6 Evaluating statistical procedures using different signal sources: a case study with Alternative-based thresholding.[1]

Abstract Statistical inference in cognitive neuroscience focuses on stringent control of false positives, accepting the concomitant sacrifices in sensitivity. However, this is accompanied by a risk of false negatives, which can be detrimental, for example, in clinical settings where false negatives may lead to surgical resection of vital brain tissue. We have recently presented a new hypothesis thresholding procedure that incorporates information on both false positives and false negatives [2]. The result is a layered statistical map, marked by voxels exhibiting (i) strong evidence against the null hypothesis, (ii) evidence against the null but at practically insignificant effect sizes, (iii) responses where activation cannot be confidently excluded and finally (iv) responses where activation can be rejected.
Statistical significance testing can be evaluated by assessing the overlap between functional activations and structural connectivity. To compare our procedure with classical significance testing, we assess the difference between alternative-based testing (ABT) and classical hypothesis testing (CHT) using cross-correlations and overlap between activation and structural connectivity profiles [3]. The approach is exemplified in a patient undergoing presurgical mapping and tractography.

6.1 Introduction

When surgically resecting brain tumors, one wants to minimize risk of resecting brain tissue involved in important cerebral functions. Pre-surgical fMRI probes such functions to localize eloquent brain tissue. Statistical inference in cognitive neuroscience focuses on control of false positives. The scientific discipline deems stringent control of false positives necessary, accepting

[1] This chapter represents collaborative work with following authors: Durnez J., Homola G., Jbabdi S., Nichols T., Moerkerke B. and Bartsch A.

the concomitant sacrifices in sensitivity. In a clinical setting, a loss in power means true activation is not discovered, and this might result in the resection of vital brain tissue. This asymmetrical way of penalising errors in statistical inference is undesirable in this context. We therefore presented a new hypothesis thresholding procedure that incorporates information on both false positives and false negatives and thus is ideally suited for pre-surgical fMRI (Durnez et al., 2013).

When we test hypotheses, we test $H_0 : \Delta = 0$ against $H_a : \Delta = \Delta_1$, where Δ is BOLD effect of interest in units of percent BOLD change and Δ_1 the non-zero effect magnitude expected under activation. In classical hypothesis testing, the evidence against H_0 is measured with the p-value, the null hypothesis probability of data as or more extreme than that observed. Thresholding a p-value at α produces a statistical test that controls the false positive rate at α. To allow direct control of false negative risk, we present a symmetrical measure p_1 which quantifies evidence against the H_a. Thresholding this probability measure at β ensures control of the false negative rate at β.

We measure the evidence against H_0 with $p_0 = P(T \geq t|H_0)$ and the evidence against H_a with $p_1 = P(T \leq t|H_a)$. We don't expect a single magnitude of true activation, but $\Delta 1$ also follows a distribution: $\Delta_1 \sim \mathcal{N}(\mu, \tau^2)$. Consequently:

$$T_i \sim \mathcal{N}\left(\frac{\mu}{\mathrm{SE}(\widehat{\Delta}_i)}, \frac{\mathrm{SE}(\widehat{\Delta}_i)^2 + \tau^2}{\mathrm{SE}(\widehat{\Delta}_i)^2}, \right) |H_a \tag{6.1}$$

When thresholding p_0 and p_1 at respectively level α and β, with for given value of μ (expected activation) and τ (its uncertainty), the result is a layered statistical map, marked by voxels exhibiting (i) strong evidence against the null hypothesis, (ii) evidence against the null but at practically insignificant effect sizes, (iii) responses where activation cannot be confidently excluded and finally (iv) responses where activations can be rejected. To compare our procedure with classical significance testing, we assess the difference between alternative-based testing (ABT) and classical hypothesis testing (CHT) using cross-correlations and overlap between activation and connectivity (Homola et al., 2012).

The approach is exemplified in a patient undergoing presurgical mapping and tractography.

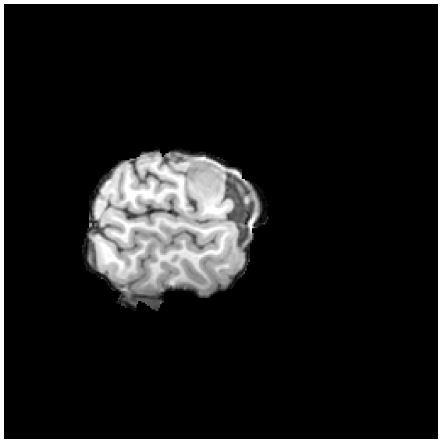

Figure 6.1: T_1-weighted scan of the patient.

