Power and sample size calculations for fMRI based on the prevalence of active peaks.

Joke Durnez[1,2], Jasper Degryse [3], Beatrijs Moerkerke [3], Jean-Baptiste Poline [5], Ruth Seurinck [3], Vanessa Sochat [1], Russell A. Poldrack[1], Thomas E. Nichols [4]

[1]Stanford University, Stanford, CA, USA

[2]Parietal team, INRIA, Neurospin, Saclay, Gif-sur-Yvette, France

[3]Ghent University, Ghent, Belgium

[4]University of Warwick, Coventry, United Kingdom

[5] ?

Corresponding author:
Joke Durnez
Department of Psychology
Stanford University
Stanford, CA
USA
Email: Joke.Durnez@gmail.com

Highlights

- The manuscript presents a method to calculate sample sizes for fMRI experiments
- The power analysis is based on the estimation of the mixture distribution of null and active peaks
- The methodology is validated with simulated and real data.

1 Abstract

Mounting evidence over the last few years suggest that published neuroscience research suffer from low power due to the use of small sample sizes. Larger sample sizes increase both the chance of detecting a true effect and the chance that a significant result indicates a true effect. As such, a (prospective) power analysis is a critical component of any paper. In this work we present a simple way to characterize the spatial signal in a fMRI study with just two parameters, and a direct way to estimate these two parameters based on an existing study. Specifically, using just (1) the proportion of the brain activated and (2) the average effect size in activated brain regions, we can produce closed form power calculations for given sample size, brain volume and smoothness. This procedure allows one to minimize the cost of an fMRI experiment, while preserving a predefined level of statistical power. The method is evaluated and illustrated using simulations and real neuroimaging data from the Human Connectome Project. The procedures presented in this paper are made publicly available in a cloud-based toolbox available at www.neuropowertools.org.

Keywords: power, fMRI, neuroimaging, sample size, effect size, statistical power

2 Introduction

In a scientific study, one typically aims for a statistical power of 80%, a quantity proposed by Cohen (1988), implying that a true effect in the population is detected with a 80% chance. Power computations allow researchers to compute the minimal number of subjects to obtain the desired statistical power. As such, power calculations avoid spending time and money on studies that are futile, and also prevent wasting time and money adding extra subjects when sufficient power was already available.

Analyses prior to fMRI experiments can optimise power in two ways. One is the optimisation of the experimental design to ensure maximal statistical power for a given scanning duration and various constraints of behavioural paradigms. Methods have been developed to find the optimal number and arrangement of stimuli over the duration of the experiment for each subject (Henson, 2007; Wager and Nichols, 2003; Friston et al., 1999; Smith et al., 2007). The second use is to find the necessary number of subjects (Desmond and Glover, 2002; Mumford and Nichols, 2008).

While it is straightforward to compute power for a single, univariate response, determining the power of an fMRI study is a formidable task. An array of parameters must be specified such as the within- and between-subject variance, the first and second level design, the temporal autocorrelation and the size of the hypothesized effect, all of which may vary voxel-by-voxel. Many of these parameters may be estimated based on a pilot study, independent of the study to be performed. The most difficult parameter to specify is the location and configuration of voxels where activations are expected.

In the earliest work on power for neuroimaging, Van Horn et al. (1998) used the noncentral F-distribution to visualise voxelwise power for PET data. Desmond and Glover (2002) computed sample sizes for fMRI

blocked designs, in a procedure that included within- and between-subject variability and the mean effect. This model was extended by Mumford and Nichols (2008), where arbitrary designs and temporal autocorrelation were taken into account. This work was intended for voxelwise or Region of Interest (ROI) analyses, where the multiple testing problem was accounted for by suitable adjustment of the alpha level. A more elaborate implementation by Hayasaka et al. (2007) also considered the multiple testing problem by using the non-central random field theory to control the family-wise error rate.

In this work we present a simple way to characterize the spatial signal in a fMRI study, and a direct way to estimate power based on an existing pilot study. Specifically, using (1) the volume of the brain activated and (2) the average effect size in activated brain regions, we can directly calculate power for given sample size, brain volume and smoothness. With such a basic formulation, we hope this will make power analyses prevalent, making better use of scarce research funding and better communicating the potential reproducibility of a study.

The present method is an extension of the procedure presented in Durnez et al. (2014) based on peak statistics. Peaks, local maxima in the statistic image, are particularly tractable as they are approximately spatially independent and have reliable random field theory results for their uncorrected and Familywise Error (FWE) corrected p-values (Durnez et al., 2014). In contrast, individual voxel values have complex dependency, and clusters have unreliable RFT p-values (Woo et al., 2014; Hayasaka, 2003; Durnez et al., 2014; Silver et al., 2011; Eklund et al., 2016). In Durnez et al. (2014) we have presented a method to estimate retrospective power for local maxima, using only an estimate of the prevalence of activation and no further distributional assumptions on the effect of interest. In the present procedure we use a statistic image from a pilot study, and use peaks above a threshold u to fit a mixture model, where a proportion (1-π_1) of the peaks follow a known null distribution, and the remainder follow a Gaussian distribution with unknown mean and variance. Once the alternative distribution H_a is estimated, the distribution can be transformed to account for a different sample size. As such, not only can the posthoc power of the pilot study be estimated, but also power for a new study with the same experiment and a different sample size, allowing general sample size calculations.

In the remainder of this paper, we present our procedure and evaluate it based on simulations that explore different fMRI characteristics, such as spatial extent of the signal and signal intensity. Next, we present an evaluation using 180 genetically unrelated subjects from the Human Connectome Project (HCP) (Van Essen et al., 2012). These HCP data are used for a number of reasons. First, these data are very high quality, resulting in a very high power when including all subjects and thus offer a high level of certainty about the location of the effect. Second, with 180 subjects, we can use subsamples of the data to create many smaller fMRI studies. The sampled results can then be compared to the results of the full dataset. The added value of using real data is that it possesses various unknown noise sources in fMRI that would be impossible to simulate. Third, we demonstrate the procedure on a typical example of an fMRI experiment using an fMRI dataset (Seurinck et al., 2011). Finally, we conclude with a discussion on the topic and the implementation of the procedure in a toolbox.

3 Methods

3.1 Measures of power

First we consider a single one-sided univariate test: The null hypothesis H_0 is rejected in favor of the alternative hypothesis H_a only when the test statistic Z exceeds significance threshold z_α; z_α is chosen to control the type I error rate, $\alpha = P(Z \geq z_\alpha | H_0)$, the power of this test is defined as $P(Z \geq z_\alpha | H_a)$,

which of course requires the distribution of Z under H_a.

In a 3 dimensional multiplicity context like voxelwise testing where many tests are performed simultaneously, several definitions for power exist (Dudoit et al., 2003). We will focus on average and familywise power. For voxelwise inference, let $\mathcal{I}_1$ denote the set of coordinate triplets for voxels that are truly activated. The coordinate triplets in $\mathcal{I}_1$ are characterised by the x, y and z coordinates in the 3 dimensional voxel space. Denote the test statistic for voxel with coordinates i as Z_i. The average power is simply the arithmetic mean of power over non-null voxels:

$$(1 - \beta_{z_\alpha}) = \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} P(Z_i \geq z_\alpha | H_{ai}) \tag{1}$$

where $|\mathcal{I}_1|$ is the number of truly activated voxels, and H_{ai} is the alternative hypothesis at voxel with coordinates i. Assuming a homogeneous signal (i.e. the signal in the data is constant over non-null voxels), average power has the traditional interpretation of power: the probability of a true positive at one voxel. The familywise error rate (FWER) is the probability of at least one type I error among multiple tests. Its counterpart, familywise power, is defined as $P(Z_i \geq z_\alpha \text{ for some } i \in \mathcal{I}_1)$, the probability of at least one true positive.

We can similarly define average and familywise power for peakwise tests, and as with cluster definition, the identification of local maxima depends on a neighborhood. The SPM[1] software (RRID:nif-0000-00305) uses 18-order neighborhood to define maxima, while FSL[2] (RRID:birnlex_2067) uses 26-order neighborhood. However, for sufficiently smooth data, these neighborhood definitions will converge. ~~While peaks can be defined independently of an excursion or screening threshold, it can be sensible to exclude the lowest peaks that fall below u. This excludes the highly variable peaks, and also is required if parametric distributional results are to be used. Let $\mathcal{J}$ comprise the coordinate triplets of all local maxima above u. Denote the test statistic for peak with coordinates j as Z_j^u. Let $\mathcal{J}_1 \subset \mathcal{J}$ denote the set of coordinate triplets for peaks above u corresponding to a voxel containing true signal, while $\mathcal{J}_0 \subset \mathcal{J}$ denotes the set of coordinate triplets for peaks above u corresponding to a voxel containing no true signal. Average power is then defined as~~

$$(1 - \beta^u_{z_\alpha}) = \frac{1}{|\mathcal{J}_1|} \sum_{j \in \mathcal{J}_1} P(Z^u_j \geq z_\alpha | H_{aj}) \tag{2}$$

~~while familywise power is defined as $P(Z_j^u \geq z_\alpha \text{ for some } j \in \mathcal{J}_1)$.~~

Let $\mathcal{J}$ comprise the coordinate triplets of all local maxima. Denote the test statistic for a peak with coordinates j as Z_j. Let $\mathcal{J}_1 \subset \mathcal{J}$ denote the set of coordinate triplets for peaks corresponding to a voxel containing true signal, while $\mathcal{J}_0 \subset \mathcal{J}$ denotes the set of coordinate triplets for peaks corresponding to a voxel containing no true signal. Average power is then defined as

$$(1 - \beta_{z_\alpha}) = \frac{1}{|\mathcal{J}_1|} \sum_{j \in \mathcal{J}_1} P(Z_j \geq z_\alpha | H_{aj}) \tag{3}$$

while familywise power is defined as $P(Z_j \geq z_\alpha \text{ for some } j \in \mathcal{J}_1)$.

The choice to control average power or familywise power is driven by the research hypothesis. If a researcher is only interested in finding one brain region, then controlling the familywise power suffices, as only one significant peak in that region leads to the rejection of the null hypothesis for that brain region. However, when task-related activation is expected in multiple brain regions, control of the familywise

[1] http://www.fil.ion.ucl.ac.uk/spm/

[2] www.fmrib.ox.ac.uk/fsl

power may lead to false negative regions. Within brain regions, the control of familywise power is most intuitive; between brain regions, we argue average power is more useful and in the remainder of this work, we focus on this measure of power.

