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

SUPPLEMENTARY MATERIAL

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

1 Choice of screening threshold

The figures below show the model performance, when using other screening thresholds.

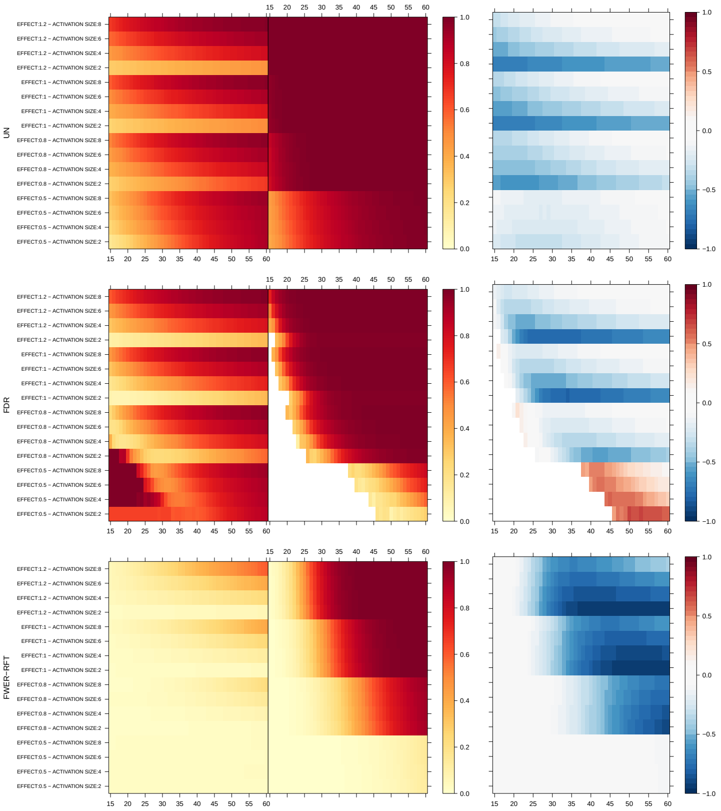

Figure 1: Plots of the peakwise average power with error rate control at 5% for different effect sizes and different amounts of activation, when using a screening threshold at 2.0.

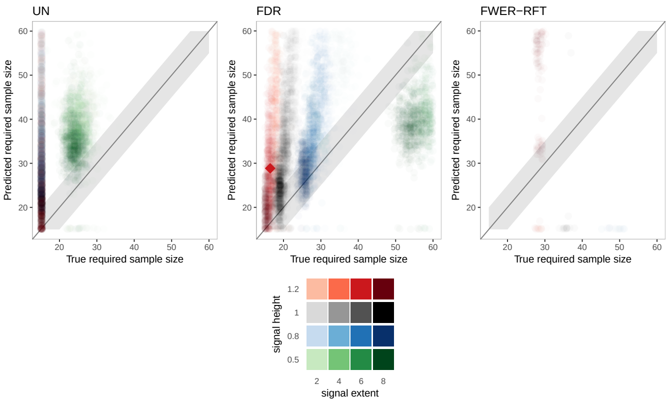

Figure 2: Plots of the predicted and true required sample size when 80% power is desired. The different plots refer to the different multiple testing procedures, when using a screening threshold at 2.0.

