Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Utilizing anatomical information for signal detection in functional magnetic resonance imaging

André Neumann[1], Norman Peitek[2], André Brechmann[2], Karsten Tabelow[3], Thorsten Dickhaus[1]

submitted: January 22, 2021

[1] Institute for Statistics, University of Bremen
P. O. Box 330 440
28344 Bremen
Germany
E-Mail: dickhaus@uni-bremen.de

[2] Combinatorial NeuroImaging, Leibniz Institute for Neurobiology
Brenneckestraße 6
39118 Magdeburg
Germany
E-Mail: npeitek@lin-magdeburg.de
Andre.Brechmann@lin-magdeburg.de

[3] WIAS Berlin
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: karsten.tabelow@wias-berlin.de

No. 2806
Berlin 2021

2020 *Mathematics Subject Classification.* 62J15, 62P10.

Key words and phrases. Aparc label, combination test, false discovery rate, family-wise error rate, mass-univariate linear model, multiple testing, program comprehension.

Financial support by the Deutsche Forschungsgemeinschaft (DFG) via grant DI 1723/3-2 is gratefully acknowledged. Brechmann's work is supported by DFG grant BR 2267/7-2.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Utilizing anatomical information for signal detection in functional magnetic resonance imaging

André Neumann, Norman Peitek, André Brechmann, Karsten Tabelow, Thorsten Dickhaus

Abstract

We are considering the statistical analysis of functional magnetic resonance imaging (fMRI) data. As demonstrated in previous work, grouping voxels into regions (of interest) and carrying out a multiple test for signal detection on the basis of these regions typically leads to a higher sensitivity when compared with voxel-wise multiple testing approaches. In the case of a multi-subject study, we propose to define the regions for each subject separately based on their individual brain anatomy, represented, e.g., by so-called Aparc labels. The aggregation of the subject-specific evidence for the presence of signals in the different regions is then performed by means of a combination function for p-values. We apply the proposed methodology to real fMRI data and demonstrate that our approach can perform comparably to a two-stage approach for which two independent experiments are needed, one for defining the regions and one for actual signal detection.

1 Introduction

Signal detection in high-dimensional data is a major topic of modern statistics. Typically, structural information like, for instance, (the degree of) sparsity of the signal is necessary for its detectability and/or (consistent) estimability; see, e. g., Figure 1 in Donoho and Jin 2004, van de Geer 2016, Chapter 7 in Wainwright 2019, as well as references therein.

Especially in the context of functional magnetic resonance imaging (fMRI), another type of structural information is localization. The primary units of fMRI measurement are volume units (voxels) of the human brain. Localization means that scattered, spread out signals (in single voxels or very small groups of voxels) are prone to be artifacts. Instead, topologically contiguous signals (forming larger groups of voxels, called regions) are much more plausible (Forman et al. 1995; Lazar 2008). This information rules out certain patterns of the signal structure a priori, and can hence be exploited to increase the statistical power for signal detection; see, among many others, Schildknecht, Tabelow, and Dickhaus 2016. There are different possibilities to define or find such regions: (i) One may refer to an atlas of the brain, like the Brodmann atlas (Brodmann 1909) and aggregate data within the regions given by this atlas. This has been the strategy of Schildknecht, Tabelow, and Dickhaus 2016. (ii) One may find the regions (of interest) in a data-driven manner, e.g., by a cluster analysis. This has been proposed, among others, by Jarmasz and Somorjai 2002. However, as emphasized for instance by Benjamini and Heller 2007, it is important that this data-driven definition of regions is "based on information outside the data that we set out to analyze", meaning that the dataset used for defining the regions should be (stochastically) independent of the dataset which is used for signal detection, to avoid selection biases. (iii) One may choose a statistical methodology which ensures statistically valid conclusions even for regions which are selected in a post-hoc manner after having seen the actual study data. This can be achieved by simultaneous inference methods which guarantee that any possible selection event is accounted for; see, e. g., Rosenblatt et al. 2018 and references therein.

DOI 10.20347/WIAS.PREPRINT.2806 Berlin 2021

	Study 1 (Siegmund et al. 2014)	Study 2 (Siegmund et al. 2017)
Participant sessions	16	14
Trials	12	30
Conditions	Bottom-up program comprehension, control (syntax), rest	Top-down program comprehension, Bottom-up program comprehension, control (syntax), rest
Scans	900	900

Table 1 – *Overview of the two related fMRI studies of program comprehension.*

```
public static void main() {
    String text = "The quick brown fox jumps";
    System.out.print(compute(text));
}

static int compute(String text) {
    int result = 0;
    boolean flag = false;

    for (int i = text.length() - 1; i >= 0; i--) {
        char c = text.charAt(i);

        if ((c >= 'a' && c <= 'z') || (c >= 'A' && c <= 'Z')) {
            flag = true;
            result++;
        } else {
            if (flag)
                break;
        }
    }

    return result;
}
```

Listing 1 – *Example code snippet in Java from Siegmund et al. 2014 that computes the length of the last word in a string. The snippet uses non-meaningful identifiers to induce bottom-up comprehension. Participants needed to figure out the output of this snippet "5".*

All of the three aforementioned strategies have their assets and their drawbacks: Strategy (i) is inexpensive and easy to implement, but the regions taken from the atlas may not be optimally aligned with the specific task at hand, and differences in the individual brain anatomies of the study participants may complicate its application. Strategy (ii) is costly (two independent experiments are needed), but is supposed to yield a more accurate definition of the regions (of interest). Strategy (iii) avoids both the (potentially suboptimal) a priori definition of regions and the need for an additional independent experiment. However, the issue of multiple testing (see, e. g., Dickhaus 2014 and Dudoit and van der Laan 2008) becomes much more severe if simultaneity over all possible selection events has to be guaranteed, and the selection of the regions is not based on a clear-cut (statistical) criterion, but on expert judgement of the study data. Therefore, it is hard to compare the results of Strategy (iii) with those of Strategies (i) and (ii).

Strategy (ii) has been followed by in a recent series of papers, namely Siegmund et al. 2014 and Siegmund et al. 2017 in which programmers comprehended program code. We provide an overview in Table 1. The authors asked participants in the first study (Siegmund et al. 2014) to understand short program code snippets, such as shown in Listing 1 (Siegmund et al. 2014). The program code did not contain any useful identifier names, which induces *bottom-up comprehension* (Pennington 1987). Siegmund et al. contrasted the bottom-up comprehension task with a syntax task, in which participants were presented with similar program code snippets, but only had to focus on syntax errors (e.g., missing semicolon). This control condition was intended to reveal only brain activation that is necessary for programmers to comprehend program code in-depth. As additional control condition,

the experiment included phases of rest in between the comprehension and syntax conditions.

