Published in final edited form as:

Neuroinformatics. 2012 October ; 10(4): 351–365. doi:10.1007/s12021-012-9152-3.

Development of PowerMap: a software package for statistical power calculation in neuroimaging studies

Karen E. Joyce¹ and Satoru Hayasaka^{2,3}

¹Department of Biomedical Engineering, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA

²Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA

³Department of Radiology, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA

Abstract

Although there are a number of statistical software tools for voxel-based massively univariate analysis of neuroimaging data, such as fMRI (functional MRI), PET (positron emission tomography), and VBM (voxel-based morphometry), very few software tools exist for power and sample size calculation for neuroimaging studies. Unlike typical biomedical studies, outcomes from neuroimaging studies are 3D images of correlated voxels, requiring a correction for massive multiple comparisons. Thus, a specialized power calculation tool is needed for planning neuroimaging studies. To facilitate this process, we developed a software tool specifically designed for neuroimaging data. The software tool, called PowerMap, implements theoretical power calculation algorithms based on non-central random field theory. It can also calculate power for statistical analyses with FDR (false discovery rate) corrections. This GUI (graphical user interface)-based tool enables neuroimaging researchers without advanced knowledge in imaging statistics to calculate power and sample size in the form of 3D images. In this paper, we provide an overview of the statistical framework behind the PowerMap tool. Three worked examples are also provided, a regression analysis, an ANOVA (analysis of variance), and a two-sample T-test, in order to demonstrate the study planning process with PowerMap. We envision that PowerMap will be a great aide for future neuroimaging research.

Keywords

statistical power; sample size; study planning; random field theory; non-central distribution

Introduction

When planning biomedical studies, investigators have to consider two seemingly paradoxical facts. It is important to have a sufficiently large number of subjects to detect

signals or effects of interest. On the other hand, it is also important to include as small a number of subjects as possible in order to avoid unnecessarily exposing them to unforeseen risks and to reduce study-associated costs. Therefore determining the appropriate number of subjects satisfying both constraints is an important step in study planning. If some preliminary data provide an estimate of the anticipated effect size, investigators can estimate statistical power and determine the sufficient number of subjects required to detect signals in their studies. In a study with a single outcome variable, the process of power analysis is straightforward (Cohen 1988). A number of software packages and tools for such scenarios are widely available. In neuroimaging studies, however, power calculation is a complicated process. This is because outcomes are 3D images with tens of thousands of correlated voxels, and the single-outcome power analysis is not appropriate for such data structure.

Although relatively little effort has been focused on power analysis in neuroimaging compared to statistical inference, there have been some attempts to calculate power. Theoretical work started as early as the application of RFT (random field theory) to neuroimaging data to obtain a corrected threshold to control family wise error (FWE) rate. Friston et al. modeled signals as a Gaussian random process, and produced a power surface by expressing power as a function of the threshold height and the signal width. This was later used to discuss power (Friston et al. 1996) and sample sizes (Friston et al. 1999) in neuroimaging data analysis from a theoretical perspective. More sophisticated power calculation methods using non-central T- and F-distributions have been suggested by others (Mumford and Nichols 2008; Van Horn et al. 1998; Zarahn and Slifstein 2001). Among these methods, the one by Mumford and Nichols has been implemented as a software tool known as fmripower (Mumford and Nichols 2008) (can be downloaded from <http://fmripower.org/> with details explained in the poster at http://fmripower.org/mumford_hbm_2007.pdf). In addition to these theoretical approaches, simulation and resampling-based methods have also been used to calculate power (Desmond and Glover 2002; Murphy and Garavan 2004). Some of these studies have been used as a guideline by neuroimaging researchers when planning the required number of subjects. Although fmripower, the only existing software tool for power calculation for neuroimaging studies, can take into account regional variability in effect sizes, it cannot correct for multiple comparisons among spatially correlated voxels.

Despite the popularity, the methods described above are unable to account for multiple comparisons among correlated voxels. To address this shortcoming, Hayasaka et al. (Hayasaka et al. 2007) developed a theoretical framework to calculate statistical power in a neuroimaging data analysis. In this context, multiple comparisons and spatial correlation among voxels work favorably in detecting signals. This is because true signals are likely to occur in areas consisting of multiple contiguous voxels due to spatial correlation, and the likelihood of detecting any part of the signals, comprising multiple voxels, is much higher than detecting a false positive occurring at an isolated voxel. Power calculation in the presence of multiple comparisons among regions-of-interests (ROIs) can be handled with a Bonferroni correction by fmripower, but not among correlated voxels. The method by Hayasaka et al. (Hayasaka et al. 2007), based on RFT for non-central T- and F-fields, can calculate statistical power to detect signals among spatially correlated voxels. In particular, this method can calculate the probability to detect signals by an RFT-based corrected

threshold (Worsley et al. 1996) controlling the family-wise error (FWE) rate. Since this method is based on a theoretical framework, power can be calculated analytically without computationally expensive simulations or resampling. Moreover, by calculating power at different locations in the brain and organizing the results as a 3D image, this method enables visualization of spatially varying power throughout the brain (Hayasaka et al. 2007).