3.2 Estimation Procedure

To calculate average peakwise power, it is essential to know the distribution of peak heights for truly active peaks Z for peak heights (distribution under H_a). To estimate this distribution, we use all peaks above a screening threshold u. The inclusion of the screening threshold allows us to use parametric distributional results. Therefore, let $\mathcal{J}^u \subset \mathcal{J}$ comprise the coordinate triplets of all local maxima above u. Similarly we define $\mathcal{J}_1^u \subset \mathcal{J}_1$ and $\mathcal{J}_0^u \subset \mathcal{J}_0$ to comprise the coordinate triplets of all local maxima above u corresponding to a voxel containing respectively true signal and no true signal. Denote the test statistic for a peak above u with coordinates j as Z_j^u. For the choice of the screening threshold, we choose a value of $u = 2.5$. We have found this value high enough for the RFT assumptions to be met for peaks (Durnez et al., 2014), but low enough to have sufficient data points to estimate a mixture model. We will further evaluate the choice of u using simulations..

Under the null, non-active peaks Z_j^u with $j \in \mathcal{J}_0^u$, have a simple distribution, asymptotically following an exponential distribution with mean $u + 1/u$ for screening threshold u as u goes to infinity (Worsley, 2007):

$$f(z_j^u | H_0, Z_j^u \geq u) \approx u \exp(-u(z_j^u - u)) \tag{4}$$

See Durnez et al. (2014) for a detailed derivation. For H_a, while a shifted exponential might work well for small signals, we found a truncated normal distribution (truncated at screening threshold u) was better at describing the distribution of active peaks Z_j^u with $j \in \mathcal{J}_1^u$:

$$f(z_j^u | H_a, \mu_1, \sigma_1, Z_j^u \geq u) = \frac{\frac{1}{\sigma_1}\varphi\left(\frac{z_j^u - \mu_1}{\sigma_1}\right)}{1 - \Phi\left(\frac{u - \mu_1}{\sigma_1}\right)}, \tag{5}$$

where φ and Φ are the density and cumulative distribution function of the standard normal distribution, respectively.

The marginal distribution of peak heights Z_j^u ($j \in \mathcal{J}^u$) can be written as the following mixture distribution:

$$f(z_j^u | \pi_1, \mu_1, \sigma_1, Z_j^u \geq u) = (1 - \pi_1) f(z_j^u | H_0, Z_j^u \geq u) + \pi_1 f(z_j^u | H_a, \mu_1, \sigma_1, Z_j^u \geq u), \tag{6}$$

where π_1 is the proportion of true positive peaks among all peaks above u. Instead of estimating all parameters at once, we take a two-stage approach. We first estimate π_1, in order to estimate μ_1 and σ_1. We found this 2 stage approach more stable than estimating all 3 parameters jointly.

There are a variety of estimation methods for π_1 that have been proposed (Benjamini and Hochberg, 2000; Storey and Tibshirani, 2003; Storey, 2002; Pounds and Morris, 2003; Pounds and Cheng, 2004) all based on the observed distribution of the p-values. A comparison of the estimators for peak inference in fMRI data analysis is discussed in Durnez et al. (2014), and here we use the preferred method from that work, the estimator of Pounds and Morris (2003).

Once $\hat{\pi}_1$ is obtained for a dataset with sample size n, we estimate the remaining parameters, μ_1 and σ_1 in Equation 6 using maximum likelihood on the same data. We show how the method is not only

applicable to one-sample or two-sample tests. In the appendix A we show how common models can be written in a form

$$Z^u = \frac{b}{\widetilde{SE}(b)}\sqrt{n}$$

where b is a random variable denoting the average experimental effect and $\widetilde{SE}(b) = SE(b)\sqrt{n}$. This relative standard error attempts to remove sample size dependence.
For a one-sample T-test $\widetilde{SE}(b)$ equals σ. The expected value of the peak height under the alternative before truncation is

$$\mu_1 = \frac{\mu}{\widetilde{SE}(b)}\sqrt{n}$$

where $\mu = E(b)$ is the (non-null) mean in effect units. We define $\delta = \mu/\widetilde{SE}(b)$ to be the unitless effect size; and for the one-sample case, this is exactly Cohen's d, $\delta = \mu/\sigma$. For more details, we refer to the appendix.
To estimate the truncated normal distribution using maximum likelihood, we use the Limited-memory Broyden-Fletcher-Goldfarb-Shanno with box constraints (L-BFGS-B) algorithm (Byrd et al., 1995). The expected peak height under the null is $z = u + 1/u$. As the expected value of the alternative distribution should exceed this value, we set $\hat{\mu}_1 = u + 1/u$ as a lower limit in the optimization algorithm. The standard deviation under the null is $1/u$ and we expect the variance under the alternative to be no less than that of the null, so we take $1/u$ as a conservative lower limit for $\hat{\sigma}_1$. We draw a random pair of starting values for $\hat{\mu}_1$ from a range between $u + 1/u$ and 10 and for $\hat{\sigma}_1$ from a range between $1/u$ and 10.
For a new sample of size n^*, we model the distribution of truly activated peaks **before truncation** as $\mathcal{N}(\mu_1^*, \sigma_1)$ with $\mu_1^* = \delta\sqrt{n^*}$. Note that we assume that the variance of the distribution σ_1^2 remains constant for different sample sizes.
To compute power, we use the normal distribution before truncation to expand the distribution below the threshold. For a given peak statistical threshold z_α, the average peakwise power is computed as

$$1 - \beta_{z_\alpha n^*} = P(Z_j \geq z_\alpha | j \in \mathcal{J}_1) = \Phi\left(\frac{z_\alpha - \mu_1^*}{\sigma_1}\right), \tag{7}$$

the complementary cumulative distribution function of the alternative distribution from Equation 5. The statistical threshold z_α can either be uncorrected or corrected for multiple testing.

3.2.1 Computing the statistical threshold z_α

We evaluate several strategies to select the statistical threshold z_α:

1. UN: Level $\alpha = 0.05$, uncorrected for multiple testing.
2. FDR: Corrected to control the false discovery rate at level 0.05 with the method of Benjamini and Hochberg (1995).
3. RFT: Corrected to control the familywise error rate at level 0.05 using a random field theory (RFT) correction (Friston et al., 2007).

The null distribution of peaks (Equation 4) is used to compute uncorrected p-values for each peak, as well as to compute the thresholds for all of the above methods. We stress that we are not encouraging the use of uncorrected thresholds (Method 1), but we need to verify the accuracy of our method in these basic settings. We use the method proposed by Benjamini and Hochberg (1995) to estimate the

significance threshold. However, the FDR method (2) is adaptive and depends on the data. With higher power, the threshold will also increase. To predict the threshold in larger sample sizes, we have derived the conditional expectation $\text{Fdr}(z)$ of the local false discovery rate given our estimates for π_1 as in Efron (2007), μ_1 and σ_1 (technical details in Appendix C). When setting $\text{Fdr}(z) = 0.05$, we obtain $\hat{z}_\alpha$, the predicted significance threshold when controlling the false discovery rate at level α. Because of the conservative nature of the FDR method (Benjamini and Hochberg, 1995), we expect our predicted significance threshold to be lower than the obtained significance threshold, leading to overestimation of power.

Method 3 assumes that the search volume and the smoothness is the same in the pilot and new study.

3.2.2 The use of a screening threshold u

Due to a substantial number of questions based on a pre-publication preprint of this paper (**?**) regarding the use of a screening threshold u, we would like to further clarify the use of u in the proposed procedure. Even though the detailed information can be found throughout the paper, we summarise the use here in more informal language for clarity.

The estimation procedure mainly consists of two steps: (1) estimating the (distribution) of the effect size in a pilot data set, using Equation 6 and (2) estimating the power to detect those effect sizes in a future data set using Equation 7.

In the first step, we use parametric results from Random Field Theory (Worsley, 2007), which requires the use of a screening threshold. We model H_a using a truncated normal distribution (truncation at u). To this end, we estimate $\hat{\pi}_1$, $\hat{\mu}_1$ and $\hat{\sigma}_1$ in the set of peaks above u. Therefore, in our evaluations in the next sections, we compare our estimates with their true values above u.

In the second step, we aim to compute power for all activation, above and below the screening threshold u. Therefore, we eliminate the screening threshold from H_a to represent a (non-truncated) normal distribution. As such, in our evaluations, we compare the power estimates with their true values without applying a screening threshold u.

3.3 Simulations

Evaluations are done with 500 simulated power analyses. For each realization, for a given sample size n we generate n statistical parametric maps, summarising evidence for activation. These represent the average effect of the design in a first-level analysis for each subject, i.e. a three-dimensional map of subject-specific b-values. We simulate n $64 \times 64 \times 64$ volumes filled with independent standard Gaussian noise, setting the voxel size to $3 \times 3 \times 3$ mm. Images are smoothed with a 3D Gaussian kernel with a full-width at half maximum (FWHM) of 8 mm. After smoothing, images are rescaled to restore unit variance. In each subject-specific b-map, we add true activation in 4 ball-shaped volumes that span either 2, 4, 6 or 8% of the total brain volume. We denote $\mathcal{I}_1$ for the set of all voxel coordinates (not only peaks) where H_a holds and $\mathcal{I}_0$ for null voxels where H_0 holds.

Within these activation regions, the effect size takes a value of 0.5, 0.8, 1 or 1.2 units for all n subjects. As such, we have 16 conditions: 4 activation sizes $\times$ 4 effect sizes. In each analysis, the activation size and effect size is held constant for all subjects. In a next step, a z image is obtained by performing an ordinary least squares group analysis with all subject-specific b-maps using FSL's second level FEAT tool. In this z image, peaks are defined and only peaks above screening threshold u are considered; we use $u = 2.5$ and we show results for $u = 2.0$ and $u = 3.0$ in the supplementary materials. The null distribution from Equation 4 thus gives peak p-values:

$$P(Z_j^u \geq z_j^u | H_0, Z_j^u \geq u) \approx \exp\left(-u(z_j^u - u)\right) \tag{8}$$

We use these simulations to study the performance of the procedure described in section 3.1.