The second study (Siegmund et al. 2017) was a follow-up study that differentiated the program-comprehension task into more nuanced conditions. The aim was to differentiate between bottom-up comprehension and *top-down comprehension* (Brooks 1983) that was induced by varying the meaningfulness of identifier names and by prior training to provide participants with necessary knowledge. As in the previous study, the syntax task served as control condition. The analysis of the influence of top-down and bottom-up program comprehension in this study built on the regions identified in the first study, thus following Strategy (ii).

In the present work, we propose a new strategy and apply it to the data from Siegmund et al. 2017: We first utilize structural information of the individual's brain scans. Here we use an automatic parcellation of the brain into so-called Aparc labels, which assign the voxels of an individual to anatomic regions. Details on this are provided in Section 3.2. This step of data analysis provides us with a significance evaluation (in terms of a p-value) for the presence of signals in anatomically defined regions for every study participant separately. Then, in a second step, we combine for each of these regions the p-values of all voxels within that region from all study participants by means of an appropriate combination function, and we evaluate the significance of the whole region by means of the resulting combined p-value. As we will demonstrate by means of a concrete example from the field of programming language comprehension, this new strategy can be similarly powerful as Strategy (ii) described above, while avoiding the additional experiment used for region definition. In our case, this additional experiment has been carried out by Siegmund et al. 2014 (and utilized in a previous data analysis), but our present methodology will not have access to the data from that additional experiment.

The rest of the paper is structured as follows. In Section 2, we describe our proposed statistical methodology. Section 3 is devoted to the detailed description of our re-analysis of the data by Siegmund et al. 2017, and the results of this re-analysis are presented in Section 4. We conclude with a discussion in Section 5.

2 Methods

In this section we describe our statistical model for fMRI data as well as the proposed data analysis workflow for detecting brain regions which are significantly associated with a certain cognitive task.

Random-effects linear model for voxel-wise multiple tests

Let Y_{ixt} denote the observed data from a functional MRI experiment at voxel x and time t for the i-th subject. Here, we adopt the common view (Lazar 2008) of a mass-univariate linear model for the data

$$Y_{ixt} = X\beta_{ix} + \varepsilon_{ixt}$$

with a design matrix X containing variables with the expected blood oxygenation level dependent (BOLD) response related to the experimental stimuli or nuisance parameters like drifts of the MR signal. The random variable ε_{ixt} is the error term with assumed zero expectation and a spatio-temporal correlation structure. Estimates $\hat{\beta}_{ix}$ of the statistical parametric map (SPM) or their contrasts $c^\top \hat{\beta}_{ix}$ and estimates of their covariance matrices $\hat{\Sigma}_{ix}$ (or the variances $\hat{\sigma}^2_{ix} = c^\top \hat{\Sigma}^2_{ix} c$) can then be obtained from a pre-whitened version of the linear model above (Lazar 2008).

The SPM then forms a random t-field (Worsley 1994) with an inherent multiple comparison problem due to the large number of local hypotheses. One common strategy is to define local p-values at each

voxel x and for each subject i based on the local values of the random t-field and to control the family-wise error rate (FWER) using accordingly adjusted thresholds (Worsley et al. 1996). However, this is known to be a very conservative approach with respect to the detectability of significant brain signals in the outlined framework. In contrast, approaches related to the control of the false discovery rate (FDR) can handle the multiple comparison problem, e.g., by the procedure developed by Benjamini and Hochberg 1995.

Parcellation of the human brain

Neuroanatomic research has found that the human brain can be parcellated into different sub-regions based on structural similarities. One of the earliest atlases is the Brodmann atlas (Brodmann 1909) which is based on the cytoarchitectural organization of the brain. The Brodmann Areas have been schematically transferred to a template brain, the so-called Talairach-Atlas (Talairach and Tournoux 1988) which is commonly used in fMRI studies to report the location of significant grand average activation, as used in (Siegmund et al. 2014; Siegmund et al. 2017). Here we chose the Harvard-Oxford brain atlas (Makris et al. 2006; Desikan et al. 2006) that provides a parcellation based on gross anatomical landmarks and delivers an Aparc label j for each voxel of each individual brain space.

Statistical inference

As outlined in the introduction, we re-used fMRI data from a program code comprehension task first analyzed by Siegmund et al. 2017 and performed a new analysis comprising the four steps outlined below. The experiment used two different levels of software program code comprehension stimuli, henceforth denote as bottom-up and top-down comprehension, to infer on the related cognitive processes. In our strategy, we combined the methods of Siegmund et al. 2017 (steps 1 and 2) and Schildknecht, Tabelow, and Dickhaus 2016 (step 3). Furthermore, we implemented our new methodological contribution of combining the evidence for activation of a given brain region across the subjects (step 4). For the first two steps, we conducted, for each subject, a random-effects linear model analysis as described above for deriving voxel-wise p-values.

Step 1: Program comprehension versus rest

In the first step, we contrasted (for each participant separately) the comprehension of program code (Siegmund et al. 2017) to the rest condition. This identifies brain areas with a positive deflection of the BOLD response. Furthermore, in order to account for the multiple comparison problem we performed the Benjamini-Hochberg test (see Benjamini and Hochberg 1995) for FDR control. Only those voxels which have been declared significant by this procedure were considered in step 2. This methodology is justified by the fact that the FDR is an established screening criterion for high-dimensional multiple test problems.

Step 2: Bottom-up comprehension versus control condition

In this step, we contrasted (again, for each participant separately) one type of program comprehension: *bottom-up comprehension*. Bottom-up comprehension is induced when program code provides no semantic cues and programmers need to comprehend each line separately and then integrate the

information in a slow, tedious process. For the significant voxels from the first step, we applied in a second step the same multiple test to the contrast bottom-up comprehension against the control condition (syntax task) on the restricted set of voxels. As a result of this step, we get for each participant i and for each considered voxel k a p-value $\tilde{p}_{ik}$.

Step 3: Aparc p-values for every participant i

This step builds upon the methodology by Schildknecht, Tabelow, and Dickhaus 2016, and it delivers for each participant i and for each Aparc label j a (confirmatory) significance evaluation with respect to the contrast specified in step 2. Hence, the evidence from all voxels of participant i in the brain region labeled by j is combined in this step of data analysis.

