Power, positive predictive value, and sample size calculations for random field theory-based fMRI inference

Dirk Ostwald (dirk.ostwald@fu-berlin.de), Sebastian Schneider, Rasmus Bruckner, Lilla Horvath
Freie Universität Berlin, Habelschwerdter Allee 45
14195, Berlin, Germany

Abstract

Recent discussions on the reproducibility of task-related functional magnetic resonance imaging (fMRI) studies have emphasized the importance of power and sample size calculations in fMRI study planning. In general, statistical power and sample size calculations are dependent on the statistical inference framework that is used to test hypotheses. Bibliometric analyses suggest that random field theory (RFT)-based voxel- and cluster-level fMRI inference are the most commonly used approaches for the statistical evaluation of task-related fMRI data. However, general power and sample size calculations for these inference approaches remain elusive. Based on the mathematical theory of RFT-based inference, we here develop power and positive predictive value (PPV) functions for voxel- and cluster-level inference in both uncorrected single test and corrected multiple testing scenarios. Moreover, we apply the theoretical results to evaluate the sample size necessary to achieve desired power and PPV levels based on an fMRI pilot study and find that minimal sample sizes of 40 to 50 participants are required for corrected cluster-level inference at medium effect sizes.

Keywords: fMRI, statistical inference, random field theory, power, positive predictive value

Introduction

A fundamental goal of task-related functional magnetic resonance imaging (fMRI) is to identify the cortical correlates of cognition. An approach routinely used to achieve this goal is mass-univariate null hypothesis significance testing in the framework of the general linear model (Cohen et al., 2017). In the recent debate on the reproducibility of research findings in the life sciences, the statistical practices of fMRI research have once again taken centre stage in the community discourse (e.g., Poldrack et al. (2017)). Here, a particular emphasis has been on statistical power and its relation to typical sample sizes in fMRI group studies (e.g., Button et al. (2013); Geuter et al. (2018)). In task-related fMRI, statistical power is broadly defined as the probability of detecting cortical activation, if this activation is indeed present. In general, statistical power and, consequently, methods for computing the sample sizes necessary to achieve desired levels of power depend on both the statistical inference framework used and assumptions about the expected cortical activation.

A prominent statistical inference framework for null hypothesis significance testing in fMRI research is based on random field theory (RFT) (e.g., Worsley et al. (1992)). RFT-based fMRI inference is a parametric framework that allows for controlling the multiple testing problem inherent in the mass-univariate approach. Technically, this framework rests on analytical approximations to the exceedance probabilities of topological features of data roughness-adapted random field null models. RFT-based fMRI inference is implemented in the two major data analysis software packages used by the neuroimaging community, namely, Statistical Parametric Mapping (SPM) and the Functional Magnetic Resonance Imaging of the Brain (FMRIB) Software Library (FSL). It encompasses up to five forms of statistical testing: uncorrected and corrected voxel-level inference, uncorrected and corrected cluster-level inference, and set-level inference (K. Friston et al., 1996). With the exception of set-level inference, all forms are routinely reported in the functional neuroimaging literature. More specifically, bibliometric analyses suggest that RFT-based fMRI inference, especially corrected cluster-level inference, accounts for approximately 70% of published task-related human fMRI studies.

Aims and scope

In light of the widespread use of RFT-based inference, previously proposed approaches for the calculation of power and sample sizes in fMRI research have a number of shortcomings. First and foremost, most previously proposed frameworks are not well aligned with the theory of RFT-based fMRI inference (e.g., Mumford and Nichols (2008); Durnez et al. (2016)), rendering them non-applicable for the most commonly employed forms of fMRI inference. Second, the framework previously proposed by Hayasaka et al. (2007) that is aligned with the theory of RFT-based fMRI inference only addresses voxel-level and not cluster-level inference. Moreover, this framework does not address the variety of power types that arise in multiple testing scenarios and thus remains imprecise with respect to the interpretation of its ensuing power and sample size values. Third, all previous frameworks assume that under the alternative hypothesis, cortical activation is expressed either in a known region of interest or over the entire cortex. Notably, neither of these assumptions necessarily reflects common intuitions of neuroimaging researchers. Finally, no previous framework allows for the necessary sample sizes to be derived based on a desired positive predictive value (PPV), a novel statistical marker for the quality of empirical research that has risen to prominence over the last decade (Ioannidis, 2005). With the current work, we address these shortcomings and report on a novel framework for power, PPV, and sample size calculations in RFT-based fMRI inference.