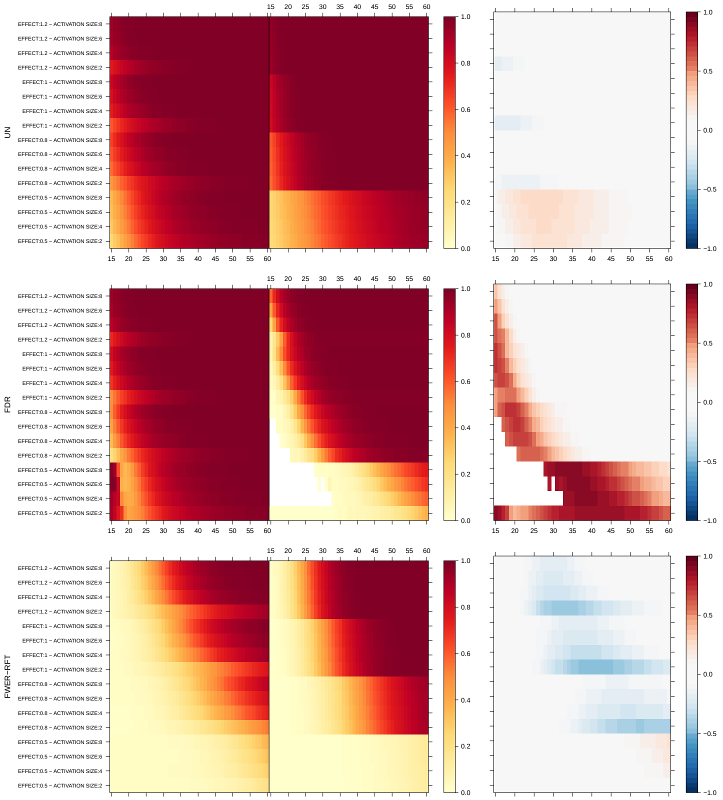

Figure 3: Plots of the peakwise average power with error rate control at 5% for different effect sizes and different amounts of activation, when using a screening threshold at 3.0.

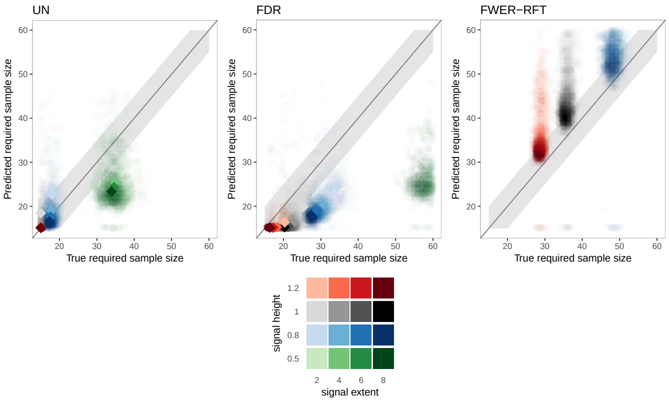

Figure 4: Plots of the predicted and true required sample size when 80% power is desired. The different plots refer to the different multiple testing procedures, when using a screening threshold at 3.0.

	FSL ($t = 2.3$)	SPM ($p = 0.001$)
5	1.03	1.38
10	0.73	0.98
15	0.59	0.80
20	0.51	0.69

Table 1: Given a screening threshold (defaults from popular software are given), how big should the Cohen's D effect be to be observable?

2 Size of the pilot study.

2.1 Pilot study sample size for different screening thresholds

One of the most important parameters for the applied researcher when performing the power analysis presented in this paper is the size of the pilot study. Pilot data is expensive and should be kept to a minimum while preserving its power estimation benefit. However, this method is based on the empirical distribution of the peaks above the screening threshold, which means only effects above this screening threshold can be observed. For example, when using the default screening threshold from SPM ($p = 0.001, t = 3.09$), the effect size in a pilot study with 10 subjects only surpasses the screening threshold with an effect size of Cohen's d of 0.98 ($= 3.09/\sqrt{10}$). The minimally detectable effects for FSL's and SPM's default screening thresholds are presented in Table 1. However, our method estimates the alternative distribution as a truncated normal distribution, and therefore allows a portion of the distribution (and even the mean) to be below the screening threshold. This fact notwithstanding, it is clear that a very small effect size that hardly surpasses the screening threshold is problematic for this estimation procedure. In this main body of this paper, we have used a screening threshold, but we present results for the simulations and HCP-validation in the next subsection.