To facilitate the use of the non-central RFT-based power calculation described above, we developed a software tool specifically designed for neuroimaging studies. In particular, it is designed for voxel-based massively univariate analyses of neuroimaging data such as fMRI (functional MRI), PET (positron emission tomography), and VBM (voxel-based morphometry). The software tool, called PowerMap, can generate power and sample size estimates for a neuroimaging study in a form of 3D images. Although PowerMap is not the first software tool to visualize statistical power as a 3D image (Mumford and Nichols 2008), it is the only tool that can account for multiple comparisons among correlated voxels and calculate power with FWE or FDR (false discovery rate) correction in mind. Power and sample size calculations discussed in this paper pertain to voxel-based tests only, and consequently PowerMap is designed for study designs involving voxel-based tests but not for cluster-based tests. This GUI (graphical user interface)-based tool allows users without advanced knowledge in imaging statistics to calculate power based on the results from a pilot data analysis. This paper is intended as an overview of how the tool works. We briefly explain some basic concepts of statistical power calculation implemented in PowerMap. Then we will present worked examples of neuroimaging studies: a T-test for the linear regression slope, a one-way ANOVA (analysis of variance), and a two-sample T-test.

Methods

Power Calculation Basics

Statistical power is defined as the probability to detect the signal or effect of interest by a statistical test when the effect is present. More formally, it is the probability to reject the null hypothesis H_0 (or accepting the alternative hypothesis H_A) when H_A is true. In PowerMap, statistical power is calculated around each voxel and visualized in the form of a 3D image. Power is calculated “around” each voxel rather than “at” each voxel, meaning that the calculated power reflects the probability of detecting H_A in a neighborhood centered at each voxel. We focus on a neighborhood because, due to spatial correlation among voxels, the signal of interest likely spans across multiple contiguous voxels rather than confined to a single voxel. The neighborhood around each voxel is defined as a sphere of radius r , where r is the smoothness parameter in terms of full-width at half-maximum (FWHM) described below (Hayasaka et al. 2007). According to the matched filter theorem, an optimally detected signal should have a width similar to the image smoothness FWHM (Worsley et al. 1996). Thus we use a sphere of radius $r = \text{FWHM}$ to calculate the effect size so that the voxel at the center is likely a part of a signal within that sphere.

In a statistical test to control the FWE rate, power is determined as the probability to detect any signal exceeding an FWE-corrected threshold (Hayasaka et al. 2007). An FWE-corrected threshold can be determined according to the desired level of FWE rate based on

RFT (Worsley et al. 1996). Interested readers can find details on power calculation for an FWE-based test elsewhere (Hayasaka et al. 2007).

Statistical power can also be calculated theoretically for a test controlling the false discovery rate (FDR). False discoveries refer to situations when a statistical test falsely concludes H_A when the underlying truth is H_0 . When a large number of multiple comparisons are performed, one may be willing to tolerate some level of false discoveries as long as the true signals can also be detected. The rationale for controlling FDR is to control the proportion of false discoveries to a small fraction of all discoveries, enabling discoveries of a large number of true positives. This is in contrast to an FWE-based test which completely eliminates false positives with a highly stringent threshold. The FDR method has been shown to control false positives at the desired level even under a large number of multiple comparisons in an imaging genetics study (Meyer-Lindenberg et al. 2008), and works even in the presence of correlations (Efron 2004). In a statistical test to control the FDR, similar to an FWE-based test described above, power can be determined as the probability of signals exceeding an FDR-corrected threshold. An FDR-threshold controls the number of false positives (R_0) among all positive findings (R) (see Table 1). In particular, the FDR-threshold u_c can be determined as the solution of the equation:

$$q = \frac{(1 - \pi)(1 - F_0(u_c))}{1 - (1 - \pi)F_0(u_c) - \pi F_1(u_c)} \quad (1)$$

where F_0 and F_1 are the cumulative distribution function (CDF) of the test statistic under H_0 and H_A respectively, π is the proportion of the true signal (M_1) among all the statistical tests (M), and q is the desired level of FDR. While there are various methods to determine F_0 and F_1 empirically (Benjamini and Hochberg 1995; Genovese et al. 2002; Efron et al. 2001), in our power calculation framework we assume F_0 is a central distribution (e.g., T-distribution, F-distribution, etc.) whereas F_1 is the corresponding non-central distribution with a certain level of non-centrality. Once the FDR-threshold u_c is determined, power γ resulting from the FDR method can be calculated as

$$\gamma = \frac{R_1}{M_1} = 1 - F_1(u_c) \quad (2)$$

where R_1 refers to the number of true positives (see Table 1).

Overview of PowerMap

The PowerMap tool generates power and sample size images based on a test statistic image from a pilot study. First, the user selects a statistic image, and provides additional information necessary for power calculation. One of the input parameters is the image smoothness. This information is not readily known unless the user has analyzed the pilot data with a software package that calculates the image smoothness in terms of full-width at half-maximum (FWHM), such as the SPM (statistical parametric mapping) package. If it is unknown, PowerMap estimates the smoothness from a statistic image provided by the user. The other input parameters required for power calculation are the degrees of freedom and the sample size(s) of the statistic image from the pilot study. Once all the necessary

information is available, PowerMap calculates statistical power for various combinations of non-centrality parameter values (describing the effect to be detected) and degrees of freedom (describing the number of subjects required) (Hayasaka et al. 2007). The results from this calculation are saved as a power surface, to be used later. Based on the power surface, a power image or a sample size image will then be generated as a 3D image with each voxel describing statistical power or a necessary sample size, respectively, at that particular voxel. The resulting power or sample size image can be viewed by an imaging viewing software tool of the user's choice. Figure 1 shows the overall processing stream of the PowerMap tool.