6.2 Methods

6.2.1 Data

We consider data from a patient with left frontal grade II oligodendroglioma. The lesion is an intra-axial space-occupying lesion of 26 (A-P) $\times$ 29 (L-R) $\times$ 32 (V-D) mm extension behind the coronary suture entered to the left percentile sulcus. Figure 6.1 shows the tumor in the left hemisphere of the brain.The patient is right-dominant for hand, foot and eye. Data are obtained with a 3T TimTrio (Siemens, Erlangen, Germany) , 32 channel head coil.

6.2.2 Preprocessing and statistical analysis of fMRI data

In order to locate anterior and posterior language regions, the patient underwent two fMRI experiments: (1) reading nonfinal embedded clause sentences versus rest (semantic language-comprehension task and (2) detection of words from pseudowords versus blocks of tones (phonological language task. The first contrast aims to detect the posterior language area, while the second contrast refers to the anterior language area. The data are processed using FSL 5.0.6 (http://www.fmrib.ox.ac.uk/fsl/). As preprocessing, the data are motion-corrected using mcflirt, pre whitened using film and smoothed with a Gaussian kernel with 5mm full width at half maximum (voxelsize $\times$ 3 $\times$ 3.45 mm). The data are transformed to T_1-space before analysis using FSL's FLIRT tool, where the transformation matrix is computed based on the transformation from the mean EPI-image to the T_1-image. First, regions-of-interest (ROI) are defined. For the posterior language area (1st paradigm),

we restrict analysis to the anterior devision of the middle and superior temporal gyrus and the supramarginal gyrus. The analysis of the anterior language area (2nd paradigm) is restricted to the inferior and middle frontal gyrus. The first-level analysis is carried out by applying a GLM within FEAT within these regions-of-interest. From the GLM we derive T-statistic images and a p-value for each voxel. We apply the testing procedure presented above, in which four layers of activation are defined. For the α-parameter, we choose the cut-off that controls the false discovery rate within the ROI. The beta-parameter is set to 0.20, to provide an average statistical power of 80%. For the alternative distribution, we aim at effects of size $\mu = 0.50$ (0.5 % BOLD change) with variation $\tau = 0.01$. After analysis with the alternative-based procedure, four voxels are labeled according to the following labeling scheme:

- **NON-ACTIVE LABEL:** $p_0 > \alpha \cap p_1 < \beta$**:** Activation can be confidently excluded.
- **ACTIVE LABEL:** $\boldsymbol{p_0 \leq \alpha \cap p_1 \geq \beta}$. The voxels show strong evidence against the null hypothesis.
- **UNCERTAINTY LABEL:** $p_0 \geq \alpha \cap p_1 \geq \beta$**:** Voxels cannot be confidently declared inactive.
- **PRACTICAL INSIGNIFICANT LABEL:** $p_0 \leq \alpha \cap p_1 \leq \beta$**:** These voxels show evidence against the null but at practically insignificant effect sizes.

To compare classical hypothesis testing (CHT) with alternative-based testing (ABT), we define the significant result based on these layers. Classical testing results comprise the **active** layer and the **practically insignificant** layer (which just corresponds to using a FDR corrected threshold at 5%. The *significant area* for the alternative-based thresholding procedure is the **active** layer and the **uncertain** layer.

6.2.3 Preprocessing and statistical analysis of DWI data

The diffusion weighted images are taken in 2x160 directions, with resolution $120 \times 120 \times 60$. DWI data are corrected for eddy-currents (including motion) and geometric distortions and brain-extracted. The DWI data are processed using FSL's FDT toolbox. Two fiber orientations are modeled and the probabilistic distributions of diffusion parameters are built up at each voxel (using bedpostx, part of FDT). After computing the fiber orientations, the fiber anisotropy results are transformed to T_1-space using FSL's FLIRT