Pilot data We first simulate a statistical map from pilot data from a one-sample t-test on 15 subjects. We compute from the z image all local maxima z_j^u above u with coordinate triplets j. Let $\mathcal{J}_1^u \subset \mathcal{I}_1$ be the coordinates of truly active peaks, and $\mathcal{J}_0^u \subset \mathcal{I}_0$ be the coordinates of truly inactive peaks.
A peak z_j^u is truly active ($j \in \mathcal{J}_1^u$) if $j \in \mathcal{I}_1$ and truly inactive ($j \in \mathcal{J}_0^u$) if $j \subset \mathcal{I}_0$. In these data, we estimate $\hat{\pi}_1$ using the procedure presented by Pounds and Morris (2003) and then the alternative truncated normal distribution $N(\hat{\mu}_1, \hat{\sigma}_1)$ as described above. With $\hat{\pi}_1$, $\hat{\mu}_1$, $\hat{\sigma}_1$, we are able to predict power of future studies in function of sample size n^*.
For each pilot data set, we predict power for a new study with $n = 15, ..., 35$ using equation 7.
To validate the model estimation, we compare the estimated $\hat{\pi}_1$ with the true underlying π_1, which is obtained by calculating the percentage of peaks that are located in activated areas, $|\mathcal{J}_1^u|/|\mathcal{J}^u|$. The screening threshold has the goal to eliminate a large portion of null peaks. As such π_1 will be higher than the voxelwise proportion of activation in the brain (i.e. 2, 4, 6 or 8 %). We also calculate the average peak effect size in truly active areas in the simulated data:

$$\widetilde{E}(Z_j^u | H_a) = \frac{1}{|\mathcal{J}_1^u|} \sum_{j \in \mathcal{J}_1^u} Z_j^u.$$

We compare $\widetilde{E}(Z_j^u | H_a)$ not with $\hat{\mu}_1$, but with the estimated expected peak height of the labeled active peaks, accounting for the screening threshold u:

$$\widehat{E}(Z_j^u | H_a) = \hat{\mu}_1 + \hat{\sigma}_1 \hat{\tau}, \tag{9}$$

where $\hat{\mu}_1$ and $\hat{\sigma}_1$ are estimated from the $n = 15$ training data, and

$$\hat{\tau} = \frac{\varphi\left(\frac{u - \hat{\mu}_1}{\hat{\sigma}_1}\right)}{1 - \Phi\left(\frac{u - \hat{\mu}_1}{\hat{\sigma}_1}\right)}.$$

The mixture distribution is estimated and evaluated in 500 simulated pilot studies of sample size n.

Study data To validate the power predictions for the final experimental study, we simulate study data for $n = 15, ..., 60$ as described above. We now compute from the z image all local maxima z_j **without threshold** u with coordinate triplets j, with $j \in \mathcal{J}_1$. We apply the significance thresholds described in section 3.2 to the identified peaks. The predicted power $(1 - \hat{\beta}_{z_\alpha})$ using the pilot data and Equation 7 is compared to the empirically derived peakwise average power $(1 - \tilde{\beta}_{z_\alpha}^u)$ in the simulated images for which the underlying truth is known:

$$(1 - \tilde{\beta}_{z_\alpha}) = \frac{|z_j^u > z_\alpha; j \in \mathcal{J}_1|}{|\mathcal{J}_1|}$$

Hence, $(1 - \hat{\beta}_{z_\alpha})$ for a sample size n^* is estimated in each of 500 simulated pilot studies of sample size n and the average over these simulations is compared to the average of the empirically derived power in 500 simulated studies of sample size n^*.

3.4 HCP data

We use data from the Human Connectome Project (HCP) (Van Essen et al., 2012) to evaluate our method with complex signal and noise structure. We use results from analysed task fMRI data with 5mm smoothing, 'level 2' models (where data from different runs on the same task are combined). For each of 180 unrelated subjects we have 47 unique contrasts.

We use a working assumption that any group analysis of 100 or more participants gives high powered results and can reflect population results.

The procedure used to analyze these data is shown in Figure 1 and is described below. The validation mechanism we use can be related to cross-validation for supervised statistical learning applications: we evaluate the generalization error by splitting our working data in multiple subsamples. One subsample is used to make predictions about larger samples. On the other larger, disjoint subsample, we evaluate those predictions. This gives us confidence about how accurately our algorithm is able to predict the power for previously unseen data.

We start with 180 subject-specific b-maps. We repeat the following resampling strategy 500 times. We split the 180 subject-specific b-maps in three subsamples: the held-in pilot data ($n = 15$), the held-in study data ($n = 15, 16, ..., 65$) and the held-out reference data ($n = 100$).

Analysis A. Held-out reference data Unlike the simulated data, we cannot observe or control which voxels are truly activated. With $n_{\text{Ref}} = 100$, we will use the set of FWER-significant voxels as working set of **empirically active** voxels. We denote $\tilde{\mathcal{I}}_1$ the set of coordinate triplets for all significant voxels, with $Z_i \geq z_\alpha$, and $\tilde{\mathcal{I}}_0$ for the remaining voxels. The threshold z_α will be set differently according to the desired multiple testing strategy (see section 3.2.1)

Analysis B. Held-in pilot data The subsample of 15 subjects serves as the held-in pilot data. On these data we perform a Ordinary Least Squares (OLS) group analysis using FSL's FEAT and define peaks based on the resulting z image. We apply a screening threshold $u = 2.5$. We derive peaks and compute peak p-values (Equation 8) and apply our estimation procedure. As in the simulations, we transform the mean of the alternative distribution to the truncated expected effect size as $\widehat{E}(Z_j^u)$ (see Equation 9).

Analysis C. Held-in study data Next, we take a subsample of n subjects, $n = 15, 16, ..., 65$. These data represent the experimental data for which the pilot analysis has served. We refer to these data as the held-in study data. With the same procedure as with the pilot data, we compute peak p-values and we perform a peakwise analysis with the thresholding procedures described in section 3.2.

Analysis D. Validation of model parameters To validate the estimation procedure, we combine the held-in pilot data and the held-out data to validate the estimation of $\hat{\pi}_1$ and the effect size as follows: when a peak from the held-in pilot data corresponds to an empirically active voxel in the powerful held-out data, we consider this as an empirically active peak and when a peak corresponds to an empiricially inactive voxel in the held-out data, it is considered empirically inactive; consistent with previous notation, $\tilde{\mathcal{J}}_1^u$ are the coordinates of empirically active peaks ($j \in \tilde{\mathcal{J}}_1^u$ if $z_j > u$), $\tilde{\mathcal{J}}_0^u$ are coordinates of empirically inactive peaks ($j \in \tilde{\mathcal{J}}_0^u$ if $z_j > u$).

We define **empirically derived** $\widetilde{\pi_1}$, as the ratio of the number of empirically activated peaks and the total number of peaks:

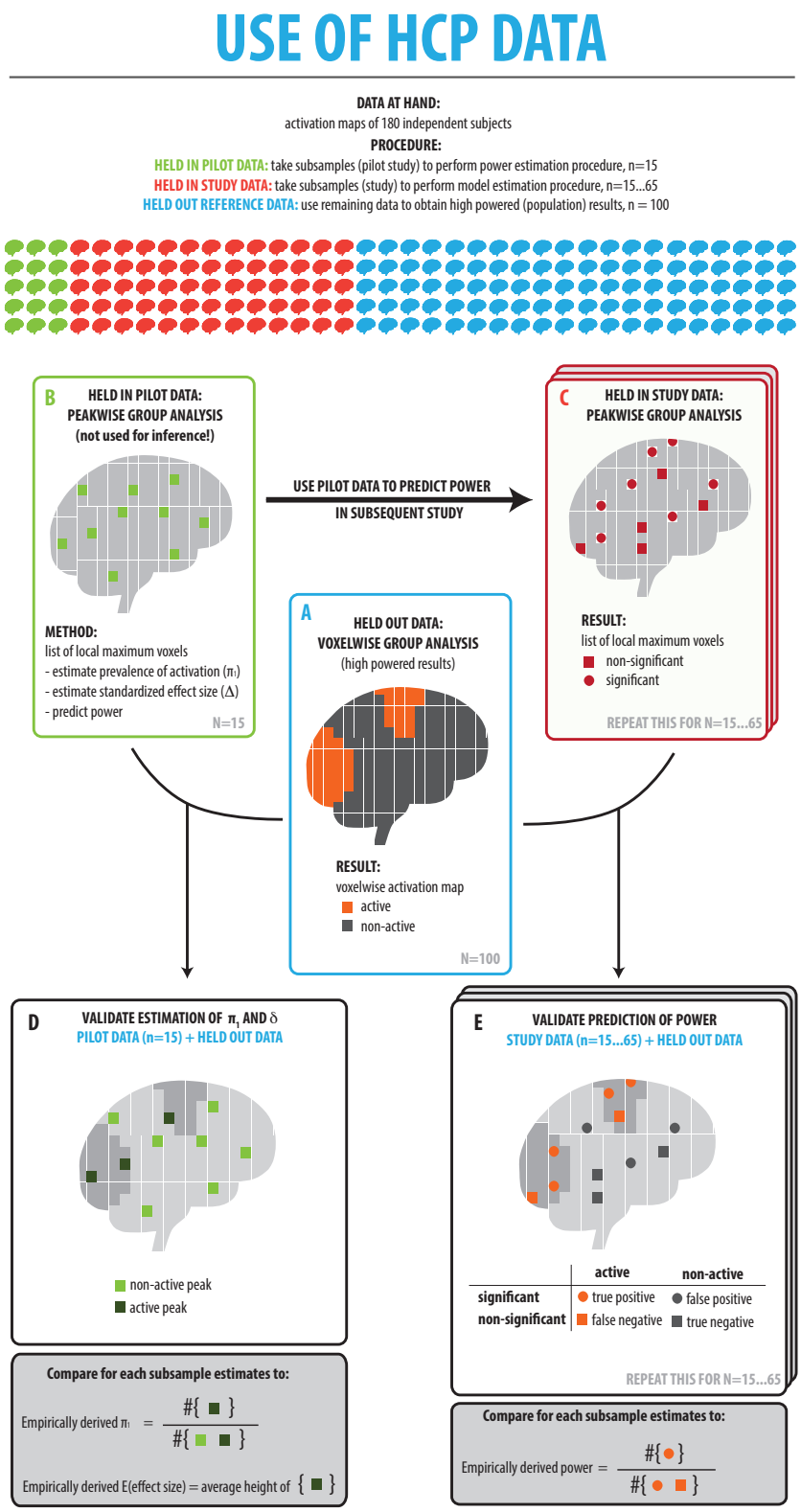

Figure 1: Overview of the procedure used to evaluate power calculations on the HCP-data. The panel labels (A-E) correspond to the labels of the different steps for the procedure in the main text.

$$\widetilde{\pi_1} = \frac{|\tilde{\mathcal{J}}_1^u|}{|\mathcal{J}^u|} \tag{10}$$

The **empirically derived expected peak height**, $\widetilde{E(z_j^u)}$, is the average peak height of all peaks that are located in empirically activated area:

$$\widetilde{E}(z_j^u|H_a) = \frac{1}{|\tilde{\mathcal{J}}_1^u|} \sum_{j \in \tilde{\mathcal{J}}_1^u} z_j. \tag{11}$$

Analysis E. Validation of power predictions Finally, we combine the held-in study data with the held-out data to validate the power estimation: **empirically derived power** is defined as the ratio of the number of significant empirically active peaks for the held-in study data (for a given thresholding procedure) and the total number of empirically activated peaks:

$$(1 - \widetilde{\beta}_{z_\alpha}) = \frac{|Z_j \geq z_\alpha|}{|Z_j|}, \text{ for } j \in \tilde{\mathcal{J}}_1 \tag{12}$$

The empirically derived power is computed for subsamples with $n = 15, ..., 65$ subjects; each power prediction based on n subjects is compared to empirical power on these $n = 15, ..., 65$ subjects. This resampling procedure is repeated 500 times for each of the 47 contrasts.