PowerMap is designed to plan studies with simple statistical hypotheses listed below. If users are interested in detecting effects or signals significantly greater than zero, they can use a one-sample T-test for their analysis. If users are interested in detecting a significant correlation between voxel values and a continuous variable of interest, then they can use a regression analysis. If users are interested in detecting any significant differences among two or more groups, then they can use either a two-sample T-test (for two groups) or an ANOVA (for more than two groups).

Input Parameters

The user input for PowerMap includes a statistical image from a pilot study, as well as some parameters associated with the image. The statistic image can be either a T-statistic image or an F-statistic image. PowerMap can handle statistic images generated by simple study designs, including a one-sample T-test (T-statistic image), a two-sample T-test (T-statistic image), a linear regression analysis (T-statistic image), and a one-way ANOVA (F-statistic image). Along with the statistic image, the user is also asked to provide a binary mask image identifying brain voxels. The mask image can include the whole brain, or a subset of the brain voxels of interest if there is an a priori hypothesis for particular brain regions (e.g., left hemisphere, temporal lobe, etc.). When the input images are specified, the user is asked to identify the study design associated with the statistic image, as well as the degrees of freedom associated with the statistic image. At this point, if the image smoothness is known, then the user is asked to enter that information as well.

Smoothness Estimation

In massively univariate analyses of neuroimaging data, image smoothness refers to the width of a Gaussian convolution filter that needs to be applied to a 3D white noise image in order to produce the same level of spatial correlation observed in the data (Worsley et al. 1992). This is often an unknown parameter, but it can be estimated from the data (Kiebel et al. 1999; Worsley 1994). If the image smoothness is unknown, then PowerMap estimates it based on the statistic image provided. In a typical statistical parametric mapping analysis, the smoothness is estimated based on the covariance matrix Λ of the gradient of the residual images Z (Hayasaka et al. 2004; Kiebel et al. 1999). However the residual images from the pilot data analysis may not be always available. Thus we provide an improvised smoothness calculation method as part of the PowerMap package. If an F-statistic image F with degrees of freedom m and n is provided by the user, then the covariance matrix of the gradient of F , $\text{var}(\nabla F)$, is calculated. This is done by calculating the numerical gradient at each voxel

location, then calculating the variance of the gradients across all the voxel locations. Once $\text{var}(\nabla F)$ is obtained, then we can estimate Λ matrix by the following formula:

$$\Lambda = \left(\frac{m}{n}\right)^2 \frac{\text{var}(\nabla F)}{\sigma_x^2 \sigma_y^2 + \sigma_x^2 \mu_y^2 + \sigma_y^2 \mu_x^2 + \mu_y^2 \mu_x^2} \quad (3)$$

where m and n are the numerator and denominator degrees of freedom of F , respectively, and

$$\mu_x = (m+n-2) \left\{ \Gamma\left(\frac{m+1}{2}\right) \Gamma\left(\frac{n-3}{2}\right) \left(\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right)\right)^{-1} \right\}$$

$$\sigma_x^2 = (m+n-2)(m+n-4) \frac{m}{(n-2)(n-4)(n-6)} - \mu_x^2$$

$$\mu_y = \frac{1}{\sqrt{2}} \frac{\Gamma\left(\frac{m+n-1}{2}\right)}{\Gamma\left(\frac{m+n}{2}\right)} \quad \sigma_y^2 = \frac{1}{m+n-2} - \mu_y^2$$

In the formula above, Γ denotes the gamma function. The derivation of (3) is given in Appendix A. Similarly, if a T-statistic image T with degrees of freedom m is provided by the user, then Λ matrix can be estimated by:

$$\Lambda = \frac{\text{var}(\nabla T)}{\sigma_x^2 \sigma_y^2 + \sigma_x^2 \mu_y^2 + \sigma_y^2 \mu_x^2 + \mu_y^2 \mu_x^2} \quad (4)$$

where

$$\mu_x = m^{1/2} \left(\frac{m-1}{m-2}\right) \quad \sigma_x^2 = \frac{2m(m-1)}{(m-2)^2(m-4)}$$

$$\mu_y = \frac{2^{1/2} \Gamma(m/2)}{\Gamma\left(\frac{m+1}{2}\right)} \quad \sigma_y^2 = \frac{1}{m-1} - \mu_y^2$$

Again, the derivation of (4) is described in Appendix A. Based on Λ , the smoothness in terms of FWHM can be calculated as $FWHM = (4 \log 2)^{3/2} (\det(\Lambda))^{-1/6}$ (Hayasaka et al. 2004).

With the smoothness parameter, either entered by the user or calculated by the procedure above, PowerMap calculates the RESEL counts for the brain volume as well as the signal volume. RESEL stands for resolution elements (Worsley et al. 1992; Worsley et al. 1996),

and it is the smoothness-adjusted volume. More specifically, the RESEL counts describe the topological characteristic of a volume by a 4-element vector: the Euler characteristic (the number of connected sets minus the number of holes), the linear diameter, the surface area, and the volume (Worsley 1996). The RESEL counts for the brain volume are calculated for the brain voxels identified in the mask image by the functions based on the *fmrstat* package (Worsley 2005). The anticipated signal volume is defined as a sphere around each voxel with radius = FWHM (Hayasaka et al. 2007), and the RESEL counts of this sphere is constant, namely $[1 \ 4 \ 2\pi \ 4\pi/3]$. This is because the sphere has the constant diameter of 2 (in terms of FWHM), and the corresponding 4-element vector for the RESEL counts (as described above) remains constant. While the RESEL counts for the brain volume is used to calculate the FWE-corrected threshold (Worsley et al. 1996), the RESEL counts for the signal volume will be used to represent the anticipated signal during power calculation.