tool, where the transformation matrix is computed based on the translation from the mean functional isotropy image to the T_1-image. For the tractography in T_1-space, we use probabilistic modelling of multiple fiber orientations using probtrackx (part of FDT). The goal of the DWI analysis is to track which voxels in the anterior language area have tracts to the posterior language area and vice versa. To that end, two masks are defined. The anterior language area is defined as the intersection of the fMRI results for the words-tones-contrast and the anterior anatomical mask described above (inferior and middle frontal gyrus). The posterior language area is defined as the intersection of the fMRI results for reading nonfinal embedded clause sentences and the posterior anatomical map used for fMRI analysis (anterior devision of middle and superior temporal gyrus and the supra marginal gyrus). Probabilistic streamlines are seeded from one of these masks. A total of 5000 samples is sent out from each tracking point. Stop masking is used to exclude indirect routes. The result is a map containing for each voxel the number of samples seeded from that voxel reaching the relevant target mask. These probabilistic pathways are thresholded at $\geq 1\%$ connecting samples passing through each voxel. We measure connectivity score as the number of connecting samples divided by the total number of samples.

6.2.4 Combining fMRI and DWI

We compare the connectivity scores and the activation scores (p-values). One way to compare is a minimum intersection map. The idea is that peaks in the structural connectivity profile should predict peaks in the functional activation profile.To generate voxel-wise minimum intersection maps (see Figure 6.2), the distributions of activation probabilities and average connectivity scores are shifted to zero minima and normalized to their robust maximum values (i.e. the 95th percentile).

6.3 Results

6.3.1 fMRI results

In Figure 6.1, we show the results of the fMRI data analysis in the four different layers. There is an indication for activation in the tumor based on the deviation from the null hypothesis of no activation. However, adding information on the alternative makes clear that the result might be statistically significant, but is not of practical significance: the effect size in these voxels is too small to represent real activation. Electrocortical stimulation mapping

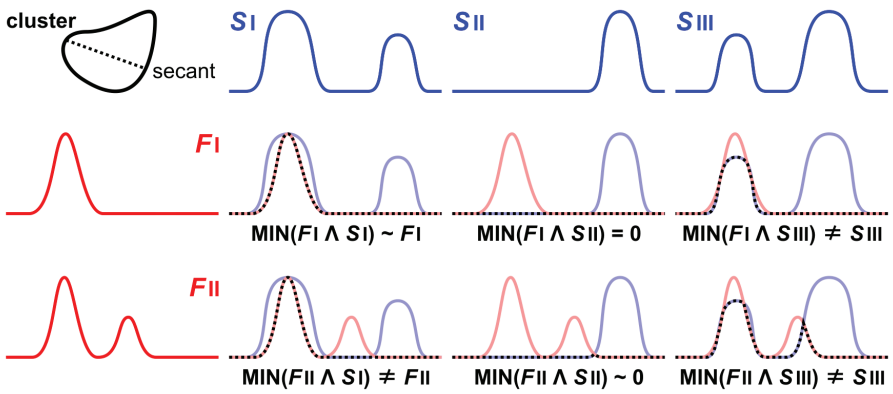

Figure 6.2: From Homola (2012) *Schematic illustration of minimum intersection maps. Minimum intersection maps are generated between different profiles of functional activation (red) and structural connectivity (blue). The profiles are normalized, i.e. scaled to the same min/max range. To build the minimum intersection (dotted), the minimum (MIM) of the two is considered at each point along the profile. Minimum intersection peaks indicate different degrees of spatial correspondence between high structural connectivity (S) and functional probability (F) values: Minimum intersection maps resembling F signify concordant presence of F- and S-peaks (left and upper right minimum intersects). Note that when F and S are too dissimilar, the minimum intersection is flat (middle). A non-flat minimum intersect with a sharp peak and displaced compared to F indicates a close but out-of-center overlap of F- and S-peaks (bottom right).*

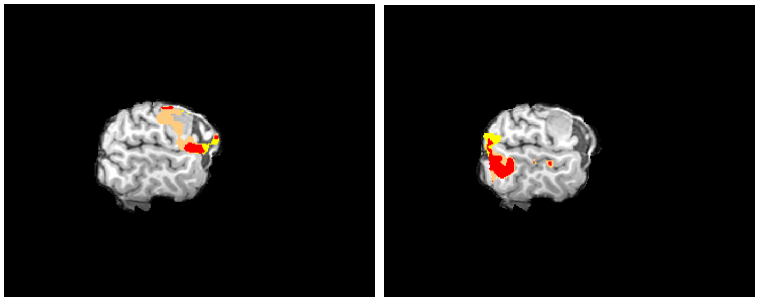

Figure 6.1: The results from the fMRI analysis with alternative based thresholding. The results for the anterior language area is shown in the left panel, the right panel represents the experiment for the posterior language area. The copper color refers to the practically insignificant voxels, red refers to active voxels, the yellow voxels show uncertainty.

