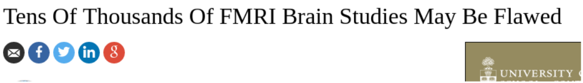
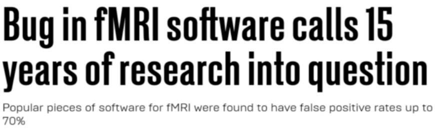
Cluster Failure: fMRI's Big Shake-Up

Chris Hammill

September 13, 2016

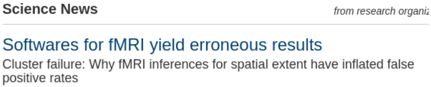


Tens Of Thousands Of FMRI Brain Studies May Be Flawed



Bug in fMRI software calls 15 years of research into question

Popular pieces of software for fMRI were found to have false positive rates up to 70%

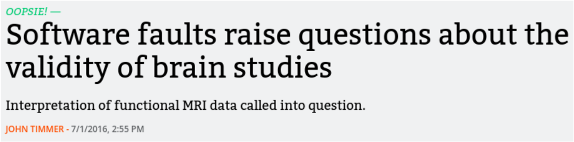


Science News

from research organiz

Softwares for fMRI yield erroneous results

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



OOPSIE! —

Software faults raise questions about the validity of brain studies

Interpretation of functional MRI data called into question.

JOHN TIMMER - 7/1/2016, 2:55 PM

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]

So What Happened

- Eklund, Nichols, and Knutsson demonstrated standard fMRI statistical inference has badly inflated false positives rates
- Makes you wonder if exciting brain region X responding to stimulus Y finding was just a cherry-picked false positive.
- Highlighted that due to non-reproducible workflows, and poor data sharing, many of these finding could never be repeated with valid inference.

How Did We Get Here

- fMRI is challenging to analyze
- Preprocessing steps widely used as black boxes
- Desire to use spatial information to determine signal significance
- Improperly specified models of spatial noise

About Group Task-Based fMRI

- Most fMRI seeks to measure brain activity by blood flow
- Blood oxygen level dependent (BOLD) contrast
- A time-series of volumes are acquired for each subject
- Stimuli are presented to the subject throughout the time series
- The BOLD signaled is modelled as a function of the stimuli
- The statistical association of the BOLD to the stimuli is compared across groups

Why Is This Tough

- Subjects move:
  - within subject each fMRI volume must be aligned to each-other
  - these must be aligned to a corresponding anatomical scan
  - these must be registered to a common space
- BOLD signal is sluggish
  - $\sim$ 2 seconds to start
  - $\sim$ 4-6 to peak
  - $\sim$ 10 to return to baseline so the stimulus time series is convolved with a function to match this behaviour
- Analyzing time series comes with it's own statistical challenges
  - how do we model temporal autocorrelation

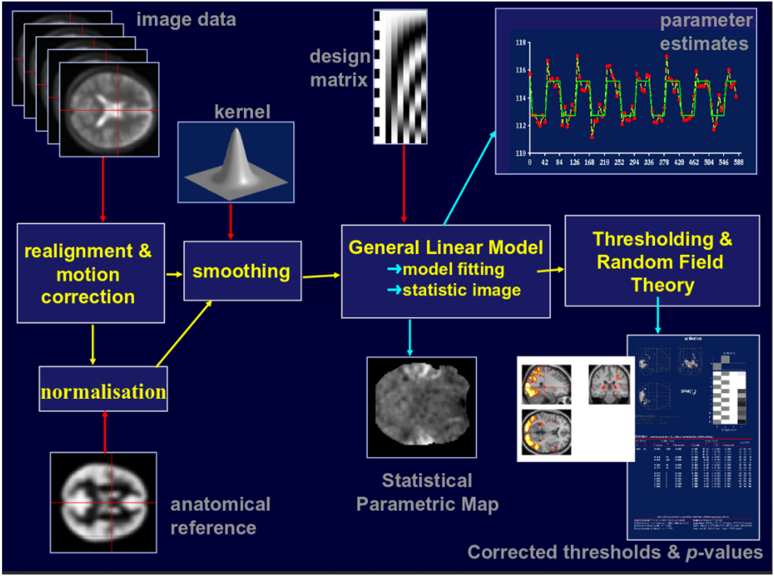


image data

design matrix

kernel

parameter estimates

realignment & motion correction

smoothing

General Linear Model
→model fitting
→statistic image

Thresholding & Random Field Theory

normalisation

anatomical reference

Statistical Parametric Map

Corrected thresholds & *p*-values [1]

