian random field theory
  - Permutation testing
  - Bootstrapping

Notation

$\boldsymbol{T_{m_1}(v)}$ - Test statistic image for test of $H_0(v)\colon\ \beta(v) = 0$

Can be T, Z, F, or Chi-square

$\boldsymbol{m_1}$ - dimension of $\beta(v)$ for all v

$\boldsymbol{p(v)}$ - p-value image computed from $T_{m_1}(v)$

Hypothesis Testing (Inference)

We define particular hypotheses of interest such as:

$$H_0(v): \beta(v) = 0$$

(no effect of diagnosis on brain activation at location v)

or (more specifically)

No effect of diagnosis on assumed brain response profile of brain activation to positive versus negative emotions during *Despicable Me* clip at location v.

Hypothesis Testing

However, hypotheses like:

$$H_0(v): \beta(v) = 0$$

Are more complex than they look.

For example, are we asking:

$$H_0(v): \beta(v) = 0 \text{ for all } v = 1, \ldots, V$$

Or

$$H_0(v_0): \beta(v_0) = 0 \text{ for a specific } v_0 = 1, \ldots, V?$$

And what precisely are the alternatives we are interested in?

Hypothesis Testing: Voxel-wise inference

Suppose we wish to test $H_{0v}: \beta(v) = 0$ for each $v = 0, \ldots, V$. This is an exceedingly common set of hypotheses to test.

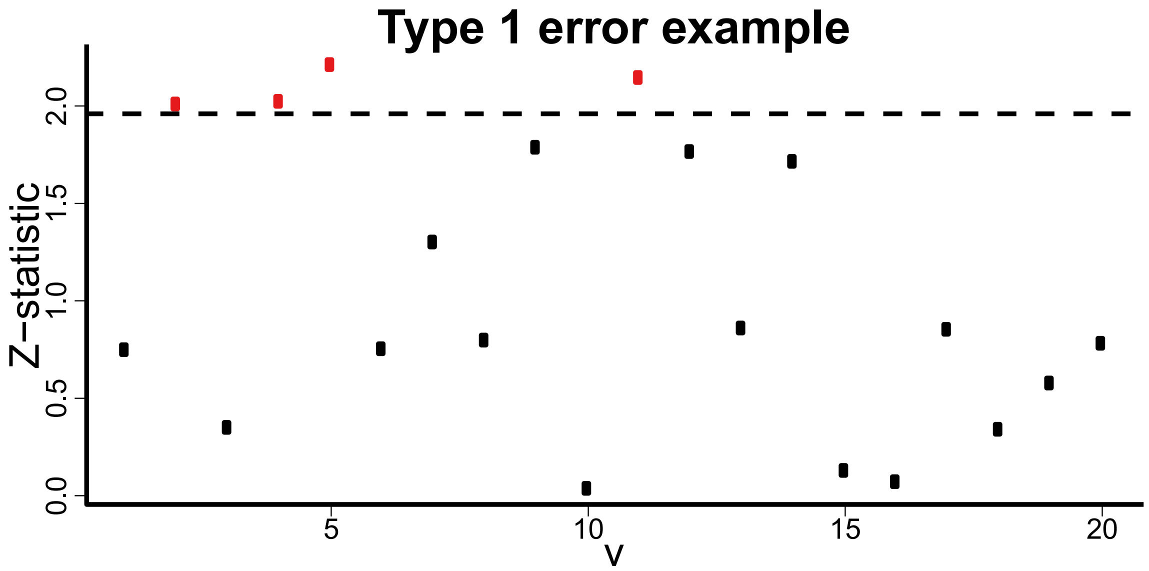
Usual process: define the type I error rate (α = 5%) and find a test statistic. If we have a good test statistic, then (under the null) approximately 5% of the time we will reject $H_{0v}: \beta(v) = 0$.

This is called **voxel-wise** inference – we are testing every voxel.

False positive rate in high dimensions

dataset-to-dataset variability

- For medical imaging data $V \approx 100{,}000$
- Performing the test of $H_0(v)$ for $v = 1, \dots, V$ leads to increased false positive rate



- $P(|Z(v)| > 1.96) = 0.05$
- $P\left(\cup_{v \in M_0} |Z(v)| > 1.96\right) = 1 - 0.95^{20} = 0.64$

Hypothesis Testing Errors

For each hypothesis test:

	$H_0(v)$ True	$H_0(v)$ False
Do Not Reject $H_0(v)$	Correct (probability 1-α)	Type II Error (probability β)
Reject $H_0(v)$	Type I Error (probability α)	Correct (probability 1-β)

Across tests:

	$\#H_0(v)$ True	$\#H_0(v)$ False	Total #
Not Rejected $H_0(v)$	U (True negative)	T (False negative)	$V - R$
Rejected $H_0(v)$	W (False positive)	S (True positive)	R
	V_0	$V - V_0$	V

Hypothesis Testing Error Control

How do we control these errors? There are two common approaches:

	$\#H_0(v)$ True	$\#H_0(v)$ False	Total #
Not Rejected $H_0(v)$	U	T	$V - R$
Rejected $H_0(v)$	W	S	R
	V_0	$V - V_0$	V

Family-wise Error Rate (FWER)

$$FWER = P(W \geq 1)$$

False Discovery Rate (FDR)

$$FDR = E\left(\frac{W}{R} \mid R > 0\right) P(R > 0)$$

Hypothesis Testing Error Control

How do we control these errors? There are two common approaches:

Family-wise Error Rate (FWER)

More conservative.

False Discovery Rate (FDR)

Less conservative.

Controlling the FWER - Bonferroni

The Bonferroni correction is the most common FWER correction.

It is notoriously conservative (high type II error rate).

Applying the Bonferroni correction is very simple: for each test H_{0v}, estimate a p-value p_v. Then, reject any hypothesis H_{0v} where

$$p_v \leq \frac{\alpha}{V}.$$

Controlling the FWER - Bonferroni

Reject any hypothesis H_{0v} where

$$p_v \leq \frac{\alpha}{V}.$$

But why? First, we need a lemma:

Lemma: Under $H_0 \colon T_{m_1} \sim F_0$, we have that

$$p_t = P(\, T > t \mid T \sim F_0 \,) \sim U(0,1)$$

Proof: $P(p_t < u) = \mathrm{P}\{\, \mathrm{P}(\, \mathrm{T} > t \mid T \sim F_0 \,) < u \mid t \sim F_0 \,\} = \mathrm{P}(1 - \mathrm{F_0(t)} < u \mid t \sim F_0)$

$$= P(\, t > F_0^{-1}(1-u) \mid t \sim F_0 \,) = 1 - F_0\{F_0^{-1}(1-u)\} = u.$$

From lecture notes by Mark van de Wiel (VU/VUmc)

Controlling the FWER - Bonferroni

Then, using Boole’s inequality, and noting by the lemma that under H_{0v} we have $p_v \sim U(0,1)$,

$$FWER \; = \; P\left(\cup_{v=1}^{V_0} \left\{p_v \le \frac{\alpha}{V}\right\}\right) \le \sum_{v=1}^{V_0} P\left(p_v \le \frac{\alpha}{V}\right) \le \frac{V_0 \alpha}{V} \le \alpha.$$

Note that this argument did not depend on how many hypotheses were true, nor did it depend on the distribution of the test statistics; in particular, it did not depend on any correlation between the test statistics.

Controlling the FWER – Boole's

Another relatively common technique for controlling the FWER is referred to as Holm's procedure.

- Order the p-values such that $p_{(1)} \leq \cdots \leq p_{(V)}$.
- Adjust the p-values:

$$p_j^* = \min\{(V - j + 1)p_j, 1\}$$

Which is equivalent to finding the largest k such that

$$p_{(k)} > \frac{\alpha}{V - k + 1}$$

And rejecting $H_{0(1)} \leq \cdots \leq H_{0(k-1)}$ but not $H_{0(k)} \leq \cdots \leq H_{0(V)}$.

Note that the Holm procedure also controls the FWER under arbitrary dependence, invoking Boole's inequality.

Controlling the FDR

- Controlling the FDR results in more liberal inference, and higher statistical power, at the cost of FWER.
- There are several common methods for controlling the FDR, but the most common is called the Benjamini-Hochberg procedure.
  - Controls the FDR under independence.
  - Controls the FDR under positive dependence.
- A modification of this procedure is also available for more general dependencies and is known at the Benjamini-Yekutieli procedure.
  - It is known to provide much less power, and so is less common.

Controlling the FDR - BH

To control the FDR, Benjamini and Hochberg (1995) suggested that we can control the FDR at level δ by:

- Ordering the p-values such that $p_{(1)} \leq \cdots \leq p_{(V)}$.
- Finding the largest k such that

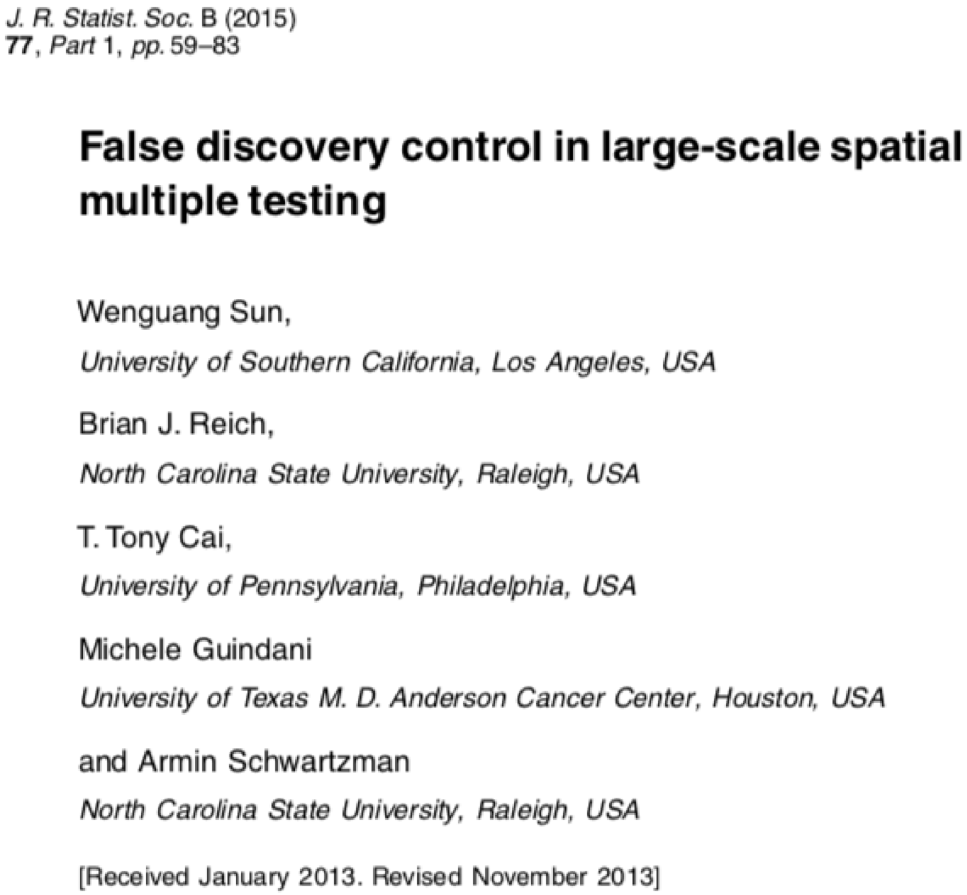
$$p_{(k)} \leq \frac{k}{V} \delta$$

And rejecting $H_{0(1)} \leq \cdots \leq H_{0(k)}$ but not $H_{0(k)} \leq \cdots \leq H_{0(V)}$.

These methods ignore joint distribution. What else can we do?

- Another (less common) approach was pioneered by Westfall and Young (1993) and is based on permutation:
- We permute the rows of the design matrix (X_i) repeatedly and generate the distribution of the maximum test statistic (or min p-value).
- Then we compare this empirical distribution of the maximum test statistic with the observed distribution of the test statistics.

Many spatial FDR methods



J. R. Statist. Soc. B (2015)
77, *Part* 1, *pp.* 59–83

False discovery control in large-scale spatial multiple testing

Wenguang Sun,
University of Southern California, Los Angeles, USA

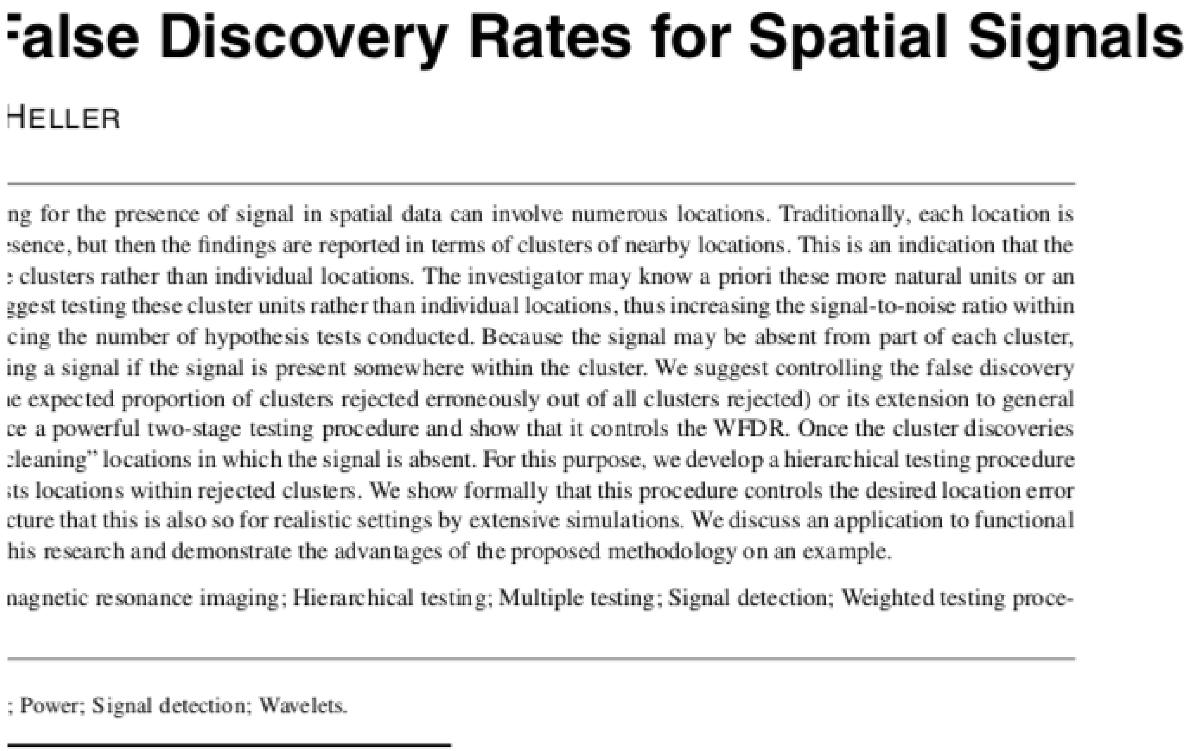
Brian J. Reich,
North Carolina State University, Raleigh, USA

T. Tony Cai,
University of Pennsylvania, Philadelphia, USA

Michele Guindani
University of Texas M. D. Anderson Cancer Center, Houston, USA

and Armin Schwartzman
North Carolina State University, Raleigh, USA

[Received January 2013. Revised November 2013]



alse Discovery Rates for Spatial Signals

HELLER

ng for the presence of signal in spatial data can involve numerous locations. Traditionally, each location is
sence, but then the findings are reported in terms of clusters of nearby locations. This is an indication that the
clusters rather than individual locations. The investigator may know a priori these more natural units or an
ggest testing these cluster units rather than individual locations, thus increasing the signal-to-noise ratio within
cing the number of hypothesis tests conducted. Because the signal may be absent from part of each cluster,
ing a signal if the signal is present somewhere within the cluster. We suggest controlling the false discovery
e expected proportion of clusters rejected erroneously out of all clusters rejected) or its extension to general
ce a powerful two-stage testing procedure and show that it controls the WFDR. Once the cluster discoveries
leaning" locations in which the signal is absent. For this purpose, we develop a hierarchical testing procedure
ts locations within rejected clusters. We show formally that this procedure controls the desired location error
cture that this is also so for realistic settings by extensive simulations. We discuss an application to functional
his research and demonstrate the advantages of the proposed methodology on an example.

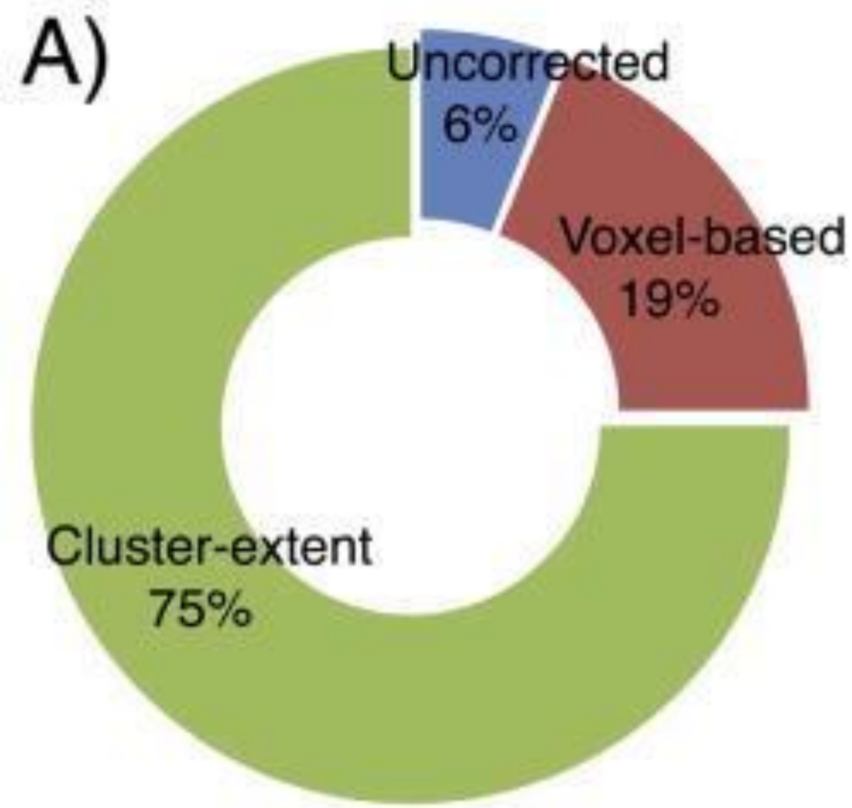
nagnetic resonance imaging; Hierarchical testing; Multiple testing; Signal detection; Weighted testing proce-

; Power; Signal detection; Wavelets.

