RESEARCH ARTICLE

More Specific Signal Detection in Functional Magnetic Resonance Imaging by False Discovery Rate Control for Hierarchically Structured Systems of Hypotheses

Konstantin Schildknecht[1]*, Karsten Tabelow[1], Thorsten Dickhaus[2]

1 Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany, **2** Institute for Statistics, University of Bremen, Bremen, Germany

* Konstantin.Schildknecht@wias-berlin.de

Abstract

Signal detection in functional magnetic resonance imaging (fMRI) inherently involves the problem of testing a large number of hypotheses. A popular strategy to address this multiplicity is the control of the false discovery rate (FDR). In this work we consider the case where prior knowledge is available to partition the set of all hypotheses into disjoint subsets or families, e. g., by a-priori knowledge on the functionality of certain regions of interest. If the proportion of true null hypotheses differs between families, this structural information can be used to increase statistical power. We propose a two-stage multiple test procedure which first excludes those families from the analysis for which there is no strong evidence for containing true alternatives. We show control of the family-wise error rate at this first stage of testing. Then, at the second stage, we proceed to test the hypotheses within each non-excluded family and obtain asymptotic control of the FDR within each family at this second stage. Our main mathematical result is that this two-stage strategy implies asymptotic control of the FDR with respect to all hypotheses. In simulations we demonstrate the increased power of this new procedure in comparison with established procedures in situations with highly unbalanced families. Finally, we apply the proposed method to simulated and to real fMRI data.

Introduction

Modern research is increasingly concerned with large-scale experiments and complex experimental designs. From a statistical perspective the analysis of such experiments often involves the issue of multiple testing of a large number (say m) of individual hypotheses. The development of methods to deal with this issue is a very active field of research with many sophisticated procedures emerging, e. g., taking a specific structure in the set of hypotheses into account; see, for example, Sections 3.3 and 12.2 in [1].

One example is the analysis of functional magnetic resonance imaging (fMRI) data; see [2] for an overview. At each unit of measurement (voxel) on a regular grid a statistical test is to be performed for the null hypothesis of no activation versus the alternative hypothesis of activation of the voxel (a signal detection problem). In such an application, the number m is often of the order of magnitude of several hundreds of thousand hypotheses.

The family-wise error rate (FWER) and the false discovery rate (FDR) are two established notions for measuring the type I error of a multiple test. The FWER denotes the probability of at least one false rejection among the m individual tests, and a multiple test is said to control the FWER (in the strong sense), if the latter probability is bounded by a pre-defined significance level α over the whole parameter set of the statistical model. One simple way to control the FWER is to carry out every individual test at the adjusted level α/m, commonly referred to as the Bonferroni correction. However, this ignores the spatial correlations of the data (cf. [3]), and can often be improved by multivariate methods. Another strategy for fMRI signal detection with FWER control incorporating the spatial dependencies of the hypotheses is based on the geometry of random fields, see [4] and [5].

On the contrary, the FDR is defined as the expected proportion of type I errors among all rejections of the multiple test φ, and φ is said to control the FDR at a given level $\alpha \in (0, 1)$ if this expected proportion is smaller than α for all parameter values of the considered statistical model. Applying this criterion leads to more liberal multiple tests, meaning that on average more null hypotheses can be rejected. The Benjamini-Hochberg procedure (or linear step-up (LSU) test φ^{LSU}, see [6]) for FDR control has become very popular in fMRI research, cf. [7]. Meanwhile, FDR control is an established criterion for the analysis of high-dimensional data, and is agreed upon to provide a suitable interpretation of the results.

When structural information regarding the hypotheses is at hand, it is often possible to incorporate this external knowledge into the statistical methodology in order to improve the test procedures with respect to power or specificity. In the fMRI context, weighted variants of φ^{LSU} considered in previous work incorporate different aspects of the spatial structure of the activation areas, which are typically organized as clusters of activation rather than as singular spots. Furthermore, the functional organization of the brain defines specific regions of interest related to specific functions that are accessible by suitable experimental paradigms, see [8]. A very old example for such a functional atlas based on cytoarchitecture is the Brodmann atlas (cf. [9]). Clustering techniques to define regions of interest and to incorporate the (in general) heterogeneous cluster sizes into φ^{LSU} were employed in [10] and [11]. Relatedly, in [12] and [13] a case was studied in which the set of hypotheses can be divided into disjoint groups with potentially different proportions of activated voxels by means of a-priori knowledge. The authors demonstrated higher power of their proposed weighted φ^{LSU} tests in comparison with the standard LSU procedure if the fraction of true null hypotheses differs between the groups.

Another class of weighted FDR-controlling multiple tests introduces a second layer of hypotheses which are added to the original set of the m individual hypotheses. Namely, each of the considered disjoint groups is associated with the group-specific null hypothesis of no activation of the whole group. This leads to a hierarchical hypotheses structure with two levels. One level consists of all the group hypotheses and the other of all the m individual hypotheses. In such a context, hierarchical multiple test procedures consist of two stages: First, the group hypotheses are tested, and families for which the group hypothesis cannot be rejected are excluded from the analysis. This strategy relaxes the (remaining) multiplicity for the second stage, where the individual hypotheses are tested. This situation was investigated, for instance, in [14, 15] and [16], and is also often encountered in other application fields like genetic association studies (cf. [17]), gene expression analyses (cf. [18]), or in electroencephalography research (cf. [19]).

In this paper we develop a new two-stage method for FDR control in the fMRI context that takes into account an a-priori partition of the brain into disjoint families of voxels. The main innovation is that non-linear critical values or rejection curves, respectively, are utilized in the second stage. To this end, we make use of the approach in [20] and [21] for implicit adaptation of FDR-controlling multiple test procedures to the amount of signals. While these papers only considered the individual hypotheses, we apply their reasoning within every group which is still under consideration in the second stage of the hierarchical two-stage test. This leads to high sensitivity regarding the voxels within such a group. This is combined with a Bonferroni-type multiplicity adjustment in the first stage, implying a good specificity during the detection of active regions (testing of the group hypotheses). We prove that this procedure controls the FWER on the set of the family hypotheses, as well as, asymptotically as $m \to \infty$, the FDR within each family and the global FDR (gFDR), which is the FDR with respect to all individual hypotheses.

The remaining sections are structured as follows. In Section "Methods" the mathematical notation is set up, some known results about FDR control are reported, the considered two-stage procedures are introduced and their statistical properties are analyzed. To evaluate the proposed new procedure we perform a number of simulations, and we analyze real fMRI data. To this end, the experimental setups and the most important results are explained and reported in Section "Results". We conclude with a discussion in the subsequent Section "Discussion". Lengthy mathematical derivations are deferred to S1 Appendix. For the sake of completeness, additional experimental results are provided in S1 and S2 Tables as well as in the figures in S1, S2, S3, S4, S5, S6 and S7 Figs.

Methods

Notation and preliminaries

We denote the number of families of hypotheses by k and the families themselves by $\mathcal{H}_1, \ldots, \mathcal{H}_k$. Each set $\mathcal{H}_\ell$ is assumed to consist of $m_\ell > 0$ individual hypotheses $H_{\ell 1}, \ldots, H_{\ell m_\ell}$, $1 \leq \ell \leq k$. In addition, for each of the k groups we consider a screening (or family) hypothesis H_ℓ^f, $1 \leq \ell \leq k$, which we will formally define in Definition 4. The aims of the statistical analysis are (i) FDR control in each family $\mathcal{H}_\ell$ separately, (ii) FDR control with respect to all individual hypotheses pooled together, denoted by the global FDR, (iii) FWER control on the group level, i. e., with respect to $(H_\ell^f)_{1 \leq \ell \leq k}$. We assume that for each hypothesis a (marginal) p-value is available, which we identify by the same sub- and / or superscript as the corresponding hypothesis.

Definition 1 (Linear step-up test φ^{LSU}) *Denote by $p_{1:m} \leq p_{2:m} \leq \ldots \leq p_{m:m}$ the ordered p-values for a collection $\mathcal{H}_m = \{H_i, i \in I = \{1, \ldots, m\}\}$ of null hypotheses at hand. Furthermore, let $H_{1:m}, \ldots, H_{m:m}$ denote the re-ordered null hypotheses in $\mathcal{H}_m$, according to the ordering of the p-values. Then, the linear step-up test φ^{LSU} at FDR level $\alpha \in (0, 1)$ rejects exactly the hypotheses $H_{1:m}, \ldots, H_{i^*:m}$, where*

$$i^* = \max\{i \in I : p_{i:m} \leq i\alpha/m\}. \tag{1}$$

If the maximum in Eq (1) does not exist, then no hypothesis is rejected.

The linear step-up test belongs to the broad class of step-up-down (SUD) multiple tests, introduced in [22].

Definition 2 (Step-up-down test of order λ in terms of p-values, cf. [21]) *Let $p_{1:m} \leq p_{2:m} \leq \ldots \leq p_{m:m}$ and α be defined as in Definition 1. For a tuning parameter $\lambda \in \{1, \ldots, m\}$ a step-up-down test $\varphi^\lambda = (\varphi_1, \ldots, \varphi_m)$ (say) of order λ based on some critical values $\alpha_{1:m} \leq \cdots \leq \alpha_{m:m}$ is defined as follows: If $p_{\lambda:m} \leq \alpha_{\lambda:m}$, set $i^* = \max\{j \in \{\lambda, \ldots, m\} : p_{i:m} \leq \alpha_{i:m} \text{ for all } i \in \{\lambda, \ldots, j\}\}$,*

whereas for $p_{\lambda:m} > \alpha_{\lambda:m}$, *put* $i^* = \sup\{j \in \{1, \ldots, \lambda - 1\} : p_{j:m} \leq \alpha_{j:m}\}$ (sup $\emptyset = -\infty$). *Define* $\varphi_i = 1$ *if* $p_i \leq \alpha_{i^*:m}$ *and* $\varphi_i = 0$ *otherwise* ($\alpha_{-\infty:m} = -\infty$).