Effect Size Calculation

Based on the statistic image selected by the user, an effect size image is generated. To do so, first the statistic image provided by the user is convolved with a sphere of radius = FWHM. This is equivalent to averaging voxel values within the sphere centered at each voxel. This step is necessary for both FWE- and FDR-corrections since PowerMap calculates power to detect a signal in a sphere around each voxel. Each voxel value in the effect size image corresponds to the effect size, either in terms of Cohen's d (for a T-statistic image) or Cohen's f^2 (for an F-statistic image), at that particular voxel location (Cohen 1988). In a one-sample T-test or a regression analysis, Cohen's d can be calculated as $d = T / \sqrt{N}$ where T is the T-statistic and N is the total sample size. In our Cohen's d estimate, T is divided by $\sqrt{N}$ rather than $\sqrt{N - 1}$ or $\sqrt{N - 2}$ corresponding to the appropriate degrees of freedom for a one-sample T-test or a regression analysis. This is done in order to produce a slightly conservative estimate of the effect size. In study planning, researchers often over-recruit a few additional subjects in their studies in order to account for drop outs and uncertainty in power calculation. Thus, PowerMap is designed to safeguard against under-recruiting subjects by slightly underestimating the effect size. In a two-sample T-test, d can be calculated as $d = T \sqrt{(N_1 + N_2) / (N_1 N_2)}$ where N_1 and N_2 are the sample sizes in the two groups to be compared. For a one-way ANOVA, Cohen's f^2 can be calculated as $f^2 = F(df_1 / df_2)$ where F is the F-statistic, and df_1 and df_2 correspond to the numerator and denominator degrees of freedom, respectively. Only the brain voxels (identified by the mask image) with positive effect sizes are used in power and sample size image generation because negative effect sizes mean the opposite direction from the effect the statistical test is designed to detect. In such cases, power is likely very small, smaller than the significance level of the test. This is equivalent to calculating power for a signal decrease when the test is designed to detect a signal increase.

While statistical power can be calculated for each voxel value in the effect size image separately, a more computationally efficient approach is to pre-calculate power for a number of combinations of df and non-centrality (nc) and to use the results to interpolate power at different voxel locations. This approach is preferred since some of the calculations involved in non-central random field theory do not have a closed form (Hayasaka et al. 2007). Since

the RESEL counts for the brain volume and the signal volume (i.e., a sphere around each voxel) are known, statistical power with an FWE-corrected threshold can be calculated for each combination of df and nc, producing a surface (see Figure 2). The power surface is displayed on the monitor as a visual indicator that the power surface calculation has been successfully completed, signifying that the remaining steps in power / sample size calculation can be undertaken. Since it serves as a visual indicator of the calculation process, the power surface figure produced by PowerMap is not designed to be interpreted by users. Power maps and sample size maps (see below) are provided for that purpose. The resulting power surface is used in power and sample size image generation described below. Likewise, if an FDR-corrected threshold is to be used, then the threshold can be calculated for each combination of df and nc by (1) assuming M_I and M correspond to the signal and brain volumes respectively. The resulting threshold can then be used in (2) to obtain FDR-corrected power. This process is repeated for all combinations of df and nc to produce a power surface.

Power and Sample Size Image Generation

An effect size image can be converted to a power image if the relationship between the effect size and power is known for certain df. This relationship is a one-to-one relationship, and can be numerically evaluated by extracting a curve at particular df from the power surface described above (see Figure 2, pink plane). In particular, the curve describes power as a function of nc, which can be easily converted to the effect size if df is known. For each voxel value of the effect size image, the corresponding power is determined by a numerical interpolation. This process is repeated for all the voxels in the effect size image, resulting in a 3D power image.

A similar process can be used to generate a sample size image. An effect size image can also be converted to a sample size image if the relationship between the effect size and df is known for a particular value of power. This relationship can be described numerically as a curve extracted from the power surface at the particular level of power (see Figure 2, green plane). The curve describes df as a function of the effect size. For each voxel value from the effect size image, df can be obtained to achieve the desired level of power. The resulting df can be easily converted to the sample size depending on the study design. While the sample size image describes the total number of required subjects in the one-sample T-test and the regression analysis, the number of subjects in each group is described in the power image for the two-sample T-test and the one-way ANOVA.

System Requirements and Compatibility

PowerMap has been tested on MATLAB R2009 software (MathWorks; Natick, MA) on the Linux and Windows platforms. Since power calculation involves a number of statistical functions in a series form, the Statistics Toolbox is also required in addition to the base MATLAB. PowerMap can handle image data in the NIfTI (Neuroimaging Informatics Technology Initiative) format as well as the ANALYZE format with a NIfTI-compatible header. Although PowerMap is self-contained and does not require any additional software tools besides MATLAB, it does not have an image viewer as a part of the software package. Thus users are encouraged to use their favorite image viewing software tools to examine

power and sample size images generated by PowerMap. PowerMap can be downloaded from the SourceForge website (<https://sourceforge.net/projects/powermap/>). After installing PowerMap, the install folder must be added to MATLAB's default path.

Worked Examples

To highlight the use of PowerMap as a study planning tool, we demonstrate three worked examples in neuroimaging studies.