To this end, let κ be a tuning parameter with values in the interval $[0,1]$ (i.e. in per cent) and let m_j be the number of voxels contained in the brain region labeled by Aparc label j. (To keep the notation feasible, we implicitly assume here that for each participant i the same number m_j of voxels belong to the brain region labeled by j.) We consider the null hypothesis H_{ij} of no relevant differential activation of the region labeled by j for participant i during the two tasks mentioned in step 2, together with its two-sided alternative hypothesis K_{ij}. We call H_{ij} the Äparc null hypothesis"for the brain region labeled by j for participant i. We formalize H_{ij} as a so-called partial conjunction hypothesis (see Schildknecht, Tabelow, and Dickhaus 2016 and the references therein for a formal mathematical description), meaning that we consider the differential activation in region j for participant i relevant, if it contains at least $u_j := \kappa \cdot m_j$ significant voxels. For testing H_{ij}, we calculate the Äparc p-value"p_{ij}^{APARC}, given by

$$p_{ij}^{\text{APARC}} := \min_{1 \le \ell \le m_j - u_j + 1} \left\{ \frac{m_j - u_j + 1}{\ell} \tilde{p}_{i,(u_j - 1 + \ell):m_j} \right\},$$

where the voxel-wise p-values $\tilde{p}_{i,1:m_j}, \ldots, \tilde{p}_{i,m_j:m_j}$ for participant i in region j are ordered from smallest to largest (see Benjamini and Heller 2008).

In order to achieve family-wise error rate (FWER) control, we have to choose the tuning parameter κ smaller than or equal to $1/J$, where J is the number of Aparc labels. Choosing $\kappa = 1/J$ corresponds to the so-called Bonferroni multiplicity correction. The choice of κ is discussed further in Appendix S1 of Schildknecht, Tabelow, and Dickhaus 2016.

Step 4: Combined Aparc hypothesis tests by Fisher's method

In this final step, we combine for each Aparc label j the Aparc p-values calculated in step 3 over all participants $i = 1, \ldots, n$.

In order to do this, we apply the so-called Fisher method to combine p-values. Namely, the Fisher test statistic T_j for region j is given by

$$T_j := -2 \sum_{i=1}^{n} \log \left(p_{ij}^{\text{APARC}} \right).$$

Under independence of the data with respect to the participants, T_j is asymptotically χ^2_{2n}-distributed (chi squared) with $2n$ degrees of freedom under the null. The latter independence assumption is justified, because the participants have been included in the study independently from each other.

Finally, we can reject the (over all participants i combined) Aparc hypothesis H_j (i.e., the respective partial conjunction hypothesis, but now with respect to the population, not with respect to a single

participant) if and only if Fisher's test statistic T_j is larger than the $(1-\alpha\kappa)$-quantile of the χ^2_{2n}-distribution with $2n$ degrees of freedom, where the tuning parameter κ has been introduced in step 3. This parameter addresses the multiplicity of the test problem with respect to the J Aparc labels which are simultaneously under consideration.

3 Data analysis

We re-use the data from Siegmund et al. 2017 and compare our results with those obtained by additionally utilizing Siegmund et al. 2014 as pre-study in the sense of Strategy (ii) outlined in the introduction.

3.1 Previous findings

In the two previous studies mentioned before, Siegmund et al. used similar analysis processes. They used BrainVoyager™ QX 2.8.4.[1] The anatomical scans were transformed into the Talairach brain to account for differences in brain size (Talairach and Tournoux 1988). They preprocessed the functional data of both studies with a standard pipeline of: 3-D motion correction, slice-scan-time correction, and temporal filtering. In addition, they applied a spatial smoothing with a Gaussian filter (FWHM=4 mm).

In the first study (Siegmund et al. 2014), the random-effects GLM revealed five brain areas (BAs 6, 21, 40, 44, 47) with significant activation with the contrast Bottom-Up Comprehension > Control condition, i.e. syntax task. In the second study, the same contrast revealed no significant areas anymore, likely due to the reduced statistical power of five instead of twelve bottom-up comprehension tasks per session. Thus, the authors ran a regions of interest analysis restricted to the identified activation clusters of the first study on the data of the second study. This resulted in a significantly stronger activation for Bottom-Up Comprehension versus Syntax in BAs 21, 40, and 44.

3.2 Data export and preparation for re-analysis

In Figure 1, we illustrate the overall process for the re-analysis of the data from Siegmund et al. 2017. For the purpose of our re-analysis, we exported the already preprocessed data from the second study. We did not use the data of the first study as our method does not need prior definitions of regions of interest. We used BrainVoyager to access the data from Siegmund et al. We exported the statistical values (i.e., t-scores, p-values) of the obtained brain activation on a voxel basis for each participant. The voxel resolution is the same BrainVoyager uses for its internal computation (i.e., on a 1 mm interpolated resolution).

In addition, we used FreeSurfer to segment and parcellate the brain of each participant based on their anatomical scan (Fischl et al. 2002; Fischl et al. 2004). We used the Destrieux' cortical atlas to assign Aparc labels on an individual participant basis (Destrieux et al. 2010). Next, we used Nipype (Gorgolewski et al. 2011; Esteban et al. 2020) to convert Freesurfer labels to a BrainVoyager-readable format.

Our last step annotated the exported functional data with the individual anatomical labels for each participant. We removed all functional voxels for coordinates that had no assigned Aparc label, which