A step-up-down test of order $\lambda = 1$ or $\lambda = m$, respectively, is called step-down (SD) or step-up (SU) test, respectively. If all critical values are identical, we obtain a single-step test.

In case of φ^{LSU}, $\lambda = m$ and $\alpha_{i:m} = i\alpha/m$ for all $1 \leq i \leq m$. In general, the choice of the order λ and of the critical values employed in an SUD test for FDR control depends on model assumptions; cf. Table 5.1 in [1].

Definition 3 (AORC-based critical values, cf. [20] and [21]) *Under the assumptions of Definitions 1 and 2, we denote by ϖ_{λ}^{AORC} the SUD test with critical values*

$$\alpha_{i:m} = \frac{i\alpha}{m - i(1 - \alpha)}, \ 1 \leq i \leq m. \tag{2}$$

The critical values in Eq (2) correspond to the so-called asymptotically optimal rejection curve (AORC) introduced in [20]. For suitable choices of λ and under the assumption of stochastically independent p-values, ϖ_{λ}^{AORC} has been shown to exhaust the FDR level α asymptotically as $m \rightarrow \infty$, while φ^{LSU} is not exhausting α if the number of true null hypotheses is smaller than m.

In a two level situation with group hypotheses and individual hypotheses, a two-stage procedure can be employed. In our case we are interested in testing the hypotheses within a family $\mathcal{H}_\ell$ only if this family has been declared active, meaning that H_ℓ^f has been rejected in the first stage of analysis. In the fMRI context a family consists of many individual hypotheses and we consider a single activation in a family (an isolated signal) rather as noise than as evidence for activation of the family. Therefore we employ a criterion which defines a family as active if there is at least a certain proportion of activated voxels in the family. This proportion has to be predefined in advance. A useful tool to formalize activity of families in this context is the partial conjunction hypothesis introduced in [23].

Definition 4 *For a given integer $1 \leq u_\ell \leq m_\ell$, the u-partial conjunction hypothesis H^{u_ℓ/m_ℓ} for family $\mathcal{H}_\ell$ is defined as the set of parameters such that $\mathcal{H}_\ell$ contains less than u_ℓ false null hypotheses, with corresponding alternative given by the set of parameters such that the number of true alternatives in $\mathcal{H}_\ell$ is at least equal to u_ℓ. Based on this, we let $H_\ell^f = H^{u_\ell/m_\ell}$. According to [23] a valid p-value for testing H_ℓ^f, under the assumption of positive regression dependency on subsets (PRDS) on the joint distribution of the m_ℓ individual p-values, can be defined as*

$$p^{u_\ell/m_\ell} = \min_{1 \leq i \leq m_\ell - u_\ell + 1} \left\{ \frac{m_\ell - u_\ell + 1}{i} p_{u_\ell - 1 + i:m_\ell} \right\}. \tag{3}$$

In general, a critical issue in connection with FDR control is the dependency structure among the p-values. The LSU test controls the FDR under the PRDS assumption regarding the joint distribution of the p-values, see [24] and [25]. It was shown in [26] that φ^{LSU} cannot be improved uniformly if the dependency among the p-values is completely unknown. Other procedures as the one introduced in [27] assume weak dependency in the sense of Definition 5.

Definition 5 (Weak dependency) *Let $p_1, \ldots, p_m$ denote (random) marginal p-values for a collection $\mathcal{H}_m = \{H_i, i \in I = \{1, \ldots, m\}\}$ of null hypotheses at hand. Let $I_N \subseteq I$ ($I_A \subseteq I$) with $|I_N| = m_N$ ($|I_A| = m_A$) denote the index set of true (false) null hypotheses in I. Then, $p_1, \ldots, p_m$ are*

called weakly dependent, if $q_N = \lim_{m \to \infty} m_N/m$ exists and

$$\hat{F}_{Nm_N}(t) = m_N^{-1} \sum_{i \in I_N} \mathbb{I}_{[0,t]}(p_i) \to F_N(t), \quad m \to \infty \tag{4}$$

$$\hat{F}_{Am_A}(t) = m_A^{-1} \sum_{j \in I_A} \mathbb{I}_{[0,t]}(p_j) \to F_A(t), \quad m \to \infty, \tag{5}$$

where $\mathbb{I}_S$ denotes the indicator function of the set S, convergence in Eqs [(4)](#) *and* [(5)](#) *is uniform for $t \in [0, 1]$ and almost sure, and F_N and F_A are continuous cumulative distribution functions with $0 < F_N(t) \le t$ for all $t \in (0, 1]$*

Throughout this work, we assume that the *p*-values within each family are PRDS and weakly dependent. While one might argue against the weak dependency assumption in the fMRI context (cf. [28]), the validity of weak dependency for *p*-values corresponding to voxel data has been discussed in [29] on the basis of simulation studies for different magnitudes of positive correlation among the voxels. No situation militating against the assumption was found. The FDR behaviour of AORC-based multiple test procedures under the weak dependency assumption regarding the joint distribution of the *p*-values was investigated in Chapter 4 of [30].

Considered two-stage multiple tests

In [16] a general method to design procedures coping with the selection of families has been provided. For a comparison with our proposed procedure φ^{HO} we make use of one of the so-called "simple selection adjusted procedures" proposed in [15], which is based on φ^{LSU} and is denoted throughout the remainder by φ^{Bog}. Under suitable assumptions, this procedure achieves control of the FDR on the average over the selected families, FDR control within each family, and FDR control on the level of the families, see [15]. A simulation study in [31] suggests that global FDR control of φ^{Bog} holds in multi-phenotype genome-wide association studies which exhibit similar characteristics as the fMRI studies considered here.

Algorithm 1 (The procedure φ^{Bog})

1. *Test the k families with the LSU procedure at level α applied to $(p^{1/m_\ell})_{1 \le \ell \le k}$, see* [Eq (3)](#). *Obtain R rejections.*

2. *In the case of $R > 0$, apply in each of the R rejected families φ^{LSU} at level $R\alpha/m_\ell$, where ℓ denotes the index of a rejected family.*

We propose to apply the following procedure which harnesses the advantages of the AORC approach and exploits the structural information.

Algorithm 2 (The procedure φ^{HO}) *Let $\lfloor x \rfloor$ denote the largest integer smaller than or equal to x.*

1. *For a given tuning parameter $\kappa > k$, let $u_\ell = \lfloor \kappa^{-1} \cdot m_\ell \rfloor + 1$ for $1 \le \ell \le k$. Reject all families $\mathcal{H}_\ell$ for which*

$$p^{u_\ell/m_\ell} \le \frac{\alpha}{\kappa}.$$

   Obtain R rejections.

2. *In the case of $R > 0$, apply in each of the R rejected families φ_λ^{AORC} at level α, with $\lambda = u_\ell$, where ℓ denotes the index of a rejected family.*

Under standard assumptions which are typically made in FDR theory, all three aims of the statistical analyses (i. e., FDR control in each family $\mathcal{H}_\ell$ separately, global FDR control, and FWER control on the group level) are achieved by φ^{HO}, at least asymptotically as $\min_{1 \leq \ell \leq k} m_\ell \to \infty$; see S1 Appendix for details.

Experiments

We will compare the two hierarchical procedures φ^{HO} and φ^{Bog} with AORC-based SUD tests regarding the empirical power on the combined set of hypotheses in Sections "Computer simulations regarding the power of φ^{HO}" and "Power simulation". In the simulations regarding fMRI data presented in Section "fMRI—Data" and "fMRI—Results", we will make comparisons of φ^{LSU} with the hierarchical procedures on the combined set of voxels by means of their empirical FDRs. When evaluating real fMRI experiments, we compare the respective numbers of detections, i. e., rejections. The procedure φ^{LSU} and the AORC-based SUD tests will be applied to the combined set of voxels, neglecting the hierarchical structure.

Computer simulations regarding the power of φ^{HO}

In this section we consider the performance of the procedures in terms of power of a multiple test. A standard notion of power of a multiple test procedure $\varphi_{(m)}$ for m hypotheses is given in Definition 1.4 of [1] as

$$\text{power}_m(\varphi_{(m)}) = \mathbb{E}\left[\frac{S_m}{m_A \vee 1}\right],$$

where S_m denotes the number of correct rejections and the expectation $\mathbb{E}$ refers to the true underlying measure. The global power of a multiple test procedure $\varphi_{(m)}$ that operates on a structured family of hypotheses as considered in Section "Methods" is given by

$$\text{gpower}_m(\varphi_{(m)}) = \mathbb{E}\left[\frac{S_m}{m_A \vee 1}\right] = \mathbb{E}\left[\frac{\sum_{\ell=1}^{k} S_\ell}{\sum_{\ell=1}^{k} m_{A\ell} \vee 1}\right],$$

where $m_{A\ \ell}$ and S_ℓ are the number of false null hypotheses and the number of correct rejections in family ℓ. For a given number B of Monte Carlo repetitions, the power of $\varphi_{(m)}$ is estimated by the average value

$$\widehat{\text{power}}_m(\varphi_{(m)}) = \frac{1}{B}\sum_{b=1}^{B}\frac{s_{m,b}}{m_A},$$

where $s_{m,\ b}$ denotes the realization of S_m in the b-th simulation run. In our simulations, we set $B = 10{,}000$ and $m = 2{,}500$.

The simulations refer to the one-sided normal means problem with $\Omega = \mathbb{R}^m$, an observable random vector $T = (T_1, \ldots, T_m)^\top$ with values in Ω such that $\mathcal{L}(T) = \mathcal{N}_m(\mu, I_m)$, where $\mu = (\mu_1, \ldots, \mu_m)^\top$, and hypotheses

$$H_j : \{\mu_j = 0\} \text{ vs. } K_j : \{\mu_j > 0\}, j \in \{1, \cdots, m\}.$$

The p-value for a hypothesis H_j is then given by

$$p_j(t_j) = \mathbb{P}_{H_j}(T_j > t_j) = 1 - \Phi(t_j),$$

where t_j denotes the observed value of T_j and Φ denotes the cumulative distribution function of the standard normal distribution.