[1] Borrowed from Nichols (2010)

Multiple Comparisons

- As with most imaging analysis, multiple comparisons is significant concern
- Solutions:
  1. Bonferroni: control your type one error rate by multiplying your results by the number of tests. This is equivalent to setting your type one error rate to α/n
  2. FDR (Benjamini-Hochberg): Order your p-values lowest to highest and accept or reject with increasing stringency α/i.
- But in low power situations this decreases sensitivity an unacceptable amount

Enter Spatial Models

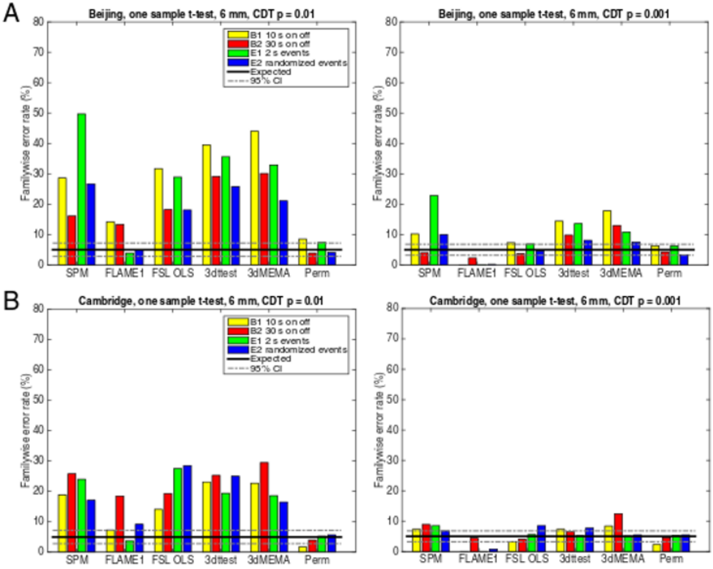
- Signals with large spatial extent are probably more likely to be real than individual high intensity
- Question becomes, how do we analyze spatial extent, and how do we correct for multiple comparisons?
- Main Idea: threshold your data and use random field theory (RFT) results to assign a p-value to clusters based on their size
- Or: Assume some properties of the spatial distribution and generate a randomization distribution of cluster sizes, assign p-values from this.

The Problems

- When statistics maps aren't smooth enough, RFT p-values are biased (2003)
- RFT typically assumes a stationary noise distribution (uniform noise over the brain) which is often invalid (2004)
- Together these problems can lead to 70% FWE rates in single subject analyses (2012)

The Paper

- In order to assess how much these problems matter for group comparisons, check the null distribution
- The authors took a large open data set with a pool of neurotypical subjets, and randomly sampled groups to compare
- If after processing and multiple comparison correction any clusters in the brain were significant that test was a false positive (error).
- The distribution for a two group difference should be Student's t distribution, and after bonferroni correction, the expected proportion of errors should be 5%
- Higher error rates imply the multiple comparison correction is insufficient.
- Five analysis functions from the three most popular fMRI software packages were compared to their non-parametric alternative


A
Beijing, one sample t-test, 6 mm, CDT p = 0.01
Beijing, one sample t-test, 6 mm, CDT p = 0.001
B
Cambridge, one sample t-test, 6 mm, CDT p = 0.01
Cambridge, one sample t-test, 6 mm, CDT p = 0.001
Familywise error rate (%)
0
10
20
30
40
50
60
70
80
SPM
FLAME1
FSL OLS
3dttest
3dMEMA
Perm
B1 10 s on off
B2 30 s on off
E1 2 s events
E2 randomized events
Expected
95% CI

{

width=80% }

The results

- All parametric tools produce FWE higher than 5%
- Situation is more extreme when cluster defining thresholds are high (FWEs ∼20-40)
- Different data sets are affected differently (Beijing less affected than Cambridge)