We show that for a sample size of 10 subjects for small effects, the procedure leads to large overestimation of the power. The reason is the following: when our procedure predicts that 90% power will be reached at sample size n^*, this means that 90% of the effects that are large enough to be detectable in the pilot study will become significant with n^* subjects. As such, our procedure ignores the effects that weren't detectable in the pilot data. These undetectable effects are by definition small and harder to deem significant with n^* subjects, leading to the observed overestimation in power calculations for a small pilot study sample size. This effect can be seen both for simulated data and the HCP data. Another problem that arises with a small pilot study sample size is that the FDR controlling procedure hardly ever finds a threshold in the pilot study, and as such power calculations are almost impossible except for high effect sizes.

2.2 Results of simulation with pilot sample size of 10 subjects

Figure 7 and 8 show the results of the validation procedure for the simulations with a pilot dataset with sample size 10. The results follow the same patterns of the results with a pilot sample size of 15, but all errors are largely blown up. That is, the effect size estimation is still problematic when only few activations are present, but this effect remains for very large effects too. The consequence is a large underestimation for the power calculations in all conditions. The effect for FDR control, when a pilot study is often impossible due to an estimation of $\pi_1 = 0$ is repeated in all different conditions, except for very high values of power.

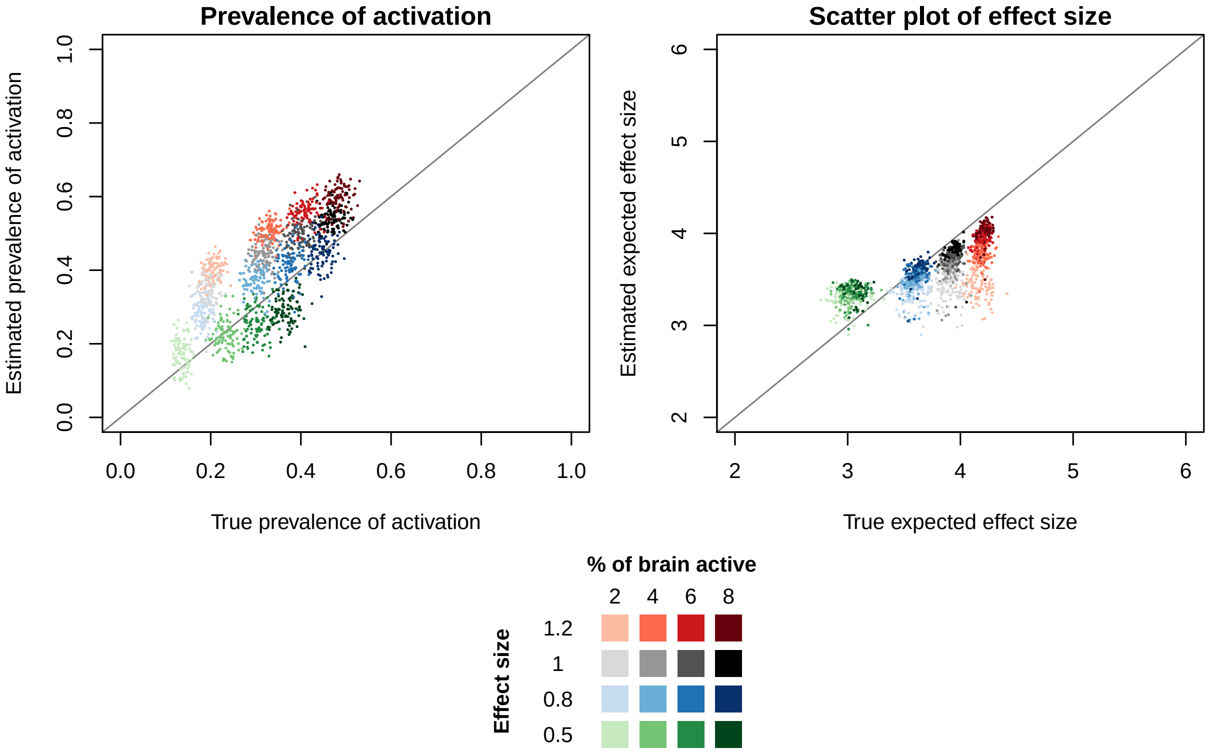

Figure 5: Simulation results. Left: Plot of estimated $\hat{\pi}_1$ against true π_1 for different sample sizes and different values for μ_1. Each dot represents a different simulation, as such there are 500 dots for each condition. Right: Plot of estimated expected peak height $\hat{\mu}_1$ against true expected peak height μ_1 for different effect sizes. The estimations are the result for a pilot dataset with $n = 10$.