Theoretical background

Power functions

In single test scenarios, such as testing for the activation of a single voxel, two types of errors can occur: the test may reject the null hypothesis when it is in fact true, referred to as a Type I error, and the test may not reject the null hypothesis when in fact the alternative hypothesis is true, referred to as a Type II error. From a frequentist perspective, Type I and Type II errors are associated with their probabilities of occurrence, denoted α and $1-\beta$, respectively, and commonly referred to as Type I and Type II error rates. The complementary probability of a Type II error, i.e., the probability rejecting the null hypothesis if the alternative hypothesis is true, is referred to as the power β of a test. A fundamental aim of test construction is to maintain low Type I and Type II error rates. To this end, a desired Type I error rate is usually selected first by defining a test significance level α', ensuring a Type I error rate of at most α'. For many commonly used tests, the power at a fixed significance level α' can then be shown to be a function $\beta(n,d)$ of an effect size measure d and the sample size n. An often recommended approach in research study design is calculating the necessary sample size n for which, under the assumption of a fixed effect size d, the power reaches a desirable level, such as $\beta(n,d)=0.8$.

Minimal and maximal power functions

In multiple testing scenarios, such as simultaneously testing for cortical activation over many voxels, a Type I or a Type II error may occur for each of the individual tests involved, inducing a variety of Type I and Type II error rates. For example, commonly considered Type I error rates in fMRI research are the *family-wise error rate* (FWER), defined as the probability of one or more false rejections of the null hypothesis, and the *false discovery rate* (FDR), defined as the expected proportion of Type I error among the rejected null hypotheses. Classically, the FWER has been the prime target for Type I error rate control in fMRI research. The prevalence of FWER control derives from the fact that the FWER can be efficiently controlled using maximum statistic-based procedures, which were at the centre of the early developments of RFT-based fMRI inference (Friston, Frith, Liddle, & Frackowiak, 1991; Worsley et al., 1992; K. Friston et al., 1994). Maximum statistic-based multiple testing procedures allow the FWER to be controlled using a family-wise error significance level α'_{FWE}. Just as the multiplicity of statistical tests in multiple testing scenarios induces a variety of Type I error rates, it also induces a variety of Type II error rates and hence power types. Power types commonly considered in multiple testing are *minimal power*, defined as the probability of one or more correct rejections of the null hypothesis, and *maximal power*, defined as the probability of correctly rejecting all false null hypotheses. When calculating the sample sizes necessary for desired power levels in Type I error rate-controlled multiple testing scenarios, it is hence essential to explicate the power type of interest. As RFT-based fMRI inference naturally lends itself to the evaluation of the minimal and maximal power functions $\beta_{min}(n,d)$ and $\beta_{max}(n,d)$, respectively, we focus on these power types in the current work.

PPV functions

In recent discussions, studies with low power have been related to high probabilities of the claimed effects to be false positives (Ioannidis, 2005; Button et al., 2013). This relationship is not inherent in classical frequentist test theory in which Type I and Type II error rates are conceived independently. Instead, the dependency of Type I error rates on Type II error rates, and hence power, arises in the context of a probabilistic model that assigns probabilities to the null hypothesis of being either true or false and the ensuing concept of a test's PPV. A test's PPV, denoted here by ψ, is defined as the probability of the null hypothesis being false given that the test rejects the null hypothesis. The PPV depends on both the Type I error rate and the *prior hypothesis parameter* $\pi \in [0,1]$, which represents the prior probability of the alternative hypothesis being true. For a constant Type I error rate and prior hypothesis parameter, the PPV is a function of the test's power and, similar to power, a function $\psi(n,d)$ of the effect and sample sizes. Moreover, in multiple testing scenarios, such PPV functions can be generalized to minimal and maximal PPV functions $\psi_{min}(n,d)$ and $\psi_{max}(n,d)$ by substitution of the respective minimal and maximal power functions. Similar to power functions, single test and multiple testing PPV functions allow finding the sample size n for which, at a given effect size d, the PPV function reaches a desirable level, such as $\psi(n,d)=0.8$.