[1] Brain Innovation BV, Maastricht, The Netherlands, `http://brainvoyager.com`

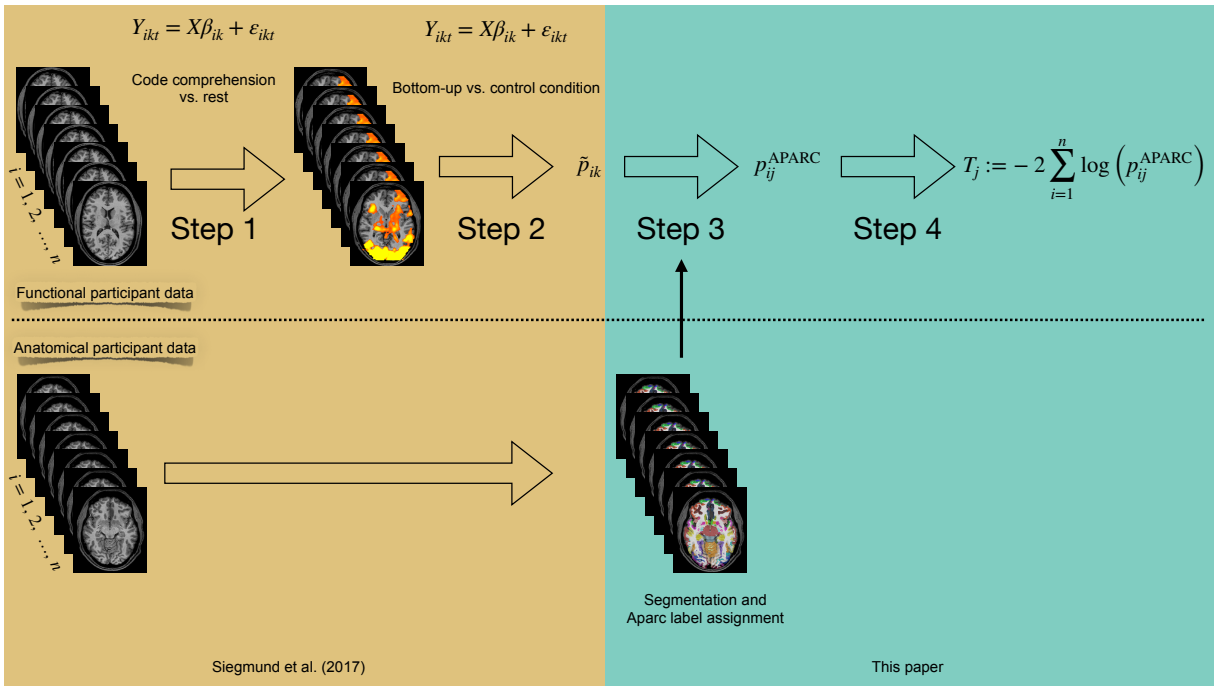

Figure 1 – *Illustration of processing of the experimental data. The box in Harvest Gold indicate analysis steps that have already been performed in Siegmund et al. 2017. The box in Monte Carlo indicates the processing steps proposed in this paper.*

typically are voxels that are not considered as grey matter.

Data sharing

Siegmund et al. provide replication packages for their studies.[2] The raw brain data, exported voxel-based activation data with their Aparc labels, and our analysis scripts will be shared upon request.

4 Results

Figure 2 displays the six brain regions which have been declared as significant (at FWER level $\alpha = 5\%$) associated with the task at hand by our described methodology. They are in general agreement with the results obtained by Siegmund et al. 2017 in which they utilized prior knowledge. We visualize the confirmed network of brain activation from Siegmund et al. in Figure 3 and our identified network of significantly activated Aparc labels in Figure 4. Figures 3 and 4 show overlapping results with regard to the Brodmann area 21 that covers the middle and inferior temporal gyrus (separated by the inferior temporal sulcus). However, we observed differences regarding smaller brain regions. We compare Siegmund's replication efforts to our results in Section 5.1.

Our method found three Aparc labels in the inferior and middle temporal gyrus in the left hemisphere. Siegmund et al. found their largest and most robust activation cluster in BA21 of the left hemisphere, which covers several gyri in the temporal lobe. These left temporal gyri are often associated with semantic processing of natural language, which is typically left-lateralized for right-handed participants.

[2] https://www.infosun.fim.uni-passau.de/se/janet/fMRI/index.php, https://github.com/brains-on-code/paper-esec-fse-2017