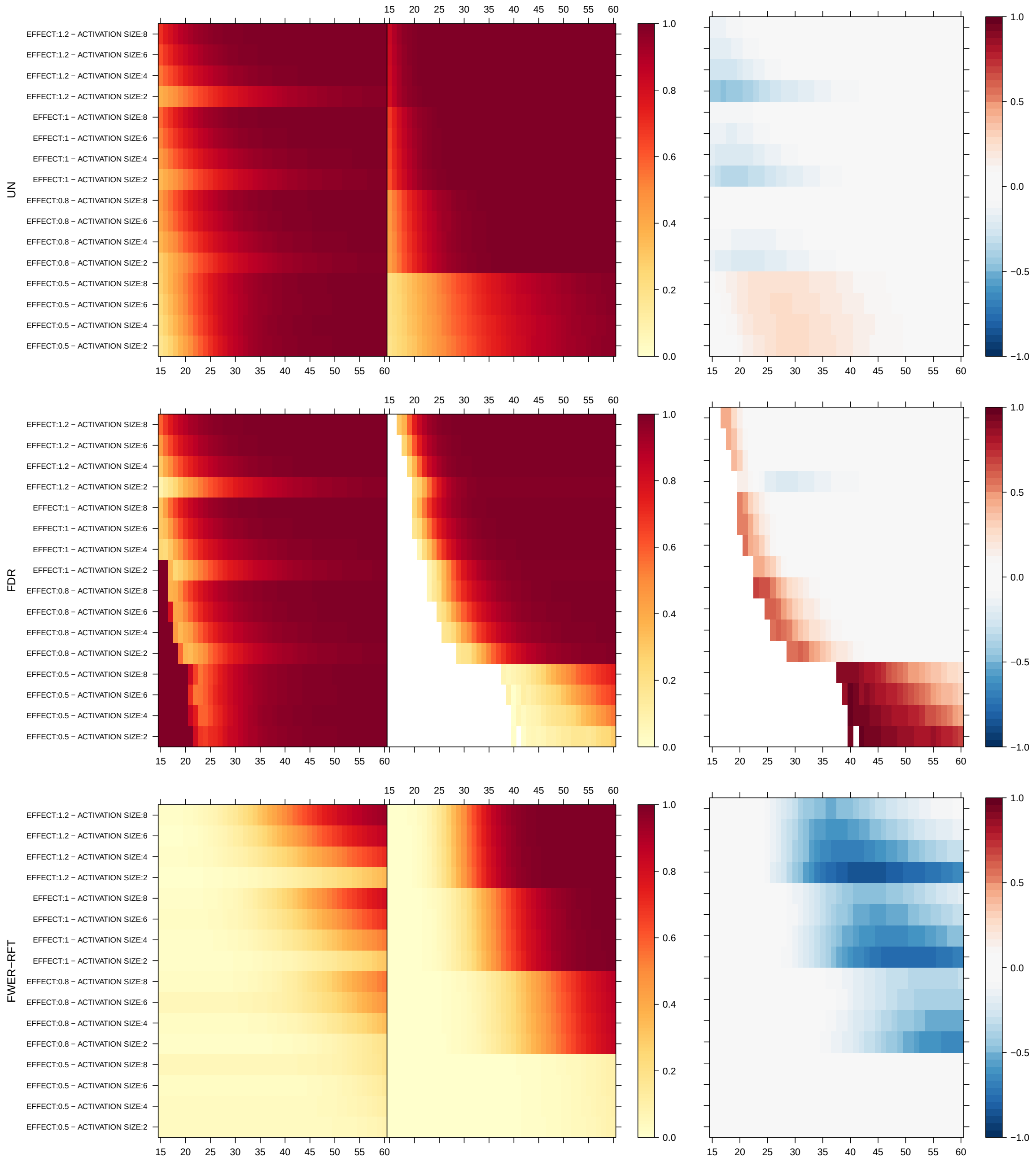

Figure 6: Simulation results. Plots of the peakwise average power with error rate control at 5% for different effect sizes and different amounts of activation. The left column shows the estimated power curves, the middle column shows the true power and the right column shows the bias. Bias is defined as the estimated power minus the true power. The peakwise average power is estimated from a pilot study with 10 subjects.