Some tweaking is needed to use these real data for evaluation purposes. The held-out reference data is analyzed using familywise error rate control. As such, we can assume that the maps with the set of voxels (and consequently peaks) that we call *empirically activated* do not contain false positives. However, while we assume that the analysis is powerful enough to represent results on a population level, within the voxels (and consequently peaks) that appear *empirically inactive*, there might be false negatives. This poses a problem for our measures of empirically derived π_1, effect size δ and power. We have derived a set of corrections for these measures to account for the presence of false negatives in the held-out data described in Appendix B.

3.5 Data example

We apply our estimation procedure to data from a study of 13 subjects on the role of higher order visual areas in imagined visual motion (Seurinck et al., 2011). First level analyses were carried out with a standard General Linear Model approach using FSL's FEAT function (Jenkinson et al., 2012). The second level analysis is performed in R with OLS, resulting in a 3D T-statistic map, that is transformed to Z-values (Hughett, 2007).

For this one Z map we apply our method with $u = 2.5$, characterizing the signal with $\hat{\pi}_1$, $\hat{\mu}_1$, $\hat{\sigma}_1$, and computing prospective power for different statistical thresholds (uncorrected, FDR-corrected, FWER-corrected with a Bonferroni procedure and FWER-corrected with a Random Field Theory procedure) and sample sizes ($n^* = 13, ..., 100$).

4 Results

4.1 Simulations

Results for the accuracy of alternative distribution parameters π_1 and μ_1 are shown in Figure 2. We can see that for very small effect sizes, the prevalence of activation π_1 is unbiased, while for medium to large effect sizes π_1 is overestimated. We show in the supplementary materials that this effect can be explained by the mismatch between the modeled beta-distribution and the observed distribution of p-values for active peaks.

Effect size tends to be underestimated, especially for the sparsest activations. When only few activation areas are present, the number of activated peaks is very small in relation to the number of null peaks and the mixture model can struggle to separate distributions. With larger activation areas, the underestimation is reduced.

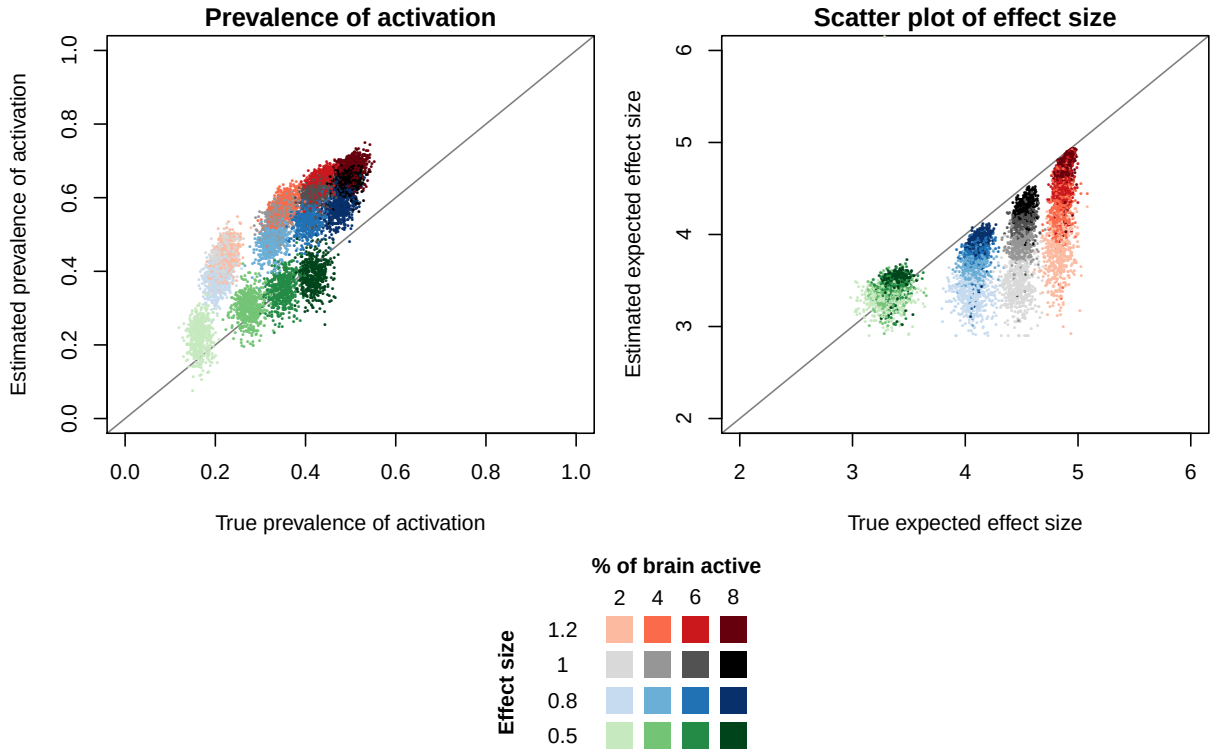

Figure 2: Left: Plot of estimated $\hat{\pi}_1$ against true π_1 for different sample sizes and different values for μ_1. Each dot represents a different simulation, as such there are 500 dots for each condition. Right: Plot of estimated expected peak height $\hat{E}(Z_j^u)$ against true expected peak height $\widetilde{E}(Z_j^u)$ for different effect sizes. The estimations are the result for a pilot dataset with $n = 15$.

Using these estimates of π_1, μ_1 and σ_1 for the case of $n = 15$, we then computed power for future studies for peak inference with 5% error rate control for the different inference procedures tested (see section 3.2.1). Figure 3 shows the plot of $(1 - \hat{\beta}_z)$ and $(1 - \widetilde{\beta}_z)$ for $n^* = 15, ..., 60$. In all conditions, but increasingly with small (local) activated brain regions, the power is largely underestimated. This leads to conservative results. For larger activated regions, the results are less conservative. As expected, the power is overestimated for FDR control.

Due to the increasing conservativism with more local effect sizes, we repeated the power calculation method using a mask that covers 28 percent of the full map. The results, shown in Figure 4 indeed indicate that the estimation of the model and power is better using a mask.

The goal of the presented method is to perform sample size calculations. Figure 5 shows the performance of these sample size calculations when 80% power is desired. While the variance is high for conditions with small effect sizes, the average bias of most conditions and multiple comparison procedures is within a range of 5 subjects. We immediately show results using the masks described in the previous paragraph.

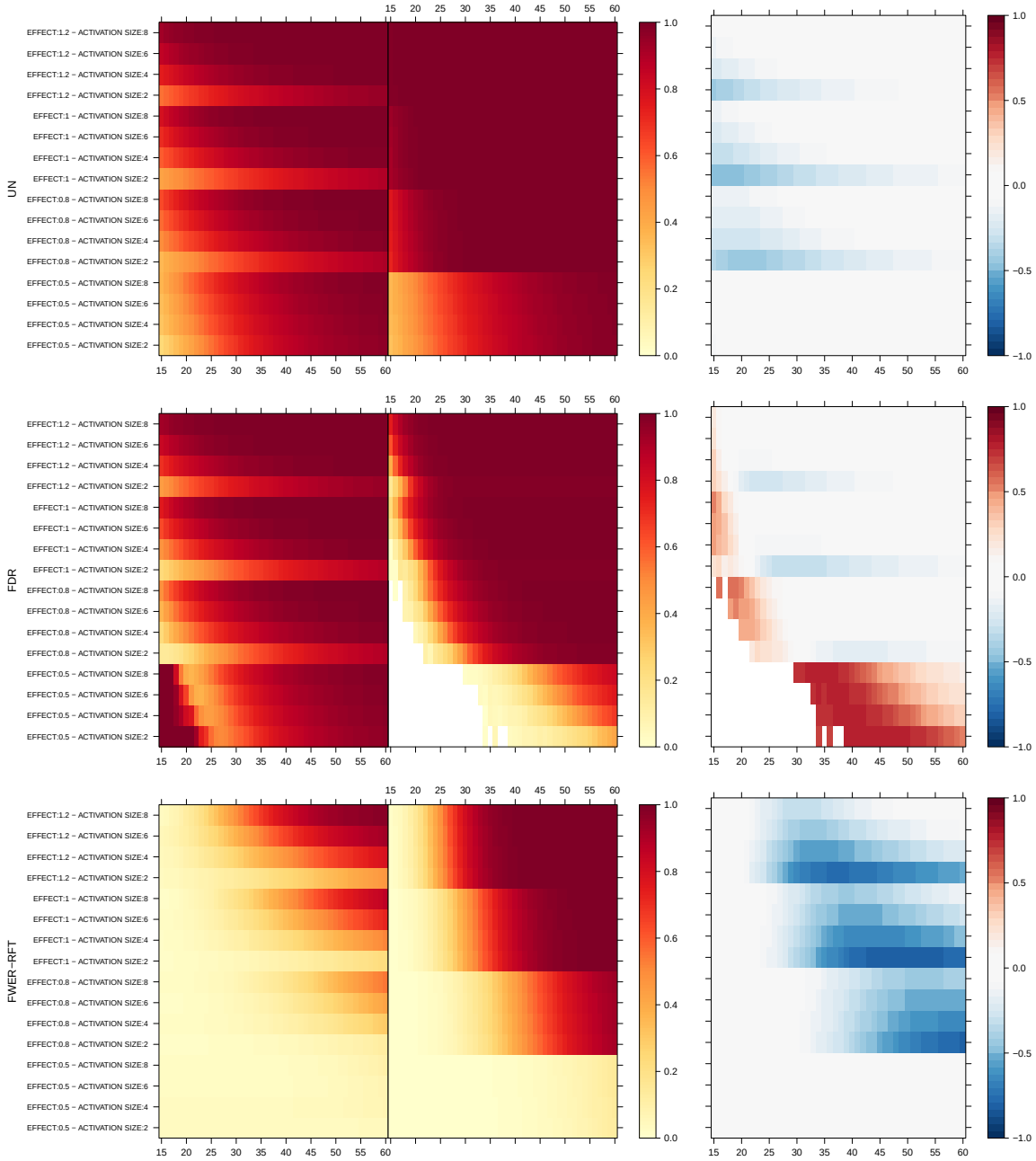

Figure 3: Plots of the peakwise average power with error rate control at 5% for different effect sizes and different amounts of activation. The left column shows the estimated power curves, the middle column shows the true power and the right column shows the bias. Bias is defined as the estimated power minus the true power. The peakwise average power is estimated from a pilot study with 15 subjects.