Example 1: Regression Analysis

This section details the process of generating a power map for an example regression analysis. Pilot study data consisted of fractional anisotropy (FA) images collected from 15 subjects aged 18 to 67 years (mean $\pm$ standard deviation: 48.5 ± 20). The hypothesis tested was that there is a decrease in FA with age due to deterioration of white matter. A pilot regression analysis was performed with age as a covariate, resulting in a T-statistic image. To generate a corresponding power image, the T-statistic image, a whole-brain mask image, and the test type (linear regression) were supplied as inputs to the PowerMap software (Figure 3). The map options were specified such that a FWE-corrected ($\alpha = 0.05$) power image would be created, with smoothness entered manually as determined from the pilot regression analysis (FWHM = 9.4 by 10 by 10.1 mm). Next, the pilot study sample size was specified as 15 subjects and the df of the pilot study was correspondingly specified as 14. Since a power image is desired, the anticipated group size must be specified. For this example, the anticipated group size was set to 30 subjects. After the PowerMap software ran, a power surface was output to the screen, and the power image was viewed separately using MRICro (McCauley Center for Brain Imaging, University of South Carolina; Columbia, SC, USA).

Several transverse slices of the power image are shown in Figure 4. The power image reveals increased power primarily in white matter regions. These regions include the genu and anterior midbody of the corpus callosum, the anterior forceps, the thalamus, and the optic radiations. For a study size of 30 subjects, there is a high ability to detect differences in FA due to age effects in these areas. These results are consistent with the findings of (Hugenschmidt et al. 2008), where a regression analysis comparing age against FA revealed changes in similar areas.

Example 2: One-Way ANOVA

This section details an example of generating a sample size image for a one-way ANOVA analysis. Pilot study data consisted of gray matter (GM) voxel-based morphometry (VBM) data for three groups of subjects, consisting of 5 young subjects (27.2 ± 8.87 years), 5 middle aged subjects (47.2 ± 9.23 years), and 5 older subjects (71.2 ± 4.08 years). The contrast was defined to test the hypothesis that GM volume is greatest in the younger group, decreased in the middle group, and lowest in the older group. A pilot one-way ANOVA analysis was performed with mean-centered intracranial volume as a covariate, resulting in an F-statistic image. This image, a corresponding whole-brain mask, and the test type (ANOVA) were supplied to the PowerMap software for generating a sample size image

(Figure 5). The output options were specified such that a FWE-corrected ($\alpha = 0.05$) sample size image would be created, with smoothness entered manually as determined from the previous pilot ANOVA ($\text{FWHM} = 9.4$ by 10 by 10.1 mm). Next, the pilot study sample size was specified as 5 subjects per group and the DF of the pilot study was correspondingly specified as [2 11]. Since a sample size image is desired, the desired level of power was specified to be 80%. After the PowerMap software ran, a power surface was output to the screen, and the sample size image was viewed separately using MRIcro.

More detailed slices of the sample size image are shown in Figure 6. For these images, the sample size image was thresholded to show sample sizes between 5 subjects per group (the size of the pilot study) and 10 subjects per group. These results indicate that lower sample sizes are needed to show significant results in areas such as the insula, the hippocampus, and the anterior cingulate.

Example 3: Two-Sample T test

This final example illustrates the process of generating a power image based on a two-sample T test statistical image with FDR correction. Pilot study data consist of two groups of older healthy subjects (69.5 ± 3.2 years), consisting of 20 control subjects and 21 intervention subjects. Perfusion data were collected for each subject to measure cerebral blood flow (CBF). The two-sample T test was designed to evaluate changes in CBF due to participation in an auditory and visual attention training program. The full details of this study are described in (Mozolic et al. 2010). The hypothesis tested was that there would be a greater increase in CBF in the intervention group than in the control group. A pilot two-sample T test was performed to test this hypothesis, resulting in a T statistic image and a mask image. These images were input to the PowerMap program (Figure 7). The map type was specified as a power map with FDR threshold correction at a significance level of 0.05. The smoothness data were specified manually as $12.8 \times 13.8 \times 16.9$ mm. Pilot study group sizes were 20 and 21 for groups 1 and 2, respectively, resulting in DF equaling 39. The full scale study group sizes were specified as 50 subjects for each group.

Slices of the resulting power map are shown in Figure 8. Regions with the greatest power include the bilateral inferior frontal cortex, medial frontal cortex, and rostral anterior cingulate cortex. These results are in agreement with the findings published in (Mozolic et al. 2010).

Discussion

In this paper, we demonstrated how our software tool PowerMap can facilitate planning of neuroimaging studies. Through the GUIs provided in PowerMap, a user can generate power and sample size images without advanced knowledge in imaging statistics. Producing power and sample sizes as images, PowerMap can visualize regional differences down to the granularity of voxels. Although there are other software packages that enable visualization of power as a brain image, PowerMap can incorporate multiple comparison corrections controlling either the FWER or FDR. To the best of our knowledge, this is the only tool that allows for multiple comparison corrections. PowerMap is designed as a self-contained

package; hence it does not require any additional image analysis packages to function (e.g., SPM, or FSL).