Benjamini & Heller, *JASA*, 2007
Shen et al., *JASA*, 2002
Sun et al., *JRSSB*, 2015

Cluster extent inference most widely used approach in medical imaging

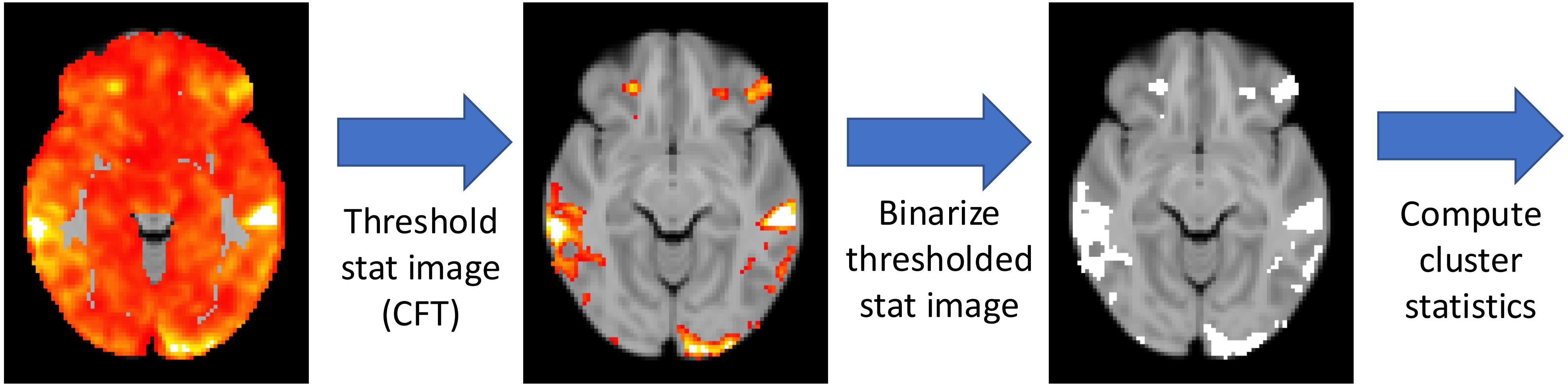
- Multiple comparisons correction
  - FWER
  - FDR
- Voxel-wise correction
- Cluster extent inference
- Other approaches
  - TFCE
  - Cluster-mass



Woo, Krishnan, & Wager, *Neuroimage*, 2014

Cluster extent inference procedure

$\boldsymbol{T_{m_1}(v)}$ - Chi-squared statistic image for test of $H_0(v)$: $\beta(v) = 0$
$\boldsymbol{p(v)}$ - p-value image computed from $T_{m_1}(v)$
m_1 - dimension of $\beta(v)$



Threshold stat image (CFT)

Binarize thresholded stat image

Compute cluster statistics

Cluster	Num voxels	Vol (mm³)
1	157	628
2	67	268
3	57	228
⋮	⋮	⋮
16	1	4

Statistic image, $T_{m_1}(v)$, for test of: $\boldsymbol{H_0(v)}$: $\boldsymbol{\beta(v) = 0}$

- Image viewed as a "random field"
- Choose **cluster forming threshold (CFT)** $p(v) < 0.001$ (i.e. $T_1(v) > 10.82$)

$\boldsymbol{T_{m_1}(v)}$ - Chi-squared statistic image for test of $H_0(v)$: $\beta(v) = 0$
$\boldsymbol{p(v)}$ - p-value image computed from $T_{m_1}(v)$
m_1 - dimension of $\beta(v)$

Cluster extent inference p-values

New set of hypotheses that are cluster specific:

- Let $E(c)$ denote the cluster extents for $c = 1, \dots, C$, where c indexes clusters in decreasing size
- $H_0(c)$: $E(c) =_d E_0(c)$, where $E_0(c)$ assumes $H_0(v)$: $\beta(v) = 0 \; \forall \; v$
- Compute cluster adjusted p-value

$$p(c) = P(E_0(1) > e(c)),$$

- where $e(c)$ is the observed cluster extent for cluster c
- **Computing $\boldsymbol{p(c)}$ requires an estimate of the joint distribution of $\boldsymbol{T_{m_1}(v)}$**

Spatial extent inference interpretation

- $p(c)$ is the probability of observing a cluster size at least as large as cluster c if the image is not associated with the covariate

Three approaches to CEI

1. Gaussian random field
2. Permutation method – randomise from FSL[1]
3. Parametric bootstrap joint (PBJ) testing procedure[2]

[1] Winkler et al., *Neuroimage,* 2014

[2] Vandekar, *Arxiv*, 2018

Gaussian Random Field Theory

- This is a fancy term describing the fact that the data across locations are correlated, and we can make assumptions about them.
- Most commonly, we assume test statistics are normally distributed but correlated across locations, which is called a Gaussian Random Field (GRF).
    - Alternative approaches assume T, F, or Chi-square distributions.
- Usually, we assume that the smoothness in the image is the same across locations (spatial stationarity) – this can get us into trouble.
- The effective number of tests is less than the number of voxels, and this can be expressed as the number of “resels” (Worsley et al., 1992). This allow us to threshold more liberally while adjusting for multiple comparisons.

Gaussian Random Field Theory

- GRF approximations require one or more parameters describing spatial smoothness. This is usually estimated from the data.
- GRF assumptions can give us the null distribution of the maximum test statistic, which we can use to do voxel-wise testing.
  - It’s fast and gives a threshold for p-values that controls FWER.
  - But it hinges on the assumptions.
- It can also be used to compute the distribution of the maximum cluster size
- What else can it give us?

Probabilities for cluster sizes

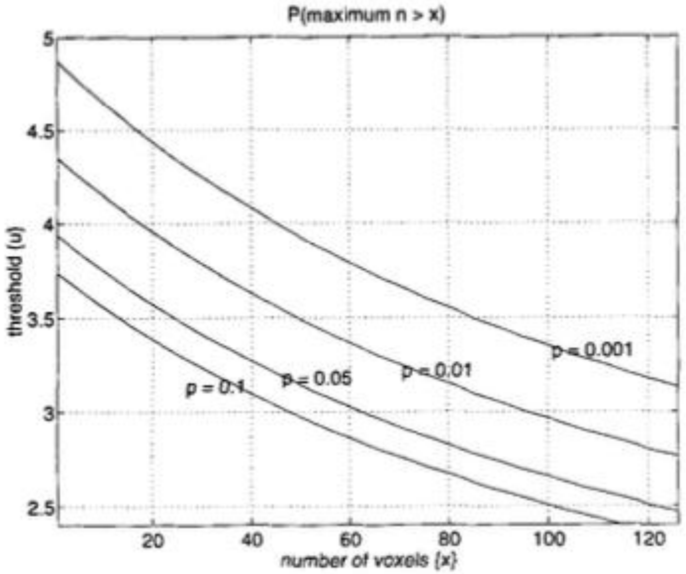


Figure 2.
Isoprobability contours calculated according to Equation 14, which give the relationship between threshold (*u*) and the number of voxels (*x*) an activation focus should contain to maintain a certain level of "probability." More strictly, P($n_{max} \geq k$) (the chance probability of obtaining one or more regions of at least *k* voxels) has been calculated over a range of voxels and thresholds and contoured at four levels (0.1, 0.05, 0.01, 0.001). In this example, *S* (volume) = 128^2, FWHM (smoothness) = 9.42, and *D* (dimensionality) = 2.

- Sophisticated inequalities based on topology yield cluster extent thresholds to compute probabilities for cluster sizes
- y-axis is the voxel-wise threshold (standard normal scale)
- x-axis is the number of voxels
- Contour lines are $P(E_0(1) > x)$,

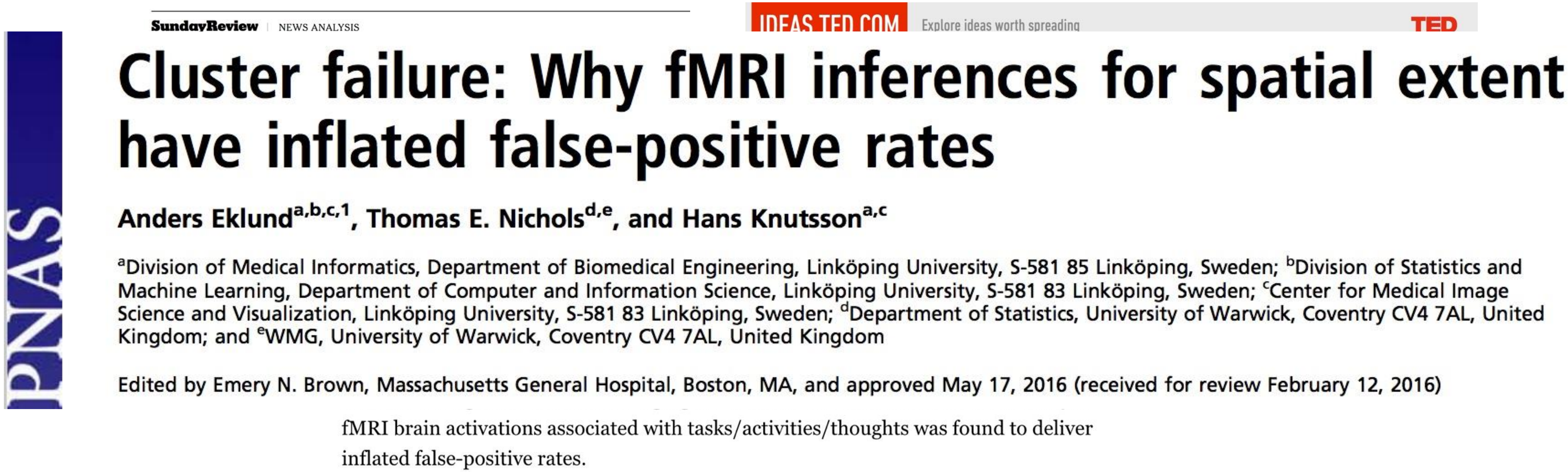
Friston et al., 1994

GRF References

- Friston et al. (1994) *Assessing the significance of focal activations using their spatial extent*
- Brett, Johnsrude & Owen (2002) *Introduction to Random Field Theory*
- Worsley et al. (1992) *Three-dimensional statistical analysis for CBF activation studies*
- Woo, Krishnan & Wager (2014) *Cluster-extent based thresholding in fMRI analyses*
- Eklund, Nichols & Knutsson (2016) *Cluster failure: Why fMRI inferences…*
- **Nonstationary RFT:** Hayasaka et al. (2004)
- **Cluster-Mass RFT:** Zhang et al. (2009)
- **Threshold-Free Cluster Enhancement (TFCE):** Smith & Nichols (2009)
- Peaks as test statistics, EC gives expected number of maxima.
- Cao & Worsley (1999) *The distribution of the maximum of Gaussian random fields on 6D correlation matrices*
- Davenport & Nichols (2020) *Selective peak inference…*
- Davenport et al. (2023) *Robust FWER control in neuroimaging using RFT*

Problem: inflated FWER in neuroimaging

- Standard mass-univariate tools have inflated FWER
- **Unrealistic assumptions about covariance structure**



SundayReview | NEWS ANALYSIS

IDEAS.TED.COM Explore ideas worth spreading TED

PNAS

Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Anders Eklund[a,b,c,1], Thomas E. Nichols[d,e], and Hans Knutsson[a,c]

[a]Division of Medical Informatics, Department of Biomedical Engineering, Linköping University, S-581 85 Linköping, Sweden; [b]Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University, S-581 83 Linköping, Sweden; [c]Center for Medical Image Science and Visualization, Linköping University, S-581 83 Linköping, Sweden; [d]Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom; and [e]WMG, University of Warwick, Coventry CV4 7AL, United Kingdom

Edited by Emery N. Brown, Massachusetts General Hospital, Boston, MA, and approved May 17, 2016 (received for review February 12, 2016)

fMRI brain activations associated with tasks/activities/thoughts was found to deliver inflated false-positive rates.

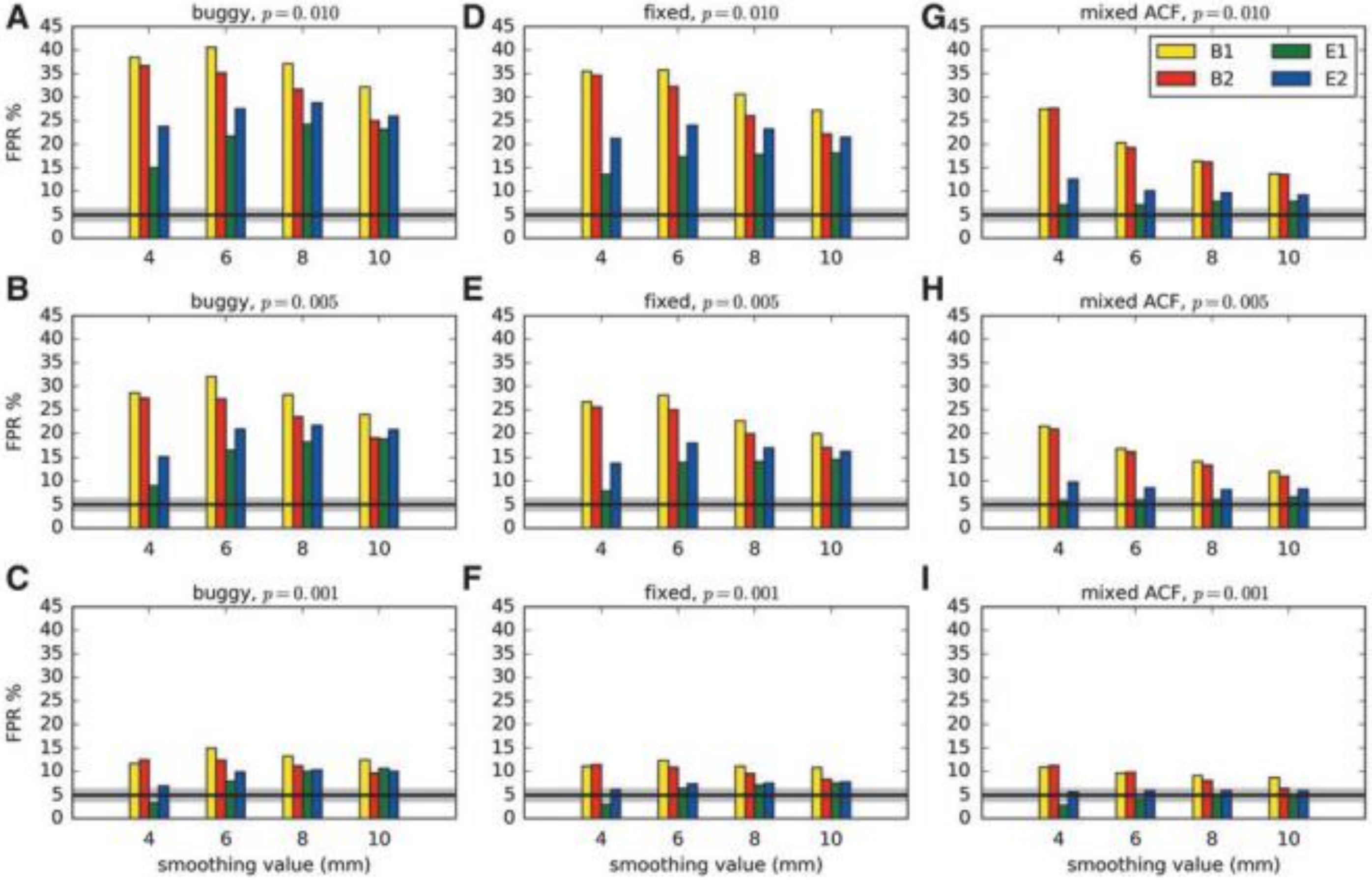
Eklund et al. 2016

- Elegant approach to evaluate the false positive rate
- Fit task models to resting state data – there should be no association
- Evaluated error rates – raises concerns about GRF error rates

Unrealistic assumptions in CEI

- Classical methods for SEI rely on smooth Gaussian random field approximations
- GRF approximations only work under very restrictive assumptions
- Many papers on this:

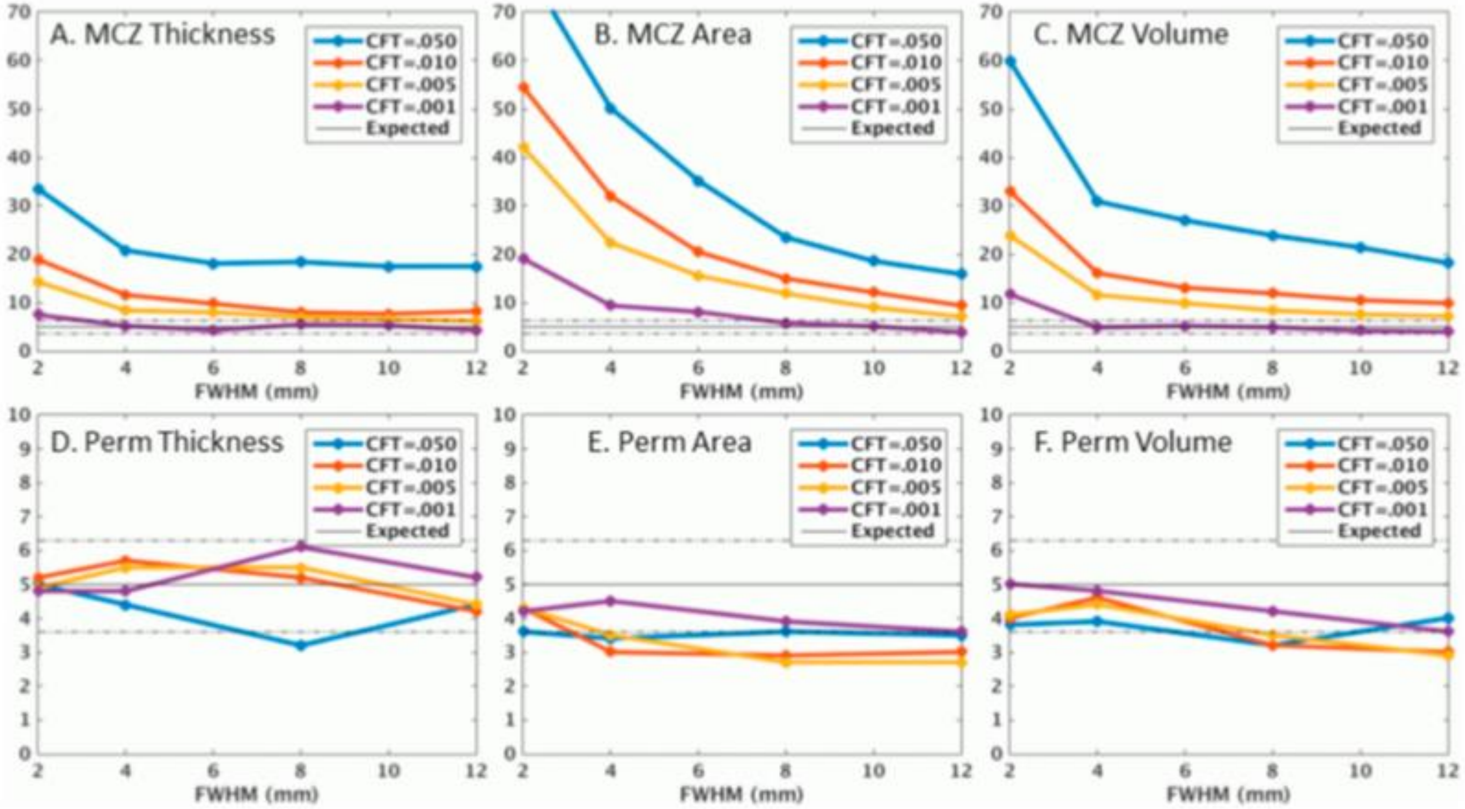
Silver 2011, Woo 2014, Eklund 2016, Flandin 2016, Slotnick 2017, Cox 2017, Kessler 2017, Mueller 2017, Greve 2018



Cox et al., *Brain Connectivity,* 2017

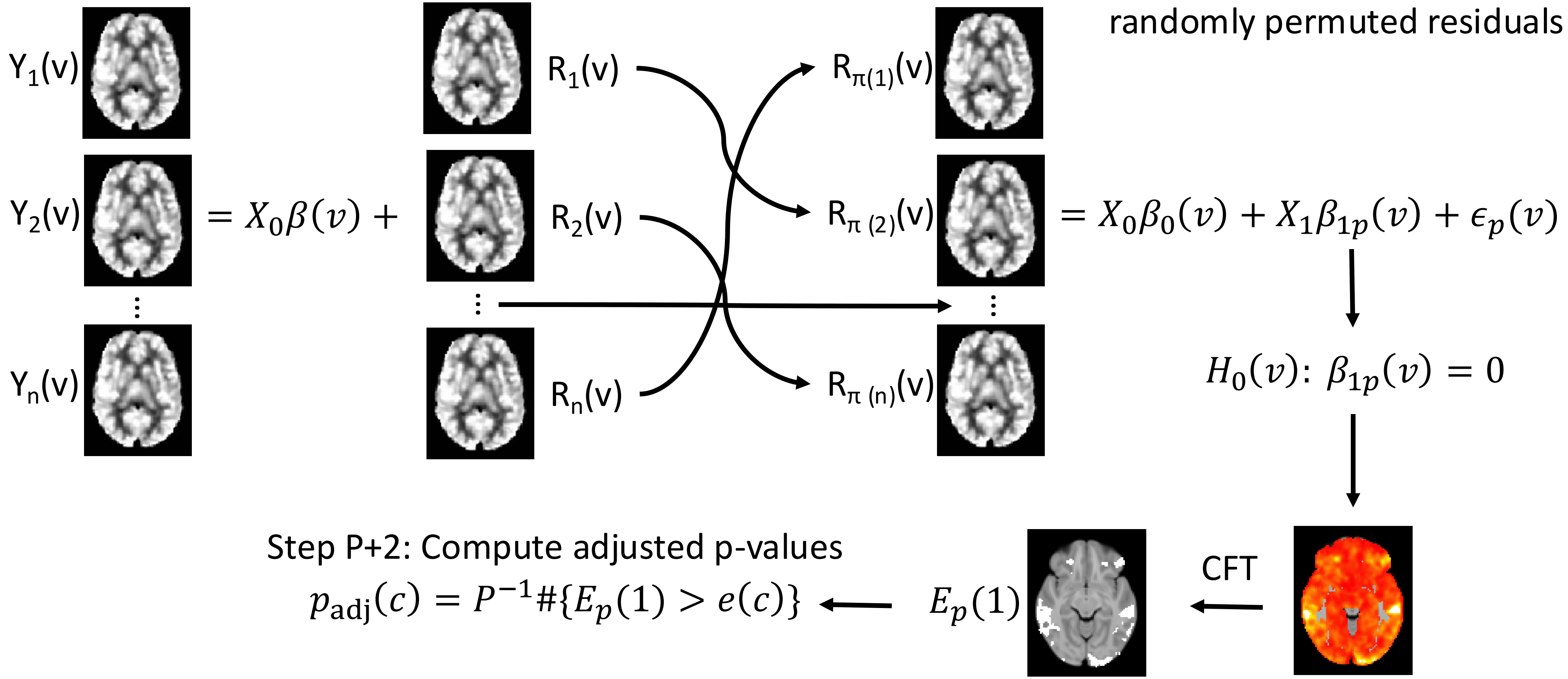
Permutation tests work quite well

- Permutation testing maintains nominal FWER
- Requires exchangeability – cannot accommodate multilevel models or heteroskedasticity



Greve & Fischl, *Neuroimage,* 2018

Permutation CEI procedure



Step 1: Get residuals from full model

$Y_1(v)$

$Y_2(v)$ $= X_0\beta(v) +$

⋮

$Y_n(v)$

$R_1(v)$

$R_2(v)$

⋮

$R_n(v)$

Steps 2 to P+1: Fit model to randomly permuted residuals

$R_{\pi(1)}(v)$

$R_{\pi(2)}(v)$ $= X_0\beta_0(v) + X_1\beta_{1p}(v) + \epsilon_p(v)$

⋮

$R_{\pi(n)}(v)$

$H_0(v)\colon\ \beta_{1p}(v) = 0$

CFT

$E_p(1)$

Step P+2: Compute adjusted p-values

$p_{\text{adj}}(c) = P^{-1}\#\{E_p(1) > e(c)\}$

Permutation procedure with covariates

- Randomise (FSL permutation procedure) uses the Freedman-Lane method
- They evaluated many in their paper

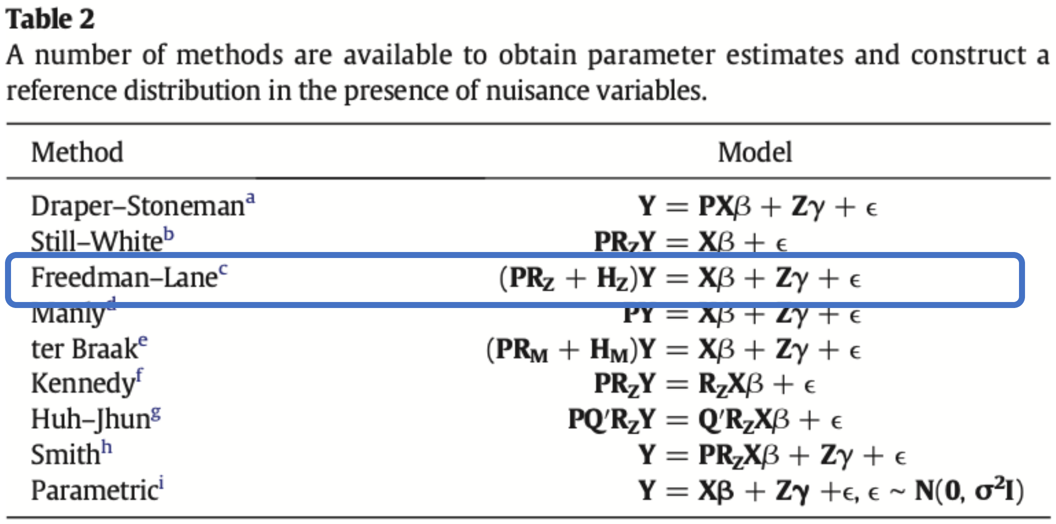
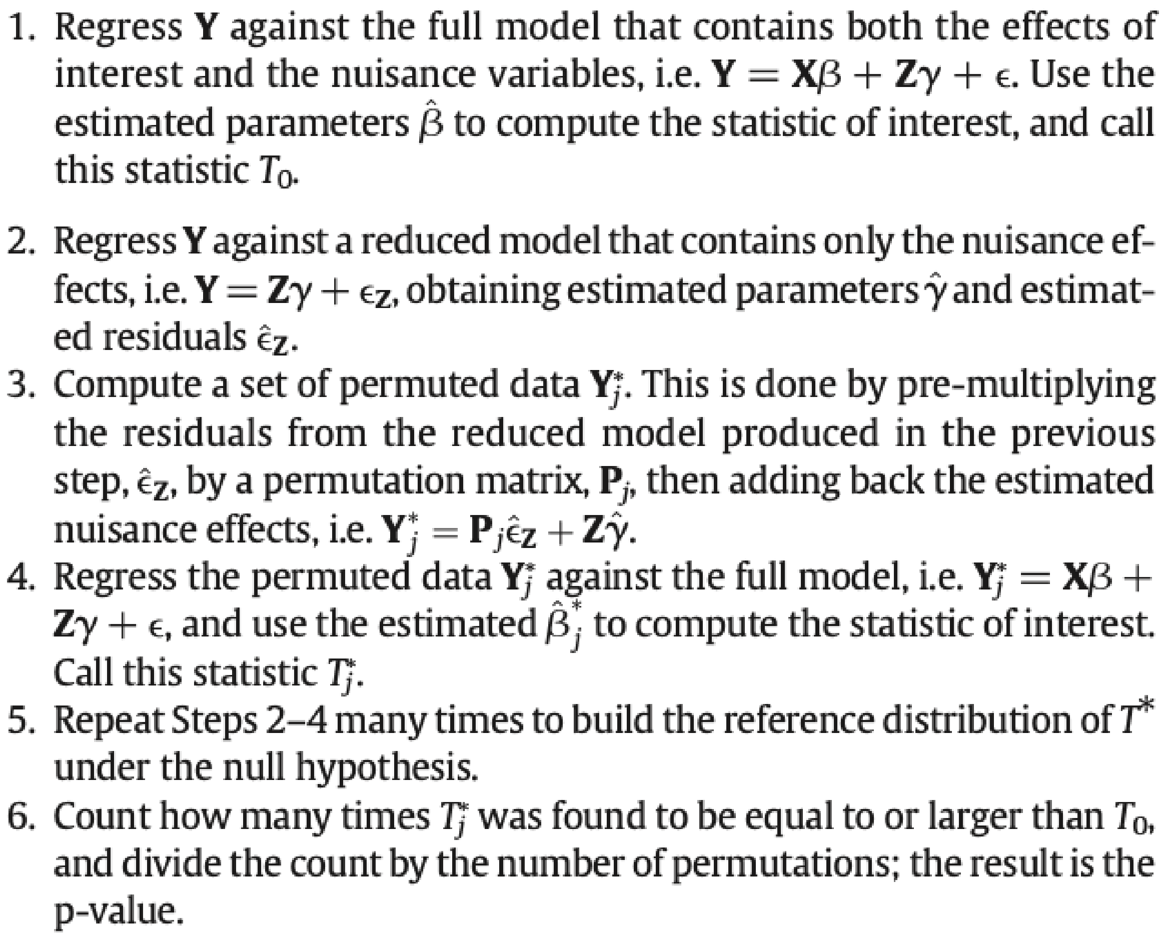


Table 2
A number of methods are available to obtain parameter estimates and construct a reference distribution in the presence of nuisance variables.

Method	Model
Draper–Stoneman[a]	$\mathbf{Y} = \mathbf{PX}\beta + \mathbf{Z}\gamma + \epsilon$
Still–White[b]	$\mathbf{PR_ZY} = \mathbf{X}\beta + \epsilon$
Freedman–Lane[c]	$(\mathbf{PR_Z} + \mathbf{H_Z})\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$
Manly[d]	$\mathbf{PY} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$
ter Braak[e]	$(\mathbf{PR_M} + \mathbf{H_M})\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$
Kennedy[f]	$\mathbf{PR_ZY} = \mathbf{R_ZX}\beta + \epsilon$
Huh–Jhun[g]	$\mathbf{PQ'R_ZY} = \mathbf{Q'R_ZX}\beta + \epsilon$
Smith[h]	$\mathbf{Y} = \mathbf{PR_ZX}\beta + \mathbf{Z}\gamma + \epsilon$
Parametric[i]	$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \epsilon,\ \epsilon \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I})$

Winkler et al., *Neuroimage, 2017*

Freedman-Lane procedure



1. Regress $\mathbf{Y}$ against the full model that contains both the effects of interest and the nuisance variables, i.e. $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$. Use the estimated parameters $\hat{\beta}$ to compute the statistic of interest, and call this statistic T_0.

2. Regress $\mathbf{Y}$ against a reduced model that contains only the nuisance effects, i.e. $\mathbf{Y} = \mathbf{Z}\gamma + \epsilon_\mathbf{Z}$, obtaining estimated parameters $\hat{\gamma}$ and estimated residuals $\hat{\epsilon}_\mathbf{Z}$.
3. Compute a set of permuted data $\mathbf{Y}_j^*$. This is done by pre-multiplying the residuals from the reduced model produced in the previous step, $\hat{\epsilon}_\mathbf{Z}$, by a permutation matrix, $\mathbf{P}_j$, then adding back the estimated nuisance effects, i.e. $\mathbf{Y}_j^* = \mathbf{P}_j\hat{\epsilon}_\mathbf{Z} + \mathbf{Z}\hat{\gamma}$.
4. Regress the permuted data $\mathbf{Y}_j^*$ against the full model, i.e. $\mathbf{Y}_j^* = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$, and use the estimated $\hat{\beta}_j^*$ to compute the statistic of interest. Call this statistic T_j^*.
5. Repeat Steps 2–4 many times to build the reference distribution of T^* under the null hypothesis.
6. Count how many times T_j^* was found to be equal to or larger than T_0, and divide the count by the number of permutations; the result is the p-value.

Permutation CEI

- For cluster extent inference, the test statistic T_0 is the maximum cluster size in the image
- The image is permuted as a whole
- A similar procedure can be used to perform FWER control as well

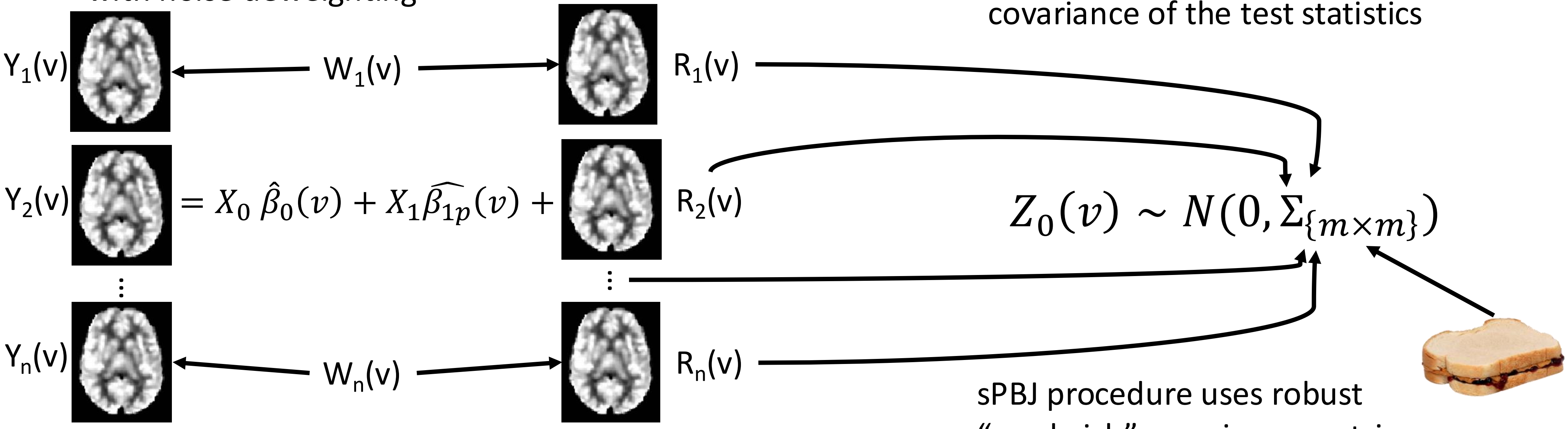
Permutation CEI procedure pros and cons

- Pros:
    - It is more robust than GRF based methods
    - It is easily implemented for linear models with the “randomise” FSL function
- Cons
    - It is not appropriate for multilevel models e.g. fMRI
    - Requires exchangeability assumption i.e. $\mathrm{Var}\{\hat{\beta}_i(v), \hat{\beta}_i(w)\} = \Sigma(v, w)$
    - It can take over a day to run

Semi-PBJ (sPBJ) CEI procedure

Step 1: Get residuals from full model with noise deweighting

Step 2: Use residuals to estimate covariance of the test statistics


$Y_1(v)$
$W_1(v)$
$R_1(v)$
$Y_2(v)$
$= X_0\,\hat{\beta}_0(v) + X_1\widehat{\beta_{1p}}(v) +$
$R_2(v)$
⋮
⋮
$Y_n(v)$
$W_n(v)$
$R_n(v)$
$Z_0(v) \sim N(0, \Sigma_{\{m\times m\}})$

sPBJ procedure uses robust “sandwich” covariance matrix

Step 3: Use computationally efficient methods to sample $Z_{0b}(v)$ for $b = 1, \dots B$

