Pitfalls of Thresholding Statistical Maps in Presurgical fMRI Mapping

K. Gorgolewski[1], M. Bastin[2], L. Rigolo[3], H. A. Soleiman[4], C. Pernet[2], A. Storkey[1], and A. J. Golby[3]
[1]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom, [2]Department of Medical Physics, University of Edinburgh, Edinburgh, United Kingdom, [3]Department of Neurosurgery, Harvard Medical School, Cambridge, MA, United States, [4]Department of Clinical Neurosciences, University of Edinburgh, Edinburgh, United Kingdom

Introduction

There is significant variability in the selection of methods used to threshold fMRI activation maps acquired for brain tumour presurgical planning. In a review of 50 recent papers, only 12% of studies claimed to use some kind of Family Wise Error (FWE) correction for multiple comparison testing. 42% of these studies used a p-value threshold lower than the standard 0.05 in an attempt to minimize the number of false positives, while 22% did not use the same threshold for all of the subjects, with the threshold being manually adjusted on per subject basis. What is more, most of the studies used simple thresholds without taking the spatial properties of the statistical maps into account; only 4% used cluster size as an additional threshold. Although thresholding methods used in neuroscience fMRI studies have greatly improved in the last few years, there still remains a lack of consensus in the clinical literature about how to identify activation boundaries accurately and objectively. As a first step towards developing robust frameworks for assessing thresholding methods for tumour resection, we investigated how several automated thresholding methods affect the distance between the activation areas determined from fMRI experiments and the tumour boundary defined on structural MRI, and how this data might potentially change surgical practice.

	SPM	FSL	Manual
Patient 1	0.23 % overlap	0.10 % overlap	8.4 mm
Patient 2	2.46 % overlap	0.18 % overlap	3.07 % overlap
Patient 3	2.82 % overlap	0.16 % overlap	2.39 % overlap
Patient 4	3.58 % overlap	0.01 % overlap	0.08 % overlap
Patient 5	11.87 mm	25.61 mm	29.18 mm
Patient 6	13.52 % overlap	0.82 % overlap	24.91 % overlap
Patient 7	1.68 % overlap	0.27 % overlap	0.48 % overlap
Patient 8	1.83 % overlap	12.4 mm	18.31 mm
Patient 9	2.38 % overlap	0.18 % overlap	.02 mm
Patient 10	7.14 mm	0.00 % overlap	7.28 mm
Patient 11	0.01 % overlap	0.00 % overlap	7.17 mm

Table 1: Distance between the edge of the activation areas and the tumour margin. Dice's similarity measures are calculated in case of an overlap.

Methods

11 patients with primary brain tumours situated near motor cortex underwent a hand clenching fMRI task. The hand used was always contralateral to the tumour location. After fitting a GLM model, t-values were calculated for every voxel. Those t-maps were thresholded in three ways: (i) manually by an expert rater with additional cluster extent threshold of 10 voxels, (ii) using SPM8 with a cluster forming threshold of 0.05 FWE corrected and False Discovery Rate (FDR) with Random Field Theory (RFT) of clusters size lower than 0.05 (SPM)[1], and (iii) using FSL and Spatially Regularized Mixture Models with a 0.5 probability threshold of belonging to the activation class (FSL)[2]. For each subject, the distance between the activation area and the manually segmented tumour was determined by an image analyst and verified by a neurologist. In the case where the tumour margin and the activation area overlapped, a Dice's similarity coefficient was calculated (see Table 1).

Determining what the appropriate distance from the tumour margin is for surgical resection based on fMRI activation maps is not straightforward [3]. We therefore determined how a hypothetical 'safety margin' of 5, 10, 15 and 20 mm would influence the clinical procedure. Specifically, if the activation region is further from the tumour than the safety margin then full resection is recommended, otherwise a partial resection should be performed. Additionally, we have calculated the theoretical statistical properties of the thresholds generated by the expert rater.