Partial alternative hypothesis scenarios

Previous approaches to the evaluation of power in fMRI inference have typically relied on the assumption that the experimental effect of interest is expressed in a known cortical region of interest, i.e., single test scenarios (e.g.,Mumford and Nichols (2008)), or in multiple testing scenarios, across the entire cortical volume (e.g., Hayasaka et al. (2007)). While there are situations in which prospective power analyses are reasonable under these assumptions, we here suggest that the evaluation of necessary samples sizes may often be desired although neither the precise location of an expected activation nor the activation of the entire cortical sheet is reasonably assumed. To this end, we propose to parameterize the power, PPV, and sample size calculations in multiple testing scenarios with a *partial alternative hypothesis parameter* $\lambda \in [0,1]$, which describes the assumed proportion of activated brain volume. Intuitively, for example, $\lambda=0.1$ corresponds to the assumption that 10% of the cortex is truly activated. Formally, λ corresponds to the continuous spatial generalization of the alternative hypotheses ratio of multiple testing scenarios. Note that if $\lambda=0$, the minimal and maximal power are necessarily identically zero, as there are no true activations. Equivalently, if $\lambda=1$, the FWER is necessarily zero, as there are no null activations.

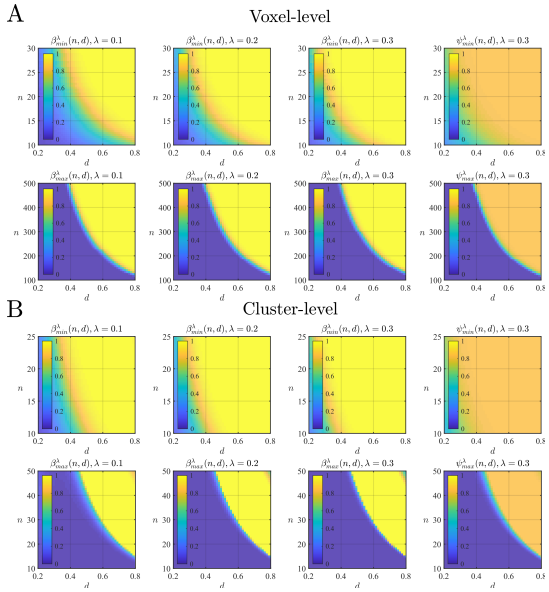

Figure 1: Minimal and maximal power and PPV functions for voxel- and cluster-level inference in the corrected multiple testing scenario. (**A**) Minimal and maximal power and PPV functions for corrected voxel-level inference for a given sample size n, effect size d, and partial alternative hypothesis parameter λ (first three columns). The fourth column depicts the corrected voxel-level minimal and maximal PPV functions for a prior hypothesis parameter of $\pi = 0.2$. (**B**) Minimal and maximal power and PPV functions for corrected cluster-level inference for a given sample size n, effect size d, and partial alternative hypothesis parameter λ (first three columns). The fourth column depicts the corrected cluster-level minimal and maximal PPV functions for a prior hypothesis parameter of $\pi = 0.2$. All cluster-level power functions were evaluated for a CDT of $u = 4.3$, and all voxel- and cluster-level power and PPV functions were evaluated for an exemplary resel volume set of $R_0 = 6$, $R_1 = 33$, $R_2 = 354$, and $R_3 = 705$.

Results

RFT-based power and PPV functions

Based on the theoretical considerations above and the mathematical theory of RFT-based fMRI inference as recently reviewed in Ostwald et al. (2018), it is possible to develop a set of power and PPV functions that are well-aligned with the RFT-based inference framework. In the following, we highlight the power and PPV functions $\beta^{\lambda}_{min}(n,d)$, $\beta^{\lambda}_{max}(n,d)$, $\psi^{\lambda}_{min}(n,d)$, and $\psi^{\lambda}_{max}(n,d)$ for corrected voxel- and cluster-level inference with fixed family-wise error significance levels α'_{FWE} and with fixed partial alternative hypothesis parameters λ. A full and mathematically detailed account of this work is provided in Ostwald et al. (2019).

