Challenges in specifying effect sizes for hypothesis testing in fMRI

Jasper Degryse[a], Ruth Seurinck[a], and Beatrijs Moerkerke[a]

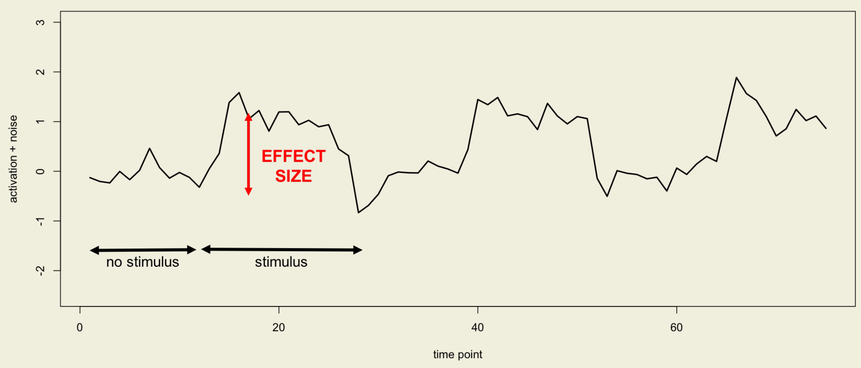
[a] Department of Data Analysis, Ghent University, Belgium

UNIVERSITEIT GENT

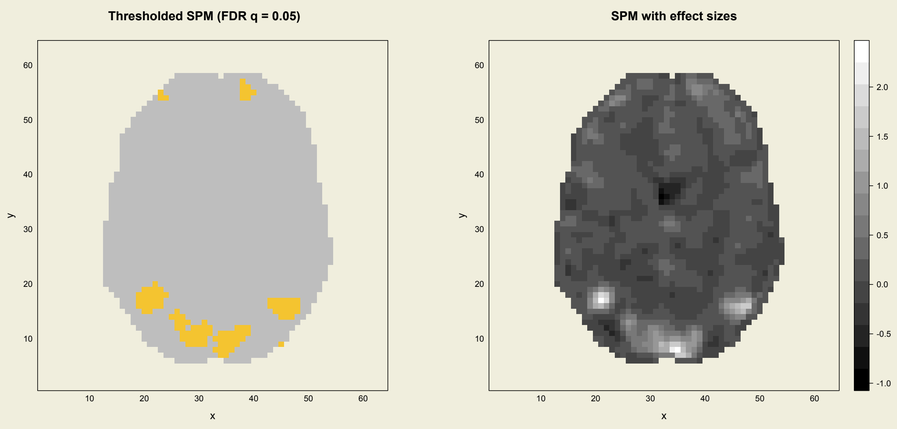
fwo Research Foundation Flanders Opening new horizons

1. Introduction

- fMRI studies are commonly evaluated using classical statistical null hypothesis testing (NHST)
  - NHST: only focus on testing against null of no activation (H_0)
  - Statistical significance $\neq$ functional relevance
- Growing awareness of importance of effect sizes (ES; magnitude of effect):
  - represents how active the voxel is during the task
  - often expressed as % BOLD signal change
  - related to underlying neurological process



- A priori specification of an expected ES is essential for **power** calculations, but this ES can also be incorporated into testing to increase sensitivity (e.g., alternative-based thresholding[1]).
- Other methods use the data to estimate an appropriate ES to include in testing (non-exhaustive list):
  - **Likelihood ratio (LR) testing:**[2] ES used for testing estimated as a specific percentile of observed ESs over voxels (e.g. 95th percentile).
  - **Amplitude thresholding:**[3] the functionally relevant ES is the magnitude that, when used as a threshold, results in the equal number of voxels as an analysis with thresholding using NHST (e.g. uncorrected with $p < 0.001$).
  - **Regions in limbo:**[4] ES used for testing is that of the voxel with the smallest ES in a cluster-thresholded SPM.
- In the current study we evaluate the influence of the ES estimation method on the performance of a promising alternative for NHST, the LR testing method[2].