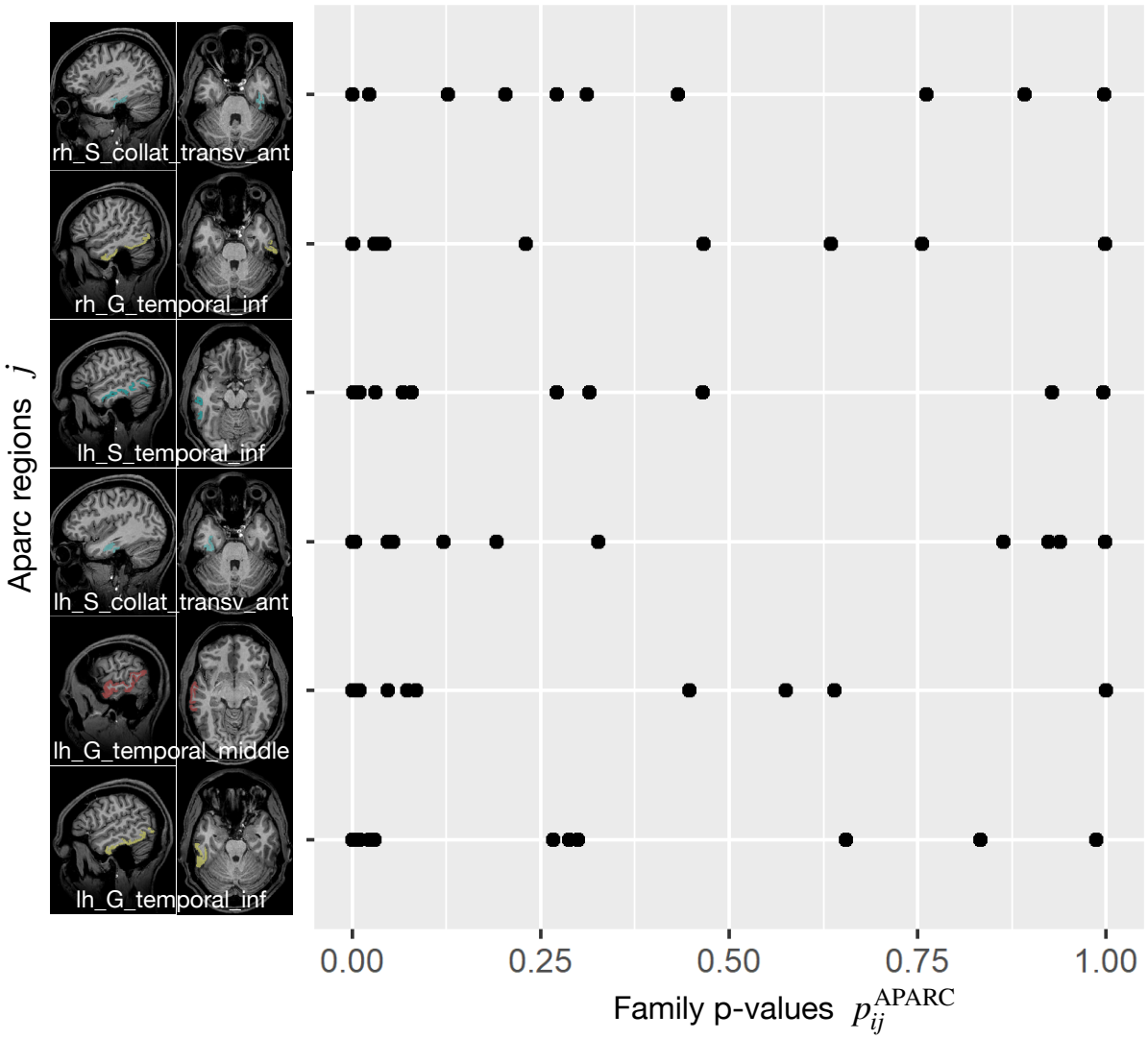

Figure 2 – *The six significant brain regions (at FWER level $\alpha = 5\%$). In each row, each point corresponds to the Aparc p-value p_{ij}^{APARC} of one study participant i, where the index j refers to the area indicated by the code at the beginning of the row.*

In the context of programming, the activation is believed to be responsible for extracting the meaning of individual identifiers and symbols during program comprehension. We found two further Aparc labels bilaterally in the inferior temporal gyrus and the anterior collateral sulci, which are both in the temporal lobe as well.

For each of these six regions indexed by j, we display all subject-specific Aparc p-values $\{p_{ij}^{\text{APARC}} : 1 \leq i \leq n\}$, where n is the number of study participants. For each j considered in Figure 2, it can clearly be observed that not a single extreme outlier (one very small p-value corresponding to one individual subject) is responsible for the statistical significance with respect to the combination test statistic T_j, but that the combined information contributed by all n subjects supports our statistical conclusions. Aparc p-values $p_{ij}^{\text{APARC}} \equiv 1$ can occur, if none of the voxels belonging to region j has been selected in Steps 1 and 2 described before for a certain subject i. By construction of T_j, this essentially means that the ëffective sample size"for such a region is reduced, while the number of degrees of freedom for the null distribution of T_j remains unchanged.

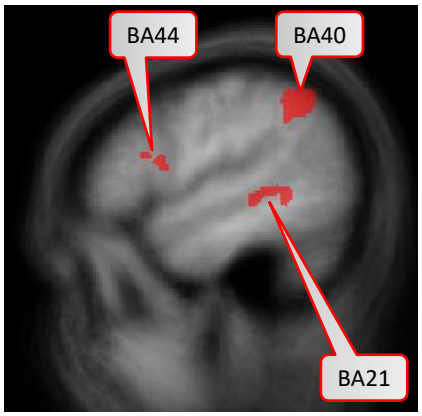

Figure 3 – *Network of left-lateralized confirmed brain areas (i.e., BAs 21, 40, 44) activated during program comprehension found in Siegmund et al. 2017.*

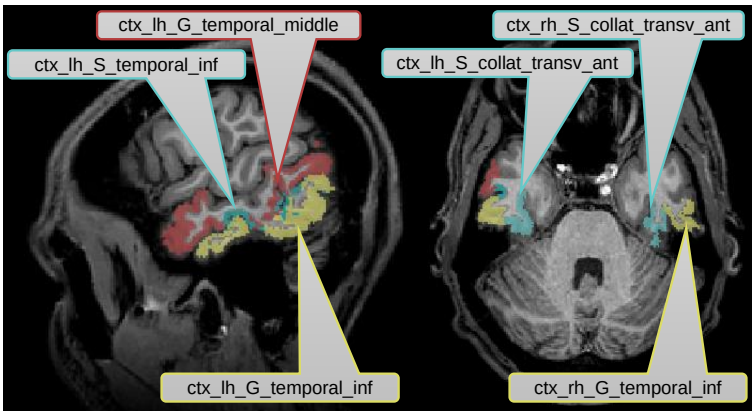

Figure 4 – *Results of our analysis with significantly activated Aparc labels, particularly in the middle and inferior temporal lobe.*

5 Discussion

5.1 Statistical sensitivity

This paper is introducing a new strategy to analyze fMRI data and is demonstrating its performance by drawing a comparison to two studies by Siegmund et al. In their second study, they investigated a new research question which was based on a design with reduced statistical power regarding the network of brain areas activated during bottom-up program comprehension. Only by using prior knowledge of the location of identified clusters from the first study and conducting a regions-of-interest analysis with increased statistical power, they were able to exceed standard thresholds of statistical significance but restricted to the subset of areas identified in the first study.

Research that similarly aims to detect small differences between cognitive processes with state-of-the-art methods would need to increase statistical power, e.g. by conducting two studies: first, to identify the network of brain areas involved in the overarching cognitive process, and second, to identify differential effects within the activated brain areas.

The method of utilizing individual anatomically defined regions of interest presented in this paper is a candidate for a more sensitive analysis as compared to relying on statistical testing of single voxels only. We demonstrate that our method is able to find significantly activated brain areas without relying on prior knowledge. Moreover, additional brain areas were identified which have been described in two fMRI studies of programmers and interpreted as being involved in visuo-spatial processing. One study investigated manipulating data structures, which shares similarities to spatial rotation (Huang et al. 2019) and one writing program code (Krueger et al. 2020). Since the study by Siegmund et al. 2014 was the first study on program comprehension they used a rather small FDR corrected significance level ($p < 0.01$ as compared to the more common $p < 0.05$). Possibly this is one reason why the two areas identified by the current approach were not identified there and as a consequence could not emerge in Siegmund et al. 2017. Thus, our current approach seems valuable in cases where new research questions are explored and little to no prior work is available and which thus would require careful statistical hypothesis testing to initially minimize false positives.