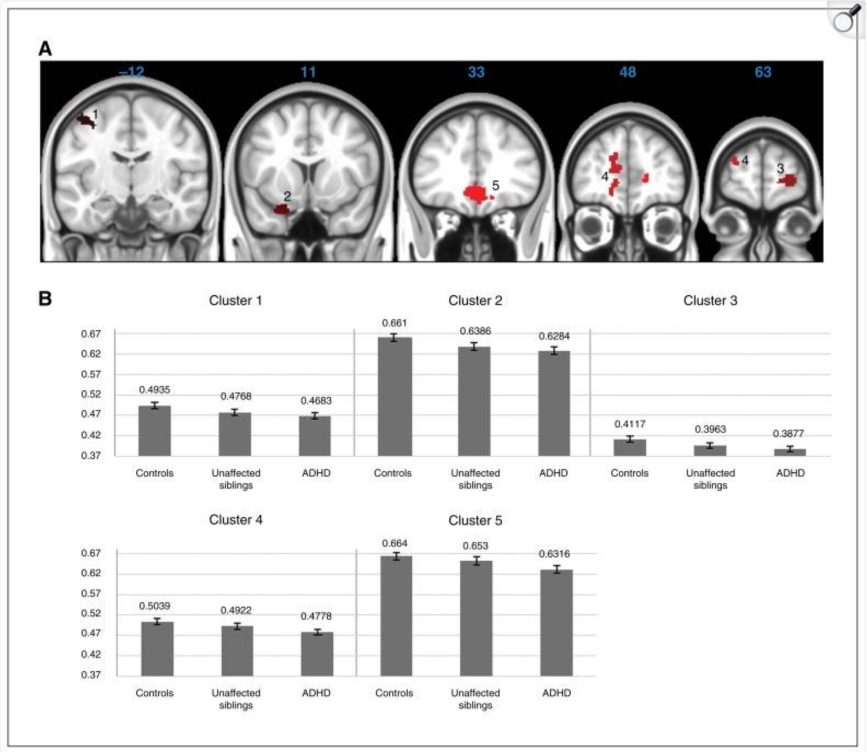
Step 4: Compute adjusted cluster p-values

Vandekar et al., *Arxiv*, 2018
Long & Ervin, *Am. Stat.*, 2000
MacKinnon & White, *J. Econo.*, 1985

Typical statistical reporting

Bratlen et al., *J Psychiatry Neurosci*, 2015

Voxel-based morphometry analysis reveals frontal brain differences in participants with ADHD and their unaffected siblings

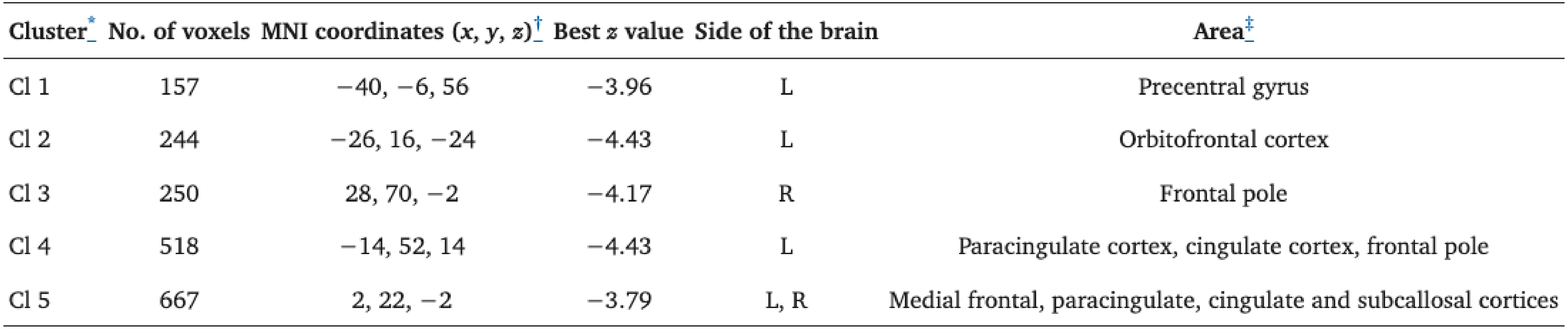


(**A**) Whole-brain significant clusters for case–control differences. Five clusters were identified: cluster 1 = precentral gyrus; cluster 2 = orbitofrontal cortex; cluster 3 = frontal pole; cluster 4 = paracingulate and cingulate cortices, frontal pole; and cluster 5 = medidal frontal, paracingulate, cingulate and subcallosal cortices. (**B**) Mean voxel differences for the identified cluster between participants with attention-deficit/hyperactivity disorder (ADHD; n = 307), their unaffected siblings (n = 169) and typically developing controls (n = 196).

- They compared structural differences at the voxel-level between ADHD, sibling, and HC
- Questions:
  - What is the interpretation of the results?
  - What kind of test was performed at each voxel? What is the contrast?
  - What kind of spatial inference was used?
- ***Methods:*** *We considered differences to be significant if they survived cluster-mass thresholding with the easythresh option in FSL (www.fmrib.ox.ac.uk/fsl), using an initial cluster forming threshold of z > 3.1. Subsequently, we estimated each cluster's significance level based on Gaussian random field theory, and those clusters surviving a family-wise error (FWE)–corrected significance threshold of p < 0.05 showing volume differences > 0.1 mL were reported.*
- ***Results:***

Bratlen et al., *J Psychiatry Neurosci*, 2015

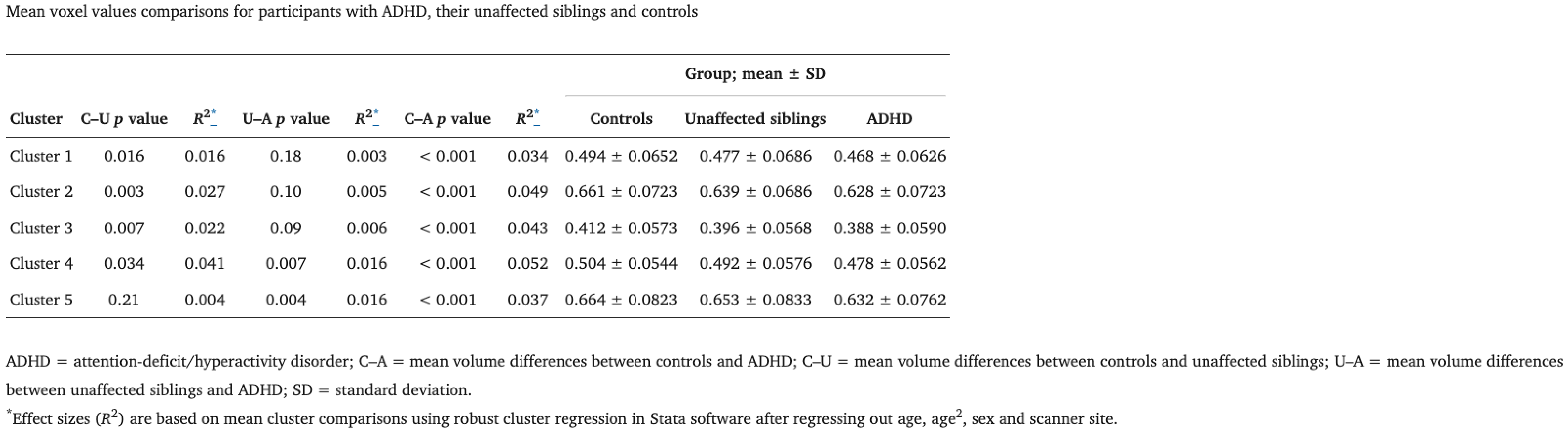
- Cluster summary table



Cluster[*]	No. of voxels	MNI coordinates (x, y, z)[†]	Best z value	Side of the brain	Area[‡]
Cl 1	157	−40, −6, 56	−3.96	L	Precentral gyrus
Cl 2	244	−26, 16, −24	−4.43	L	Orbitofrontal cortex
Cl 3	250	28, 70, −2	−4.17	R	Frontal pole
Cl 4	518	−14, 52, 14	−4.43	L	Paracingulate cortex, cingulate cortex, frontal pole
Cl 5	667	2, 22, −2	−3.79	L, R	Medial frontal, paracingulate, cingulate and subcallosal cortices

Posthoc comparisons

- Analyses in significant clusters.
- What is different from the C-A p-values shown here?



Mean voxel values comparisons for participants with ADHD, their unaffected siblings and controls

							Group; mean ± SD		
Cluster	**C–U *p* value**	***R*^{2*}**	**U–A *p* value**	***R*^{2*}**	**C–A *p* value**	***R*^{2*}**	**Controls**	**Unaffected siblings**	**ADHD**
Cluster 1	0.016	0.016	0.18	0.003	< 0.001	0.034	0.494 ± 0.0652	0.477 ± 0.0686	0.468 ± 0.0626
Cluster 2	0.003	0.027	0.10	0.005	< 0.001	0.049	0.661 ± 0.0723	0.639 ± 0.0686	0.628 ± 0.0723
Cluster 3	0.007	0.022	0.09	0.006	< 0.001	0.043	0.412 ± 0.0573	0.396 ± 0.0568	0.388 ± 0.0590
Cluster 4	0.034	0.041	0.007	0.016	< 0.001	0.052	0.504 ± 0.0544	0.492 ± 0.0576	0.478 ± 0.0562
Cluster 5	0.21	0.004	0.004	0.016	< 0.001	0.037	0.664 ± 0.0823	0.653 ± 0.0833	0.632 ± 0.0762

ADHD = attention-deficit/hyperactivity disorder; C–A = mean volume differences between controls and ADHD; C–U = mean volume differences between controls and unaffected siblings; U–A = mean volume differences between unaffected siblings and ADHD; SD = standard deviation.

[*] Effect sizes (R^2) are based on mean cluster comparisons using robust cluster regression in Stata software after regressing out age, age^2, sex and scanner site.

Loitfelder et al., *PLOS One*, 2014

Brain Activity Changes in Cognitive Networks in Relapsing-Remitting Multiple Sclerosis – Insights from a Longitudinal fMRI Study

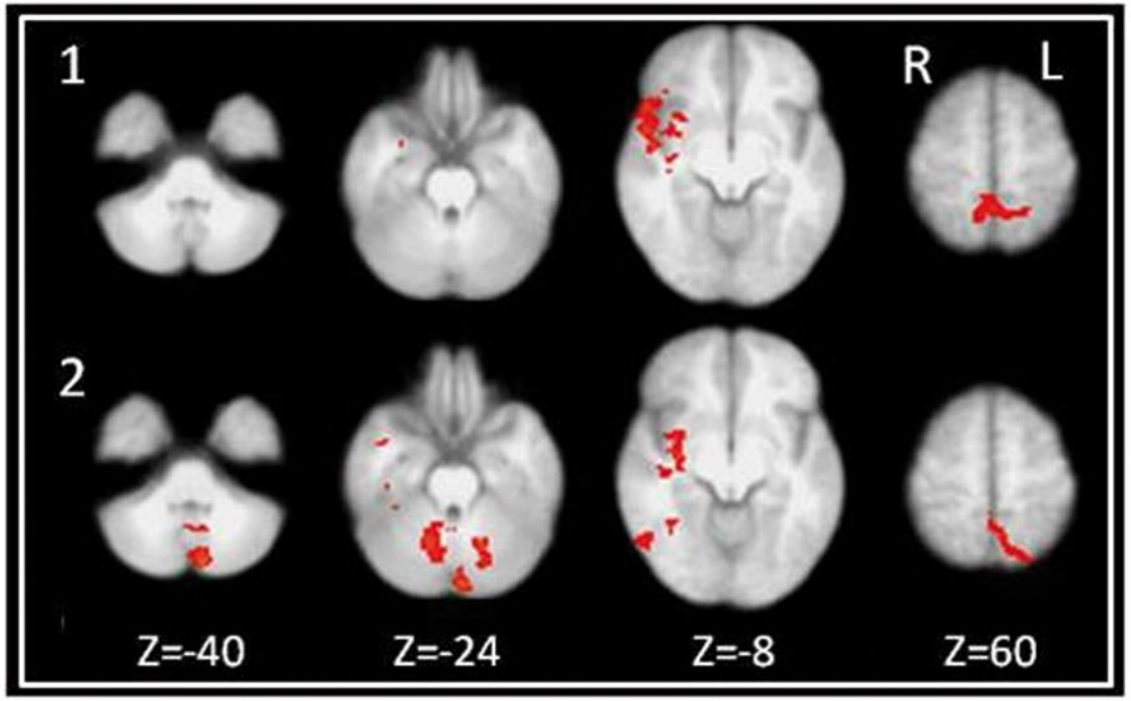
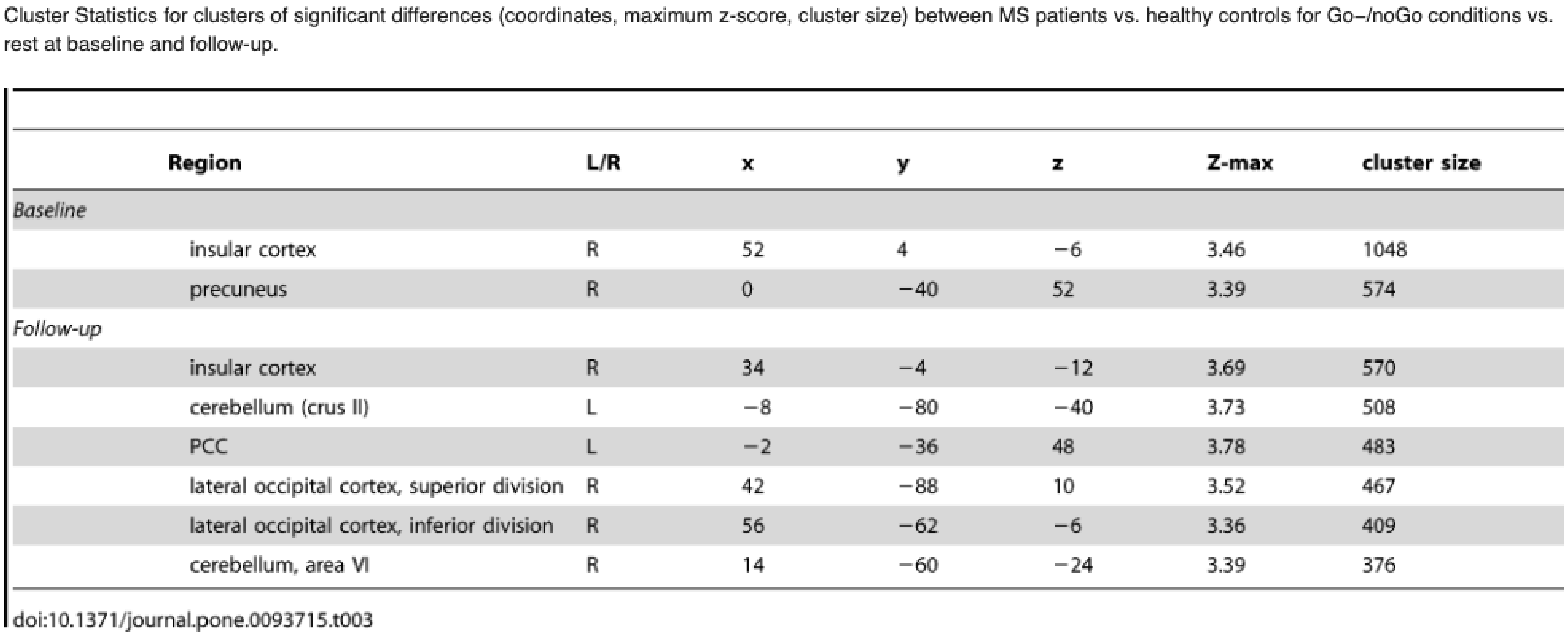


Figure 2. Areas with increased brain activation in MS patients vs. controls.
Clusters of significant activation difference (mixed effects higher level analyses; Z>2.3; corrected cluster significance threshold p=0.05) in contrasts for Go-/noGo task vs. rest for MS patients compared to controls at baseline (1) and follow-up (2).

- They compared functional changes at the voxel-level between MS and HC during a Go/No-Go task
- Questions:
  - What model was used?
  - What kind of test was performed at each voxel? What is the contrast?
  - What kind of spatial inference was used?
- ***Methods:*** *Higher-level analysis was done using FLAME stage 1 (FMRIB's Local Analysis of Mixed Effects). Z (Gaussianised T/F) statistic images were thresholded using clusters determined by Z>2.3 and a (corrected) cluster significance threshold of p=0.05 (for further details see [4]). At higher level, contrasts were calculated within-groups over time and across groups over time.*
- ***Results:*** *Patients demonstrated increased activation compared to HC in the insular cortex and precuneus at BL. At FU, they additionally activated the posterior cingulate cortex (PCC), the cerebellum (left crus II, right lobule VI), and the lateral occipital cortex (superior and inferior division, Figure 2, Table 3).*

Loitfelder et al., *PLOS One*, 2014



Cluster Statistics for clusters of significant differences (coordinates, maximum z-score, cluster size) between MS patients vs. healthy controls for Go–/noGo conditions vs. rest at baseline and follow-up.

Region	L/R	x	y	z	Z-max	cluster size
Baseline						
insular cortex	R	52	4	−6	3.46	1048
precuneus	R	0	−40	52	3.39	574
Follow-up						
insular cortex	R	34	−4	−12	3.69	570
cerebellum (crus II)	L	−8	−80	−40	3.73	508
PCC	L	−2	−36	48	3.78	483
lateral occipital cortex, superior division	R	42	−88	10	3.52	467
lateral occipital cortex, inferior division	R	56	−62	−6	3.36	409
cerebellum, area VI	R	14	−60	−24	3.39	376

doi:10.1371/journal.pone.0093715.t003

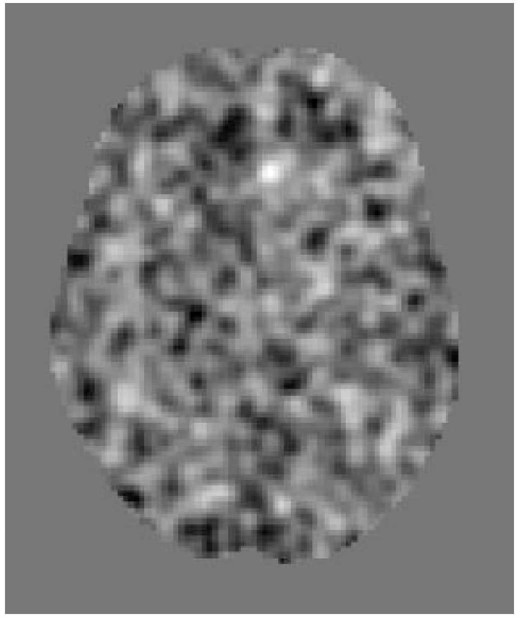
Extras after this

FSL slides on GRF and permutation

Distribution of Max Cluster Size

So, just as was the case for the z-values, we now have a distribution $f(R)$ that allows us to calculate a Family Wise threshold $u(R)$ pertaining to cluster size.