Although PowerMap simplifies power calculation in neuroimaging studies tremendously, there are some limitations. Most notably, since the FWE-corrected power calculation algorithm is based on RFT, it suffers from similar limitations as the SPM-type analyses. Namely, the statistic image has to be smooth, with its smoothness being uniform throughout the brain space. Voxel values in the statistic image are assumed to follow a known distribution (T- or F-distribution). Moreover, there are some statistical issues unique to the power calculation setting. The estimation of the effect size needs to be carried out with a very small sample size in a typical pilot study; thus the resulting effect size estimates have large errors associated with them. In addition to the effect size, PowerMap also provides a means to estimate the image smoothness from a statistic image rather than from a collection of residual images. This may further introduce variability into the parameters necessary for power calculation. However, our qualitative comparison indicates the impact of the improvised smoothness estimation implemented in PowerMap may not be substantial (Figures 9 and 10). Another limitation of PowerMap is that it can only handle simple study designs. We focused on very basic scenarios of neuroimaging analyses in PowerMap, and consequently it is unable to handle complex designs such as unequal inter-subject variances in different groups as done in the FSL package (Smith et al. 2004). Thus, users need to be aware that the results from PowerMap may incorrectly underestimate the number of subjects required for their studies. Indeed, PowerMap can only handle simple study designs, and consequently study designs PowerMap cannot handle are too numerous to list. Users of PowerMap should be aware of this limitation when they plan studies with complex designs. In addition, PowerMap can handle only two types of multiple comparison corrections, namely RFT-based FWE-corrections and FDR-corrections. There are other types of multiple comparison correction methods exist, such as Bonferroni corrections and permutation-based FWE corrections (Nichols and Hayasaka 2003). Thus, users planning on using these correction methods cannot use PowerMap for their study planning. Finally, we would like to point out that statistical power calculation can rarely account for all the variability present in data yet to be collected. This is true even for non-imaging data. Even though we are confident that the algorithm and processing functions included in PowerMap are theoretically valid, we also would like to emphasize that power calculation results generated by PowerMap are never perfect and users need to interpret the results very cautiously.

Users of PowerMap should also be aware of statistical issues associated with massively univariate analyses of neuroimaging data. First, although PowerMap can estimate power and sample sizes associated with FWE- and FDR-corrected data, it is only designed for voxel-based tests (Nichols and Hayasaka 2003) and is unable to account for power in cluster-based tests (Hayasaka and Nichols 2003). This is because cluster-based tests operate on a different statistical principle than voxel-based tests (Friston et al. 1994). Thus, users should not plan their study using PowerMap if they intend to use cluster-based tests. Secondly, the two multiple comparison correction methods, FWE-corrections and FDR-corrections, operate in two fundamentally different principles. While FWE-corrections protect against any false positives in an analysis, FDR-corrections tolerate some level of false positives (Nichols and Hayasaka 2003). Consequently, FWE-corrected tests are more conservative and less

powerful compared to FDR-corrected tests. Users may be tempted to choose an FDR-correction when they calculate power and/or the sample size for their study because of this reason. However, if users plan to analyze their data using an FWE-correction, power and the sample size should be calculated using the FWE-correction in PowerMap, rather than misrepresenting inflated power by selecting an FDR-correction.

In summary, PowerMap enables simple statistical power calculation for average neuroimaging researchers. There may be some theoretical or technical limitations associated with this tool, but the results generated by PowerMap can provide a convincing support for a research proposal, as a picture is far easier to interpret than a paragraph justifying sample size with statistical jargon. Accurate study planning enables a researcher to scan a sufficiently large number of subjects to detect the effect of interest, but not exceedingly many subjects to limit the study-related costs within the budget. PowerMap can determine just the right number of study subjects while accounting for a large number of parameters and voxel values in neuroimaging studies. Thus we envision that this will be a great tool for future neuroimaging research.

Acknowledgements

This work is supported by the National Institute of Neurological Disorders and Stroke (NINDS) (NS059793). The authors would like to thank Ms. Malaak Moussa for reviewing our analysis results.

Appendix A

For three independent random variables X , Y , and Z , (means μ_X , μ_Y , and μ_Z , and variances σ_X^2 , σ_Y^2 , and σ_Z^2 , respectively), the variance of the product XYZ is

$$Var(XYZ) = \sigma_X^2 \sigma_Y^2 \sigma_Z^2 + \mu_X^2 \sigma_Y^2 \sigma_Z^2 + \mu_Y^2 \sigma_X^2 \sigma_Z^2 + \mu_Z^2 \sigma_X^2 \sigma_Y^2 + \mu_X^2 \mu_Y^2 \sigma_Z^2 + \mu_X^2 \mu_Z^2 \sigma_Y^2 + \mu_Y^2 \mu_Z^2 \sigma_X^2 \quad (A1)$$

Another useful result is that, for a chi-square random variable W with $df=v$, then the expected value of W^r for real-valued r is given by

$$E[W^r] = \frac{2^r \Gamma(\frac{v}{2} + r)}{\Gamma(\frac{v}{2})} \quad (A2)$$

where Γ denotes the gamma function (Johnson et al. 1995). Worsley has shown that the gradient of a 3D T-random field T , ∇T , follows the distribution

$$\nabla T = m^{\frac{1}{2}} \left(1 + \frac{T^2}{m} \right) S^{-\frac{1}{2}} Z$$

where T is the value of a T-random random field with $df=m$ at the location where the gradient is calculated, S is a chi-square random variable with $df=m+1$, and Z is a 3×3-dimensional normal random variable with mean zero and variance defined by a matrix Λ , with T , S , and Z are all independent (Worsley 1994). The gradient of T depends on the value of T itself. The variance matrix Λ needs to be estimated based on an observed T-statistic

image. Since ∇T can be seen as the product of three independent random variables, namely,