However, inferior frontal gyrus (BA 44) and inferior parietal lobule (with BA 40), shown to be significant

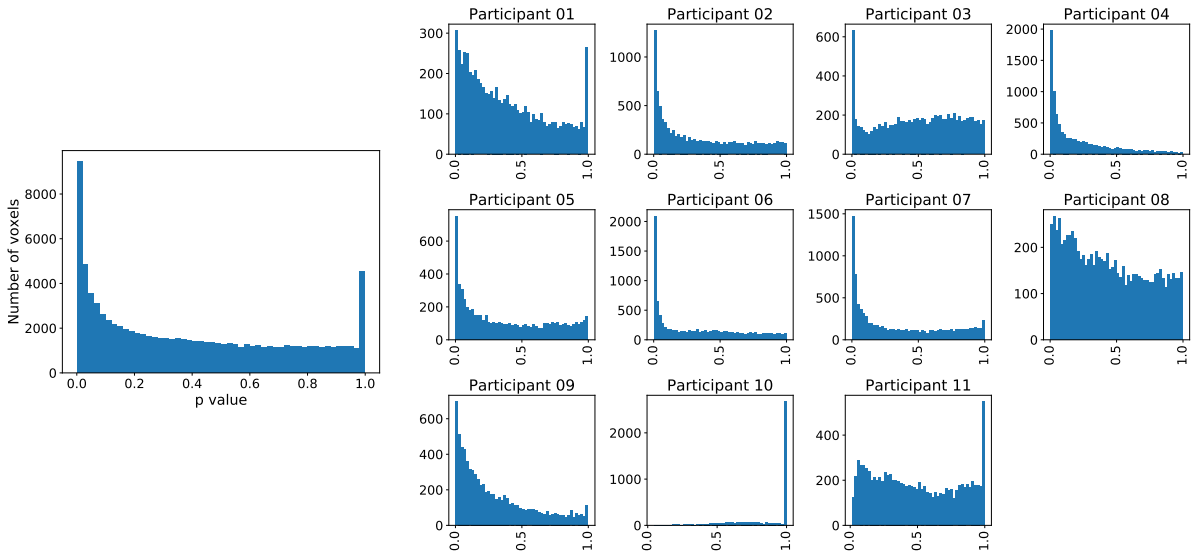

Figure 5 – *Histogram of p values of* `ctx_lh_G_temporal_middle` *label that is evaluated as significantly activated. The left plot is across all participants, while the right side shows the distribution across individual participants. The majority of participants is contributing to the significance.*

	# Voxels Overlapping with BA21	in %	# Voxels of Entire Aparc Label	in % in BA21
Left-Cerebral-White-Matter	634	22%	210620	0,3%
ctx_lh_G_temporal_middle	617	22%	7515	8,2%
ctx_lh_S_temporal_sup	512	18%	9379	5,5%
ctx_lh_S_temporal_inf	85	3%	2499	3,4%
...	...	...	...	...

Table 2 – *The region of interest in BA21 identified by Siegmund et al. 2014 consists of 2844 voxels. Only a subset of these voxels is assigned to Aparc labels. However, the assigned Aparc label are larger and only a smaller section overlaps with the activation cluster.* ***Bolded Aparc labels*** *are evaluated as significant. Only Aparc labels with at least 75 overlapping voxels are included.*

in Siegmund et al. 2017, were not significant with our method. Therefore, we further investigated this difference exemplarily with the activation cluster in BA 40. In Figure 5, we display a histogram of p-values for the (statistically significant) Aparc label in the middle temporal gyrus, in which a majority of the participants contribute with small p-values. In contrast, in Figure 6, we display a histogram of p-values for the inferior parietal lobule. Across the entire group, we also observe an accumulation of very small p-values, but only three of eleven participants contribute to this. Unlike the regions-of-interest analysis done by Siegmund et al., our method relying on combining Aparc p-values by Fisher's method does not reject hypotheses with such a p-value distribution.

By transforming the Aparc labels into the Talairach space, we observed that the voxels included in activation cluster of BA 40 do not perfectly align with the Aparc label that should cover this brain area, i.e. the inferior angular gyrus of the parietal lobe. Table 3 shows that the overlap to the activation cluster in BA40 is less than 50% and that only around 5% of its voxels are assigned to an Aparc label at all. In contrast, Table 2 shows that at least 75% of the voxels of the BA 21 cluster are aligned to Aparc labels of which two are significant with our method. Thus, a small activation cluster within a large anatomical region may get lost with our approach. We further discuss this potential drawback of the used anatomical segmentation and possible remedies in Section 5.3.

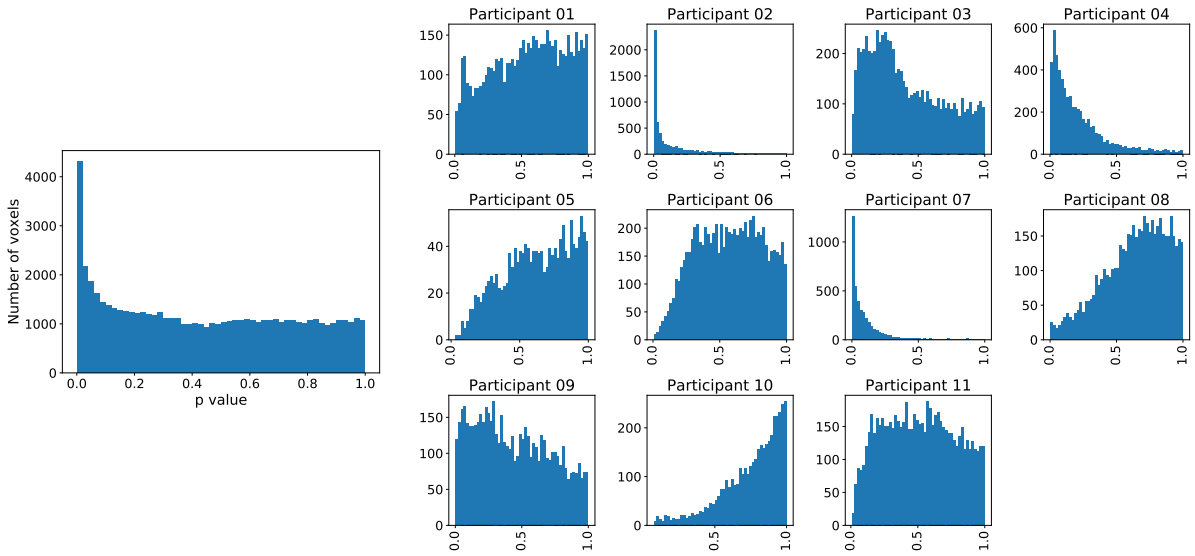

Figure 6 – *Histogram of p values of* `ctx_lh_G_pariet_inf-Angular` *that is not evaluated as significantly activated. The left plot is across all participants, while the right side shows the distribution across individual participants. While there are many small p values, only a few participants contribute these small values.*

	# Voxels Overlapping with BA40	in %	# Voxels of Entire Aparc Label	in % in BA40
ctx_lh_G_pariet_inf-Supramar	304	17,1%	6318	4,3%
ctx_lh_G_pariet_inf-Angular	290	16,3%	4975	5,1%
Left-Cerebral-White-Matter	157	8,9%	178875	0,1%
...	...	...	...	...