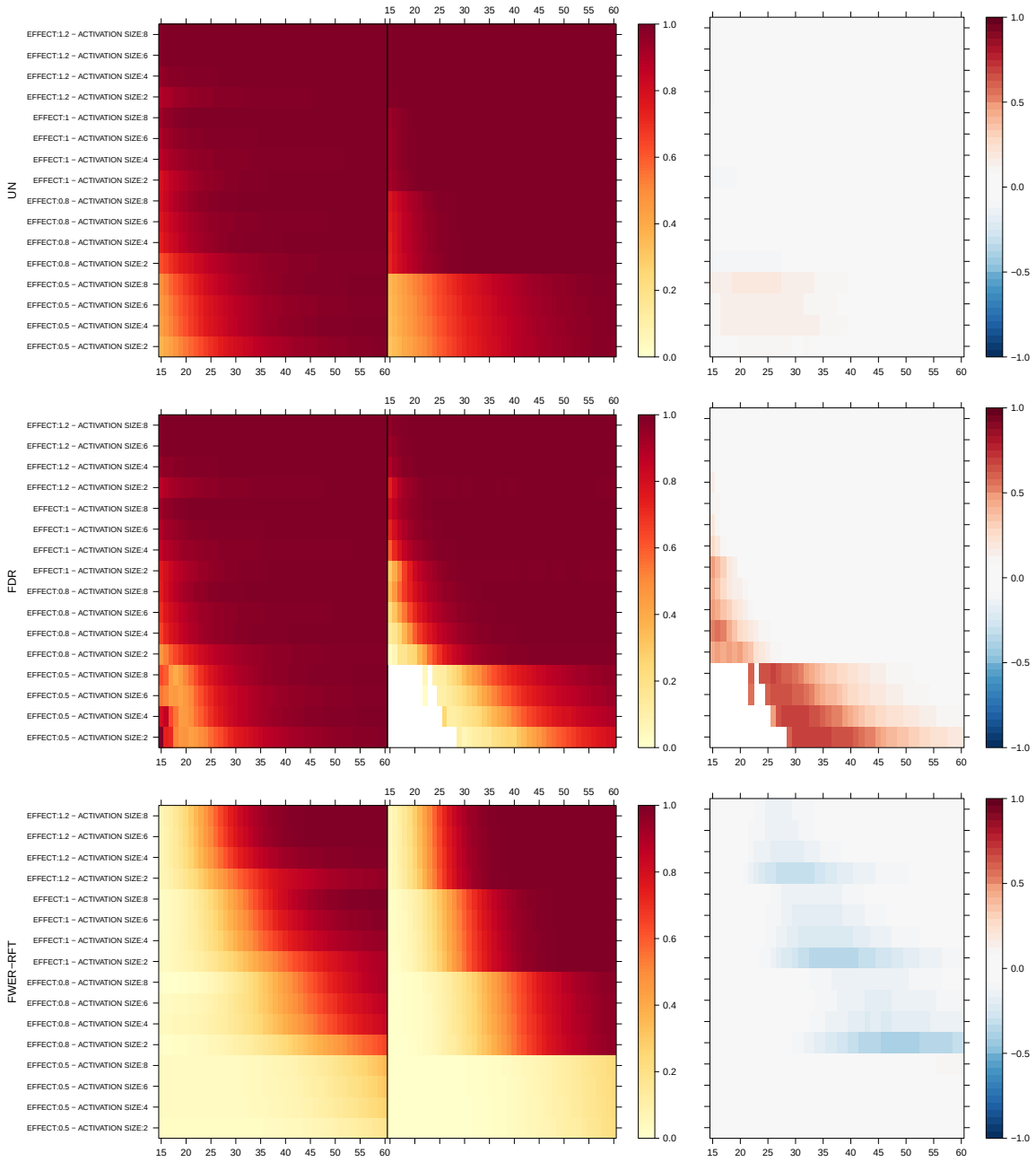

Figure 4: Plots of the peakwise average power with error rate control at 5% for different effect sizes and different amounts of activation, using a mask covering about 1/4th (28%) of the original map. The left column shows the estimated power curves, the middle column shows the true power and the right column shows the bias. Bias is defined as the estimated power minus the true power. The peakwise average power is estimated from a pilot study with 15 subjects.

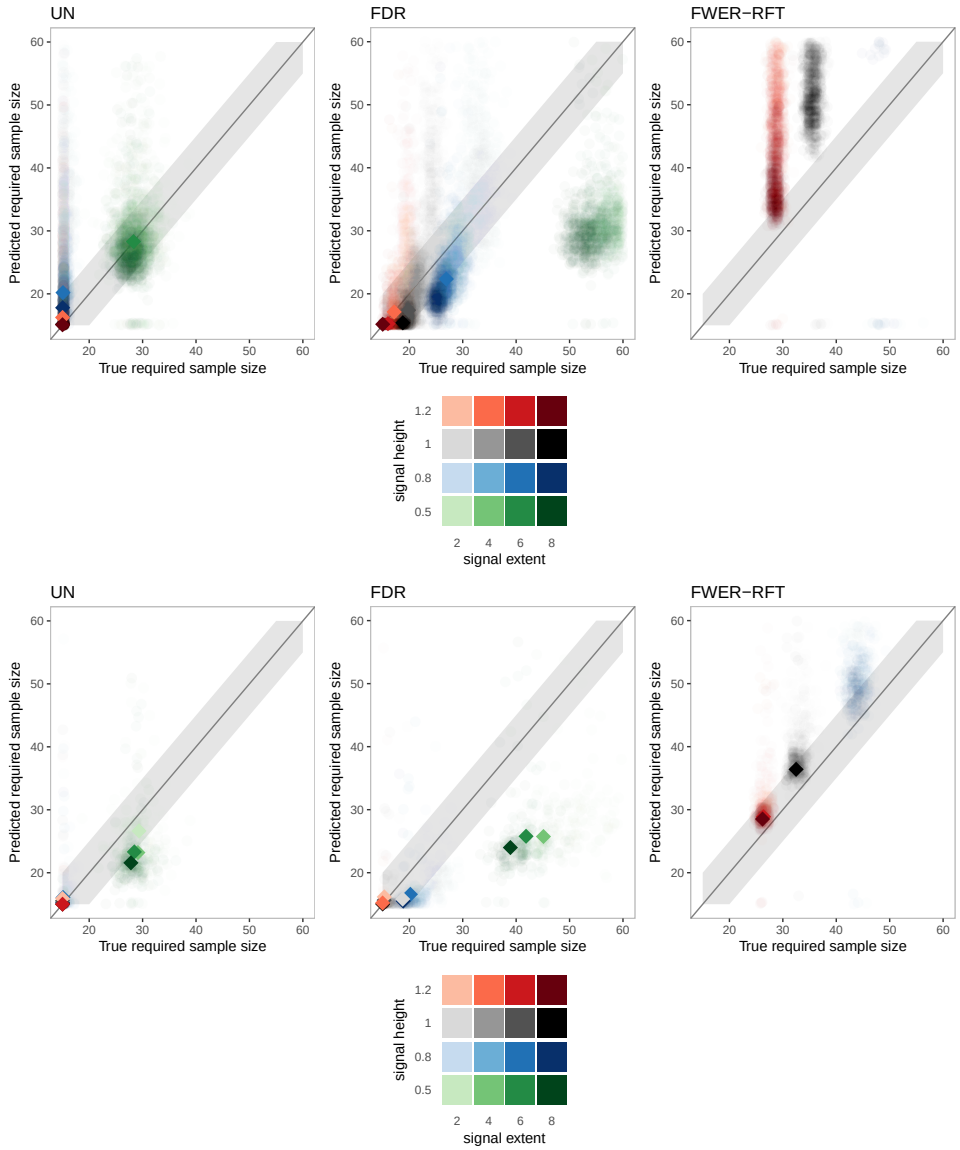

Figure 5: Plots of the predicted and true required sample size when 80% power is desired. The upper plot shows the results without applying a mask, the lower plot shows the results with mask. The different plots refer to the different multiple testing procedures. Points inside the grey area identify points with a maximum bias of 5 subjects. Each semi-transparent dot represents a different simulation, as such there are 500 dots for each condition. The fully colored dots present the average per condition. The estimated sample size results from a pilot study with 15 subjects.

4.2 HCP data

A power analysis was successful for all contrasts. Results of the estimation procedure for the prevalence of activation π_1 and the effect size based on a pilot dataset of about 15 subjects are shown in Figure 6. We show that for a range of contrasts and tasks we find overestimated estimates for the prevalence of activation. The estimations are good with low variance and the effect sizes suffer from modest overestimation. For the effect sizes, we find - as in the simulations - underestimation for smaller (more local) effect sizes.

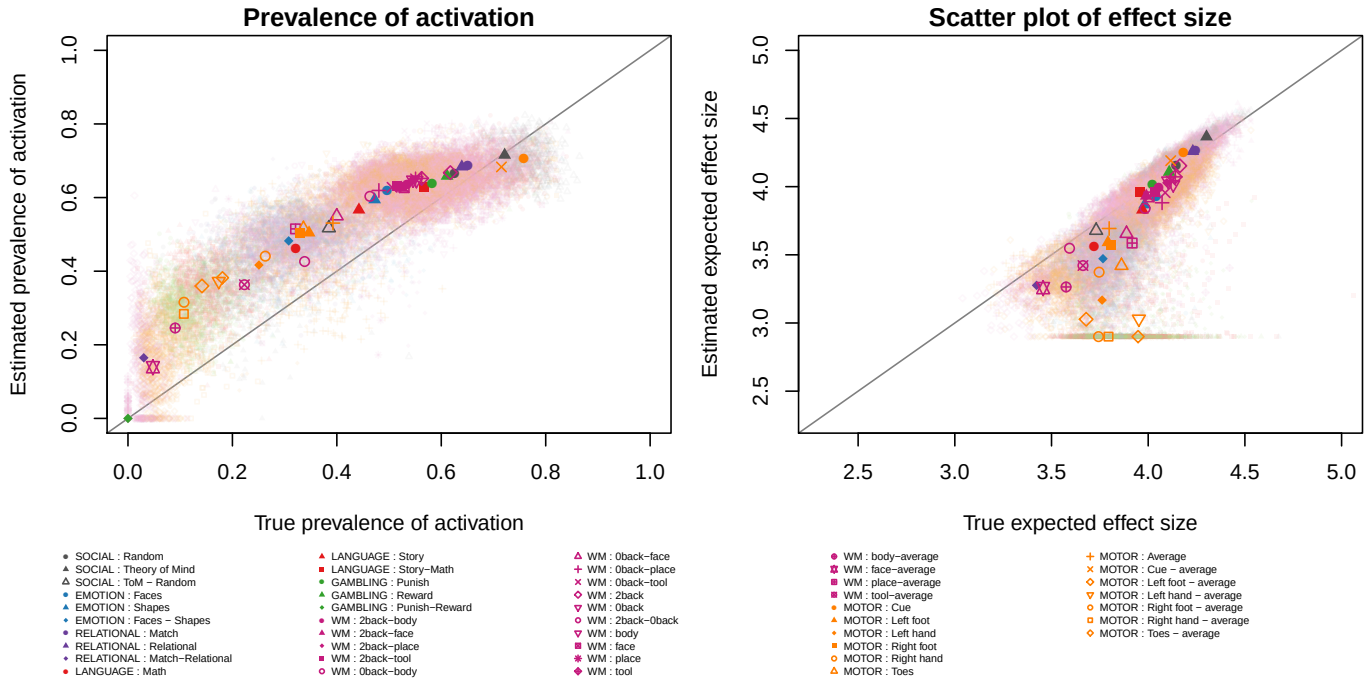

Figure 6: An evaluation of the estimation of π_1 (left) and the effect size (right). Each pastel colored plotting symbol corresponds to one random sub-sample taken from the data, for one particular sample size, experiment and contrast. The average estimation for each contrast is plotted in a darker color.