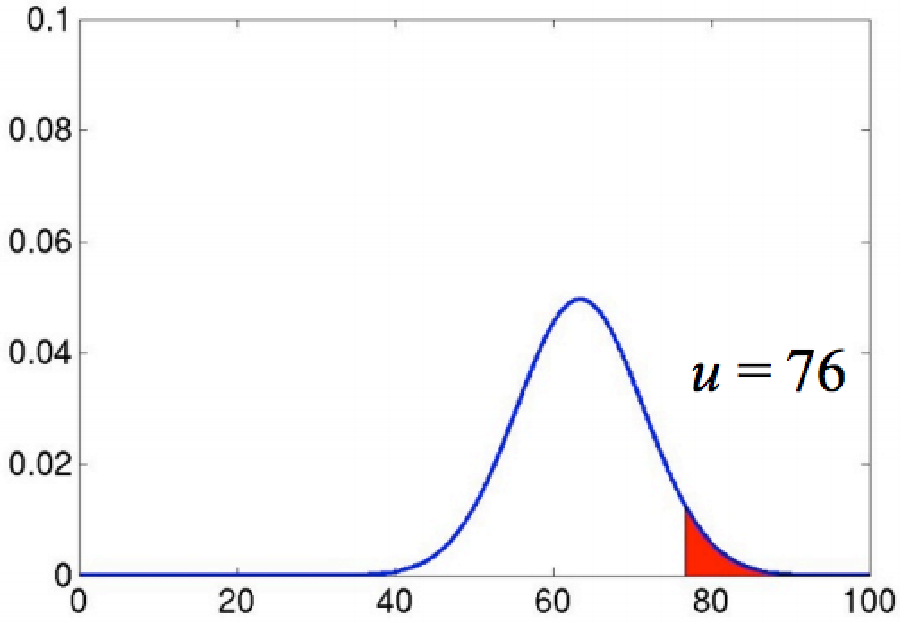
$f(R)$ depends crucially on the initial "cluster-forming" threshold?



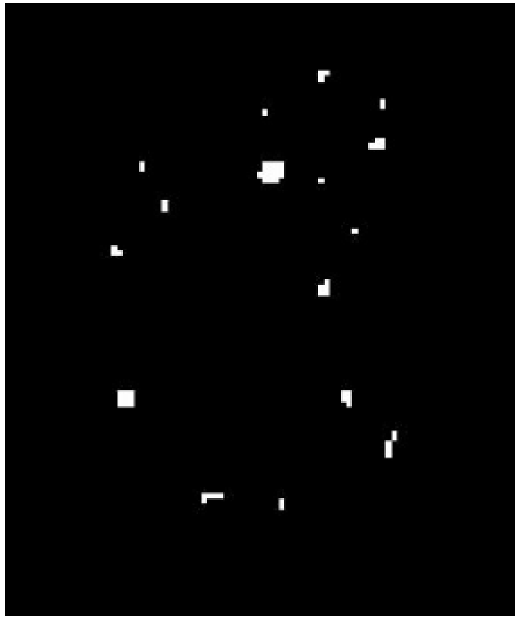
$z = 2.3$

Distribution of Max Cluster Size

So, just as was the case for the z-values, we now have a distribution $f(R)$ that allows us to calculate a Family Wise threshold $u(R)$ pertaining to cluster size.



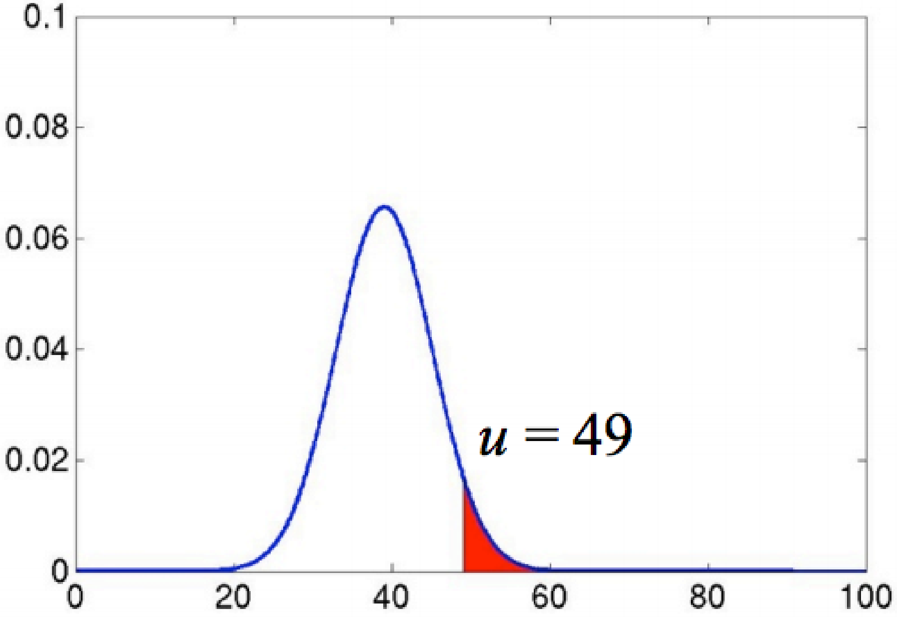
$f(R)$ depends crucially on the initial "cluster-forming" threshold?



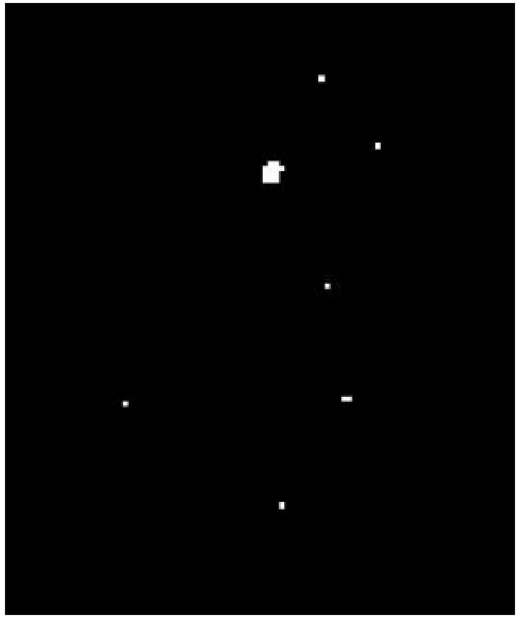
$z = 2.3$

Distribution of Max Cluster Size

So, just as was the case for the z-values, we now have a distribution $f(R)$ that allows us to calculate a Family Wise threshold $u(R)$ pertaining to cluster size.



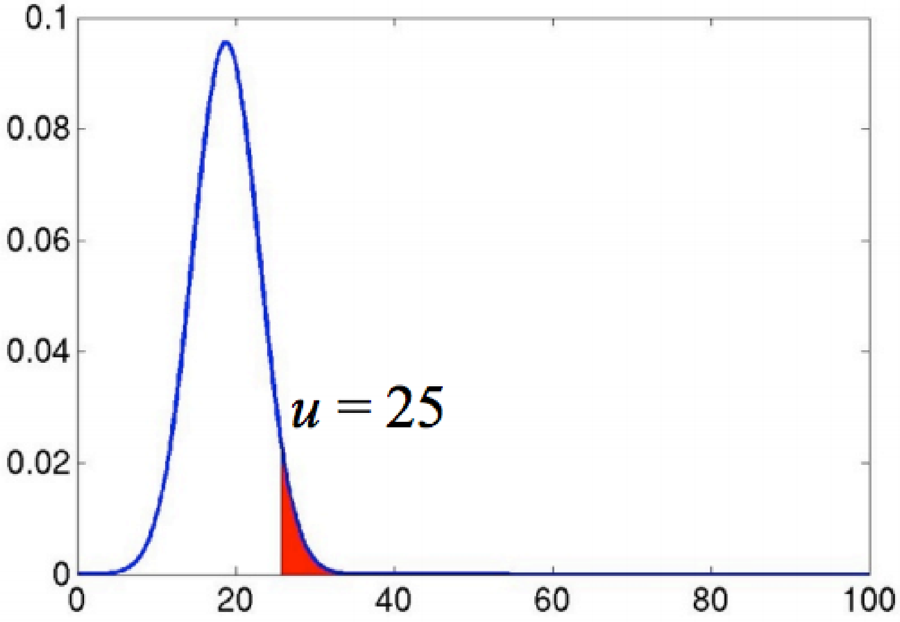
$f(R)$ depends crucially on the initial "cluster-forming" threshold?



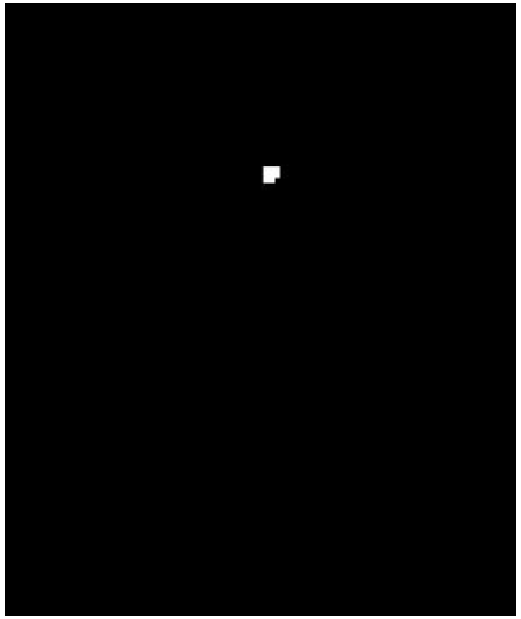
$z = 2.7$

Distribution of Max Cluster Size

So, just as was the case for the z-values, we now have a distribution $f(R)$ that allows us to calculate a Family Wise threshold $u(R)$ pertaining to cluster size.



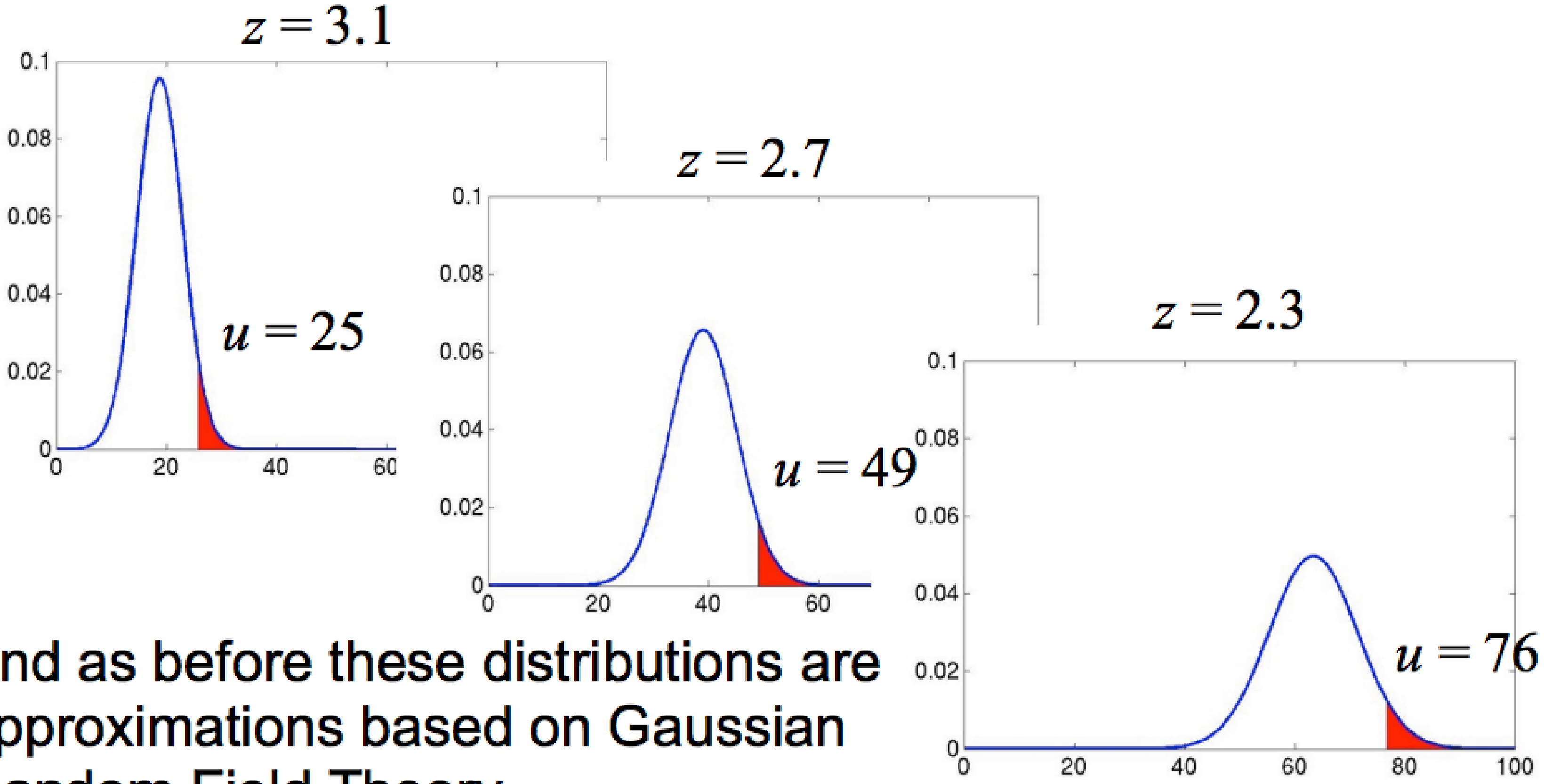
$f(R)$ depends crucially on the initial “cluster-forming” threshold?



$z = 3.1$

Distribution of Max Cluster Size

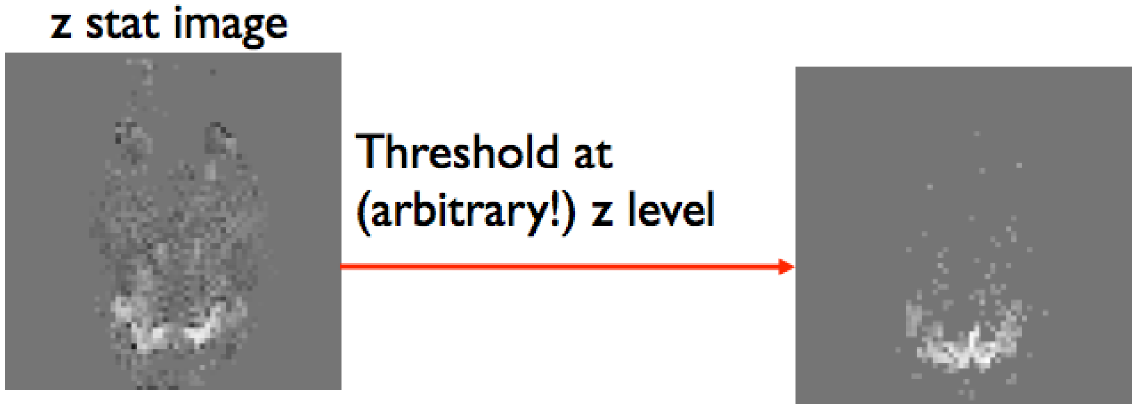
Hence the distribution for the cluster size should really be written $f(R,z)$ and the same for $u(R,z)$



And as before these distributions are approximations based on Gaussian Random Field Theory.

Clustering cookbook

Instead of resel-based correction, we can do clustering:

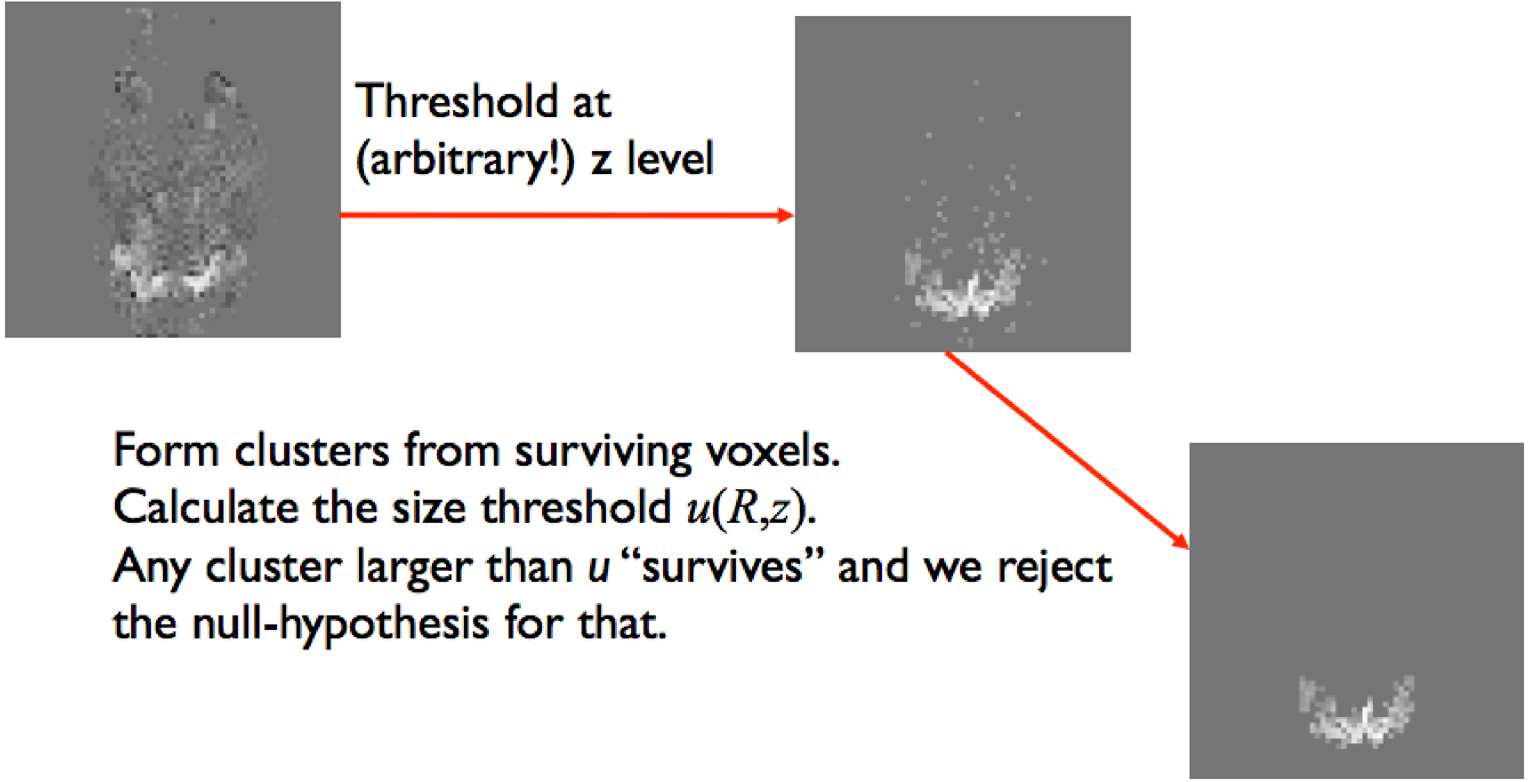


z stat image

Threshold at
(arbitrary!) z level

Clustering cookbook

Instead of resel-based correction, we can do clustering



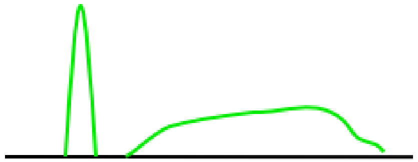
z stat image

Threshold at (arbitrary!) z level

Form clusters from surviving voxels.
Calculate the size threshold $u(R,z)$.
Any cluster larger than u “survives” and we reject the null-hypothesis for that.