For convenience, we set all $\mu_j, j \in I_A$, to the same value $\mu^* > 0$. The power of the different procedures will be investigated for different effect sizes μ^*. The effect size μ^* will range from 0.5 up to 5 in steps of 0.5. Furthermore, we assume that the family $\mathcal{H}_m = (H_1, \ldots, H_m)$ is structured into two subfamilies $\mathcal{H}_{m_1}$ and $\mathcal{H}_{m_2}$. The parameter κ is set to 1,000, see Section "Power simulations" for justification. We let $\pi_\ell = m_\ell/m$ and $q_{N\ell} = m_{N\ell}/m_\ell$, $\ell = 1, 2$, where $m_{N\ell}$ denotes the number of true null hypotheses in family ℓ. Table 1 lists the considered parameter configurations. The FDR level was set to $\alpha = 5\%$ in all simulations.

fMRI—Data. Simulations and analysis of experimental data were all performed within the **R** language and environment for statistical computing and graphics, cf. [32]. The **R**-scripts for the creation of the simulated data and their analysis are available from http://www.wias-berlin. de/preprint/2127/codeANDdata_2127.zip. *Simulated fMRI data.* We simulated fMRI data using the **R**-package **neuRosim** (cf. [33]) described in detail in [34]. The simulated data consisted of 105 volumes of size $20 \times 20 \times 20$ isotropic voxels. The simulated stimulus had onset times at the 16-th, 46-th and 76-th volume, a duration overlapping 15 volumes and a repetition time of two seconds. The expected hemodynamic response to this block design was created using a convolution of the task indicator function with the standard "double-gamma" hemodynamic response function, see [35]. The "activation" region in this data was set to a sphere of radius 3 voxels. The center of the sphere was set in voxel coordinates (5, 5, 5) for Simulation A and in voxel (10, 10, 10) for Simulation B. Noise was added using a Rician distribution including spatial and temporal correlations.

We then analyzed the data with a standard general linear model (GLM) approach using the **R**-package **fmri** (cf. [36] and [37]) including corrections for temporal autocorrelations and quadratic signal trends. From the resulting statistical parametric map we determined local p-values.

We defined an arbitrary partition of the spatial domain into eight families of voxels corresponding to the eight "corners" of the data cube. For both simulation datasets we then applied the hierarchical testing procedures φ^{HO} and φ^{Bog}, as well as φ^{LSU} at an FDR level of 0.05.

Statistical Parametric Mapping (SPM) auditory fMRI test data. For validation of our new inference method on experimental fMRI data we used a publicly available single subject fMRI dataset with an auditory stimulus design. The data can be downloaded at http://www.fil.ion. ucl.ac.uk/spm/data/auditory/ together with details on its acquisition.

The number of volumes at a repetition time of 7 seconds was 96 with alternating blocks of rest and auditory stimulus, starting with rest, each lasting for six volumes. Echo planar imaging (EPI) data was acquired on a modified 2T Siemens MAGNETOM Vision system. The spatial dimension of the data was $64 \times 64 \times 64$ isotropic voxels of length 3mm. Calculation of local p-values was performed as described for the simulated fMRI data.

To define suitable families of voxels we normalized AFNI's (cf. [38]) EPI template (`TT_EPI-tlrc`) in Talairach space with Brodmann labels to the functional data using the

Table 1. Parameter configurations in the one-sided normal means problem.

	$\pi = (\pi_1, \pi_2)$	$q_N = (q_{N1}, q_{N2})$
1	(0.5, 0.5)	(0.5, 0.5)
2	(0.5, 0.5)	(0.8, 0.1)
3	(0.8, 0.2)	(0.8, 0.1)
4	(0.5, 0.5)	(0.99, 0.01)
5	(0.8, 0.2)	(0.99, 0.01)

doi:10.1371/journal.pone.0149016.t001

normalization toolbox of SPM8. Thus each voxel in the functional data was assigned a label according to the Brodmann atlas. Any other suitable atlas or definition of families could have been used here. We then applied the procedures φ^{HO}, φ^{Bog}, and φ^{LSU} to all voxels that had been assigned any label by the atlas matching described above, restricting analysis to the labelled cortex areas only.

fMRI dataset using a sports imagination task. We also re-used an fMRI dataset from [37] originating from an experiment performed with one healthy adult female subject. The data are publicly available under http://www.jstatsoft.org/v44/i11. The alternating design of rest and task blocks, starting with rest, was identical to the one of the simulated fMRI data and resulted in 105 volumes. The rest and task blocks had a duration of 30 seconds, the repetition time was 2 seconds. The task was imagination of playing tennis. The spatial dimension of the data cube was $64 \times 64 \times 30$ with an in-plane resolution of 3.75mm and a slice thickness of 4mm. The echo time of the EPI sequence was 40ms and the flip angle was 80 degrees. Before the first rest block six dummy scans were discarded to allow for T_1 saturation.

We repeated the analysis described for the SPM auditory fMRI test data, i.e., normalizing the Brodmann labels to the functional data using SPM8 and performing a standard GLM analysis with the **R**-package **fmri** to calculate local *p*-values. Signal detection was performed using the procedures φ^{HO}, φ^{Bog}, as well as φ^{LSU}.

Other fMRI datasets. We also analyzed two more fMRI scans of another subject in a finger tapping task within the same task protocol as described for the sports imagination dataset. One of the datasets had a doubled in-plane resolution. The analysis yielded very similar results (with respect to the performance of the signal detection procedure) as the sports imagination dataset, which is why we decided not to show the results of the analysis here.

Results

Power simulations

The five panels in Fig 1 refer to the five parameter configurations from Table 1 with the choice of $\kappa = 1,000$. This choice was made to ensure that the partial conjunction hypotheses coincide with the intersection hypotheses, for comparative purposes with the other procedures. For specific values of the proportion of true null hypotheses, the influence of κ on the performance of the procedure φ^{HO} is demonstrated in S1 Appendix.

In the second panel row of Fig 1 (comprising panels 4–5), the ratios $q_{N\ell}$, $\ell = 1, 2$, are highly unbalanced. It can clearly be observed that this leads to an improvement in terms of power of the proposed procedure φ^{HO} over the existing multiple tests φ^{Bog} and $\varphi^{AORC}_{u_\ell}$, at least for $\mu^* \in [2, 3]$. In the first panel row (comprising panels 1–3), however, the empirical power of $\varphi^{AORC}_{u_\ell}$ is uniformly higher than that of φ^{Bog} and φ^{HO}, respectively.

We may remark that a more detailed analysis of the decision patterns of the three concurring multiple tests (not shown here) revealed that the higher power of $\varphi^{AORC}_{u_\ell}$ in panels 2 and 3 is mainly due to the fact that φ^{Bog} and φ^{HO} discard the first family $\mathcal{H}_{m_1}$ already in the first stage of the analysis (with high probability). Often, such a behavior is wanted in practice, because few isolated signals are typically interpreted as artifacts, especially in the fMRI context.

fMRI—Results

fMRI—Simulations. We first show the results for Simulation A, where the "activation area" is fully located within one of the defined families, in Figs 2, 3 and 4. Every procedure detects all true alternatives (which are marked in yellow), but we can observe a different
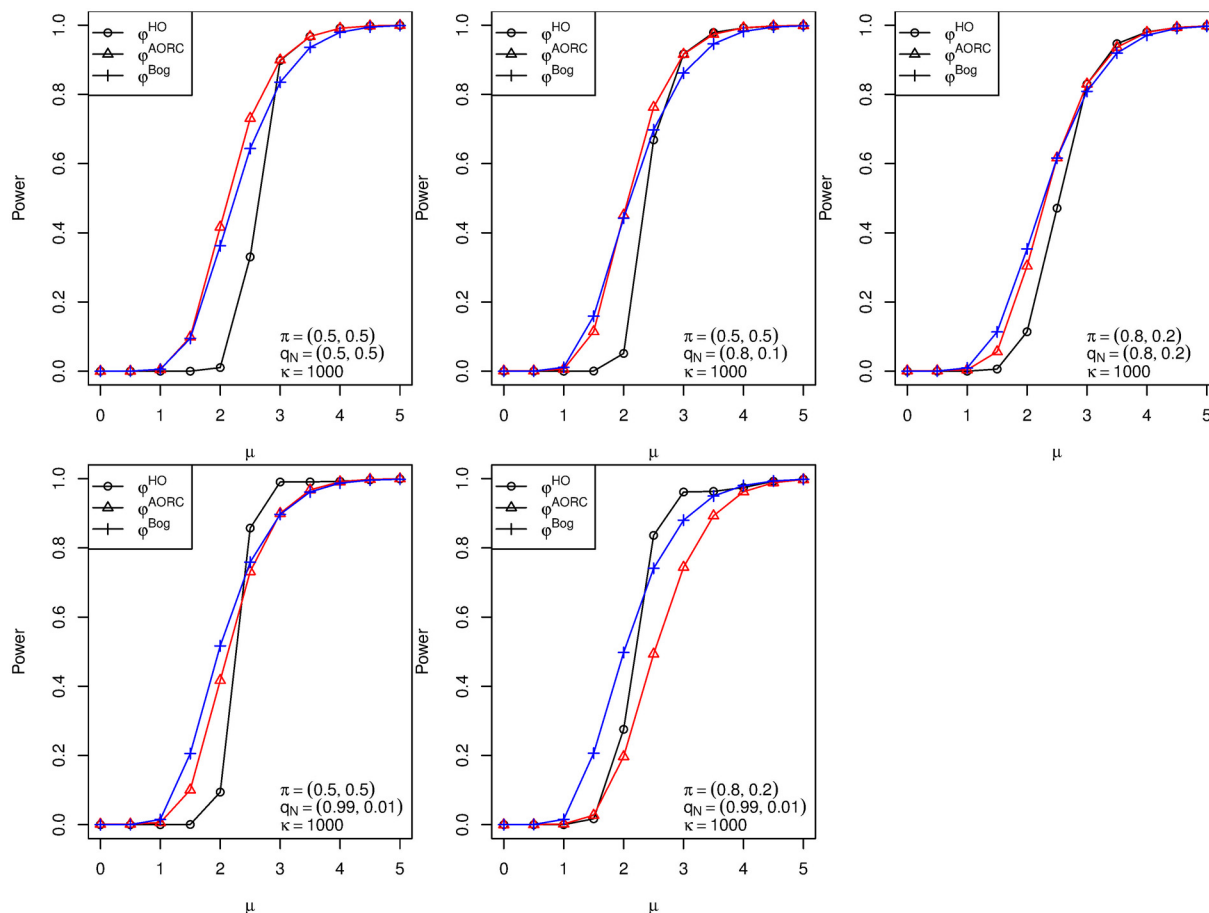
Fig 1. Empirical powers of the procedures. Empirical powers of the procedures φ^{HO} (black), φ^{Bog} (blue) and $\varphi_{u_\ell}^{AORC}$ (red) as a function of the effect size μ^* in the one-sided normal means problem. The total number of hypotheses equals $m = 2500$, and the number of groups equals $k = 2$. The parameter configurations $\pi = (\pi_1, \pi_2)$ and $q_N = (q_{N1}, q_{N2})$ are as in Table 1.

number of false discoveries (indicated in red). The hierarchical procedure φ^{HO} does not make any discoveries in families without activation.