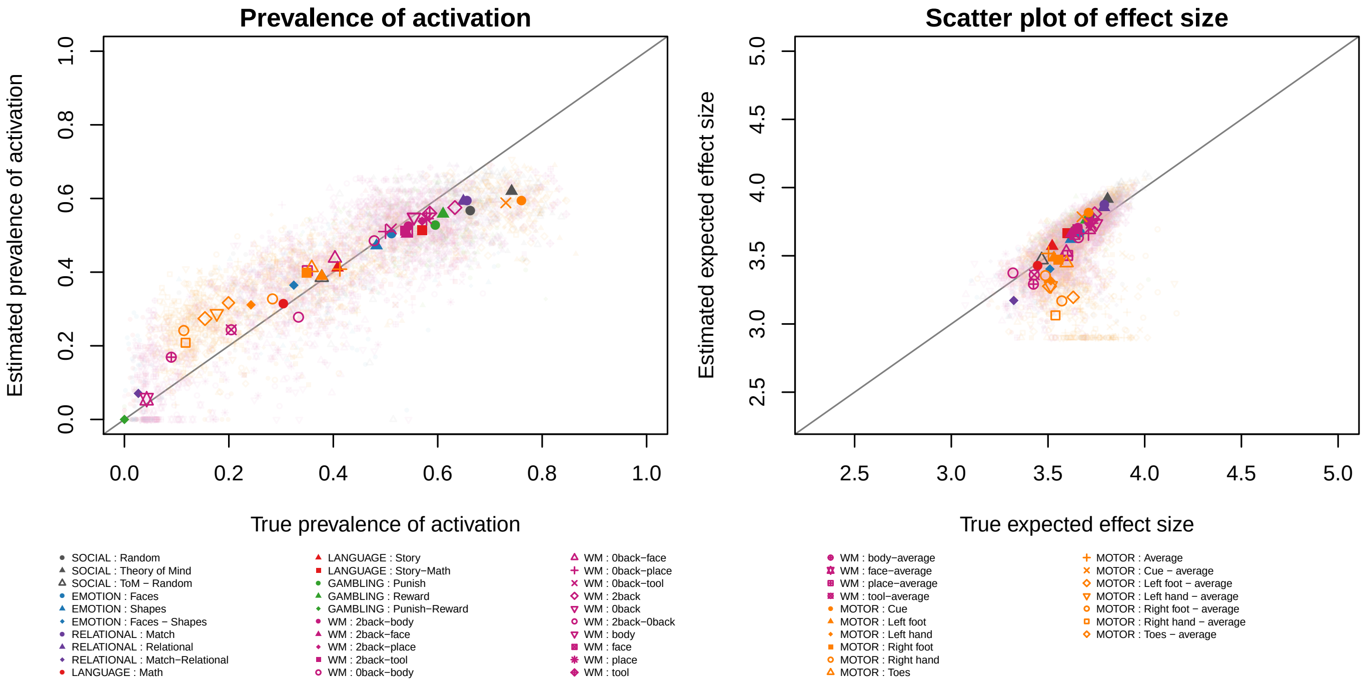

Figure 7: HCP results. Left: Plots of estimated $\hat{\pi}_1$ against true π_1 for different sample sizes and different values for μ_1. Each dot represents a different simulation, as such there are 500 dots for each condition. Right: Plot of estimated expected peak height $\hat{\mu}_1$ against true expected peak height μ_1 for different effect sizes. The estimations are the result for a pilot dataset with $n = 10$.