Figure 1A depicts maximal and minimal power and PPV functions for corrected voxel-level inference at a significance level of $\alpha'_{\mathrm{FWE}} = 0.05$. Specifically, the two leftmost panels of 1A depict the minimal and maximal power functions $\beta^{\lambda}_{min}(n,d)$ and $\beta^{\lambda}_{max}(n,d)$ for corrected voxel-level inference and a partial alternative hypothesis parameter of $\lambda = 0.1$. Achieving a minimal power level of $\beta^{\lambda}_{min}(n,d) = 0.8$ for a medium effect size of $d = 0.5$ requires sample sizes in the range of $n = 15$ to $n = 30$. To achieve similar levels of maximal power $\beta_{max}(n,d)$, the same effect size requires sample sizes of $n = 200$ to $n = 500$. As shown in the upper three panels of Figure 1A, increasing the partial alternative hypothesis parameter to $\lambda = 0.2$ and $\lambda = 0.3$ decreases sample sizes necessary to achieve a minimal power of $\beta^{\lambda}_{min}(n,d) = 0.8$. For maximal power, such a decrease is not observed. Intuitively, this relationship can be understood as follows: increasing the proportion of cortical activation increases the chances of detecting activation at a single cortical location (minimal power) but not of detecting activations at all locations (maximal power). Finally, for a prior hypothesis parameter of $\pi = 0.2$, PPV levels of $\psi^{\lambda}_{min}(n,d) = \psi^{\lambda}_{max}(n,d) = 0.8$ can be achieved with effect and sample sizes largely similar to those for minimal and maximal power, as depicted for $\lambda = 0.3$ in the rightmost column of Figure 1A. Figure 1B depicts maximal and minimal power and PPV functions for corrected cluster-level inference at a significance level of $\alpha'_{\mathrm{FWE}} = 0.05$. As for voxel-level inference, the leftmost panels of Figure 1B depict the minimal and maximal power functions for a partial alternative hypothesis parameter of $\lambda = 0.1$. Here, achieving a minimal power of $\beta^{\lambda}_{min}(n,d) = 0.8$ for a medium effect size of $d = 0.5$ requires sample sizes in the range of $n = 10$ to $n = 20$, while achieving a maximal power of $\beta^{\lambda}_{max}(n,d) = 0.8$ at the cluster level requires sample sizes of $n = 30$ to $n = 50$. As for corrected voxel-level inference, increasing the partial alternative hypothesis parameter to $\lambda = 0.2$ and $\lambda = 0.3$ decreases the necessary sample sizes for minimum power but not for maximum power. Finally, for a prior parameter of $\pi = 0.2$, $\psi^{\lambda}_{min}(n,d) = \psi^{\lambda}_{max}(n,d) = 0.8$ can also be achieved at the cluster level with effect and sample sizes largely similar to those for power (Figure 1B, rightmost column).

Exemplary application

The power and PPV functions presented above imply the sample sizes necessary to achieve desired power and PPV levels over a broad range of possible effect sizes. To demonstrate the practical value of these functions, we finally consider their application in the concrete scenario of determining the sample size necessary to achieve power and PPV levels of 0.8 for a single effect size estimate. To this end, we re-analysed fMRI data from the first 10 participants in a previously reported perceptual decision-making study in which the amount of visual evidence for a presented stimulus to depict a face or a car was varied. At the group level, contrasting fMRI activity levels between high and low visual evidence revealed a cluster of activity in the left medial frontal gyrus, as shown in the upper panel of 2A. Our aim was to use the effect size estimate derived from this cluster to calculate the sample sizes necessary to achieve minimal and maximal power and PPV levels of 0.8 for

corrected voxel- and cluster-level inference at a significance level of $\alpha'_{\text{FWE}} = 0.05$, a partial alternative hypothesis parameter of $\lambda = 0.1$, and a prior hypothesis parameter of $\pi = 0.2$. To this end, we evaluated the average T-values of the cluster, yielding $T = 4.65$, which translates into an effect size estimate of $\hat{d} = 4.65/\sqrt{10} = 1.47$. However, it is well known that effect size estimates resulting from the thresholding of mass-univariate statistical parametric maps exhibit biases (Poldrack et al., 2017). To correct our effect size estimate for this bias, we capitalized on recent results by Geuter et al. (2018), which are depicted in the lower panel of 2A. Specifically, using task-related fMRI data from the Human Connectome Project 500, Geuter et al. (2018) estimated the effect size bias exhibited by activations detected in random data subsets of 10 to 100 participants from the approximately 500 participants. As reported in Figure 7A of Geuter et al. (2018) and visualized in the lower panel of 2A, this effect size bias is most severe for small data subsets and decreases with increasing data subset size. For a data subset of $n = 10$, the effect size bias amounts to approximately $\Delta d = 1$. We thus used this empirically validated bias estimate to correct our effect size estimate to $\hat{d}_c = \hat{d} - \Delta d = 0.47$. Using the power and PPV functions discussed in the previous section we then obtained the following results: at the voxel level, sample sizes of $n = 19$ and $n = 374$ are required to achieve minimal and maximal power levels of 0.8, respectively (2B). At the cluster level, sample sizes of $n = 12$ and $n = 48$ are required to achieve minimal and maximal power levels of 0.8 (2C), respectively. For all testing scenarios considered and for the current parameter settings, slightly smaller sample sizes are required to achieve PPV levels of 0.8.