Comparing the detected activation areas with the known ground truth, we estimated the global and within-family false discovery rates as well as the mean FDR over the families for 1000 Monte Carlo repetitions. We can observe differences regarding the detection of false positives, see Table 2. The procedure φ^{LSU} has the most rejections, but violates the FDR in every family, except for the family in which the signal is located. All empirical FDRs are below 5% for the other two procedures.

We show the detection results for Simulation B, where true activations are located within all defined families of voxels, in Figs 5, 6 and 7. In analogy to the presentations under Simulation A, we show the slices of activated voxels determined by the three different procedures overlayed with the true activation.

A visual inspection of the figures and the table confirms the desired behaviour of the procedure. In Table 2 we clearly observe that the families without activation (i. e., families 2–8 in Simulation A) are in most of the Monte Carlo repetitions excluded from the analysis by φ^{HO}
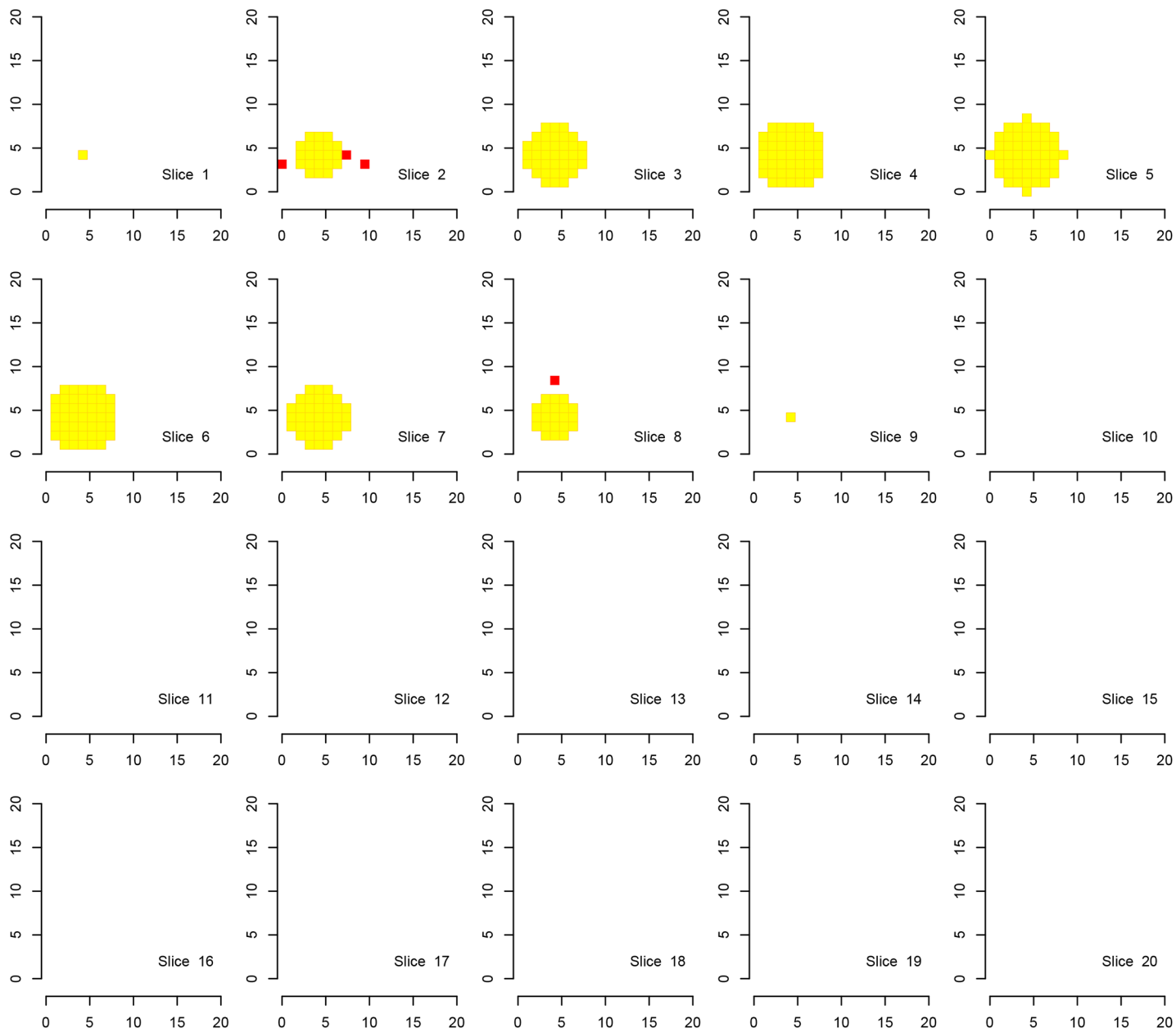
Fig 2. Discoveries of the procedure φ^{HO} in fMRI Simulation A. Discoveries of the procedure φ^{HO} in Simulation A on a cube with side length 20. There are eight disjoint families consisting of cubes with a side length of 10 voxels, each one located in one corner of the original cube. Shown are 20 slices corresponding to the third dimension. Ground activation (yellow) and the false rejections of φ^{HO} (red) are shown.