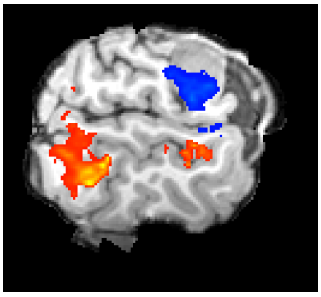

Figure 6.2: The results from the DWI analysis. The red voxels are the connectivity values in the posterior language area mask, while the blue voxels represent the connectivity values in the anterior language area.

and the postoperative patient condition after gross tumor resection confirmed that there were no essential intratumoral activations. We show how the uncertain layer is in this case mainly an extension of the width of the the active region and can be seen as a 'safe' boundary delineation.

6.3.2 DWI results

We show in Figure 6.3 how high connectivity values are indeed related with the posterior language area. The anterior language area is not correctly discovered, as the connectivity values in frontal mask are highest inside the tumor. It should be noted that the specific value of the connectivity values is dependent on the size of the masks and is therefore not interpretable.

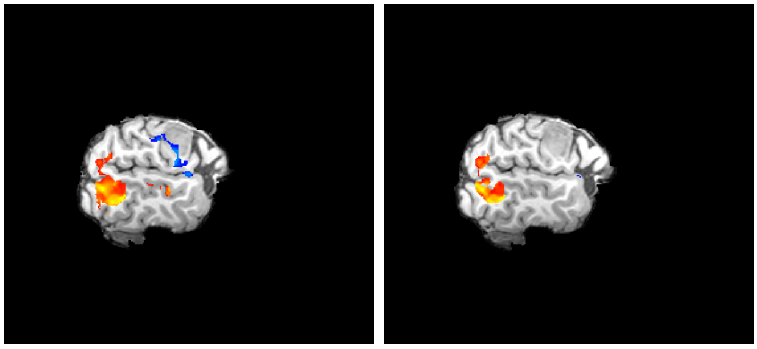

Figure 6.3: The minimum intersection maps. The red voxels represent the overlap between fMRI and DWI for the posterior language region, while the blue voxels represent the overlap for the anterior language area. The left image shows the classical testing results, the right figure are the results from the alternative based thresholding procedure.

6.3.3 Minimum intersection maps

For the posterior language area, we find reasonable overlap between connectivity measures and fMRI results in a comparable way. For the anterior language area, we find slight overlap between the classical hypothesis testing results and connectivity **in** the tumor, whilst no overlap between alternative-based testing and classical hypothesis testing.

6.3.4 Measures of activation and connectivity scores

More insight in the connectivity scores in relation to the testing procedures are given in Figure 6.4. One important remark is that the connectivity values in the posterior language area are overall smaller than the connectivity values in the anterior language area. This can be explained by the fact that the anterior mask is smaller than the posterior mask, and thus there is a smaller chance that sent samples arrive in the anterior mask than in the posterior mask. This difference has no intrinsic meaning with respect to the connectivity in both regions.

The main finding is the apparent relation between p_0 and connectivity scores. Low connectivity scores are related to high p_0 values, as expected. But furthermore we also see that higher p_1-values are also linked with higher connectivity scores. As such we conclude that it is indeed useful to add the p_1-value to the testing criterion.