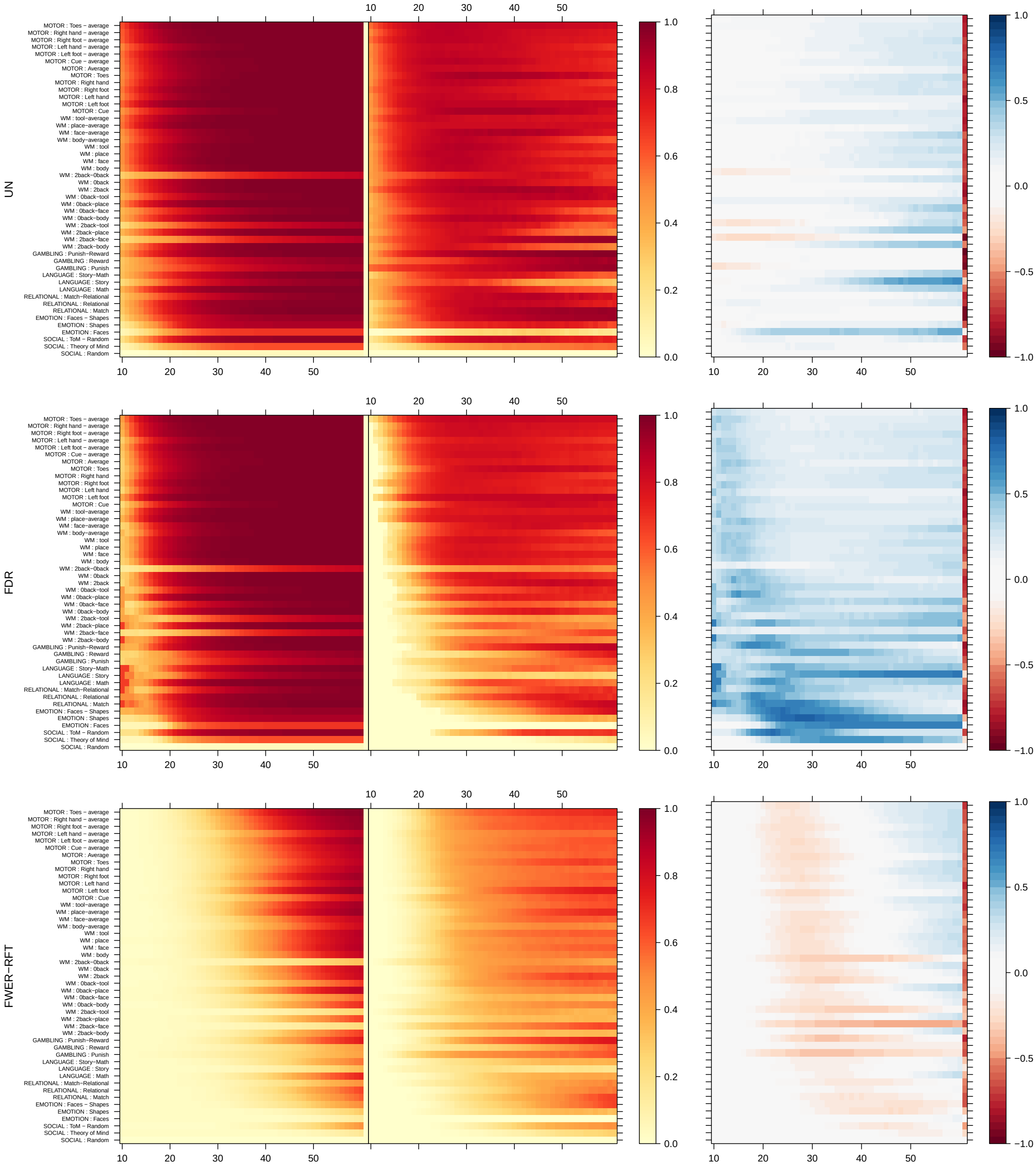

Figure 8: HCP results. Evaluation of the power estimation over different subjects for all unique HCP-contrasts for thresholding with different error rate corrections at $\alpha = 0.05$ from a pilot study with 10 subjects. The left column shows the estimated power curves, the middle column shows the true 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.