Table 3 – *The region of interest in BA40 identified by Siegmund et al. 2014 consists of 1777 voxels. Only a subset of these voxels is assigned to Aparc labels. However, the assigned Aparc label are larger and only a smaller section overlaps with the activation cluster. No Aparc label is evaluated as significant. Only Aparc labels with at least 75 overlapping voxels are included.*

From the methodological point of view, our main contribution consists in a novel way how to combine evidence: Instead of aggregating single voxel data over all participants by mapping them to a standard brain template, we define subject-specific regions and combine the evidence on the level of these regions by means of a combination function. This is a generic methodological approach which is not restricted to the specific study setup of Siegmund et al., but can be applied to essentially any fMRI study design, involving an arbitrary number of contrasts.

Furthermore, also the (final) combination step of our proposed approach is generic in the sense that instead of the Fisher combination function any other (appropriate) combination function for p-values may be used. Recently, there has been a renewed interest in p-value combination methods; see, e. g., Wilson 2019 (with discussion), Vovk and Wang 2020, and Vovk and Wang 2021.

5.2 Comparison to related fMRI analysis methods

Under the multiple testing framework, testing of grouped null hypotheses with (potential) application in fMRI is an active research topic. Heller et al. 2006 as well as Benjamini and Heller 2007 employed clustering techniques to define regions of interest, and they incorporated the heterogeneous cluster

sizes in a weighting scheme for the linear step-up test by Benjamini and Hochberg 1995. In the same vein, Hu, Zhao, and Zhou 2010 as well as Zhao and Zhang 2014 made use of the different proportions of true null hypotheses in each of the groups in their proposed weighting. A Bayesian variant of this idea has been derived by Liu, Sarkar, and Zhao 2016. Hierarchical methods, which exclude groups without strong evidence for the presence of signals in several stages of data analysis, have been worked out by Yekutieli 2008, Benjamini and Bogomolov 2014, and Schildknecht, Tabelow, and Dickhaus 2016, among others. However, these methods rely on combining the subject-specific data on the voxel level, which is a standard technique as mentioned, for instance, in Section 5 of Lindquist 2008. Also on the basis of (combined) voxel data, Shi and Guo 2016 proposed a hierarchical independent component analysis for the comparison of brain functional networks. To the best of our knowledge, our proposed method to define subject-specific regions by means of Aparc labels and to combine the resulting subject-specific Aparc p-values by means of a combination function is a novel idea.

5.3 Outlook: from anatomical to functional aggregation

Our presented method avoids potential imprecision of using common brain templates. We used the parcellation of the brain for each individual participant into Aparc labels before aggregation into the group analysis. This procedure provides more labels than a traditional Brodmann atlas. Still, some of the regions are very large and thus presumably contain several functional areas. There is ongoing research to subdivide the brain based on cytoarchitectonic details, e.g. the Jülich-Brain (Amunts et al. 2020). This will provide more detailed parcellation schemes in the future and that could easily be integrated into our proposed method. Another refinement could be to use functionally defined brain regions for our presented methodology. A study that implements functional localizers could identify participant-specific functional maps of the brain for well defined standardized tasks (e.g. see Nieto-Castañón and Fedorenko 2012). Then, in the analysis, our presented methodology can aggregate across all participant-specific brains with less imprecision than traditional methods. We would like to note that such functional localizers are currently restricted to research areas that include brain regions with specific functional specialization (Saxe, Brett, and Kanwisher 2006, Friston et al. 2006). The studies we presented in this paper are concerned with a rather complex cognitive task. Moreover, understanding program code is highly individual since different programmers rely on different comprehension strategies based on their preferences, domain knowledge, and experience (e.g., “Using a behavioral theory of program comprehension in software engineering”; Brooks 1983; Soloway and Ehrlich 1984).

Acknowledgments

We thank Jörg Stadler for his technical support for our data processing with Nipype and FreeSurfer. Financial support by the Deutsche Forschungsgemeinschaft (DFG) via grant DI 1723/3-2 is gratefully acknowledged. Brechmann’s work is supported by DFG grant BR 2267/7-2.

References

Amunts, K., H. Mohlberg, S. Bludau, and K. Zilles (2020). “Julich-Brain: A 3D probabilistic atlas of the human brain’s cytoarchitecture”. *Science* 369 (6506), pp. 988–992.

Benjamini, Y. and M. Bogomolov (2014). “Selective inference on multiple families of hypotheses”. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 76 (1), pp. 297–318.

Benjamini, Y. and R. Heller (2007). “False discovery rates for spatial signals”. *J. Amer. Statist. Assoc.* 102 (480), pp. 1272–1281.

– (2008). “Screening for partial conjunction hypotheses”. *Biometrics* 64 (4), pp. 1215–1222.

Benjamini, Y. and Y. Hochberg (1995). “Controlling the false discovery rate: A practical and powerful approach to multiple testing”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 57 (1), pp. 289–300.

Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Großhirnrinde in ihren Prinzipien dargestellt auf Grund des Zellbaues*. Leipzig: Barth.

Brooks, R. “Using a behavioral theory of program comprehension in software engineering”. In:

– (1983). “Towards a theory of the comprehension of computer programs”. *Int’l Journal of Man-Machine Studies* 18 (6), pp. 543–554.

Desikan, R., F. Ségonne, B. Fischl, B. Quinn, B. Dickerson, D. Blacker, R. Buckner, A. Dale, R. P. Maguire, B. Hyman, M. Albert, and R. Killiany (2006). “An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest”. *Neuroimage* 31 (3), pp. 968–80.

Destrieux, C., B. Fischl, A. Dale, and E. Halgren (2010). “Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature”. *NeuroImage* 53 (1), pp. 1 –15.

Dickhaus, T. (2014). *Simultaneous statistical inference with applications in the life sciences*. Springer-Verlag Berlin Heidelberg.

Donoho, D. and J. Jin (2004). “Higher criticism for detecting sparse heterogeneous mixtures”. *Ann. Statist.* 32 (3), pp. 962–994.

Dudoit, S. and M. van der Laan (2008). *Multiple testing procedures with applications to genomics.* Springer Series in Statistics. New York, NY: Springer.

Esteban, O., C. J. Markiewicz, C. Burns, M. Goncalves, D. Jarecka, E. Ziegler, S. Berleant, D. G. Ellis, B. Pinsard, C. Madison, and et al. (2020). “nipy/nipype: 1.5.0”.