$X = m^{1/2} (1 + T^2 / m)$, $Y = S^{-1/2}$, and Z , we can use (A1) to calculate its variance. Using (A2), we can calculate the means and the variances of these variables:

$$\mu_X = m^{1/2} \left(\frac{m-1}{m-2} \right) \quad \sigma_X^2 = \frac{2m(m-1)}{(m-2)^2(m-4)}$$

$$\mu_Y = \frac{2^{1/2} \Gamma(m/2)}{\Gamma\left(\frac{m+1}{2}\right)} \quad \sigma_Y^2 = \frac{1}{m-1} - \mu_Y^2$$

$$\mu_Z = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \sigma_Z^2 = \Lambda$$

Since μ_Z is zero, some terms in (A1) cancel and the variance of ∇T can be obtained as

$$\text{Var}(\nabla T) = (\sigma_X^2 \sigma_Y^2 + \mu_X^2 \sigma_Y^2 + \mu_Y^2 \sigma_X^2 + \mu_X^2 \mu_Y^2) \Lambda$$

and this can be solved for Λ to yield (4). In order to derive the similar results for ∇F , we focus on the variance of ∇G where G is a random field defined as $(m/n)F$. In other words, G is a random field defined as the ratio U/V of two chi-square random fields U (with $\text{df}=m$) and V ($\text{df}=n$). Worsley (Worsley 1994) has described that ∇G can be described as

$$\nabla G = 2G^{1/2} (1+G) W^{-1/2} Z$$

where G is the value of a G-random field at the location where the gradient is calculated, with G defined as $(m/n)F$, F is an F-random variable with $\text{df}=(m,n)$, W is a chi-square random variable with $\text{df}=m+n$, and Z is a 3×3-dimensional normal random variable as described above. The gradient of G depends on G itself. If we let $X = 2G^{1/2} (1+G)$, $Y = W^{-1/2}$, and Z as it is, then their means and variances are

$$\mu_X = (m+n-2) \left\{ \Gamma\left(\frac{m+1}{2}\right) \Gamma\left(\frac{n-3}{2}\right) \left(\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right) \right)^{-1} \right\}$$

$$\sigma_X^2 = (m+n-2)(m+n-4) \frac{m}{(n-2)(n-4)(n-6)} - \mu_X^2$$

$$\mu_Y = \frac{1}{\sqrt{2}} \frac{\Gamma\left(\frac{m+n-1}{2}\right)}{\Gamma\left(\frac{m+n}{2}\right)} \quad \sigma_Y^2 = \frac{1}{m+n-2} - \mu_Y^2$$

$$\mu_Z = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \sigma_Z^2 = \Lambda$$

Then we can use these results in (A1) to derive the variance of ∇G . As in the T-random field described above, since μ_Z is zero, some terms in (A1) cancel and the variance of ∇G can be obtained as

$$\text{Var}(\nabla G) = (\sigma_X^2 \sigma_Y^2 + \mu_X^2 \sigma_Y^2 + \mu_Y^2 \sigma_X^2 + \mu_X^2 \mu_Y^2) \Lambda$$

Finally, since $F = (n/m)G$, $\nabla F = (n/m)\nabla G$. Therefore $\text{Var}(\nabla F) = (n/m)^2 \text{Var}(\nabla G)$ and this can be solved for Λ as in the T-random field, resulting in (3).

References

- Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B*. 1995; 57:289–300.
- Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
- Desmond JE, Glover GH. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. *J Neurosci Methods*. 2002; 118(2):115–128. [PubMed: 12204303]
- Efron B. Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the American Statistical Association*. 2004; 99:96–104.
- Efron B, Tibshirani R, Storey JD, Tusher VG. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*. 2001; 96:1151–1160.
- Friston KJ, Holmes A, Poline JB, Price CJ, Frith CD. Detecting activations in PET and fMRI: levels of inference and power. *Neuroimage*. 1996; 4(3 Pt 1):223–235. [PubMed: 9345513]
- Friston KJ, Holmes AP, Worsley KJ. How many subjects constitute a study? *Neuroimage*. 1999; 10(1): 1–5. [PubMed: 10385576]
- Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, Evans AC. Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*. 1994; 1:210–220. [PubMed: 24578041]
- Genovese CR, Lazar NA, Nichols T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*. 2002; 15(4):870–878. [PubMed: 11906227]
- Hayasaka S, Nichols TE. Validating cluster size inference: random field and permutation methods. *Neuroimage*. 2003; 20(4):2343–2356. [PubMed: 14683734]
- Hayasaka S, Peiffer AM, Hugenschmidt CE, Laurienti PJ. Power and sample size calculation for neuroimaging studies by non-central random field theory. *Neuroimage*. 2007; 37(3):721–730. [PubMed: 17658273]
- Hayasaka S, Phan KL, Liberzon I, Worsley KJ, Nichols TE. Nonstationary cluster-size inference with random field and permutation methods. *Neuroimage*. 2004; 22(2):676–687. [PubMed: 15193596]
- Hugenschmidt CE, Peiffer AM, Kraft RA, Casanova R, Deibler AR, Burdette JH, et al. Relating imaging indices of white matter integrity and volume in healthy older adults. *Cereb Cortex*. 2008; 18(2):433–442. [PubMed: 17575289]