The estimation and bias on the power estimates for error rate control at 5% is given in Figure 7. In most contrasts, the bias ranges from 10% underestimation (blue) to 10% overestimation (red). In general, the underestimation occurs for lower values of power and disappears over a short range of subjects. The underestimation is much larger for contrasts with a smaller effect size. For larger values of power, the estimates are overestimated. With power analysis methods assuming a parametric distribution, there will always be a point where 100 percent power is achieved. However, in our estimation of the true power and the correction mechanism we apply, 100 percent power is never achieved. Whether this is a true effect or an artifact of the mechanism is unfortunately unknown.

Figure 8 shows the performance of the model when predicting sample sizes. We see that over all contrasts, we see underestimation for thresholding using false discovery rate. This result can be explained by the

underestimation of our calculation of the threshold for the false discovery rate. The motor contrasts show a larger overestimation of power with uncorrected thresholding. This can be due to the fact that the motor contrasts are known to be very local contrasts. This corresponds to the results from the simulated condition when the percentage of active brain regions is small.

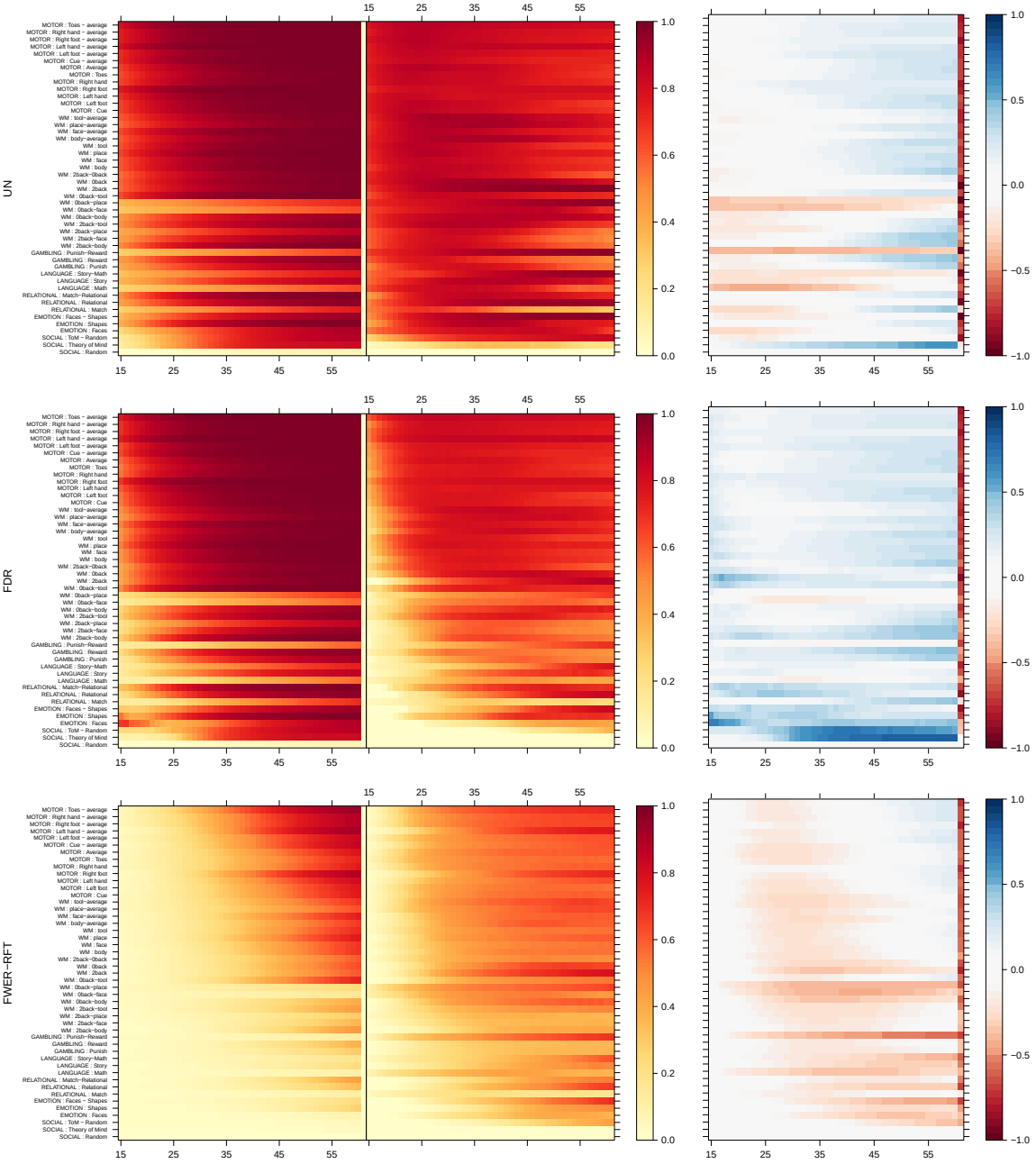

Figure 7: Evaluation of the power estimation over different subjects for all unique HCP-contrasts for thresholding with different error rate corrections at $\alpha = 0.05$ from a pilot study with 15 subjects. The left column shows the estimated power curves, the middle column shows the true power and the right column shows the bias. Bias is defined as the estimated power minus the true power. The contrasts are sorted by their average empirically derived effect size.

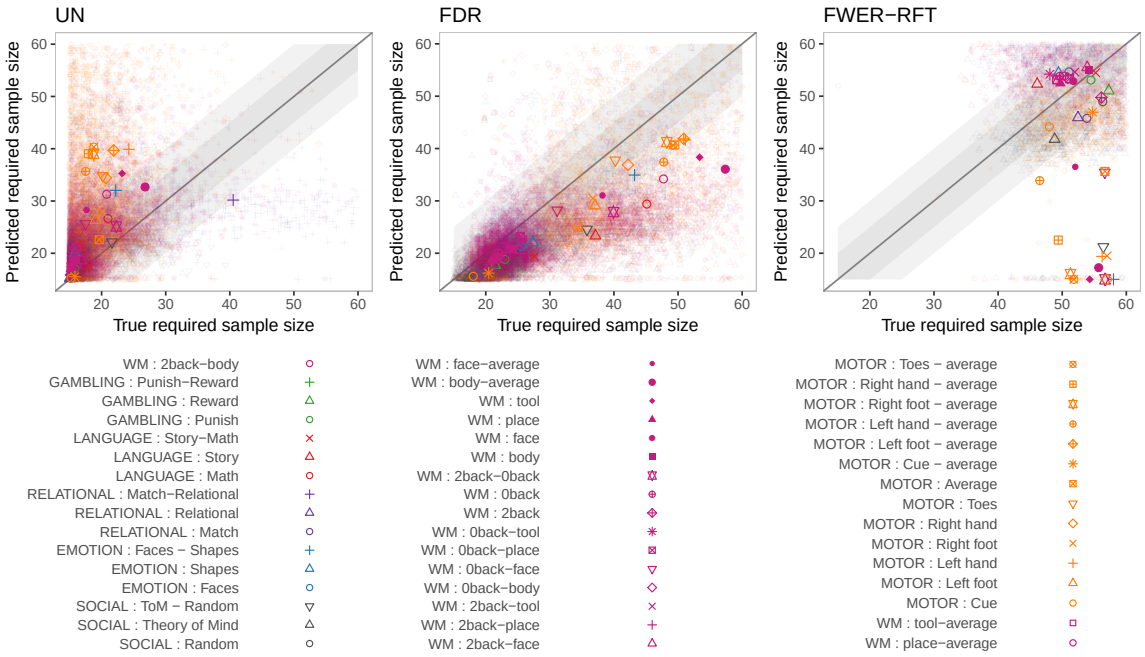

Figure 8: Plots of the predicted and true required sample size when aiming for 80% power. The different plots refer to the different multiple testing procedures. Points inside the light grey area identify poins with a maximum bias of 15 subjects, the darker grey area refers to a maximum bias of 5 subjects. Each semi-transparent dot represents a different subsample, as such there are 500 dots for each condition. The fully colored dots present the average per task. The estimated sample size results from a pilot study with 15 subjects.

4.3 Example

The results of the estimation procedure is shown in Figure 9. Based on these values of π_1, μ_1, σ_1, power predictions for different thresholding methods and a range of n^*'s are shown in Figure 10.

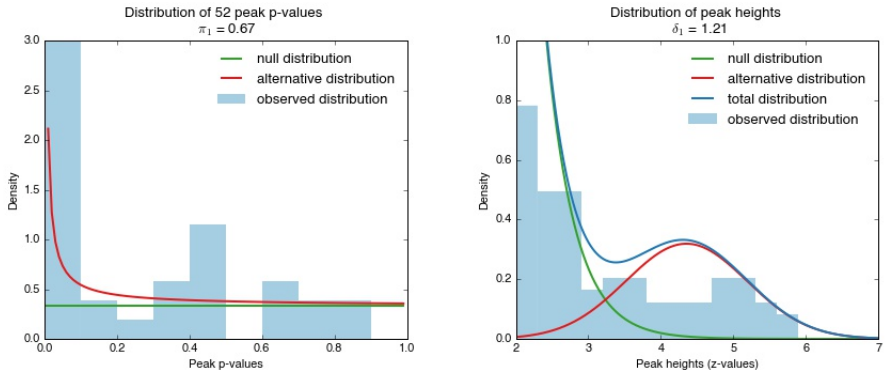

Figure 9: Left: Estimated distribution of peak p-values. The histogram of peak p-values is shown in light blue, the lines show the estimated part of the histogram stemming from the null distribution (green) and the total distribution (blue). Right: Estimated distribution of peak heights. The histogram of the peak heights is shown in light blue, the lines show the estimated distributions for the null (dark green), the alternative (light green) and the total distribution (blue)

The results of the power estimation procedure can be found in Figure 10. If the aim of the power analysis is to obtain at least 80% average power, then this study would require 15 subjects for uncorrected thresholding at $\alpha = 0.05$. False discovery rate thresholding at $\alpha = 0.05$ requires 16 subjects. For family-wise error rate control at $\alpha = 0.05$, the study would require 47 subjects with Random Field Theory thresholding.