2. Methods

- LR statistic (Kang et al., 2015):
  - l_1: likelihood of the data given the estimated functionally relevant ES
  - l_0: likelihood of the data given an ES of 0
  - LR statistic: l_1/l_0
  - Thresholded using pre-specified value $k \geq 1$. If the LR statistic > 1, there is more evidence in favor of the alternative as opposed to the null.
  - This method does not necessarily use input from an a priori thresholded SPM in contrast to the methods of de Hollander et al. (2014) and Gross & Binder (2014).
- We consider the following methods to estimate an ES to include into the LR testing method:
  1. A pre-defined (e.g., 95) percentile of all ES as proposed in the original LR testing method of Kang et al. (2015).
  2. The ES used for amplitude thresholding as in Gross & Binder (2014).
  3. The mean ES of the active voxels after thresholding with different traditional methods: uncorrected $p < 0.001$, FDR-control, FWER-control and cluster extent corrected (circular). We compare with the various traditional thresholding methods mentioned above (4).

3. Results of the LR test with k=8

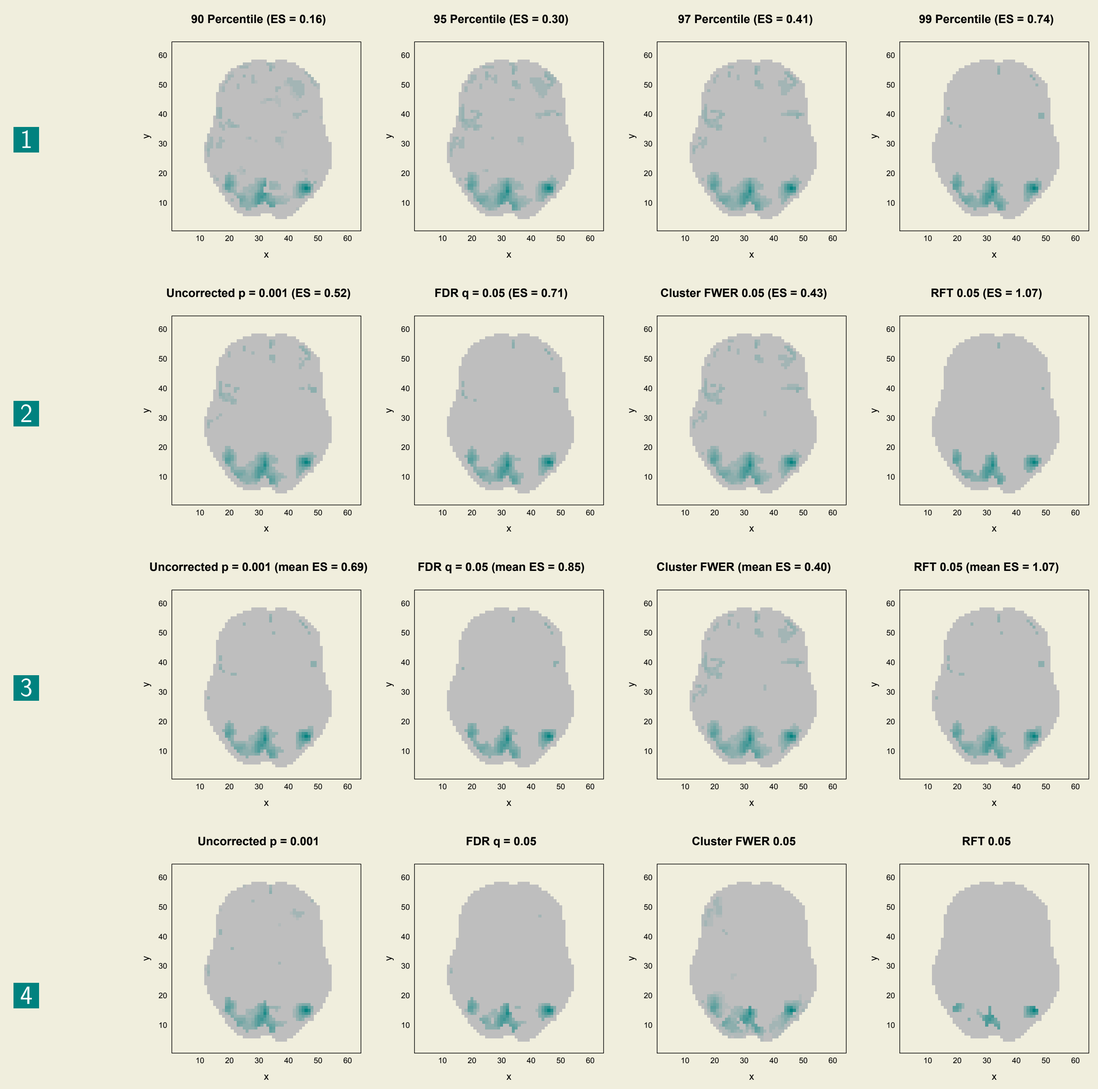


Figure: Data from a motion localizer task to idenitify hMT/V5+[5]. The greener the voxel, the larger its ES.

Conclusions

- Applying the LR testing method with the mean ES after various traditional thresholding procedures results in an SPM with different voxels that have larger, more functionally relevant ESs as compared to the thresholded methods as shown in (4).
- The benefit of the LR testing method is its independence from thresholding prior to estimating the functionally relevant ES.