- Johnson, NL.; Kotz, S.; Balakrishnan, N. Continuous Univariate Distributions. 2nd ed.. Vol. 2. New York: John Wiley & Sons, Inc; 1995.
- Kiebel SJ, Poline JB, Friston KJ, Holmes AP, Worsley KJ. Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model. *Neuroimage*. 1999; 10(6):756–766. [PubMed: 10600421]
- Meyer-Lindenberg A, Nicodemus KK, Egan MF, Callicott JH, Mattay V, Weinberger DR. False positives in imaging genetics. *Neuroimage*. 2008; 40(2):655–661. [PubMed: 18201908]
- Mozolic JL, Hayasaka S, Laurienti PJ. A cognitive training intervention increases resting cerebral blood flow in healthy older adults. *Front Hum Neurosci*. 2010; 4:16. [PubMed: 20300200]
- Mumford JA, Nichols TE. Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage*. 2008; 39(1):261–268. [PubMed: 17919925]
- Murphy K, Garavan H. An empirical investigation into the number of subjects required for an event-related fMRI study. *Neuroimage*. 2004; 22(2):879–885. [PubMed: 15193618]
- Nichols T, Hayasaka S. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat Methods Med Res*. 2003; 12(5):419–446. [PubMed: 14599004]
- Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*. 2004; 23(Suppl 1):S208–S219. [PubMed: 15501092]
- Van Horn JD, Ellmore TM, Esposito G, Berman KF. Mapping voxel-based statistical power on parametric images. *Neuroimage*. 1998; 7(2):97–107. [PubMed: 9558642]
- Worsley KJ. Local Maxima and the Expected Euler Characteristic of Excursion Sets of chi-squared, F and t Fields. *Advances in Applied Probability*. 1994; 26:13–42.
- Worsley KJ. The geometry of random images. *Chance*. 1996; 9:27–40.
- Worsley KJ. An improved theoretical P value for SPMs based on discrete local maxima. *Neuroimage*. 2005; 28(4):1056–1062. [PubMed: 16125975]
- Worsley KJ, Evans AC, Marrett S, Neelin P. A three-dimensional statistical analysis for CBF activation studies in human brain. *J Cereb Blood Flow Metab*. 1992; 12(6):900–918. [PubMed: 1400644]
- Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC. A unified statistical approach for determining significant signals in images of cerebral activations. *Human Brain Mapping*. 1996; 4:58–73. [PubMed: 20408186]
- Zarahn E, Slifstein M. A reference effect approach for power analysis in fMRI. *Neuroimage*. 2001; 14(3):768–779. [PubMed: 11506549]

Information Sharing Statement

PowerMap can be freely downloaded by the general public from the SourceForge website (<https://sourceforge.net/projects/powermap/>).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

An overview of power and sample size calculation by the PowerMap tool.

Figure 2.

Extracting power or sample size information from a power surface. A power image may be generated by extracting a curve (pink plane) from the power surface at the desired df. Similarly, a sample size map may be generated by extracting this curve at the desired level of power (green plane).

Figure 3.

Using the PowerMap tool to generate a power image from a linear regression analysis. The GUI input screens are shown in order (blue box), along with the resulting power surface and power image (green box).

Figure 4. Power images resulting from a linear regression analysis comparing age and fractional anisotropy with a desired sample size of 30 subjects.

Figure 5.

Using PowerMap to generate a sample size image from a one-way analysis of variance (ANOVA). The GUI input screens are shown in order (blue box), along with the resulting power surface and sample size image (green box). Note that the group size input screen is shown for group 3 only. A total of 3 group size input screens are presented to the user, but the screens for groups 1 and 2 are omitted for space.

Figure 6.

Sample size images resulting from an analysis of variance comparing voxel-based morphometry across 3 groups: younger, middle, older. The desired power was specified at 80%. Images were thresholded to show sample sizes in the range of 5 to 10 subjects per group.

Figure 7.

Using PowerMap to generate a power image from a two-sample T statistic image using FDR correction. The GUI input screens are shown in order (blue box), along with the resulting power surface and power image (green box). Note that the group size input screen is shown for group 2 only. A total of 2 group size input screens are presented to the user, but the screen for group 1 is omitted for space.

Figure 8.

Power images resulting from a two-sample T statistic image using FDR correction comparing changes in CBF in an attention training group to a control group. The sample size of the full scale study was specified to be 50 subjects per group.

Figure 9.

Power images (a) and sample size images (b) based on two different types of image smoothness estimates. These images are generated on the same data set as Figure 3. Users can input the known image smoothness calculated from an image analysis software tool, or alternatively, PowerMap can employ the improvised method to calculate the estimated smoothness based on the statistics image as described in Appendix A. In each of the panels (a) and (b), the image on the left is based on manually input smoothness and the image on the right is based the smoothness estimated by PowerMap. Although there are subtle differences in power images (a), the sample size images appear very similar (b). For power images the desired sample size was assumed to be 30, and for sample size images the desired power was assumed to be 80%.

Figure 10.

The effect of image smoothness that is specified by the user and image smoothness that is estimated by PowerMap on power images (a,c) and sample size images (b,d). Each scatter plot depicts the power (a) or sample size (b) using input smoothness versus estimated smoothness for any given brain matter voxel. The images depict the absolute value of the difference between the power (c) and sample size (d) images shown in Figure 9.

Table 1

Outcomes from multiple hypothesis testing. Each comparison is classified according to the underlying truth and the outcome of the hypothesis test as the null (H_0) or alternative (H_A) hypothesis.

		Statistical Test Outcome		
		H_0	H_A	Total
Underlying Truth	H_0	A_0	R_0	M_0
	H_A	A_1	R_1	M_1
Total		A	R	M