5 Discussion

In neuroscience, results are often based on fMRI studies that suffer from a lack of power. In order to save costs and effort while preserving sufficient power for detecting important effects, we presented a method to predict power for different sample sizes. While other methods for power calculations in fMRI often require the estimation of many different parameters that are often difficult to estimate (Mumford and Nichols, 2008; Desmond and Glover, 2002), our method is based on only peak values that require Random Field Theory assumptions for the computation of p-values. On the other hand, with the lack of need for a specific hypothesis of location comes the disadvantage of dimished local specificity: a power analysis is performed for **the average** activation in the brain (or mask).

Our results indicate that the method works well when the size of activated brain regions is reasonably large. When the activated region is small, the presented method underestimates the power largely and applying a restrictive mask helps the estimation. This can be seen in the simulations, where bias is larger when only 2% of the brain is active than when 6% is active. In the HCP data, we also see larger bias for the motor tasks, a contrast known for larger effect sizes but very local effects. This indicates further that bias arises when the activated region is small, irrespective of the effect size. Applying the method to small activated brain regions results in mostly conservative results for sample size calculations, i.e.

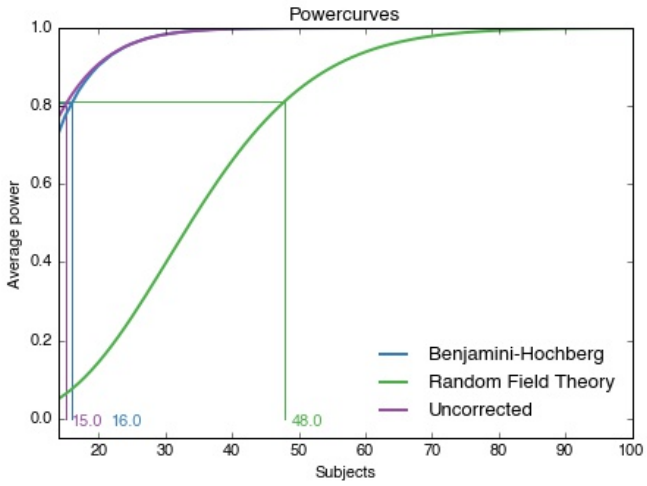

Figure 10: Estimated peakwise average power, $u = 2.5$ for different multiple testing procedures as a function of sample size. The vertical lines show for each multiple testing procedure the required number of subjects to obtain a power of at least 80%.

overestimation of required sample size. However, to confidently use the method, it is best to restrict the search region (i.e. apply a ROI mask based on previous findings or anatomical masks, independent from the results from the pilot study), an advice that generally applies to fMRI analysis.

Furthermore, the pilot study should be sufficiently high for the power analyses to work. In the supplementary materials, we show the results for a smaller sample size ($n = 10$), which results in larger biases for small effect sizes and small activation foci. Although a sample size of 15 is much larger than the pilot data that is often used to validate an experiment, it is crucial to provide sufficient degrees of freedom, not only for this power analysis, but for any type of power analysis.

In our validations, we find upward bias when estimating prevalence of activation, and downward bias when estimating the effect sizes. A possible source for this bias is the smoothing of the data. In our simulations, we generate maps with a strict separation between null and non-null voxels. However, the spatial smoothing of the data averages over null and effect voxels. This lowers the effect sizes in the activated peaks, especially close to the border of the activated region and for small effect sizes. Cheng and Schwartzman (2015) denoted these voxels as the transition region. This could be an alternative explanation why our procedure performs worse for small activated regions, where voxels are always spatially close to the border between null and non-null, and for small effect sizes.

We focus on peak level inference for several reasons. First, the use of peak level inference is increasingly being used in the fMRI literature. Often, peak heights are the only measure reported that can be related to a standardised effect size. Automated paper extraction tools such as NeuroSynth[3] and BrainSpell[4] have large databases with peak data, which can in turn be used for meta-analyses. Our power procedure, while not directly applicable to reported maxima, is a first step towards power analysis using reported effect sizes. Second, we have shown in previous work that the assumption of a uniform distribution of the p-values under the null is attained with peak p-values, but not with cluster p-values (Durnez et al., 2014). As this is an assumption crucial for the procedure presented here, we opt for peak inference, but

[3]http://www.neurosynth.org
[4]http://www.brainspell.org

not cluster inference. Moreover, problems with localisation and stability have been reported with cluster inference (Roels et al., 2014; Woo et al., 2014; Eklund et al., 2016). However, when a user wants to infer power for cluster inference, this procedure on peaks can be used as a lower bound, as the power of cluster inference should be generally higher than peak inference (Friston et al., 2007). Lastly, we did not create a voxelwise power analysis tool as power analysis for voxelwise inference is already developed (Hayasaka et al., 2007; Mumford and Nichols, 2008).

In this paper, we focused on power analysis for null hypothesis significance testing (NHST). However, in the field of neuroimaging, different analysis strategies like machine learning and bayesian analysis are increasingly being used for signal localisation. For those analysis types, the question of power is as relevant as it is for NHST: can we detect what we aim to detect. However, given that the measured and/or optimised outcome of the significance procedures of these methods are different quantities (prediction accuracy / the bayes factor), this method can not be used for other analysis modalities than NHST. However, increasing the sample size for all procedures will result in a better separation of null an alternative hypotheses, but the rate with which depends on the goal of the analysis, whether this is optimising the prediction accuracy, controlling the bayes factor or controlling the false positive rate.

We have evaluated the procedure using simulated data. The data represent a simplified fMRI experiment but we still vary a number of parameters, like the effect size and the thresholding procedure to ensure that the findings are generalizable to a range of different possible fMRI experiments. In our simulations, we have used a constant effect size of activation over different subjects. We have not applied subject-specific effect sizes, as we believe this would not alter the average effect size, but rather it would inflate the total variance, leading to a smaller normalized effect size. Thus we have considered only varying the average effect size μ_1 and not separately the between subject variance.

This method is only a first step in developing a means to better predict the power of fMRI studies. Many different extensions are possible. One of these possibilities is the development of a testing procedure that would allow to use the pilot data in the final study without harming the false positive rate (see, e.g., similar ideas in genetics Skol et al. 2006). Second, the estimated effect size could incorporate other characteristics besides sample size, like intrasubject variance or scan time (Mumford and Nichols, 2008) These additional parameters would allow the optimization of future studies without the restriction that all characteristics are identical to the pilot study.

Although the evaluation on this method was performed on whole-brain analyses, it is also possible to only apply it to a certain part of the brain, when a region of interest is specified.

We have made the procedure available to the community in a toolbox which is publicly available at `www.neuropowertools.org`, for which the code can be found at `https://github.com/neuropower/neuropower`. All code used for the validations and example in this paper are available online `http://github.com/jokedurnez/neuropower-validation/`.

Acknowledgements

We would like to thank Dr. Deanna Barch and Dr. Greg Burgess for their kind help in harvesting the HCP data and comments. This work was partially supported by the Laura and John Arnold Foundation. Jasper Degryse was supported by the Fund for Scientific Research-Flanders (FWO-V). Joke Durnez has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 706561. The computational resources (STEVIN Supercomputer Infrastructure) and services used in this work were kindly provided by Ghent University, the Flemish Supercomputer Center (VSC), the Hercules Foundation and the Flemish Government and department

EWI. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that have contributed to these research results. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. URL: http://www.tacc.utexas.edu Lastly, we would like to thank the reviewers for their helpful comments.

Appendix

A Generalisation from one-sample T-test to other models

In section 3.2 we declare how all Z statistics can be written in the form $Z = \frac{b}{\widetilde{SE}(b)}\sqrt{n}$ where $\widetilde{SE}(b) = SE(b)\sqrt{n}$. For a one sample t test $\widetilde{SE}(b) = \sigma$. Here we explain how this approach can be applied to other types of models.

A.1 Two-sample T-test

For a two-sample T-test, the Z statistic is:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1/n} + \frac{\sigma_2^2}{n_2/n}}}\sqrt{n},$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means of each group, σ_1^2 and σ_2^2 are the respective variances, n_1 and n_2 are the respective sample sizes and $n = n_1 + n_2$ is the total sample size.
Here $b = \bar{X}1 - \bar{X}2$ and $\widetilde{SE} = \sqrt{\frac{\sigma_1}{n1/n} + \frac{\sigma_2}{n2/n}}$. The relative SE can be seen as only depending on the relative sample sizes, and thus the computed power set a total sample size and will be appropriate for a new study that has the same group variances relative sample sizes.

A.2 Linear regression

In linear regression, the statistic Z for a $n \times p$ design matrix X and contrast c can be written as:

$$Z = \frac{c\hat{\beta}}{\sqrt{c(X'X)^{-1}c'\sigma^2}} = \frac{c\hat{\beta}}{\sqrt{c(\frac{1}{n}\sum_i x_i'x_i)^{-1}c'\sigma^2}}\sqrt{n},$$

where x_i is the ith row of X, and thus $\widetilde{SE}(b) = \sqrt{c(\frac{1}{n}\sum_i x_i'x_i)^{-1}c'\sigma^2}$. To consider an arbitrary number of subjects, we need to define a p-dimensional distribution F to generate possible covariate values, $x_i \sim F$. The $p \times p$ term in the denominator can be seen as the uncentered second moment of this distribution:

$$\frac{1}{n}\sum_i x_i'x_i \approx \mu'\mu + \text{cov}(x)$$

where $\mu = E(x)$. Thus the sample size can be set arbitrarily, with the assumption that new observations of x can be drawn from F.

B Corrections for the mismatch between true and empirically derived π_1, effect size and power

We first introduce some notation in Table 1:

J	Total number of peaks
$\mathcal{J}_0^u$	Indices of peaks arising in null regions above u
$\mathcal{J}_1^u$	Indices of peaks arising in non-null regions above u
$\tilde{\mathcal{J}}_0^u$	Indices of peaks arising in empirically derived null regions above u
$\tilde{\mathcal{J}}_1^u$	Indices of peaks arising in empirically derived non-null regions above u
	$J_0^u = \lvert\mathcal{J}_0^u\rvert$, $J_1^u = \lvert\mathcal{J}_1^u\rvert$
	$\tilde{J}_0^u = \lvert\tilde{\mathcal{J}}_0^u\rvert$, $\tilde{J}_1^u = \lvert\tilde{\mathcal{J}}_1^u\rvert$
π_{00}^u	Proportion of $\tilde{\mathcal{J}}_0^u$ that is truly null
π_{10}^u	Proportion of $\tilde{\mathcal{J}}_1^u$ that is truly null

Table 1: Notation for correction of population level estimators of π_0 and μ_1.