doi:10.1371/journal.pone.0149016.g002

and φ^{Bog}. In contrast, activation is reported in all families when using the test φ^{LSU}. It is not surprising that in families without signal the FDR in the family is not controlled for the LSU-procedure. If the signal is found in every family (Simulation B) there is no advantage in the use of the hierarchical approach. The order of magnitude regarding the FDRs seems to be the same
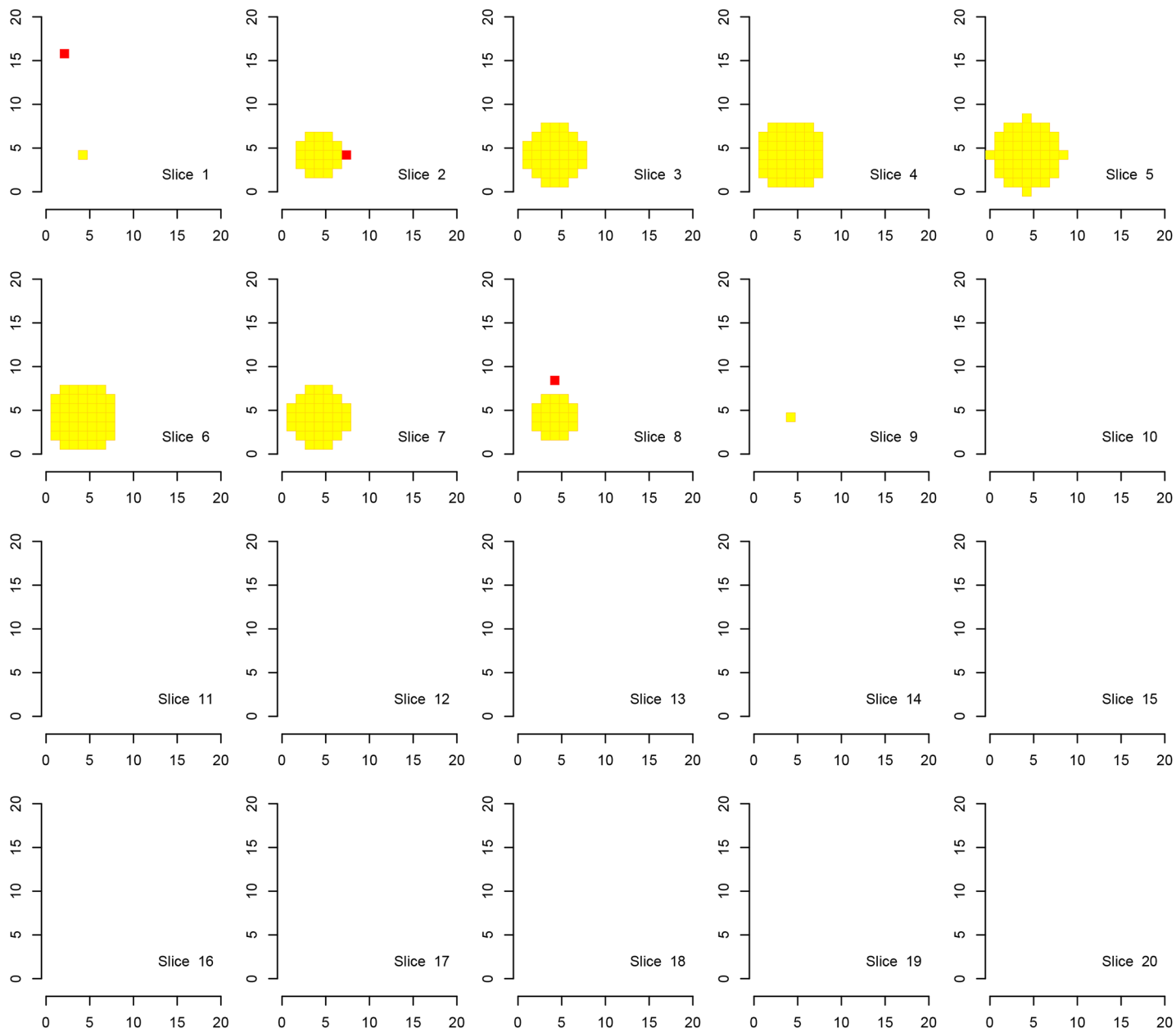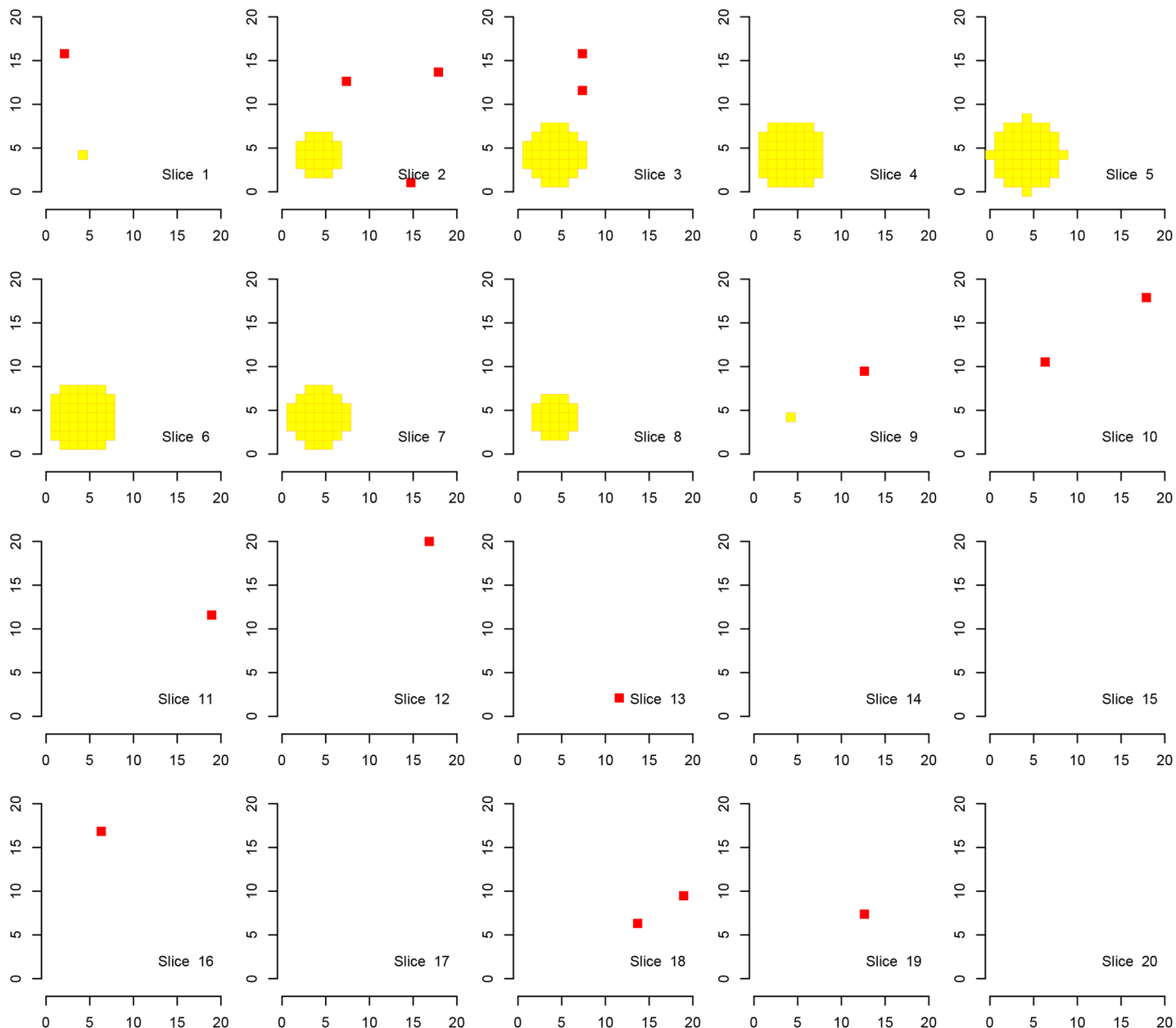
Fig 3. Discoveries of the procedure φ^{Bog} in fMRI Simulation A. Discoveries of the procedure φ^{Bog} in Simulation A on a cube with side length 20. There are eight disjoint families consisting of cubes with side length of 10 voxels, each one located in one corner of the original cube. Shown are 20 slices corresponding to the third dimension. Ground activation (yellow) and false rejections of φ^{Bog} (red) are shown.

for the two hierarchical procedures, although the attained FDR level of the procedure φ^{HO} is closer to 5%, suggesting higher power.

SPM auditory fMRI test data. We show the detection results in the auditory cortex of the proposed procedure φ^{HO} overlayed on the functional division of the brain according to the

Fig 4. Discoveries of the procedure φ^{LSU} in fMRI Simulation A. Discoveries of the procedure φ^{LSU} in Simulation A on a cube with side length 20. There are eight disjoint families consisting of cubes with side length of 10 voxels, each one located in one corner of the original cube. Shown are 20 slices corresponding to the third dimension. Ground activation (yellow) and the false rejections of φ^{LSU} (red) are shown.

Brodmann atlas and compare them with the detections found by the procedures φ^{LSU} and φ^{Bog} in Fig 8. We can see that the hierarchical procedures detect voxels mainly located in the auditory areas, while the LSU procedure finds activations all over the brain. The full figures showing all slices can be found in S2, S3 and S4 Figs.

Table 2. Global FDR, mean FDR and within family FDR in the fMRI Simulation A and B for the different procedures.

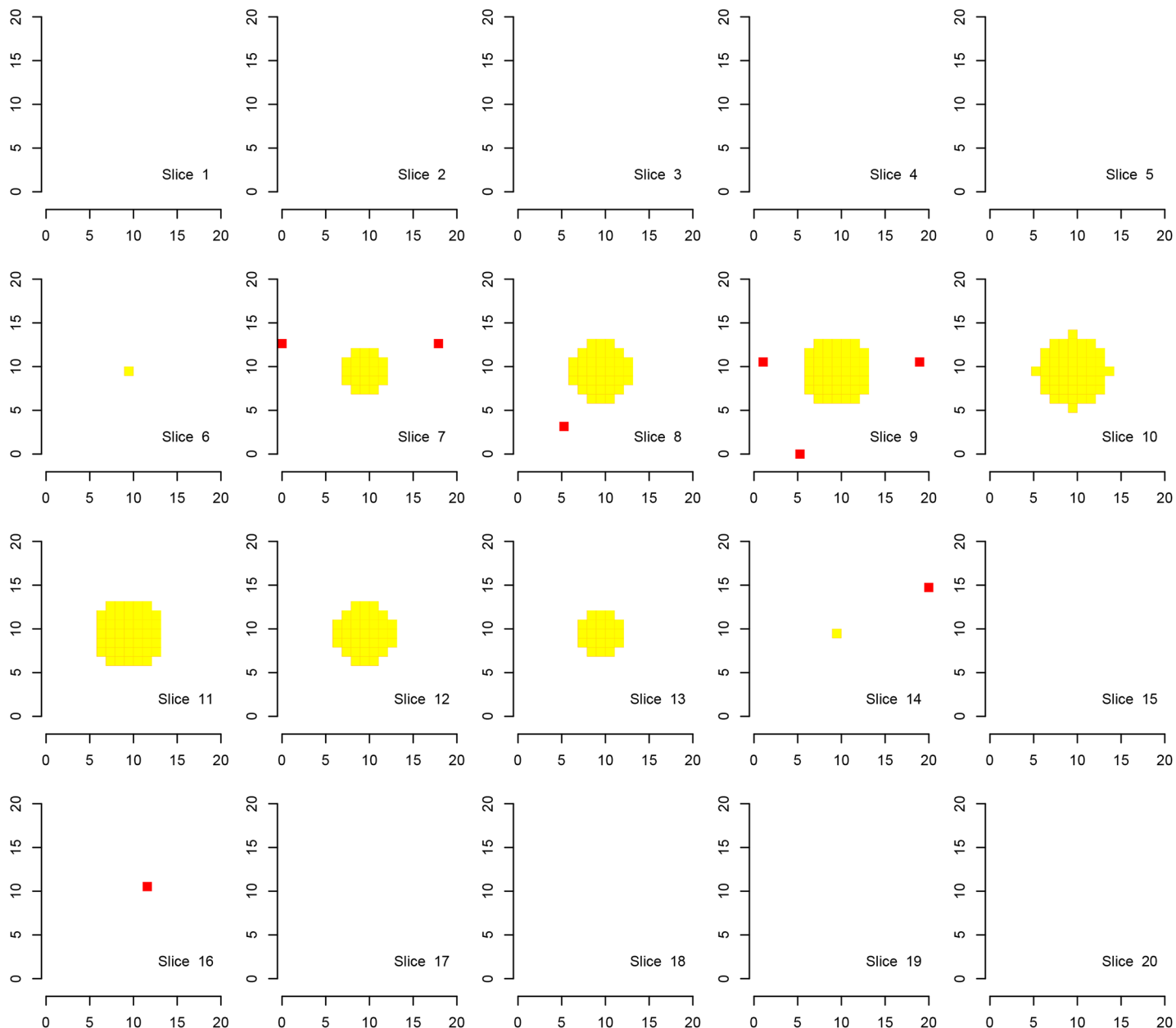|	Simulation A			Simulation B		
	φ^{HO}	φ^{LSU}	φ^{Bog}	φ^{HO}	φ^{LSU}	φ^{Bog}
gFDR	0.0352	0.0333	0.026	0.0362	0.0341	0.0347
FDR Fam. 1	0.0352	0.003	0.0257	0.0354	0.0191	0.0331
FDR Fam. 2	0	0.685	0.006	0.0357	0.0288	0.006
FDR Fam. 3	0	0.696	0.015	0.0356	0.0295	0.014
FDR Fam. 4	0	0.678	0.006	0.0331	0.0419	0.006
FDR Fam. 5	0.001	0.676	0.005	0.0354	0.0285	0.005
FDR Fam. 6	0	0.683	0.010	0.0363	0.0443	0.010
FDR Fam. 7	0	0.693	0.012	0.0355	0.0428	0.012
FDR Fam. 8	0	0.691	0.005	0.0337	0.0603	0.005
mean FDR	0.0045	0.6006	0.0106	0.0351	0.0369	0.0114

doi:10.1371/journal.pone.0149016.t002

The table in S1 Table shows the number of discoveries in the different Brodmann areas. In agreement with Fig 8, it can be seen from this table that the proposed procedure leads to a far more concentrated signal detection in areas related to the auditory stimulus.