Conclusion

In summary, we have developed power and PPV functions for RFT-based fMRI inference, which represents one of the mainstays of task-related fMRI data analysis. Further, we have demonstrated, how these functions can be used to determine the minimal sample sizes necessary to achieve desired power and PPV levels in study planning, and find that for the most commonly used approach of corrected cluster-level inference, minimal samples sizes of 40 to 50 participants are required at a medium effect size.

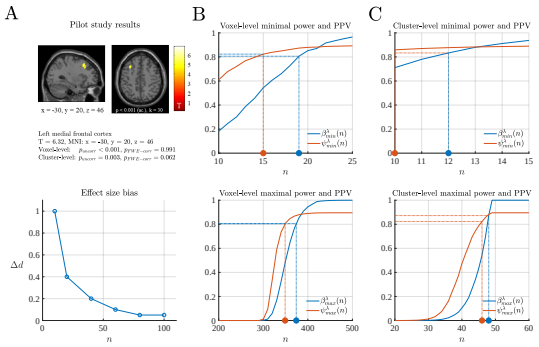

Figure 2: Exemplary application of the RFT-based power, PPV, and sample size calculation framework. (**A**) The upper panel depicts the results of a perceptual decision-making pilot study with $n = 10$ participants for contrasting perceptual choices based on high and low visual sensory evidence. The T-values from the identified cluster in the left medial frontal gyrus were averaged to obtain a raw effect size estimate, which was then adjusted based on the effect size bias estimates reported in Figure 7 of Geuter et al. (2018) and reproduced in the lower subpanel of panel (A). (**B**) Sample size calculations for voxel-level minimal and maximal power and PPV based on the effect size estimates of the pilot fMRI study. (**C**) Sample size calculations for cluster-level minimal and maximal power and PPV based on the effect size estimates of the pilot fMRI study.

References

Button, K., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Cohen, J., et al. (2017). Computational approaches to fmri analysis. *Nature neuroscience*, *20*(3), 304.

Durnez, J., et al. (2016). Power and sample size calculations for fmri studies based on the prevalence of active peaks. *bioRxiv*.

Friston, Frith, C., Liddle, P., & Frackowiak, R. (1991). Comparing functional (pet) images: the assessment of significant change. *Journal of Cerebral Blood Flow & Metabolism*, *11*(4), 690–699.

Friston, K., et al. (1994). Assessing the significance of focal activations using their spatial extent. *Human brain mapping*, *1*(3), 210–220.

Friston, K., et al. (1996). Detecting activations in pet and fmri: levels of inference and power. *Neuroimage*, *4*(3), 223–235.

Geuter, S., et al. (2018). Effect size and power in fmri group analysis. *bioRxiv*, 295048.

Hayasaka, S., et al. (2007). Power and sample size calculation for neuroimaging studies by non-central random field theory. *NeuroImage*, *37*(3), 721–730.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), e124.

Mumford, J. A., & Nichols, T. E. (2008). Power calculation for group fmri studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage*, *39*(1), 261–268.

Ostwald, D., et al. (2018). Random field theory-based p-values: a review of the spm implementation. *ArXiv*.

Ostwald, D., et al. (2019). Power, positive predictive value, and sample size calculations for random field theory-based fmri inference. *bioRxiv*.

Poldrack, R., et al. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, *18*(2), 115.

Worsley, K., et al. (1992). A three-dimensional statistical analysis for cbf activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism*, *12*(6), 900–918.