How do we choose the (arbitrary!) z-threshold?

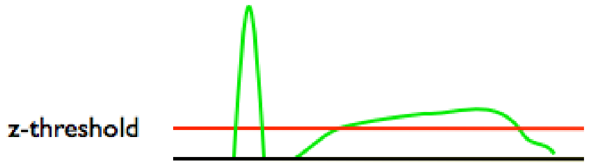
This is arbitrary and a trade-off



How do we choose the (arbitrary!) z-threshold?

This is arbitrary and a trade-off

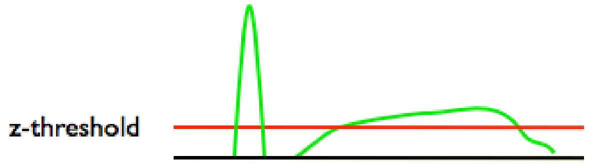
1. **Low threshold** - can violate RFT assumptions, but can detect clusters with large spatial extent and low z



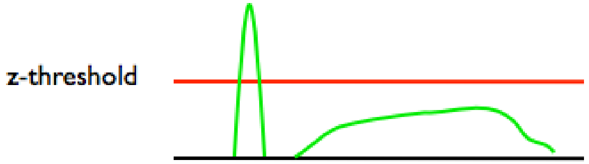
How do we choose the (arbitrary!) z-threshold?

This is arbitrary and a trade-off

1. **Low threshold** - can violate RFT assumptions, but can detect clusters with large spatial extent and low z



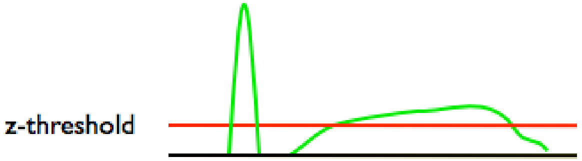
2. **High threshold** - gives more power to clusters with small spatial extent and high z



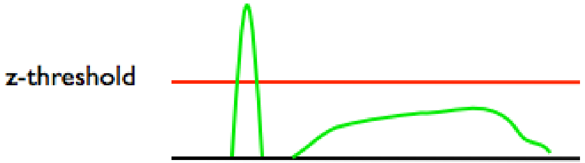
How do we choose the (arbitrary!) z-threshold?

This is arbitrary and a trade-off

1. **Low threshold** - can violate RFT assumptions, but can detect clusters with large spatial extent and low z



2. **High threshold** - gives more power to clusters with small spatial extent and high z



Tends to be more sensitive than voxel-wise corrected testing

Results depend on extent of spatial smoothing in pre-processing

GRF for cluster-wise tests

- Needs a null-hypothesis, a test-statistic and an initial cluster forming threshold.

- Calculates a (size) threshold based on number of RESELS and initial (z) threshold

+ Gives a (size) threshold such that the family-wise error is controlled.
+ Calculates that threshold very fast.

- Hinges on strict assumptions about the distribution of the data $(\sim N(0,\Lambda))$
- Inference pertains to entire cluster
- Initial threshold is arbitrary

Randomise, slow and reliable.

Impressive as the GRF based thresholding is, there are situations where we can't use it.

randomise always delivers

We may want to use a test-statistic for which the distribution is unknown.
Example: The "TFC enhanced" *t*-statistic. (A highly non-linear spatial filter)

Our data may not be normally distributed. Then our *t*-values will, paradoxically, not be *t*-distributed. (what are the chances...)
Example: VBM-style data (data whose value is probability of a certain tissue-type)

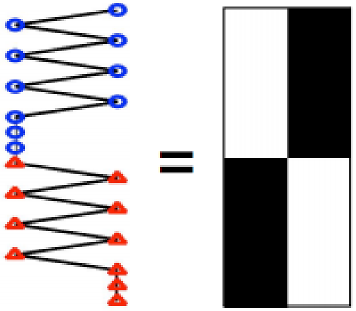
We want to restrict our analysis to a particular (small and irregularly shaped) sulcus to increase our sensitivity when we have a prior spatial hypothesis.

Example: VBM-style analysis

- Our data is segmented grey matter maps
- A voxel is either grey matter, or not.

Group #1
(Oxford students)

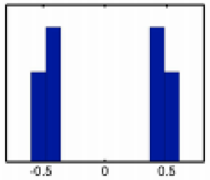
Group #2
(Train spotters)



$$\begin{bmatrix}\beta_1\\\beta_2\end{bmatrix}$$

$$\begin{bmatrix}\beta_1\\\beta_2\end{bmatrix} = \begin{bmatrix}0.4\\0.6\end{bmatrix}$$ Ok!

hist($\mathbf{e}$)



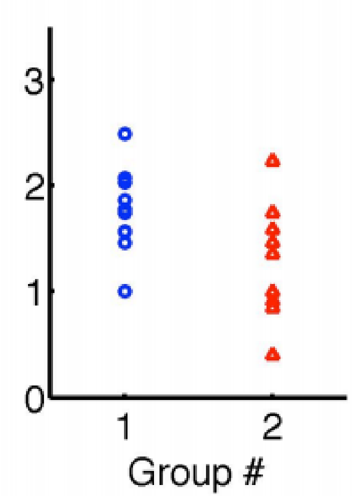
~ *N* ?

☹

Oh dear! What now?

- We could use Monte-Carlo to simulate the distributions. As I did for these slides.
  - But, hinges on lots of assumptions about the data
- We could permute the data itself.

We have performed an experiment



And calculated a statistic, e.g. a *t*-value

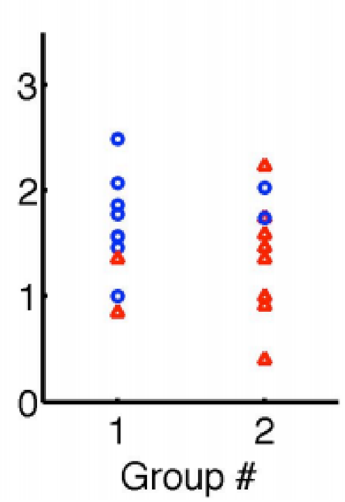
$$t = 2.27$$

If the null-hypothesis is true, there is no difference between the groups. That means we should be able to “re-label” the individual points without changing anything.

Oh dear! What now?

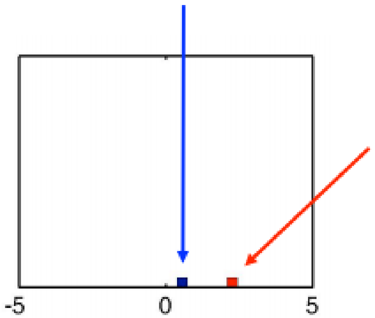
- We could Monte-Carlo simulate the distributions. As I did for these slides.
  - But, hinges on lots of assumptions about the data
- We could permute the data itself.

One re-labelling



t-value after re-labelling

$t = 0.67$



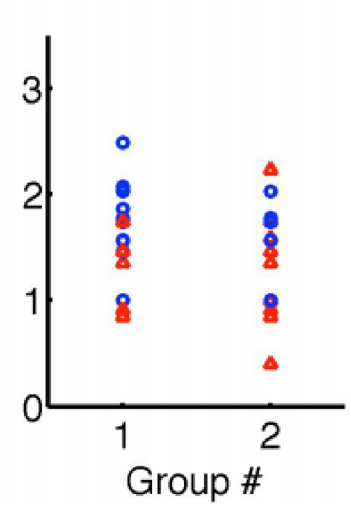
Original labelling

Let’s start collecting them

Oh dear! What now?

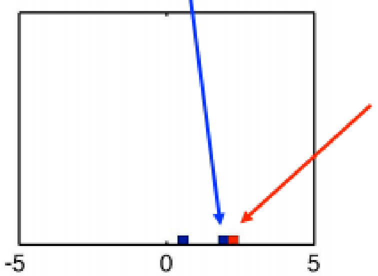
- We could Monte-Carlo simulate the distributions. As I did for these slides.
  - But, hinges on lots of assumptions about the data
- We could permute the data itself.

Second re-labelling



t-value after re-labelling

$t = 1.97$



Original labelling

And another one

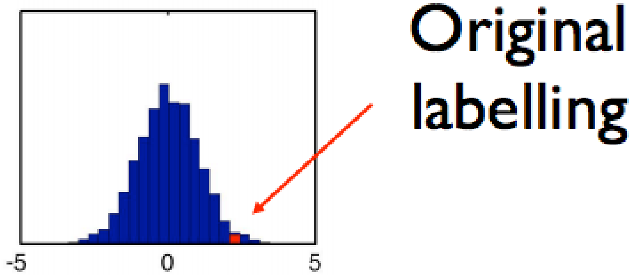
Oh dear! What now?

- We could Monte-Carlo simulate the distributions. As I did for these slides.
  - But, hinges on lots of assumptions about the data
- We could permute the data itself.

Of the 5000 re-labellings, only 90 had a t-value > 2.27 (the original labelling).

I.e. there is only a ~1.8% (90/5000) chance of obtaining a value > 2.27 if there is no difference between the groups

C.f. $p(x \geq 2.27) = 1.79\%$ for t_{18}

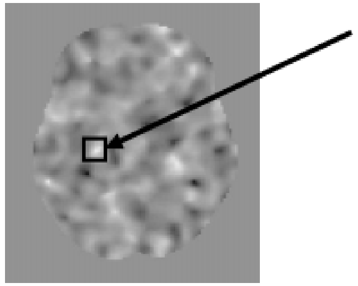


5000 re-labellings. Phew!

And now we can do the wacky stats.

This is what we got

We compared activation by painful stimuli in two groups of 5 subjects each.



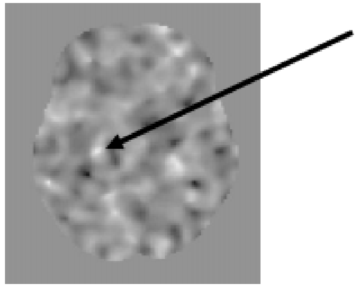
Very intriguing activation. $t_8 = 4.65$

Prof. ran to write to Science. **But**, did she jump the gun?

And now we can do the wacky stats.

We compared activation by painful stimuli in two groups of 5 subjects each.

This is what we got

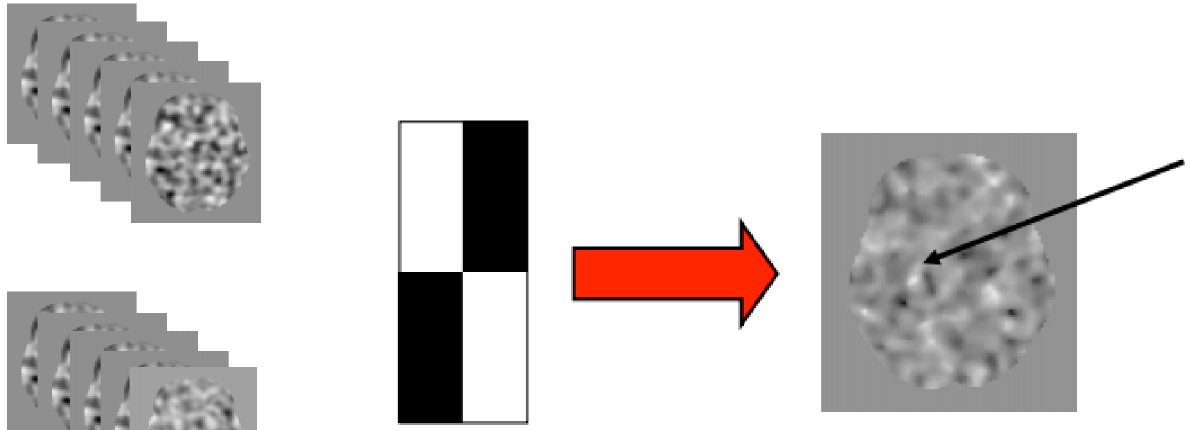


Very intriguing activation. $t_8 = 4.65$

Prof. ran to write to Science. **But**, did she jump the gun?

Group 1

Group 2



max(t)=4.65

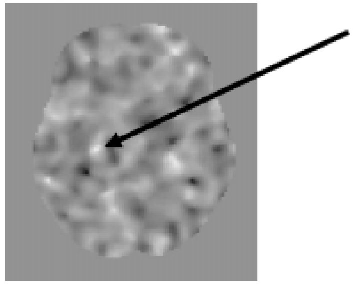
2nd level model

Our group difference map

And now we can do the wacky stats.

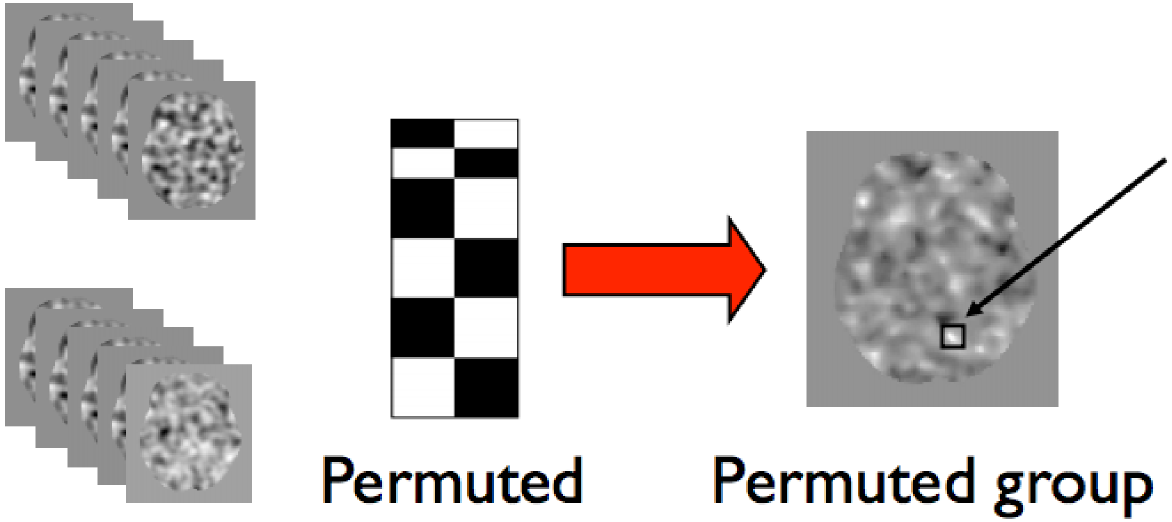
We compared activation by painful stimuli in two groups of 5 subjects each.

This is what we got



Very intriguing activation. $t_8 = 4.65$

Prof. ran to write to Science. **But**, did she jump the gun?



Group 1

Group 2

Permuted model

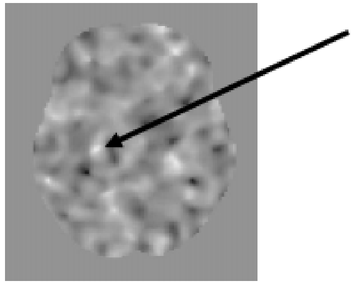
Permuted group difference map

max(t)=8.23

And now we can do the wacky stats.