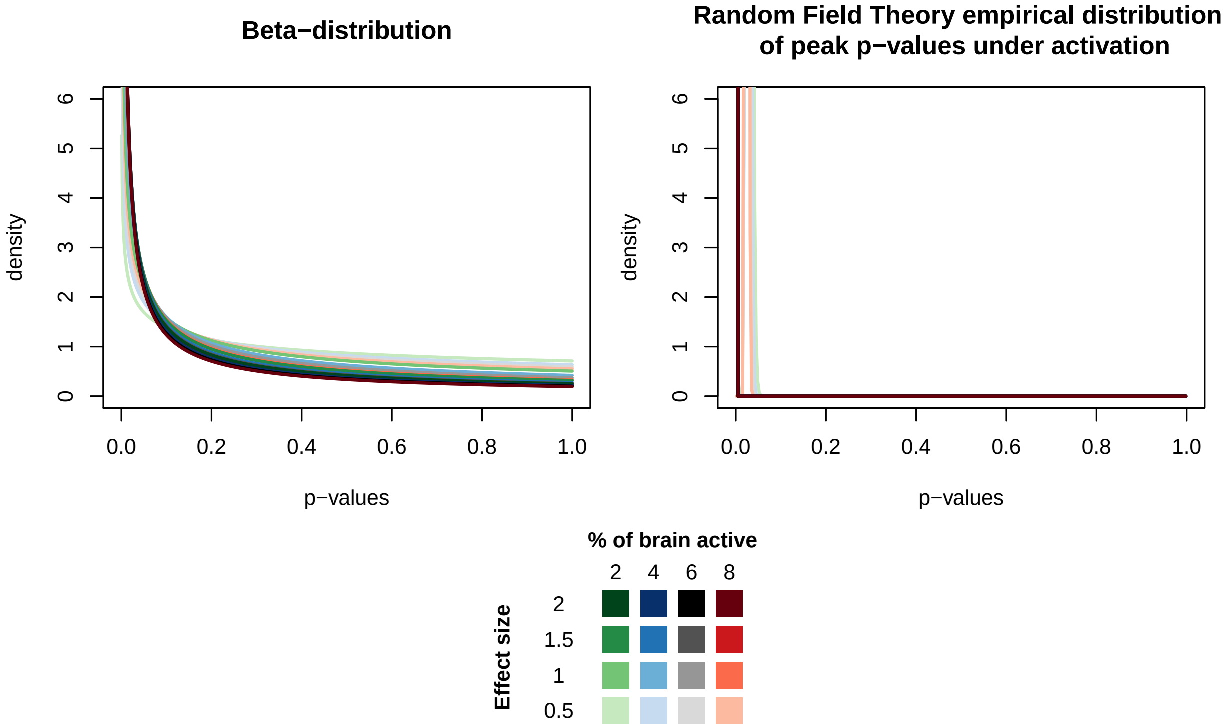

Figure 9: Plot of estimated beta-distributions and expected distributions

3 Mismatch between beta-distribution and cumulative density distribution for peaks.

The prevalence of activation of the peaks are overestimated in the simulations. This is mainly due to the mismatch between the beta distribution with which we modeled the alternative peaks, and the observed cumulative density distribution for peaks. To illustrate we have plotted in Figure 9 the estimated beta-distribution that is fitted to the alternative peaks, and the observed cumulative distribution based on the estimated effect sizes. For the former, we have calculated the average location parameter of the beta-distribution from Pounds and Cheng (2004) for the different conditions. We have plotted these beta-distributions in the left plot of Figure 9. For the latter, we have simulated 10^8 values following the estimated truncated normal distribution. For each of these values, we have calculated the p-value following Equation 2 in the main manuscript. The resulting density of these values are plotted in the right plot of Figure 9. It can be seen that the distributions are more similar for small effect sizes, but as the effect size increases, so does the mismatch. This explains why smaller sample sizes have less bias than larger sample sizes.

References

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