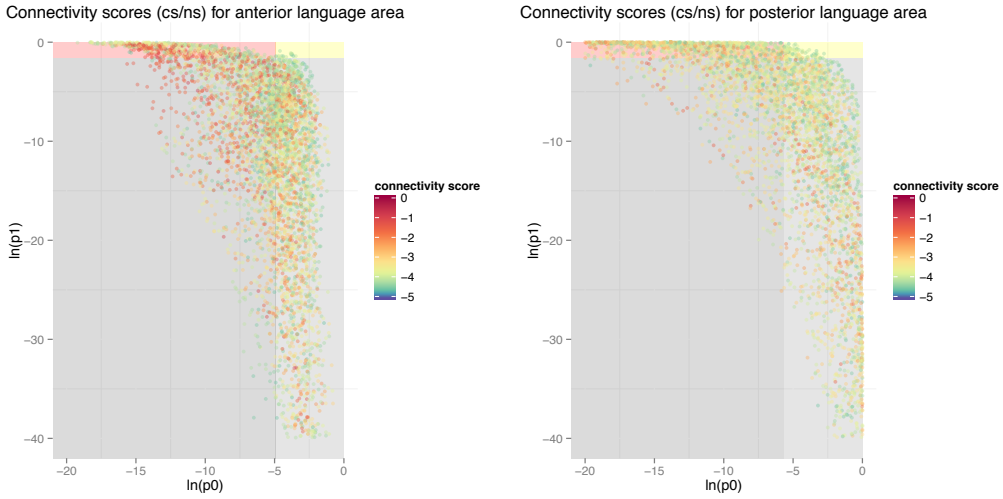

Figure 6.4: The connectivity scores with respect to the voxelwise p_0 and p_1-value. Each dot represents a voxel with a connectivity score higher than 1%. Furthermore the testing procedures are shown as background colors. The labeling is as follows: red refers to the active label, yellow represents the uncertain label, practical insignificance is shown in dark grey and light grey is for non-significant voxels.

6.3.5 Spatial cross-correlations

We show the relationship between p_0-values and connectivity scores in Figure 6.7. For the posterior language area, we show that higher p_0-values are accompanied by lower connectivity scores as expected. However, for the anterior language area, there is a clear bimodal distribution visible and therefore the smoother cannot be interpreted in a straightforward way. The same figure in relation with p_1-values instead of p_0-values is shown in Figure **??**. We expect that higher p_1-values indicate higher functional activation, and should be accompanied by higher connectivity values. This is observable in the posterior language area, but again because of the bimodality in the anterior language area, we make no interpretations of the smoother. Finally, we show the connectivity values in relation with $(1-p_0)/(1-p_1)$. Higher values of the numerator indicate more activation, higher values of the denominator indicate less activation. As such, higher values of the fraction indicate more activation. Again, we find expected results in the posterior language area but not in the anterior language area. In the posterior language area, we find a drop in connectivity values for high activation.

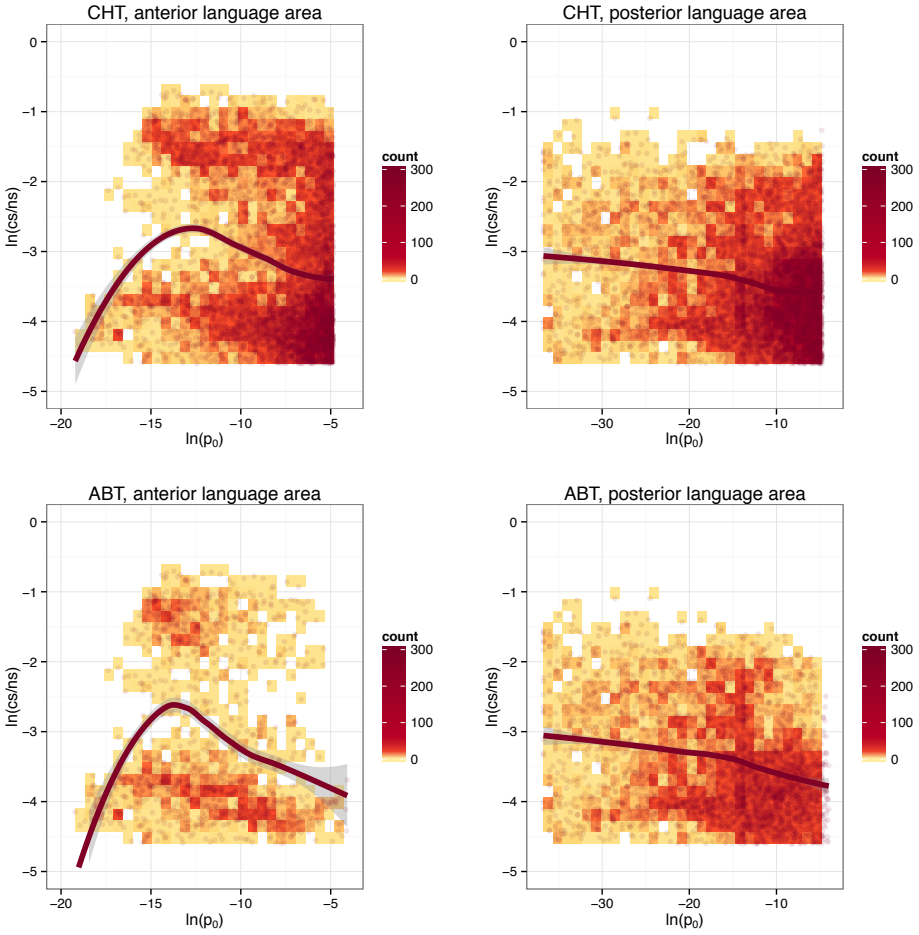

Figure 6.5: Two-dimensional histograms for the p_0-values against the connectivity values. Darker colors indicate more datapoints in the given bin. The line through the datapoints is a non-parametric loess smoother.