fMRI dataset using a sports imagination task. We show the detection results of the proposed procedures overlayed on the Brodmann atlas. A visual inspection of Fig 9 reveals activation in the whole brain. As it can be seen in the table in S2 Table, in every area of the brain many activated voxels are detected by all procedures. We might hypothesize that the stimulus of this experiment, which is an imagination task, is related to much less specific activation due to its complexity. Similar to the situation in fMRI Simulation B we do not observe that the hierarchical procedures are more specific than φ^{LSU} regarding the Brodmann areas. The full figures are provided in S5, S6 and S7 Figs.

Discussion

This work focused on the use of structural information in a new procedure to control the FDR. We provided a rigorous mathematical analysis of this new procedure and proved asymptotic control of the FDR. In simulations we studied the performance of the proposed method in situation with finite m. Furthermore, we applied it to simulated and real fMRI datasets.

For fMRI analysis our procedure bears the unique advantage of being specific to the families/regions in which brain activity is located and is highly sensitive within each family. This conclusion can be clearly drawn from Table 2 and is supported by the figures. Other FDR controlling procedures suffer from false positives in areas without signal. We first filter where strong signal can be found and continue to locate the voxels which are responsible for the strong signal, making use of the nonlinear critical values originating from the theory around the AORC. It was possible to demonstrate that when the activation is concentrated in a-priori known regions the procedure can be used to increase the specificity on the level of the families while finding a similar number of discoveries as the standard approaches within the families of interest. The hierarchical approach was demonstrated to perform close to the non-hierarchical approach if families do not differ in the number of true alternatives. However, we forfeit sensitivity for weak signals if the pre-test is not passed. The use of the Brodmann atlas for the real fMRI data is just a simple example of a division of the brain into functionally different regions, which can (and should) be replaced by more suitable selections
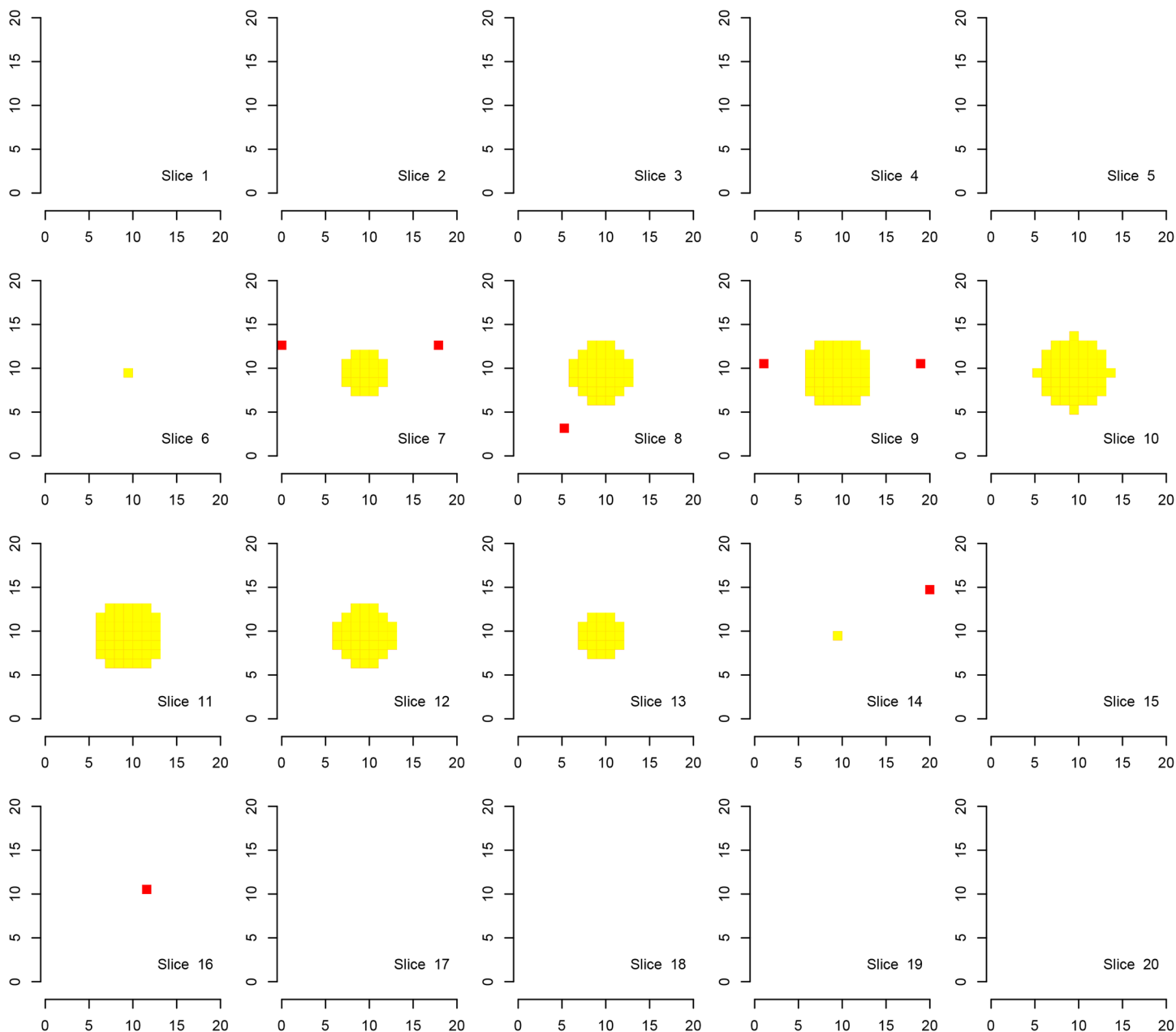
Fig 5. Discoveries of the procedure φ^{HO} in fMRI Simulation B. Discoveries of the procedure φ^{HO} in Simulation B on a cube with side length 20. There are eight disjoint families consisting of cubes with a side length of 10 voxels, each one located in one corner of the original cube. Ground activation (yellow) and the false rejections of φ^{HO} (red) are shown.

doi:10.1371/journal.pone.0149016.g005

in specific applications. In summary our procedure shows superior specificity during the detection of active regions of interest in the brain while being highly sensitive regarding the voxels within a detected region, suggesting good applicability of the FDR in signal detection in fMRI.
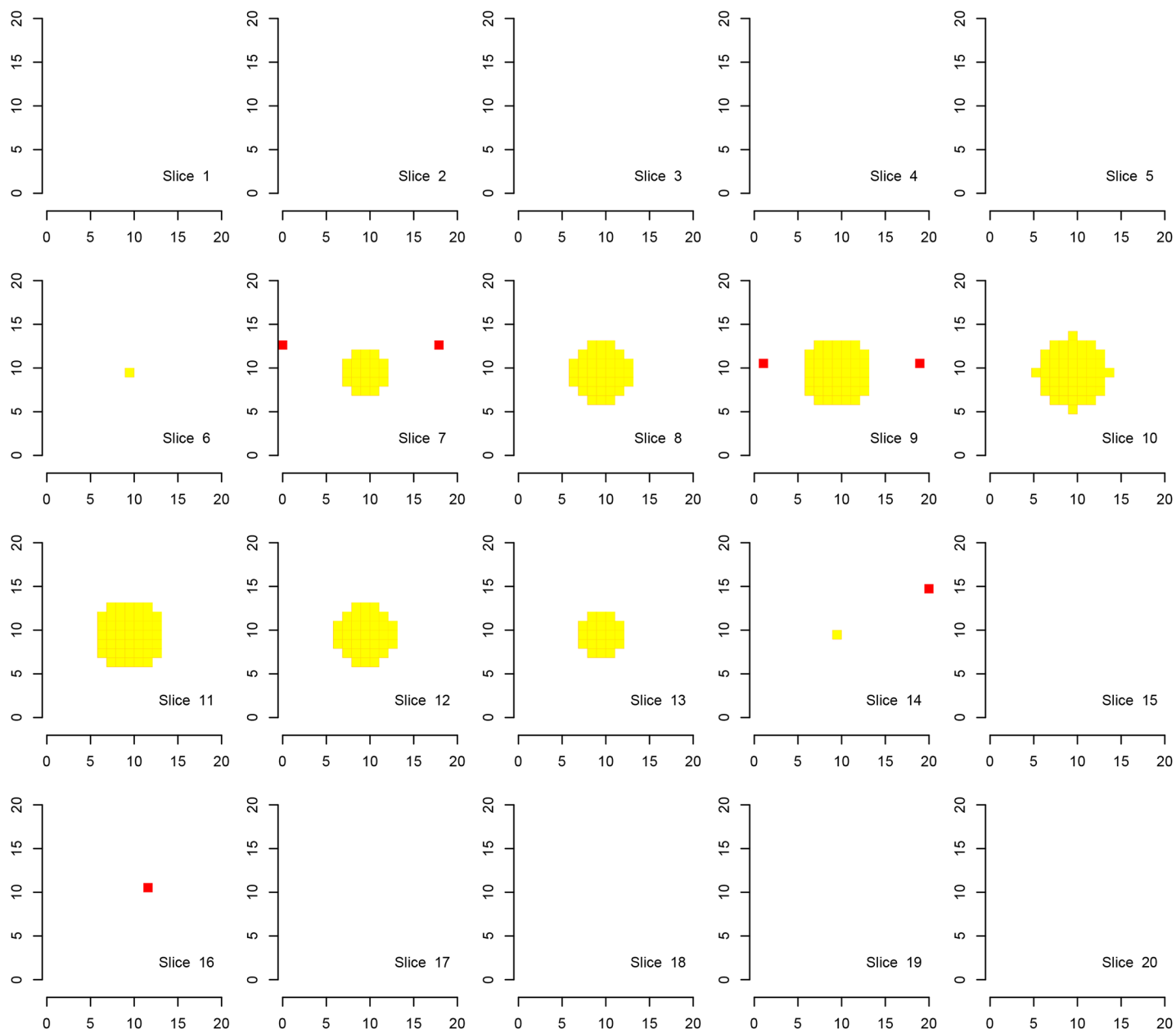
Fig 6. Discoveries of the procedure φ^{Bog} in fMRI Simulation B. Discoveries of the procedure φ^{Bog} in Simulation B on a cube with side length 20. There are eight disjoint families consisting of cubes with a side length of 10 voxels, each one located in one corner of the original cube. Ground activation (yellow) and the false rejections of φ^{Bog} (red) are shown.