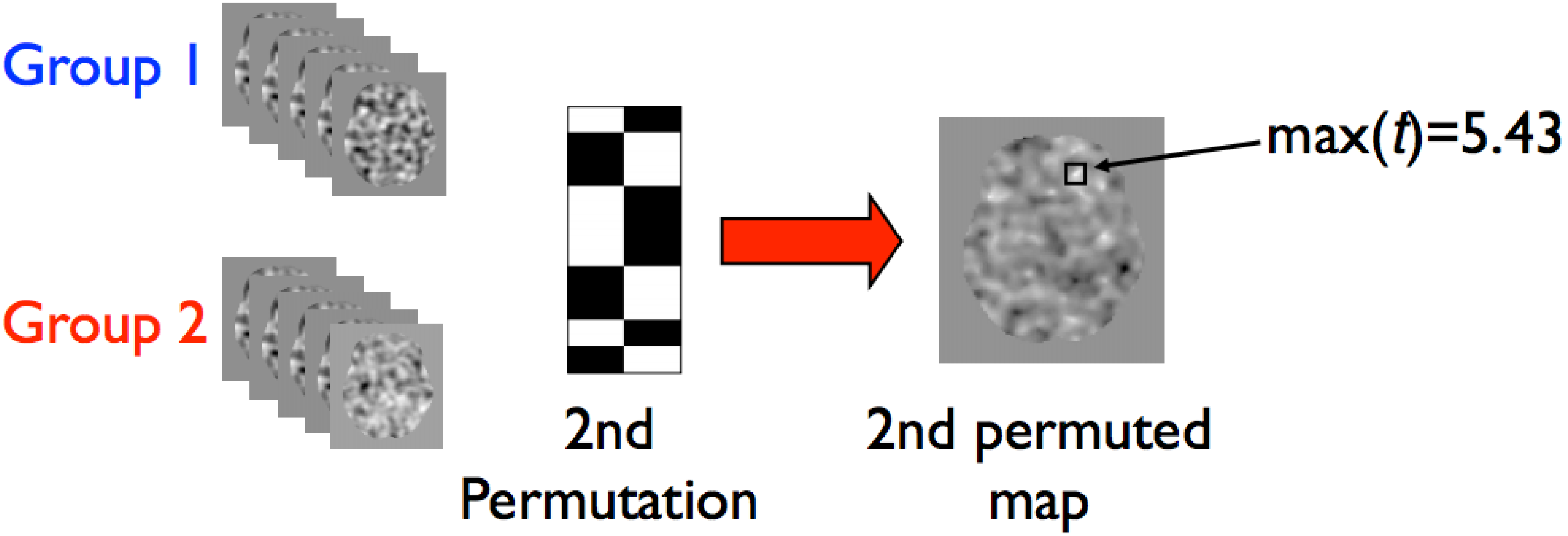
We compared activation by painful stimuli in two groups of 5 subjects each.

This is what we got



Very intriguing activation. $t_8 = 4.65$

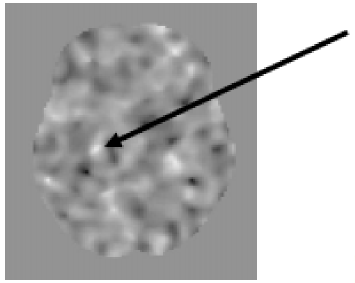
Prof. ran to write to Science. **But**, did she jump the gun?



And now we can do the wacky stats.

We compared activation by painful stimuli in two groups of 5 subjects each.

This is what we got

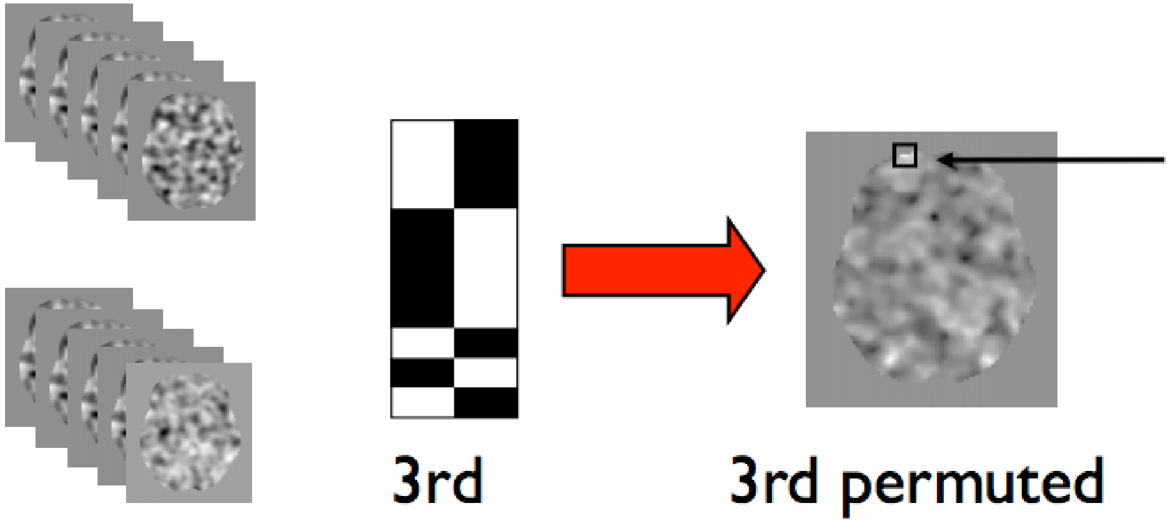


Very intriguing activation. $t_8 = 4.65$

Prof. ran to write to Science. **But**, did she jump the gun?

Group 1

Group 2



$\max(t)=5.84$

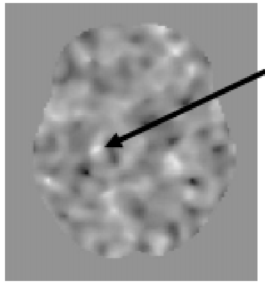
3rd Permutation

3rd permuted map

And now we can do the wacky stats.

We compared activation by painful stimuli in two groups of 5 subjects each.

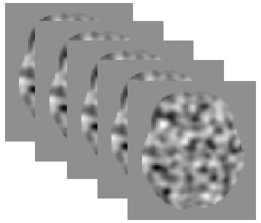
This is what we got



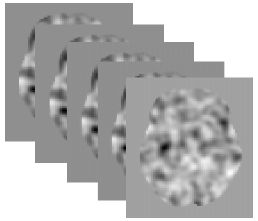
Very intriguing activation. $t_8 = 4.65$

Prof. ran to write to Science. **But**, did she jump the gun?

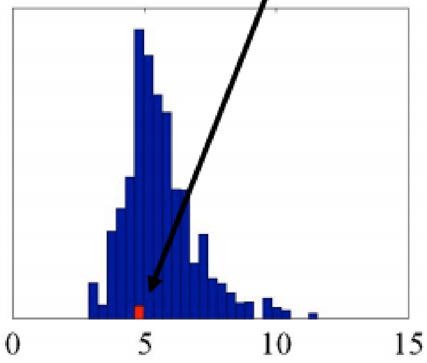
Group 1



Group 2



Original labelling



5000 permutations

3925 permutations yielded higher max(t)-value than original labelling. We cannot reject the null-hypothesis.

Permutations for dummies

- Needs a null-hypothesis and a test-statistic.

+ Builds its own "null-distribution" from your single data-set.
+ No assumptions about the data.
+ Can use any test-statistic, e.g. max(*t*), max($n_{cluster}$) etc.
+ Can use "classical" statistics (e.g. *t*-test) when data have strange distribution.

- Need to ensure exchangeability.

- Don't hold your breath.

Slides on sPBJ evaluation

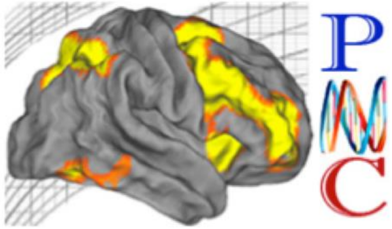
Philadelphia Neurodevelopmental Cohort (PNC)[1]

- Cohort study to investigate genetic and imaging risk factors associated with psychiatric disorders
  - Led by Raquel Gur and Hakon Hakonarson
- Cross-sectional N-back data 1,600 volunteers ages 8-23

n	% female	Age (SD)	Age range
1000	55%	14.7 (3.5)	8-23

[1] Satterthwaite et al., *Neuroimage*, 2016

N-back data analysis



- We analyzed a subset of 1000 subjects imaged as part of the Philadelphia Neurodevelopmental Cohort
- **Our scientific goal is to understand how activation associated with increasing working memory load is related to task performance**

$$Y_i(v) = \alpha_0(v) + \alpha_1(v) \times \mathrm{sex}_i + \alpha_2(v) \times \ \mathrm{age}_i + \alpha_3(v) \times \mathrm{mot}_i$$

$$+\beta(v) \ \times \ d_i' + \epsilon_i(v)$$

- Our null hypothesis at v is $H_0(v)\colon \ \beta(v) = 0$

Satterthwaite et al., *J Neuro*, 2013

Simulating realistic heteroskedasticity

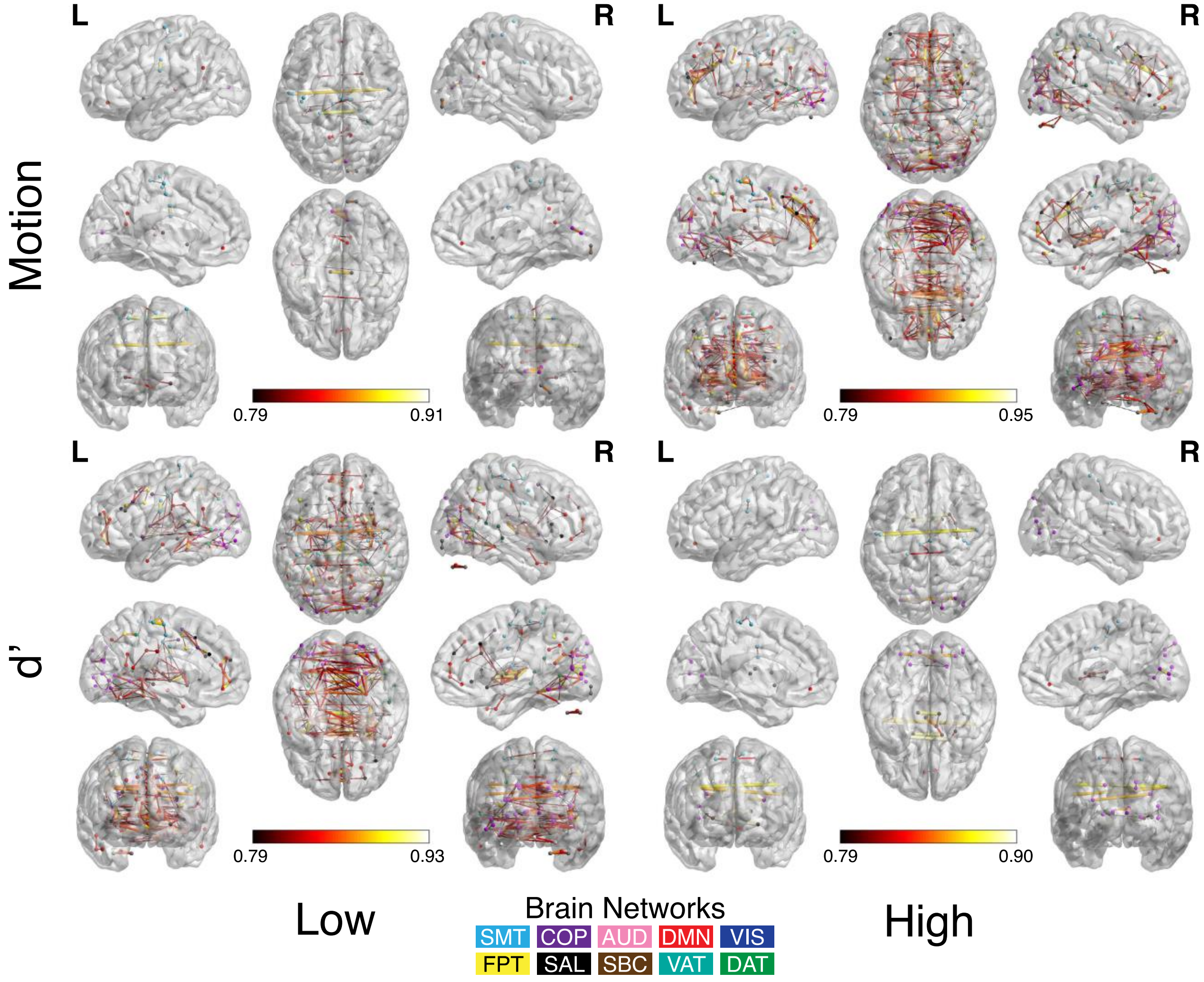
- To generate realistic heteroskedasticity we bootstrapped from N-back data
- In a subset of 1000 subject from the PNC, we fit the model

$$
\begin{aligned}
Y_i(v) = \alpha_0(v) + \alpha_1(v) \times \mathrm{sex}_i \\
+f(v, \mathrm{age}_i) + g(v, \mathrm{mot}_i) + h(v, d_i') + \epsilon_i(v)
\end{aligned}
$$

- The sample mean of $\hat{\epsilon}\ _i(v)$ is then independent of the covariates
- The covariance function $\mathrm{Cov}\{\hat{\epsilon}\ _i(v), \hat{\epsilon}\ _i(w)\}$ may be affected by covariates

The covariance function is affected by covariates in the N-back sample

- $\text{Cov}\{Y_i(v), Y_i(w) \mid X_i\} \neq \text{Cov}\{Y_j(v), Y_j(w) \mid X_j\}$ implies heteroskedasticity



Intuition behind simulation procedure

- Treat residuals $\hat{\epsilon}_{\ i}(v)$ of 1000 subjects as population under the null
- In each simulation draw a bootstrap sample and fit the model

$$\hat{\epsilon}_{\ ib}(v) = \alpha_{0b}(v) + \alpha_{1b}(v) \times \mathrm{sex}_{ib} + \alpha_{2b}(v) \times \ \mathrm{age}_{ib} + \alpha_{3b}(v) \times \mathrm{mot}_{ib}$$
$$+\beta_b(v) \ \times \ d'_{ib} + \epsilon_{ib}(v)$$

- Perform test of $H_0(v)\colon \ \beta_b(v) = 0$ and $H_0(v)\colon \ \alpha_{3b}(v) = 0$ to assess effect of heteroskedasticity on FWER of SEI procedures

Procedures we will compare

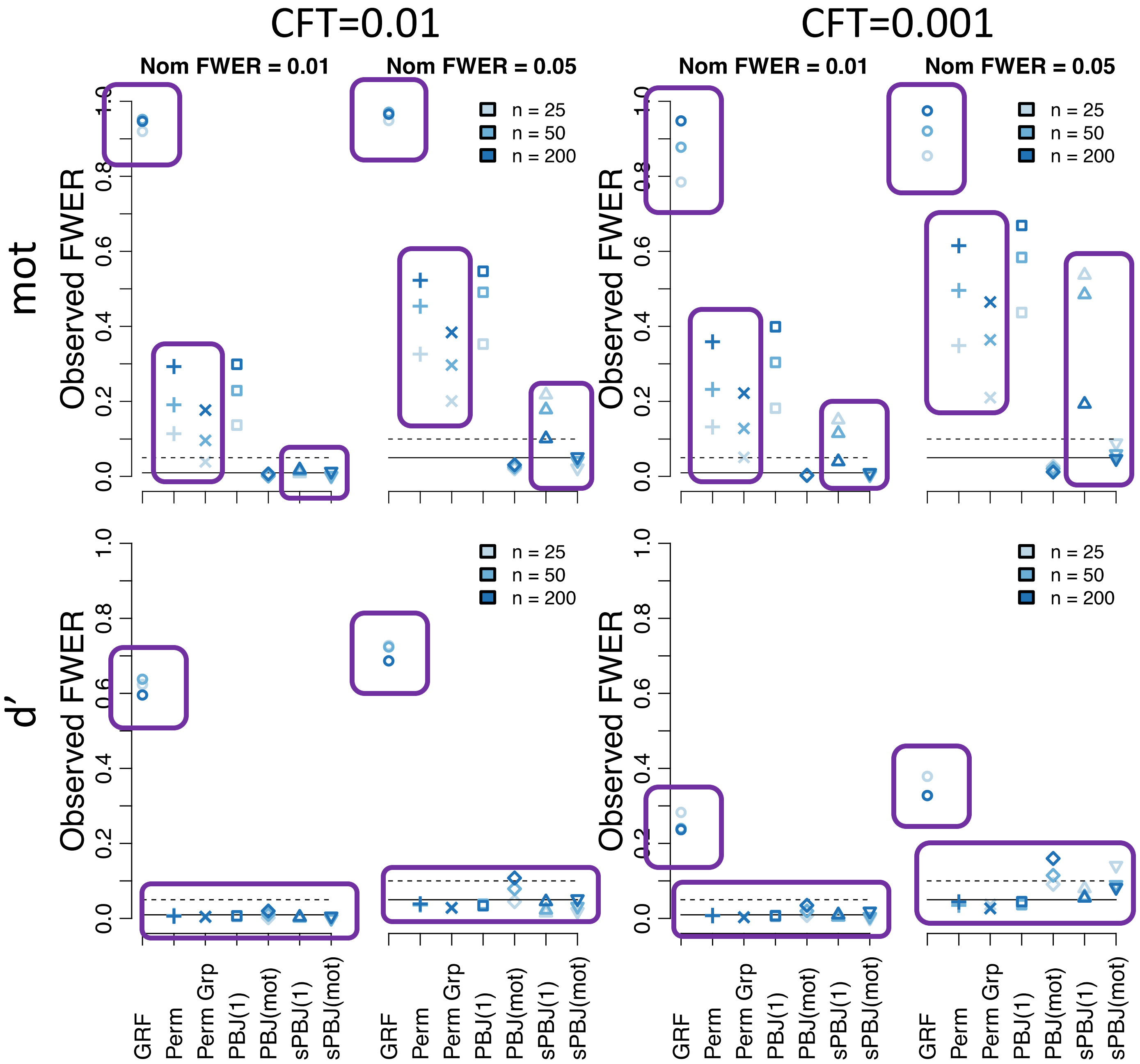
- GRF - classical GRF based method assuming local correlation in field
- Perm - permutation procedure assumes exchangeability
- Perm Grp - permutation procedure, attempts to adjust for non exchangeability
- PBJ - parametric bootstrap joint procedure [PBJ(1), PBJ(mot)]
- sPBJ - semiparametric bootstrap joint procedure [sPBJ(1), sPBJ(mot)]

Friston et al., *Neuroimage*, 1996
Winkler et al., *Neuroimage*, 2014
Winkler et al., *Neuroimage*, 2015
Vandekar et al., *Biometrics*, 2019