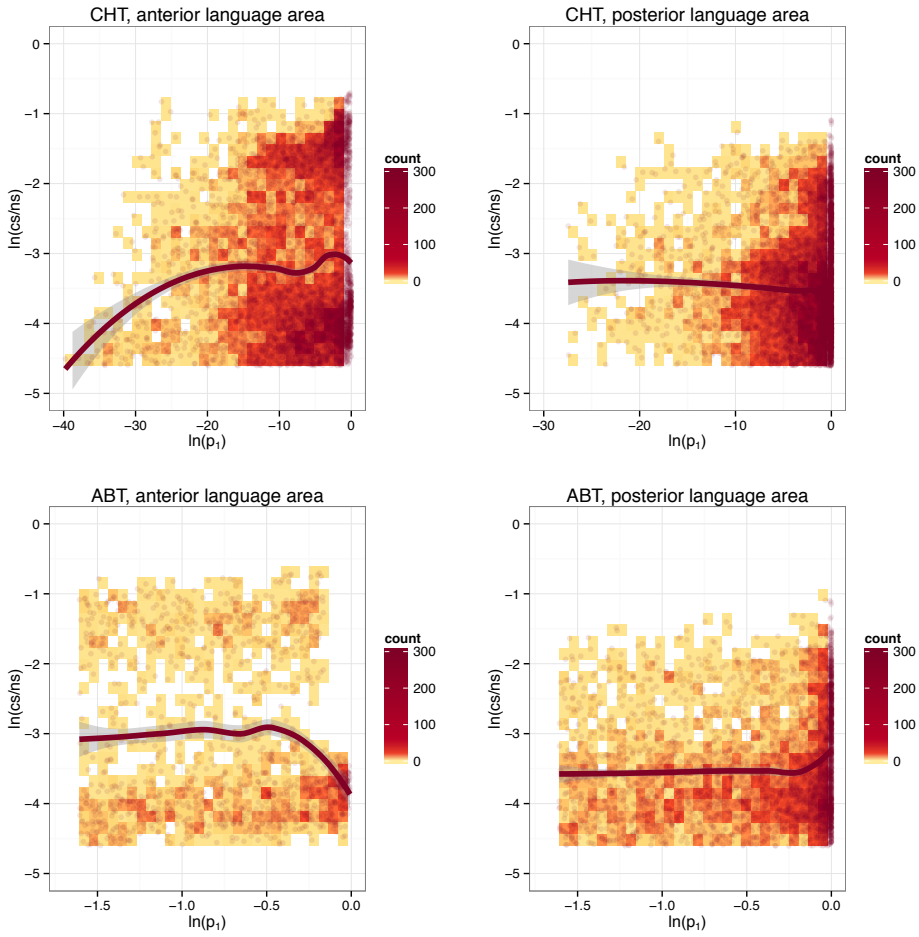

Figure 6.6: Two-dimensional histograms for the p_1-values against the connectivity values. Darker colors indicate more datapoints in the given bin. The line through the datapoints is a non-parametric loess smoother.