- However, the estimated ES is sensitive to the amount of activation as demonstrated with simulations as shown here for a single voxel:

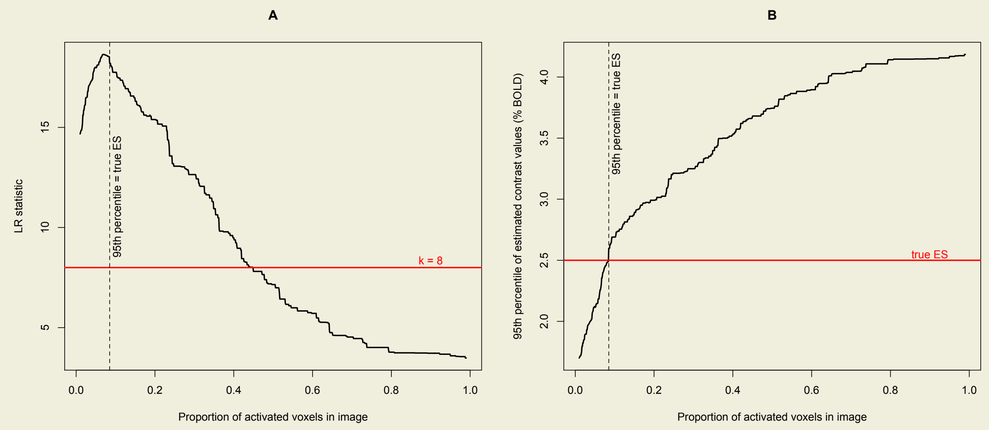


Figure: In this simulation study, the ES for active voxels was 2.5% BOLD signal change. The definition of the functionally relevant ES and the LR statistic is highly dependent on the proportion of active voxels in the brain. The LR testing method performs well in most situations, also with small judgmental errors about the amount of activation. Larger errors could result in inconclusive findings.

5. Discussion

- While including an ES into test criteria offers a more balanced view on results with respect to functional relevance, defining this is challenging and highly impacts results.
- Anatomical a priori areas as a means to define which voxels to base (independent) ES estimation on.
- The same challenges are faced in power calculations.
- High-impact and large-scale open-source projects can provide insight, effectively helping in the definition of functional relevant ESs while avoiding the need to collect data in addition to the experimental data.
- Use of independent data to provide a priori ES: more robust results?

6. References

1 Durnez, et al. (2013). *Cognitive, Affective, & Behavioral Neuroscience*
2 Kang, H., Blume, J., Ombao, H., & Badre, D. (2015). *NeuroImage*
3 Gross, W. L., & Binder, J. R. (2014). *NeuroImage*
4 de Hollander, et al. (2014). *PloS one*
5 Seurinck, et al. (2011). *Journal of Cognitive Neuroscience*