B.1 Correction of model estimates

As our reference level is analysed with family-wise error rate (FWER) control, we expect π_{10}^u to be negligable and we set it to 0. π_{00}^u can be estimated using the Beta-Uniform Model by Pounds and Morris (2003). With these definitions, we can derive the number of peaks that are falsely classified in the FWER-analysis for the **Held-in pilot data** in table 2:

	Empirically derived Null peaks	Empirically derived Active peaks	
True null peaks	$\pi_{00}^u \tilde{J}_0^u$	$\pi_{10}^u \tilde{J}_1^u = 0$	J_0^u
True active peaks	$(1-\pi_{00}^u)\tilde{J}_0^u$	$(1-\pi_{10}^u)\tilde{J}_1^u = \tilde{J}_1^u$	J_1^u
	$\tilde{J}_0^u$	$\tilde{J}_1^u$	

Table 2: Classification table of peaks after FWE thresholding in the held-in pilot data (with thresholding at u).

Thus, an uncontaminated estimate of π_0 can be found as:

$$\tilde{\pi}_0 = \frac{J_0^u}{J^u} = \frac{\hat{\pi}_{00}^u \tilde{J}_0^u}{J^u}$$

Similarly, bias-corrected versions of μ_1 are possible. First we note

$$E(Z^u|\mathcal{J}_0^u) = u + 1/u$$
$$E(Z^u|\mathcal{J}_1^u) = \mu_1$$

where the conditional expectation indicates the set of peaks under consideration. Now, similar to table 2, we can decompose the expectation over observable sets:

$$E(Z^u|\tilde{\mathcal{J}}_0^u) = \pi_{00}^u(u + 1/u) + (1-\pi_{00}^u)E(Z^u|\mathcal{J}_1^u)$$
$$E(Z^u|\tilde{\mathcal{J}}_1^u) = E(Z^u|\mathcal{J}_1^u)$$

Thus $\tilde{\mu}_1$ can be estimated:

$$\tilde{\mu}_1 = \frac{(1-\pi_{00}^u)\tilde{J}_0^u}{J_1^u}\hat{E}(Z^u|\tilde{J}_1^u) + \frac{\tilde{J}_1^u}{J_1^u}\frac{E(Z^u|\tilde{\mathcal{J}}_0^u) - \hat{\pi}_{00}^u(u+1/u)}{1-\hat{\pi}_{00}^u} \tag{13}$$

B.2 Correction of power estimates

For that for the power estimation procedure, we aim to estimate power without threshold u. Therefore, we use the same notation as in 1, but drop the u to represent the peak indices without applying a threshold. Similar to 2, we can derive the number of peaks that are falsely classified in the FWER-analysis for the **Held-in study data** in table 3:

	Empirically derived Null peaks	Empirically derived Active peaks	
True null peaks	$\pi_{00}^u\tilde{J}_0$	0	J_0
True active peaks	$(1-\pi_{00})\tilde{J}_0$	$\tilde{J}_1$	J_1
	$\tilde{J}_0$	$\tilde{J}_1$	

Table 3: Classification table of peaks after FWE thresholding in the held-in study data (without thresholding).

We again estimate $\pi_0 0$, the proportion of active peaks among all empirically derived null peaks, $\tilde{J}_0$, using the Beta-Uniform Model. Using table 3, an uncontaminated estimate of $1-\beta_{z_\alpha}$ can be found as:

$$\begin{aligned}(1-\tilde{\beta}_{z_\alpha}) &= \frac{|Z_j \geq z_\alpha|}{J_1}, \text{ for } j \in \tilde{\mathcal{J}}_1 \\ &= \frac{|Z_j \geq z_\alpha|}{\tilde{J}_1 + (1-\pi_{00})\tilde{J}_0}, \text{ for } j \in \tilde{\mathcal{J}}_1\end{aligned}$$

Note that we assume that there is no overlap between Z_j, for $j \in \tilde{\mathcal{J}}_0$ and J_1, which makes our uncontaminated estimate conservative.

C Estimation of the significance threshold when controlling the false discovery rate.

To predict the significance threshold when controlling the false discovery rate in larger sample sizes, we first redefine the mixture distribution in Equation 6 for a larger samplesize by replacing μ_1 by μ_1^*:

$$f(z_j^u|\pi_1, \mu_1^*, \sigma_1, Z_j^u \geq u) = (1-\pi_1)f(z_j^u|H_0, Z_j^u \geq u) + \pi_1 f(z_j^u|H_a, \mu_1^*, \sigma_1, Z_j^u \geq u) \tag{14}$$

Next, we can use the definition the local fdr as a Bayes posterior probability that a case is null given z, see Equation 2.6 and Equation 2.8 in Efron (2007), and its conditional expectation:

$$\text{fdr}(z) = \frac{(1-\pi_1)f_0(z)}{f(z)}$$
$$\text{Fdr}(Z) = \int_{-\text{ inf}}^{z} \text{fdr}(Z)f(Z)dZ \Big/ \int_{-\text{ inf}}^{z} f(Z) \tag{15}$$

with π_1 the proportion of active statistics, $f_0(z)$ the density function under H_0 and $f(z)$ the mixture density function. When plugging in the null density of peaks from Equation 4 and the mixture density from Equation 14, we obtain,

$$\text{fdr}(z^u) = \frac{(1-\pi_1)f(z_j^u|H_0, Z_j^u \geq u)}{f(z_j^u|\pi_1, \mu_1^*, \sigma_1, Z_j^u \geq u)}$$
$$\text{Fdr}(Z^u) = \int_{-\text{ inf}}^{z} \text{fdr}(Z)f(Z^u|\pi_1, \mu_1^*, \sigma_1, Z^u \geq u)dZ^u \Big/ \int_{-\text{ inf}}^{z} f(Z^u|\pi_1, \mu_1^*, \sigma_1, Z^u \geq u) \tag{16}$$

When numerically solving 16 for z when setting $\text{Fdr}(Z^u) = \alpha$, we obtain $E(z_\alpha|\pi_1, \mu_1^*, \sigma_1, u)$, the expected significance threshold when controlling the false discovery rate at level α.

References

Benjamini, Y. and Hochberg, Y. (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):289–300.

Benjamini, Y. and Hochberg, Y. (2000). On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics. *Journal of Educational and Behavioral Statistics*, 25(1):60–83.

Byrd, R., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208.

Cheng, D. and Schwartzman, A. (2015). On the explicit height distribution and expected number of local maxima of isotropic Gaussian Random Fields. *biorXiv*.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Earlbaum Associates, Hilsdale.

Desmond, J. E. and Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. *Journal of neuroscience methods*, 118(2):115–28.

Dudoit, S., Shaffer, J. P., and Block, J. C. (2003). Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science*, 18(1):71–103.

Durnez, J., Moerkerke, B., and Nichols, T. E. (2014). Post-hoc power estimation for topological inference in fMRI. *NeuroImage*, 84:45–64.

Efron, B. (2007). Size, power and false discovery rates. *Annals of Statistics*, 35(4):1351–1377.

Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, page 201602413.

Friston, K. J., Ashburner, J., Kiebel, S. J., Nichols, T. E., and Penny, W. D. (2007). *Statistical parametric mapping: the analysis of functional brain images.*

Friston, K. J., Zarahn, E., Josephs, O., Henson, R. N., and Dale, a. M. (1999). Stochastic designs in event-related fMRI. *NeuroImage*, 10(5):607–19.

Hayasaka, S. (2003). Validating cluster size inference: random field and permutation methods. *NeuroImage*, 20(4):2343–2356.

Hayasaka, S., Peiffer, A. M., Hugenschmidt, C. E., and Laurienti, P. J. (2007). Power and sample size calculation for neuroimaging studies by non-central random field theory. *NeuroImage*, 37(3):721–30.

Henson, R. (2007). Efficient experimental design for fMRI. In *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, pages 193–210.

Hughett, P. (2007). Accurate Computation of the F-to-z and t-to-z Transforms for Large Arguments. *Journal of Statistical Software*, 23(1):1–5.

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., and Smith, S. M. (2012). Fsl. *NeuroImage*, 62(2):782–90.

Mumford, J. A. and Nichols, T. E. (2008). Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *NeuroImage*, 39(1):261–8.

Pounds, S. and Cheng, C. (2004). Improving false discovery rate estimation. *Bioinformatics (Oxford, England)*, 20(11):1737–45.

Pounds, S. and Morris, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–1242.

Roels, S., Bossier, H., Loeys, T., and Moerkerke, B. (2014). Data-analytical stability of cluster-wise and peak-wise inference in fMRI data analysis. *Journal of Neuroscience Methods*, pages 1–11.

Seurinck, R., de Lange, F. P., Achten, E., and Vingerhoets, G. (2011). Mental rotation meets the motion aftereffect: the role of hV5/MT+ in visual mental imagery. *Journal of cognitive neuroscience*, 23(6):1395–404.

Silver, M., Montana, G., and Nichols, T. E. (2011). False positives in neuroimaging genetics using voxel-based morphometry data. *NeuroImage*, 54(2):992–1000.

Skol, A. D., Scott, L. J., Abecasis, G. R., and Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature genetics*, 38(2):209–213.

Smith, S. M., Jenkinson, M., Beckmann, C. F., Miller, K., and Woolrich, M. (2007). Meaningful design and contrast estimability in FMRI. *NeuroImage*, 34(1):127–36.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440–5.

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D. M., Behrens, T. E. J., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W., Della Penna, S., Feinberg, D., Glasser, M. F., Harel, N., Heath, a. C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S. E., Prior, F., Schlaggar, B. L., Smith, S. M., Snyder, a. Z., Xu, J., and Yacoub, E. (2012). The Human Connectome Project: a data acquisition perspective. *NeuroImage*, 62(4):2222–31.

Van Horn, J. D., Ellmore, T. M., Esposito, G., and Berman, K. F. (1998). Mapping voxel-based statistical power on parametric images. *NeuroImage*, 7(2):97–107.

Wager, T. D. and Nichols, T. E. (2003). Optimization of experimental design in fMRI: a general framework using a genetic algorithm. *NeuroImage*, 18(2):293–309.

Woo, C.-W., Krishnan, A., and Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *NeuroImage*, 91:412–9.

Worsley, K. J. (2007). Random Field Theory. In Friston, K., Ashburner, J., Kiebel, S. J., Nichols, T. E., and Penny, W., editors, *Statistical Parametric Mapping*, chapter Random Fie, pages 232–236. Academic Press, London.