doi:10.1371/journal.pone.0149016.g006

From a more general perspective, the proposed procedure φ^{HO} is designed to discard families which contain only few scattered signals. This may result in sub-optimal global power, but leads to higher specificity on the group level, compared with non-hierarchical procedures which test all m hypotheses together. Often, as in the fMRI context discussed above, the groups

Fig 7. Discoveries of the procedure φ^{LSU} in fMRI Simulation B. Discoveries of the procedure φ^{LSU} in Simulation B on a cube with side length 20. There are eight disjoint families consisting of cubes with side length of 10 voxels, each one located in one corner of the original cube. Ground activation (yellow) and the false rejections of φ^{LSU} (red) are shown.

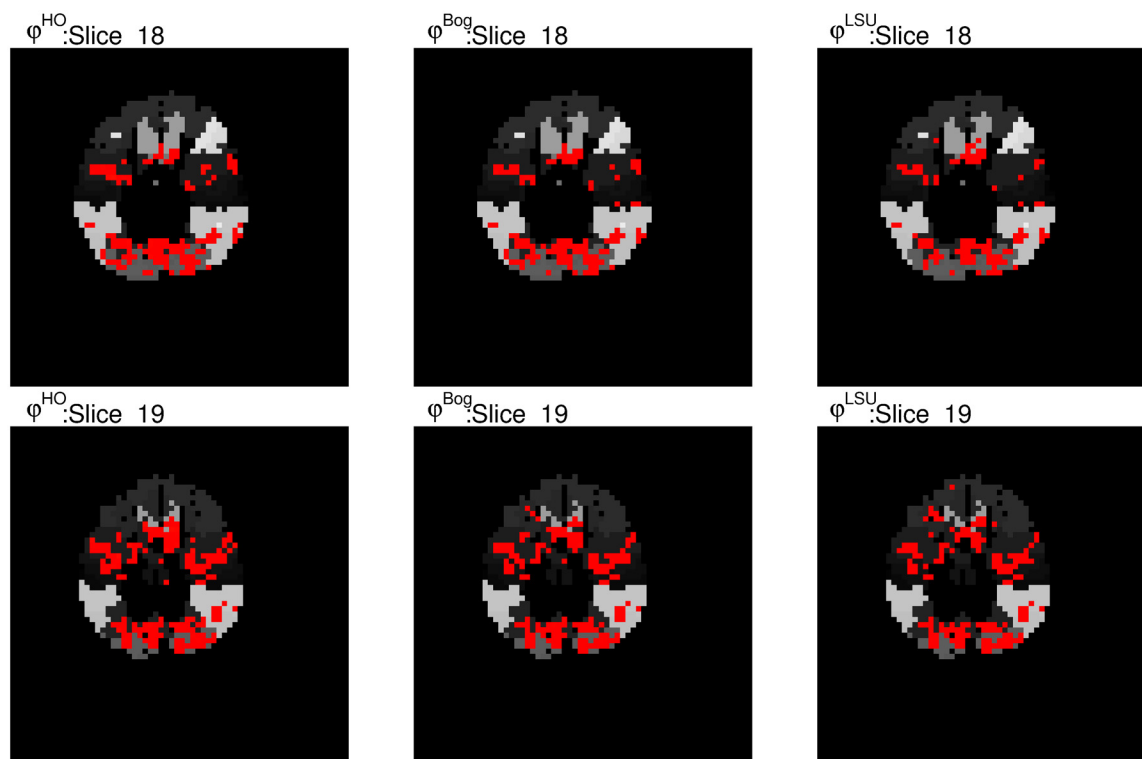doi:10.1371/journal.pone.0149016.g007
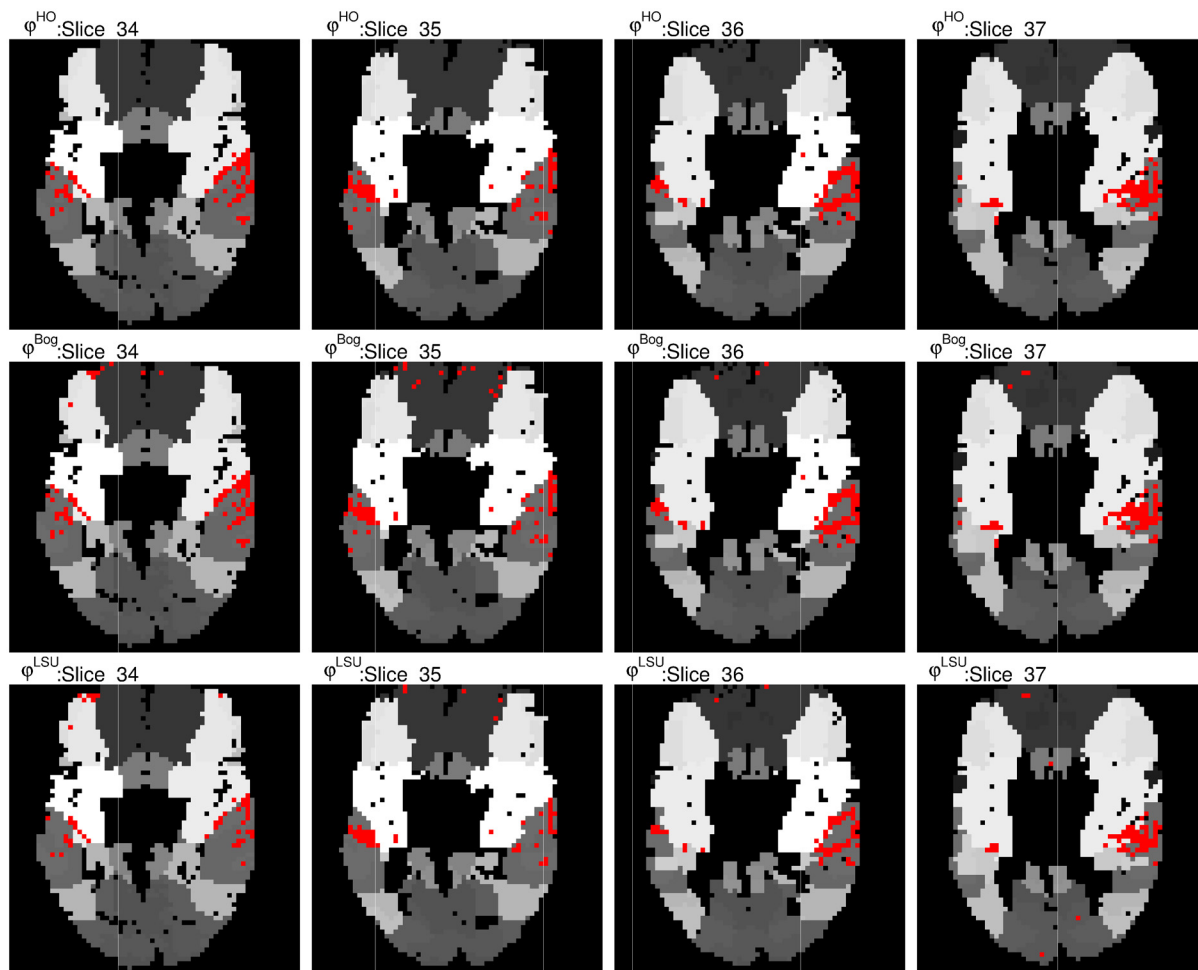
Fig 8. Discoveries on chosen slices of the brain for the SPM auditory fMRI dataset. We present chosen slices (auditory cortex visible) of the brain for the SPM auditory fMRI dataset and the discoveries proposed procedure φ^{HO} in the first row. The discoveries in the second row correspond to the procedure φ^{Bog} and in the third row the corresponding discoveries of φ^{LSU} are displayed.

doi:10.1371/journal.pone.0149016.g008

are the experimental units of interest, and in such a situation the hierarchical approach is recommendable. The test φ^{HO} depends on a tuning parameter κ, which has to be chosen by the researcher before the start of the analysis. A value $\kappa \leq m_\ell$ for a family $\mathcal{H}_\ell$ has the interpretation, that a family is declared active if there is evidence that it contains at least m_ℓ/κ true alternatives. If $\kappa > m_\ell$ the partial conjunction hypothesis becomes the intersection hypothesis.

An interesting and challenging direction for future research is the consideration of additional layers of hierarchy in FDR-controlling multiple test procedures. For example, consider a hierarchical system $\mathcal{H}_m$ of m hypotheses which is closed under intersection. In the case that FWER control at level α is targeted, the closure principle (see [39]) allows one to test all m hypotheses in $\mathcal{H}_m$ at full level α, provided that the coherence rule is adhered to (rejection of a hypothesis $H_i \in \mathcal{H}_m$ implies that all hypotheses in $\mathcal{H}_m$ which are subsets of H_i are also rejected). How this principle can be transferred to the concept of (global) FDR control will be explored in future work.