Results and Discussion

Figure 1 shows two example cases. Table 1 shows tumour distance and overlap for all subjects. Table 2 shows that if the safety margin is large enough (20 mm) both automated thresholding methods perform similarly to manual thresholding, which we assume is the 'gold standard'. However, if the safety margin is reduced to 5 mm almost one third of the cases are classified differently; automated methods tend to produce larger activation regions leading to a partial resection recommendation. Additionally statistical analysis of theoretical properties of the manual thresholds shows that the rater chooses thresholds to minimise the number of false positive voxels, i.e. FWE corrected p-values lower than 0.002 and FDR lower than 0.0003. However, the clusterwise statistics show that the FDR values in two cases are surprisingly high (Table 3). This is due to a fixed cluster size threshold of 10 voxels which does not take into account the height of the cluster forming threshold. This problem can be addressed using modern thresholding methods based on RFT which estimate the expected cluster size for a given cluster forming threshold [4].

	SPM	FSL	Manual
5 mm	81.82%	81.82%	54.55%
10 mm	90.91%	81.82%	81.82%
15 mm	100.00%	90.91%	81.82%
20 mm	100.00%	90.91%	90.91%

Table 2: Percentage of cases with partial resection recommendation using different safety margins.

The conservative approach of using very high p-values, both in the reviewed papers and presented data, results in very few if any false positives. This, however, means that the number of false negatives (falsely claiming that a piece of tissue is not involved in the particular cognitive task and can be safely removed) will also be very high. Neuroscience approaches focus on controlling the number of false positives to ensure that the reported findings, if any, are true. In neurosurgery we are in fact interested in the opposite question – which parts of the brain are not involved in a particular task and can be safely removed. The standard statistical methods for testing differences, like the t-test, do not indicate whether data sampled from two populations are the same, just whether they are different. In other words if no significant difference has been found in the sampled data we cannot claim that the two populations are the same, i.e. we do not have statistically significant proof that they are different. Further research is needed to investigate the use of equality testing in presurgical mapping using fMRI data to create "safety maps", rather than mapping functional brain regions. It is also important to note that as long as the person making clinical decisions based on thresholded maps, presumably the neurosurgeon, understands how they were prepared he or she can adjust the safety margins appropriately. In case of manual thresholding even if it is very conservative it can be a successful base for procedure planning as long as the expert thresholding it is consistent.

The problem becomes more significant when the fMRI cortical mapping is prepared by an outside centre. Such situations call for standardization in the way thresholded fMRI activation maps are prepared and data to inform the neurosurgeon about the statistical properties of the presented data, such as the expected number of false positives and false negatives.

	thr	P Bonf	P RF	FDR	cFDR
Patient 1	10	0	0	0	0
Patient 2	9	0	0	0	0
Patient 3	10.3	0	0	0	5E-06
Patient 4	10	0	0	0	0
Patient 5	6	0.000622	0.0019	4E-05	0.0045
Patient 6	13.1	0	0	0	0.803
Patient 7	17.5	0	0	0	0.7887
Patient 8	7	0.000006	4E-05	0	1E-06
Patient 9	10	0	0	0	0
Patient 10	5.7	0.0018	0.0054	0.0002	0.0029
Patient 11	9	0	0	0	0

Table 3: Statistical properties of the manual thresholds. FWE corrected voxelwise P – using Bonferronni correction (P_Bonf) or Random Field Theory based correction (P_RF), voxelwise FDR (FDR), and clusterwise FDR (cFDR).

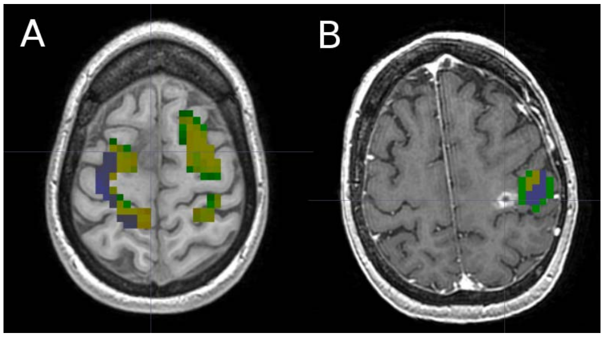

Figure 1: Patient 1 (A) and 10 (B). Voxels are colour coded by the thresholding method they were included by: FSL (green), FSL+SPM (yellow), FSL+SPM+manual (purple).

[1] 1. Chumbley et al. *Neuroimage*. 2009;44(1):62-70; [2] Woolrich et al. *Medical Imaging, IEEE Transactions on*. 2005;24(1):1-11; [3] 1. Håberg et al. *Neurosurgery*. 2004;54(4):902; [4] Friston et al. *Human Brain Mapping*. 1993;1(3):210-220;