Heteroskedasticity simulation: type 1 error rates

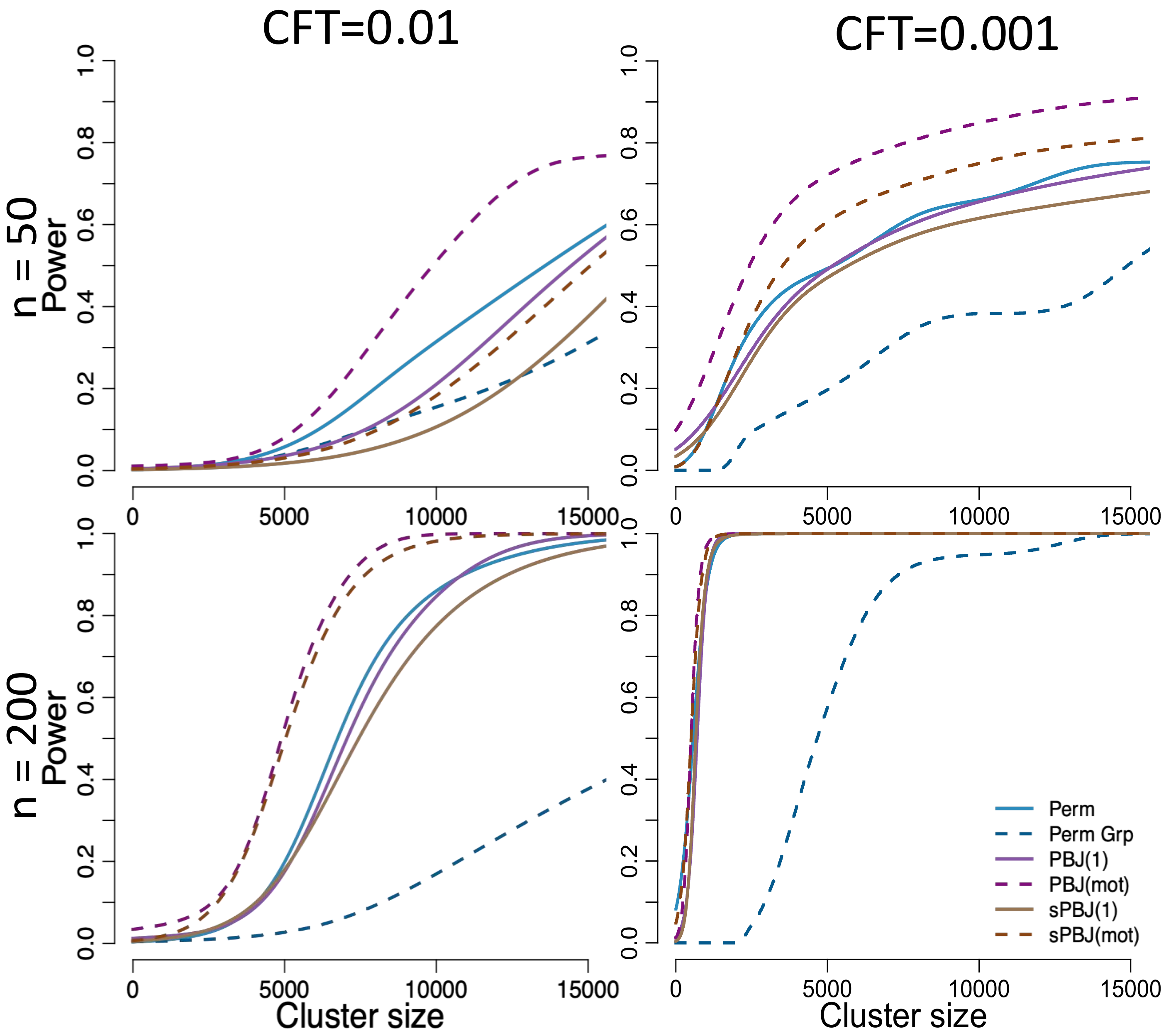
- For the test of the motion covariate
  - GRF near 100% error rate
  - Permutations and unweighted PBJ have inflated error rate
  - Robust methods have near nominal performance at less conservative CFTs
- For the test of the d' covariate
  - GRF still quite high
  - Other methods near or approach nominal level
- Next slide: power analysis

Vandekar et al., *Biometrics*, 2019



Heteroskedasticity simulation: power results

- Power results for the test of the d' covariate for a rejection threshold of 0.01
    - Type 1 error rates near the nominal level
- Motion deweighting may improve power here

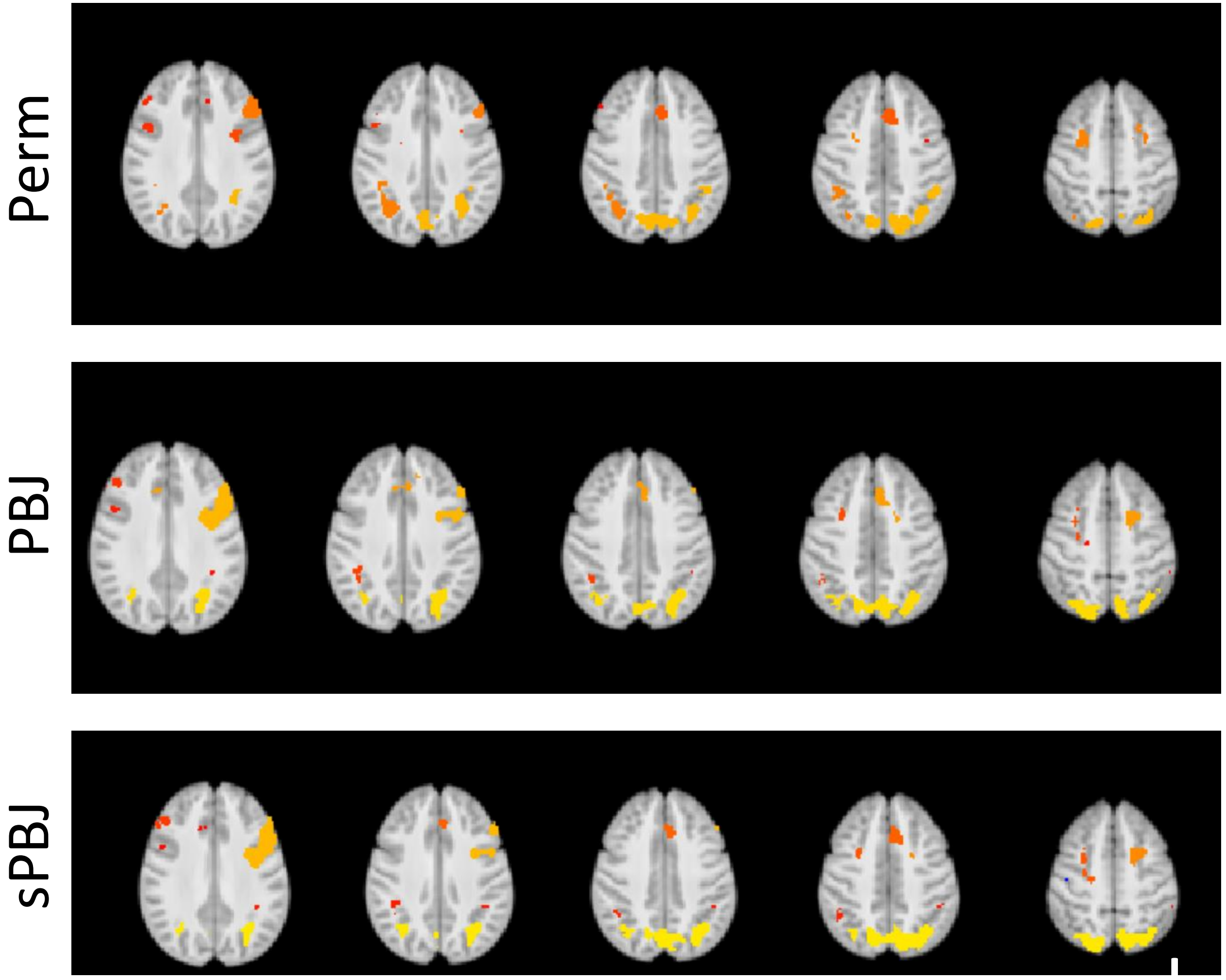


N-back data analysis

- We used a random subset of 200 subjects to evaluate the different procedures
- Recall our goal: **to identify regions where WM performance is associated with WM activation**
- The PBJ and sPBJ procedures used voxel-wise weights proportional to the inverse of the subject variance image estimates.
- Use SEI to compute p-values for each cluster

N-back data analysis results

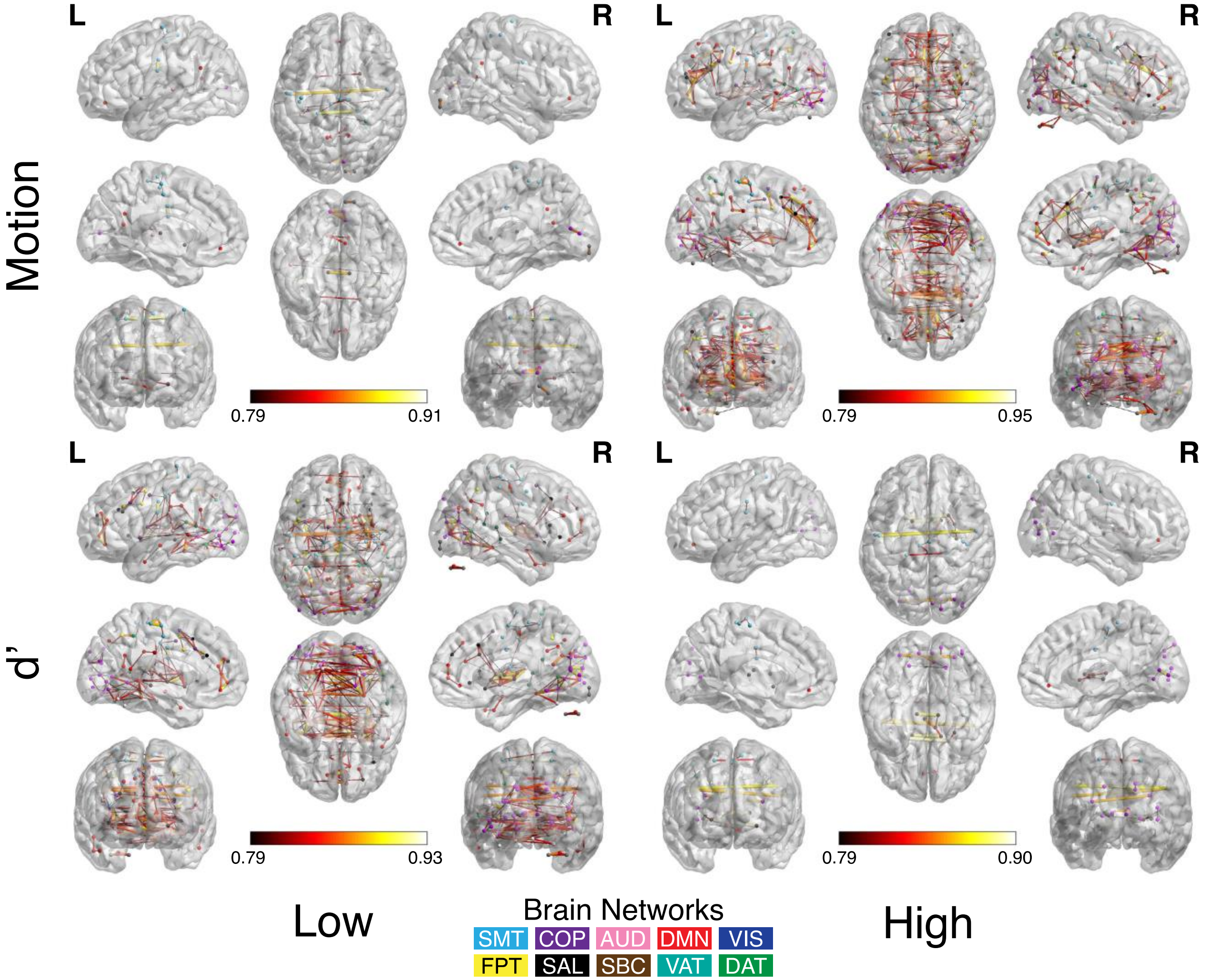
- The sPBJ have smaller p-values in some regions
- Small p-values indicate cluster sizes that are unlikely under the global null $H_0(v)$: $\beta_b(v) = 0$, for all v



-log$_{10}$(p)

The covariance function is affected by covariates in the N-back sample

- $\mathrm{Cov}\{Y_i(v), Y_i(w) \mid X_i\} \neq \mathrm{Cov}\{Y_j(v), Y_j(w) \mid X_j\}$ implies heteroskedasticity
- Functional connectivity is evidence of heteroskedasticity



Comparing assumptions

	GRF	FLAME	Permutation	sPBJ
Assumes homoskedasticity	😩	😃 / 😩	😃 / 😩	😃
Uses normal approximations	😩	😃	😃	😩 / 😃
Repeated measurements	😃	😃	😩 / 😃	😃
Robust to model misspecification	😩	😩	😩	😃
Uses some kind of approximation	😩	😃 / 😩	😩	😩

*Methods overlap software packages, e.g. permutation is available in SPM and AFNI, but not as the default

TFCE

TFCE

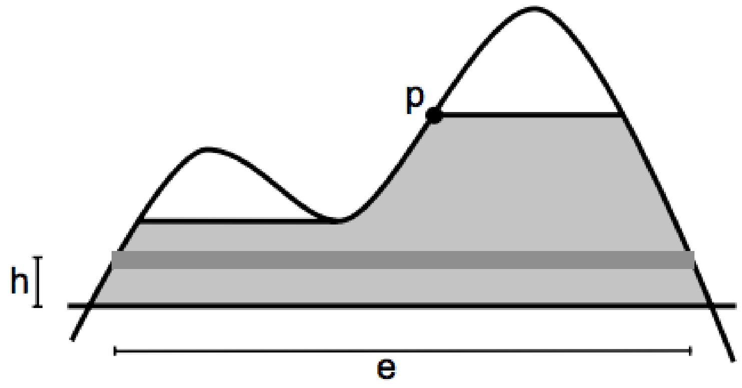
Threshold-Free Cluster Enhancement

[Smith & Nichols, NeuroImage 2009]

- Cluster thresholding:
  - popular because it's sensitive, due to its use of spatial extent
  - but the pre-smoothing extent is arbitrary
  - and so is the cluster-forming threshold

    ➡ unstable and arbitrary

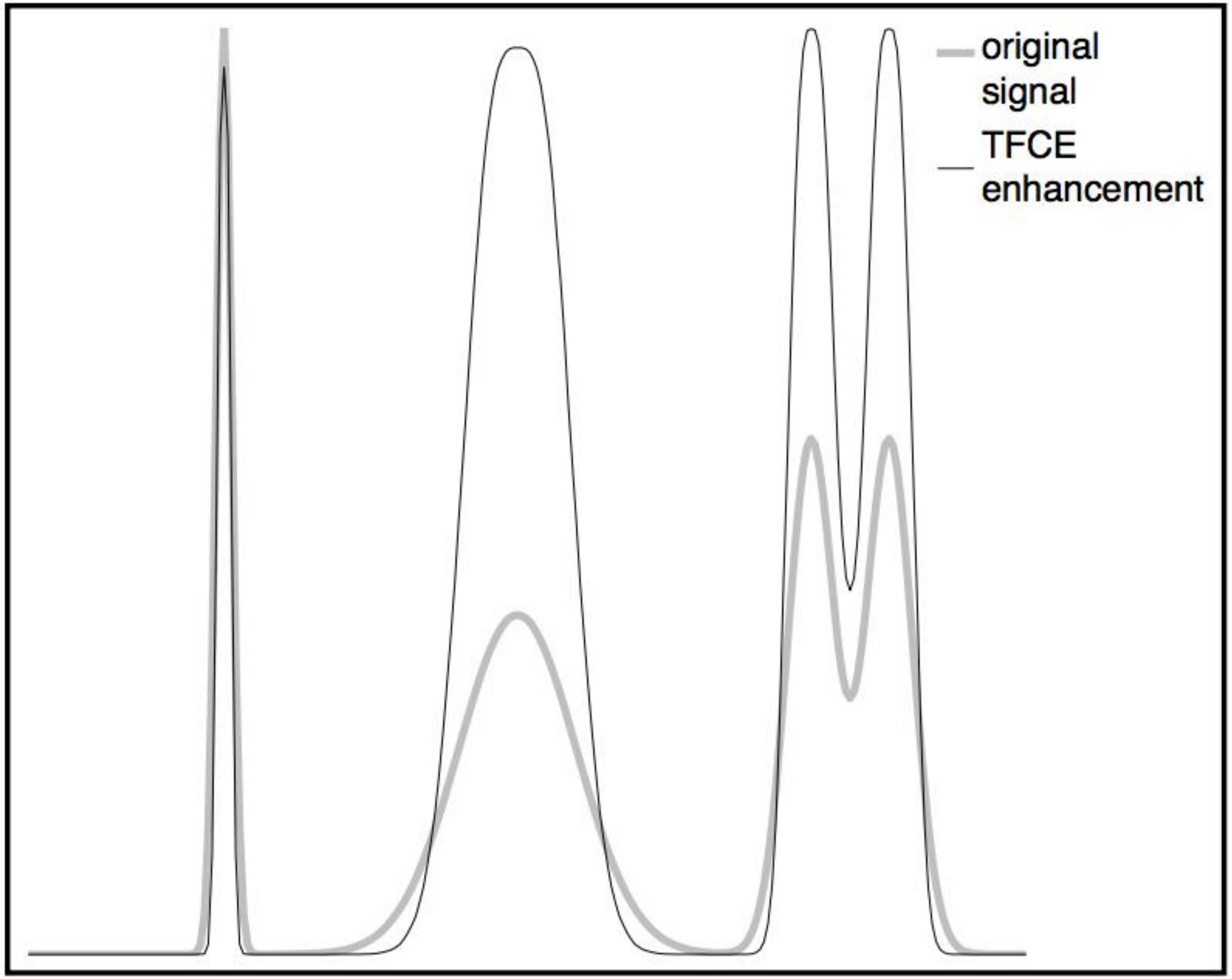
- TFCE
  - integrates cluster "scores" over all possible thresholds
  - output at each voxel is measure of local cluster-like support
  - similar sensitivity to optimal cluster-thresholding, but stable and non-arbitrary



The TFCE value at point p is given by the sum, over the shaded area, of the score from each contributing incremental section:

$$\mathrm{TFCE}(p) = \sum_h e(h)^E \cdot h^H$$

Qualitative example



TFCE for FSL-VBM