Fig 9. Discoveries on chosen slices of the brain for the sports imagination task dataset. We present chosen slices (motor cortex visible) of the brain for the sports imagination task and highlight the discoveries of the procedure φ^{HO} in the first column. The discoveries in the second column correspond to the procedure φ^{Bog} and in the third column the corresponding discoveries of φ^{LSU} are displayed.

doi:10.1371/journal.pone.0149016.g009

Supporting Information

S1 Appendix. Mathematical derivations. Mathematical proofs and investigation of the proposed procedure regarding the tuning parameter κ.
(PDF)

S1 Table. Discoveries in the SPM auditory experiment. Number of discoveries in the SPM auditory experiment overall and in each Brodmann area for the procedures φ^{HO}, φ^{Bog}, and φ^{LSU}.
(XLS)

S2 Table. Discoveries in the sports imagination task dataset. Number of discoveries in the sports imagination task dataset overall and in each Brodmann area for the procedures φ^{HO}, φ^{Bog}, and φ^{LSU}.
(XLS)

S1 Fig. Empirical power of φ^{HO}. Empirical power of the procedure φ^{HO} for two different fractions q_N of true null hypotheses, as a function of the tuning parameter κ and the signal strength μ^* in the normal means problem with variance 1.
(TIF)

S2 Fig. Discoveries of φ^{HO} for the SPM auditory dataset. Discoveries of the proposed procedure φ^{HO} for the SPM auditory fMRI dataset on the Brodmann areas of the brain for all slices.
(TIF)

S3 Fig. Discoveries of φ^{Bog} for the SPM auditory dataset. Discoveries of the procedure φ^{Bog} for the SPM auditory fMRI dataset on the Brodmann areas of the brain for all slices.
(TIF)

S4 Fig. Discoveries of φ^{LSU} for the SPM auditory dataset. Discoveries of the procedure φ^{LSU} for the SPM auditory fMRI dataset on the Brodmann areas of the brain for all slices.
(TIF)

S5 Fig. Discoveries of φ^{HO} for the sports imagination task dataset. Discoveries of the proposed procedure φ^{HO} for the sports imagination task dataset on the Brodmann areas of the brain for all slices.
(TIF)

S6 Fig. Discoveries of φ^{Bog} for the sports imagination task dataset. Discoveries of the procedure φ^{Bog} for the sports imagination task dataset on the Brodmann areas of the brain for all slices.
(TIF)

S7 Fig. Discoveries of φ^{LSU} for the sports imagination task dataset. Discoveries of the procedure φ^{LSU} for the sports imagination task dataset on the Brodmann areas of the brain for all slices.
(TIF)

Acknowledgments

Author Contributions

Conceived and designed the experiments: KS KT. Performed the experiments: KS KT. Analyzed the data: KS TD KT. Contributed reagents/materials/analysis tools: KS TD KT. Wrote the paper: KS TD KT.

References

1. Dickhaus T. Simultaneous statistical inference. With applications in the life sciences. Berlin: Springer; 2014.

2. Lazar NA. The Statistical Analysis of Functional MRI Data. Statistics for Biology and Health. Springer; 2008.

3. Worsley KJ. Detecting activation in fMRI data. Stat Methods in Med Res. 2003; 12:401–418. doi: 10.1191/0962280203sm340ra

4. Worsley KJ, Evans AC, Marrett S, Neelin P. A Three-Dimensional Statistical Analysis for CBF Activation Studies in Human Brain. J Cereb Blood Flow Metab. 1992; 12(6):900–918. doi: 10.1038/jcbfm.1992.127 PMID: 1400644

5. Adler RJ, Taylor JE. Random fields and geometry. New York, NY: Springer; 2007.

6. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol. 1995; 57(1):289–300.

7. Genovese CR, Lazar NA, Nichols T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage. 2002; 15(4):870–878. PMID: 11906227

8. Huettel S, Song AW, McCarthy G. Functional Magnetic Resonance Imaging. 3rd ed. Sinauer Associates, Inc; 2014.

9. Brodmann K. Vergleichende Lokalisationslehre der Großhirnrinde in ihren Prinzipien dargestellt auf Grund des Zellbaues; 1909. Leipzig: Barth.

10. Heller R, Stanley D, Yekutieli D, Rubin N, Benjamini Y. Cluster-based analysis of FMRI data. NeuroImage. 2006; 33(2):599–608. PMID: 16952467

11. Benjamini Y, Heller R. False discovery rates for spatial signals. J Am Stat Assoc. 2007; 102 (480):1272–1281. doi: 10.1198/016214507000000941

12. Hu JX, Zhao H, Zhou HH. False Discovery Rate Control With Groups. J Am Stat Assoc. 2010; 105 (491):1215–1227. doi: 10.1198/jasa.2010.tm09329 PMID: 21931466

13. Zhao H, Zhang J. Weighted *p*-value procedures for controlling FDR of grouped hypotheses. J Stat Plann Inference. 2014; 151–152:90–106.

14. Yekutieli D. Hierarchical false discovery rate-controlling methodology. J Am Stat Assoc. 2008; 103 (481):309–316. doi: 10.1198/016214507000001373

15. Bogomolov M. Testing of Several Families of Hypotheses. Ph. D. dissertation, Tel-Aviv University; 2011.

16. Benjamini Y, Bogomolov M. Selective inference on multiple families of hypotheses. J R Stat Soc Ser B Stat Methodol. 2014; 76(1):297–318. doi: 10.1111/rssb.12028

17. Yekutieli D, Reiner-Benaim A, Benjamini Y, Elmer GI, Kafkafi N, Letwin NE, et al. Approaches to multiplicity issues in complex research in microarray analysis. Stat Neerl. 2006; 60(4):414–437. doi: 10.1111/j.1467-9574.2006.00343.x

18. Li Y, Ghosh D. A two-step hierarchical hypothesis set testing framework, with applications to gene expression data on ordered categories. BMC Bioinformatics. 2014; 15:Article 108.

19. Singh AK, Phillips S. Hierarchical control of false discovery rate for phase locking measures of EEG synchrony. NeuroImage. 2010; 50(1):40–47. doi: 10.1016/j.neuroimage.2009.12.030 PMID: 20006711

20. Finner H, Dickhaus T, Roters M. On the false discovery rate and an asymptotically optimal rejection curve. Ann Stat. 2009; 37(2):596–618. doi: 10.1214/07-AOS569

21. Finner H, Gontscharuk V, Dickhaus T. False Discovery Rate Control of Step-Up-Down Tests with Special Emphasis on the Asymptotically Optimal Rejection Curve. Scandinavian Journal of Statistics. 2012; 39(2):382–397. doi: 10.1111/j.1467-9469.2012.00791.x

22. Tamhane AC, Liu W, Dunnett CW. A generalized step-up-down multiple test procedure. Can J Stat. 1998; 26(2):353–363. doi: 10.2307/3315516

23. Benjamini Y, Heller R. Screening for partial conjunction hypotheses. Biometrics. 2008; 64(4):1215–1222. doi: 10.1111/j.1541-0420.2007.00984.x PMID: 18261164

24. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001; 29(4):1165–1188.

25. Sarkar SK. Some results on false discovery rate in stepwise multiple testing procedures. Ann Stat. 2002; 30(1):239–257. doi: 10.1214/aos/1015362192

26. Guo W, Rao MB. On control of the false discovery rate under no assumption of dependency. J Stat Plann Inference. 2008; 138(10):3176–3188.

27. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. J R Stat Soc, Ser B, Stat Methodol. 2004; 66(1):187–205. doi: 10.1111/j.1467-9868.2004.00439.x

28. Logan BR, Geliazkova MP, Rowe DB. An evaluation of spatial thresholding techniques in fMRI analysis. Hum Brain Mapp. 2008 Dec; 29(12):1379–1389. doi: 10.1002/hbm.20471 PMID: 18064589

29. Chen S, Wang C, Eberly LE, Caffo BS, Schwartz BS. Adaptive control of the false discovery rate in voxel-based morphometry. Hum Brain Mapp. 2009; 30(7):2304–2311. doi: 10.1002/hbm.20669 PMID: 19034901

30. Gontscharuk V. Asymptotic and Exact Results on FWER and FDR in Multiple Hypotheses Testing. Ph. D. dissertation, Heinrich-Heine-Universität Düsseldorf; 2010.

31. Peterson C, Bogomolov M, Benjamini Y, Sabatti C. Many Phenotypes without Many False Discoveries: Error Controlling Strategies for Multi-Traits Association Studies; 2015. Preprint, arXiv:1504.00701v1.

32. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2015. ISBN 3-900051-07-0. Available from: http://www.R-project.org.

33. Welvaert M. neuRosim: Functions to generate fMRI data including activated data, noise data and resting state data; 2012. R package version 0.2-10. Available from: http://CRAN.R-project.org/package=neuRosim.

34. Welvaert M, Durnez J, Moerkerke B, Verdoolaege G, Rosseel Y. neuRosim: An R Package for Generating fMRI Data. J Stat Softw. 2011; 44(10):1–18. doi: 10.18637/jss.v044.i10

35. Glover GH. Deconvolution of impulse response in event-related BOLD fMRI. NeuroImage. 1999; 9:416–429. PMID: 10191170

36. Tabelow K, Polzehl J. fmri: Analysis of fMRI Experiments; 2015. R package version 1.5-1. Available from: http://CRAN.R-project.org/package=fmri.

37. Tabelow K, Polzehl J. Statistical Parametric Maps for Functional MRI Experiments in R: The Package fmri. J Stat Softw. 2011 10; 44(11):1–21. doi: 10.18637/jss.v044.i11

38. Cox RW. AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. Comput and Biomed Res. 1996; 29:162–173.

39. Marcus R, Peritz E, Gabriel KR. On closed test procedures with special reference to ordered analysis of variance. Biometrika. 1976; 63(3):655–660. doi: 10.1093/biomet/63.3.655