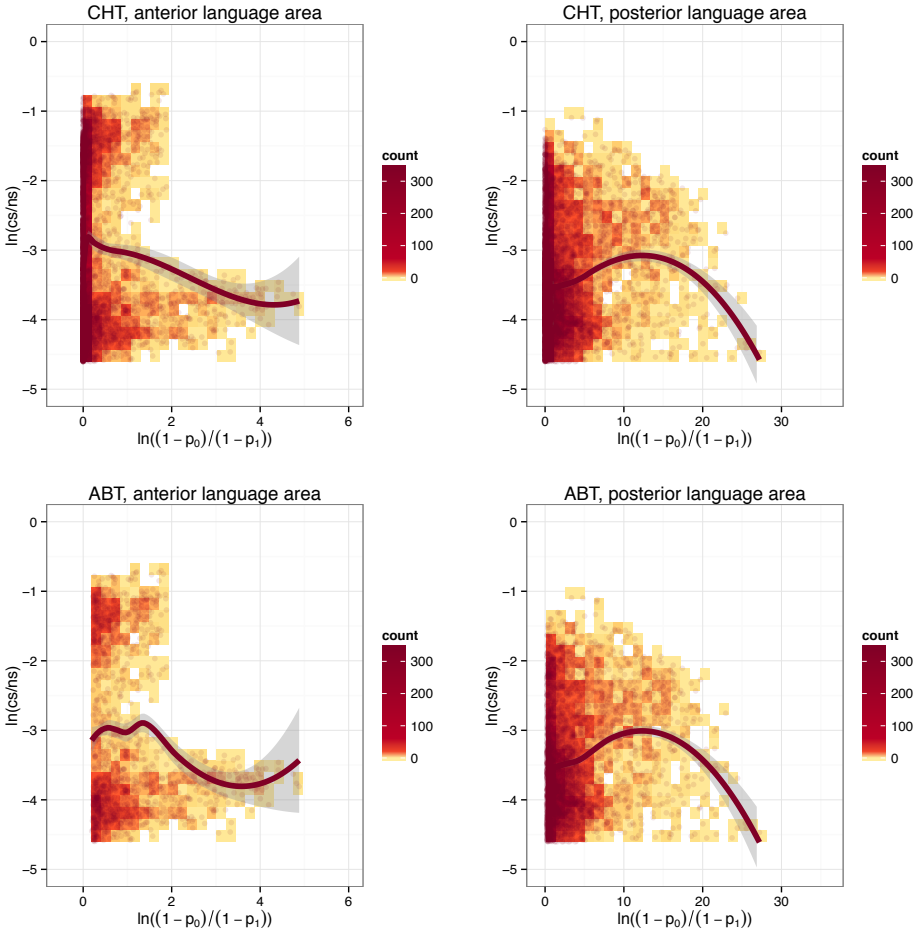

Figure 6.7: Two-dimensional histograms for $(1 - p_0)/(1 - p_1)$ against the connectivity values. Darker colors indicate more datapoints in the given bin. The line through the datapoints is a non-parametric loess smoother.

6.4 Discussion

In this work, we show how statistical procedures can be evaluated by relating the results from two distinct statistical procedures for fMRI to the structural connectivity profile. More precise, we localise the two main language areas in the brain, of which we know there is a strong connection between both. We measure voxelwise connectivity measures, indicating the strength of connection from one region to the other region, and relate these to the voxelwise testing procedure.

One important finding from the fMRI results is that only the classical testing procedure detects activation within the tumor while removing the tumor did not have any significant effect, and moreover electrocortical stimulation mapping confirmed that there were not intramural activations. However, we also found high connectivity measures in the tumor and not in the area where it is to be expected. Furthermore we find in this area an unexpected bimodal distribution of the connectivity values. There might be several reasons for these results, such as distortion because of the tumor. Another reason could be that co-registration (between fMRI - structural T_1-scan and DWI data) resulted in spatial deplacement of crucial information. Co-registration is optimal in terms of finding the global minimal deviation from the template, but around tumors, registration often fails locally. A third reason for the connectivity pattern in the tumor could be due to our mask definition. We have used large masks, which allows large pathways to be discovered. To avoid contamination between different pathways, we have used exclusion masks. However, it is still possible that there is contamination of the dorsal pathway from the ventral pathway, which could explain the high intratumoral connectivity values. Based on these data and analyses, a unique cause for the results cannot be identified.

One possible way to further inspect the relationship between functional and structural measures, could be to define unrelated regions for negative control. High connectivity values in unrelated regions (either close or far from the tumor) could disentangle possible explanations for the results.

We'd like to further remark that while we find a drop in connectivity profile for high activation values, this drop has also been observed with Homola et al. (2012) in a different setting.

This work aims to show how evaluation of statistical procedure could be validated using only real data. We showed an example of such a validation with data from a single person. However, these results are not answering all questions. It is clear that on the one hand more research should be done on the relation between fMRI and connectivity in general. On the other hand, this specific validation shows useful but to draw conclusions on the

performance of both thresholding procedures, the validation requires more data and deeper analyses.