Fischl, B., D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. Van Der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, et al. (2002). “Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain”. *Neuron* 33 (3), pp. 341–355.

Fischl, B., A. Van Der Kouwe, C. Destrieux, E. Halgren, F. Ségonne, D. H. Salat, E. Busa, L. J. Seidman, J. Goldstein, D. Kennedy, et al. (2004). “Automatically parcellating the human cerebral cortex”. *Cerebral cortex* 14 (1), pp. 11–22.

Forman, S., J. Cohen, M. Fitzgerald, W. Eddy, M. Mintun, and D. Noll (1995). “Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold.” *Magnetic Resonance in medicine* 33 (5), pp. 636–647.

Friston, K., P. Rotshtein, J. Geng, P. Sterzer, and R. Henson (2006). “A critique of functional localisers”. *NeuroImage* 30 (4), pp. 1077–1087.

Gorgolewski, K., C. D. Burns, C. Madison, D. Clark, Y. O. Halchenko, M. L. Waskom, and S. S. Ghosh (Aug. 2011). “Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python.” *Front Neuroinform* 5.

Heller, R., D. Stanley, D. Yekutieli, N. Rubin, and Y. Benjamini (2006). “Cluster-based analysis of fMRI data”. *NeuroImage* 33 (2), pp. 599–608.

Hu, J., H. Zhao, and H. Zhou (2010). “False discovery rate control with groups”. *J. Am. Stat. Assoc.* 105 (491), pp. 1215–1227.

Huang, Y., X. Liu, R. Krueger, T. Santander, X. Hu, K. Leach, and W. Weimer (2019). “Distilling neural representations of data structure manipulation using fMRI and fNIRS”. In: *Proceedings of International Conference on Software Engineering (ICSE)*. IEEE, pp. 396–407.

Jarmasz, M. and R. Somorjai (2002). “Exploring regions of interest with cluster analysis (EROICA) using a spectral peak statistic for selecting and testing the significance of fMRI activation time-series”. *Artif Intell Med* 25 (1), pp. 45–67.

Krueger, R., Y. Huang, X. Liu, T. Santander, W. Weimer, and K. Leach (2020). “Neurological divide: An fMRI study of prose and code writing”. In: *Proceedings of International Conference on Software Engineering (ICSE)*, pp. 678–690.

Lazar, N. (2008). *The statistical analysis of functional MRI data*. Statistics for Biology and Health. Springer.

Lindquist, M. (2008). “The statistical analysis of fMRI data”. *Statist. Sci.* 23 (4), pp. 439–464.

Liu, Y., S. Sarkar, and Z. Zhao (2016). “A new approach to multiple testing of grouped hypotheses”. *J. Statist. Plann. Inference* 179, pp. 1–14.

Makris, N., J. M. Goldstein, D. Kennedy, S. M. Hodge, V. S. Caviness, S. V. Faraone, M. T. Tsuang, and L. J. Seidman (2006). “Decreased volume of left and total anterior insular lobule in schizophrenia”. *Schizophr. Res.* 83 (2-3), pp. 155–71.

Nieto-Castañón, A. and E. Fedorenko (2012). “Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses”. *NeuroImage* 63 (3), pp. 1646–1669.

Pennington, N. (1987). “Stimulus structures and mental representations in expert comprehension of computer programs”. *Cognitive Psychology* 19 (3), pp. 295–341.

Rosenblatt, J., L. Finos, W. Weeda, A. Solari, and J. Goeman (Nov. 2018). “All-resolutions inference for brain imaging”. *NeuroImage* 181, pp. 786–796.

Saxe, R., M. Brett, and N. Kanwisher (2006). “Divide and conquer: A defense of functional localizers”. *NeuroImage* 30 (4), pp. 1088–1096.

Schildknecht, K., K. Tabelow, and T. Dickhaus (Feb. 2016). “More specific signal detection in functional magnetic resonance imaging by false discovery rate control for hierarchically structured systems of hypotheses”. *PLOS ONE* 11 (2), pp. 1–21.

Shi, R. and Y. Guo (2016). “Investigating differences in brain functional networks using hierarchical covariate-adjusted independent component analysis”. *Ann. Appl. Stat.* 10 (4), pp. 1930–1957.

Siegmund, J., C. Kästner, S. Apel, C. Parnin, A. Bethmann, T. Leich, G. Saake, and A. Brechmann (2014). “Understanding understanding source code with functional magnetic resonance imaging”. In: *Proceedings International Conference on Software Engineering (ICSE)*. ACM, pp. 378–389.

Siegmund, J., N. Peitek, C. Parnin, S. Apel, J. Hofmeister, C. Kästner, A. Begel, A. Bethmann, and A. Brechmann (2017). “Measuring neural efficiency of program comprehension”. In: *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. ESEC/FSE 2017. New York, NY, USA: Association for Computing Machinery, pp. 140–150.

Soloway, E. and K. Ehrlich (1984). “Empirical studies of programming knowledge”. *IEEE Transactions on Software Engineering.* 10 (5), pp. 595–609.

Talairach, J. and P. Tournoux (1988). *Co-planar stereotaxic atlas of the human brain*. Thieme.

van de Geer, S. (2016). *Estimation and testing under sparsity*. Vol. 2159. Lecture Notes in Mathematics. Lecture notes from the 45th Probability Summer School held in Saint-Four, 2015, École d’Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. Springer, [Cham], pp. xiii+274.

Vovk, V. and R. Wang (2020). “Combining p-values via averaging”. *Biometrika* forthcoming.

– (2021). “E-values: Calibration, combination, and applications”. *Annals of Statistics* forthcoming.

Wainwright, M. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

Wilson, D. (2019). “The harmonic mean p-value for combining dependent tests”. *Proceedings of the National Academy of Sciences* 116 (4), pp. 1195–1200.

Worsley, K. (1994). “Local maxima and the expected Euler characteristic of excursion sets of χ^2, F and t fields.” *Advances in Applied Probability* 26, pp. 13–42.

Worsley, K., S. Marrett, P. Neelin, K. Friston, and A. Evans (1996). “A unified statistical approach for determining significant signals in images of cerebral activation”. *Human Brain Mapping* 4, pp. 58–73.

Yekutieli, D. (2008). “Hierarchical false discovery rate-controlling methodology”. English. *J. Am. Stat. Assoc.* 103 (481), pp. 309–316.

Zhao, H. and J. Zhang (2014). “Weighted p-value procedures for controlling FDR of grouped hypotheses.” English. *J. Stat. Plann. Inference* 151-152, pp. 90–106.