ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 39 (2008) 261–268

Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation

Jeanette A. Mumford[a,*] and Thomas E. Nichols[b,c]

[a] *University of California Los Angeles, UCLA Department of Psychology, Box 951563, 1285 Franz Hall, Los Angeles, CA 90095, USA*
[b] *GlaxoSmithKline Clinical Imaging Centre, Imperial College, Hammersmith Hospital, London, UK*
[c] *FMRIB Centre, Oxford University, Oxford, UK*

Received 28 February 2007; revised 2 July 2007; accepted 16 July 2007
Available online 19 August 2007

When planning most scientific studies, one of the first steps is to carry out a power analysis to define a design and sample size that will result in a well-powered study. There are limited resources for calculating power for group fMRI studies due to the complexity of the model. Previous approaches for group fMRI power calculation simplify the study design and/or the variance structure in order to make the calculation possible. These approaches limit the designs that can be studied and may result in inaccurate power calculations. We introduce a flexible power calculation model that makes fewer simplifying assumptions, leading to a more accurate power analysis that can be used on a wide variety of study designs. Our power calculation model can be used to obtain region of interest (ROI) summaries of the mean parameters and variance parameters, which can be use to increase understanding of the data as well as calculate power for a future study. Our example illustrates that minimizing cost to achieve 80% power is not as simple as finding the smallest sample size capable of achieving 80% power, since smaller sample sizes require each subject to be scanned longer.

Introduction

When designing a new scientific study a common practice is to perform a power calculation to evaluate whether the study design maximizes power: the probability of detecting an effect if it is present. Power calculations can prevent an investigator from spending time on an experiment that is underpowered. In other words, even if there truly is an effect there will not be enough power to detect it. Power calculations can also prevent from collecting data that will only cause a slight increase in power. It could be possible to save money, by scanning each subject for a shorter amount of time, while still achieving a high level of power. The goal of this work is to develop methods for carrying out power calculations for group fMRI experiments.

A fMRI power calculation must be flexible, allowing investigators to study a variety of study designs while properly incorporating the variance of the effect they wish to detect. The variance of a two-level group fMRI analysis is composed of the within-subject variance, which is a combination of the first level design, temporal autocorrelation and variance and the between-subject variance introduced at the second level.

Although methods have been developed for finding an efficient first level study designs, which impacts the within-subject variance, the most efficient first level design does not necessarily ensure the group level study will have sufficient power (Wager and Nichols, 2003; Friston et al., 1999; Smith et al., 2007; Josephs and Henson, 1999; Dale, 1999; Liu et al., 2001). Limited research on group fMRI power analysis has been done in the past; Desmond and Glover (2002) present an approach for calculating power for a group fMRI model accounting for both within- and between-subject variability. Simulation-based calculations, which can be time consuming, are one limitation of this method. The study designs are also limited, only allowing for block design study paradigms analyzed by a paired *t*-test at the first level and estimation of a single group mean at the second level. Additionally, the temporal autocorrelation is not incorporated into the within-subject variance estimate, potentially leading to incorrect variances.

We have improved upon the power calculation of Desmond and Glover (2002) by creating a more flexible method. Our results will illustrate the need for a flexible power method by showing that when the model used in a power analysis does not closely match the model to be used to analyze the future data, power can be over- or underestimated. Our power model follows the general two-stage summary statistics model, making it easy to adapt to current popular fMRI software packages using the same model such as Statistical Parametric Mapping (SPM2) and the FMRIB Software Library (FSL) (Mumford and Nichols, 2006). Highlights of our method include non-simulation-based calculations, allowing quick power calculations; incorporation of temporal autocorrelation, improving upon the accuracy of within-subject variance; flexibility of first level design, block or event-related study designs can be studied; and flexibility of second level design beyond a simple one-

* Corresponding author.
E-mail address: mumford@ucla.edu (J.A. Mumford).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2007.07.061

sample *t*-test. Specification of the variance in a power analysis is typically difficult, especially in the case of fMRI analysis as the variance parameters are rarely reported or studied and covariance structures differ between software packages and can often consist of upwards of 10 parameters per voxel. Typically, when a power analysis is carried out it is helpful to compare multiple similar studies to gain intuition about the variance that should be used in the power analysis. We use a simple 3-parameter covariance summary of the covariance supplied by the fMRI software which is easy to incorporate into a power calculation, helps build intuition about the covariance associated with fMRI studies and allows for the comparison of studies regardless of what software was used.

Theory

The model

In the two-stage summary statistics model, the purpose of the first level is to study each subject independently and the second level combines first level results from all subjects to obtain group results. Let each subject's first level model be given by

$$Y_k = X_k\beta_k + \epsilon_k, \tag{1}$$

where $k=1, \ldots, N$ is a subject indicator, Y_k is the $T_k \times 1$ vector of fMRI response data, $\mathbf{X}_k$ is the $T_k \times p$ design matrix, β_k is a vector of p parameters and the error vector of length T_k is Gaussian distributed with variance σ_k^2 and correlation $\mathbf{V}_k$, $\epsilon_k \sim N(0, \sigma_k^2 \mathbf{V}_k)$. Note that while each subject can have differing number of scans (T_k), all of the design matrices, $\mathbf{X}_k$, must have the same number of columns, each column expressing the same effect in each subject's data. The first level estimate for each subject, k, using generalized least squares (GLS) with known $\mathbf{V}_k$ is given by

$$\hat{\beta}_k = (X_k' V_k^{-1} X_k)^{-1} X_k' V_k^{-1} Y_k$$
$$\mathrm{Var}(\hat{\beta}_k) = \sigma_k^2 (X_k' V_k^{-1} X_k)^{-1}.$$

Consider a single contrast of these parameter estimates from each subject $c\hat{\beta}_k$, where $\boldsymbol{c}$ is a vector and let the vector of contrast estimates be denoted $\hat{\beta}_{\mathrm{cont}} = [c\hat{\beta}_1, \ldots, c\hat{\beta}_N]'$. Using the subscript g to denote group level parameters, the group level model is given by

$$\hat{\beta}_{\mathrm{cont}} = X_g \beta_g + \varepsilon_g, \tag{2}$$

where $\mathbf{X}_g$ is the $N \times p_g$ design matrix and $\epsilon_g \sim N(0, \mathbf{V}_g)$, with $\mathbf{V}_g = \mathrm{diag}(\sigma_k^2 c(\mathbf{X}_k' \mathbf{V}_k^{-1} \mathbf{X}_k)^{-1} \boldsymbol{c}') + \sigma_g^2 \mathbf{I}_N$, where $\mathrm{diag}(a_i)$ is a diagonal matrix with the elements $a_1, \ldots, a_N$ along the main diagonal, σ_g^2 is the between-subject variance and $\mathbf{I}_N$ is an $N \times N$ identity matrix.

At the second level, we can consider a contrast vector, $\boldsymbol{c}_g$. The estimate of a contrast $\boldsymbol{c}_g\beta_g$ and its variance are given by

$$c_g \hat{\beta}_g = c_g (X_g' V_g^{-1} X_g)^{-1} X_g' V_g^{-1} \hat{\beta}_{\mathrm{cont}}$$
$$\mathrm{Var}(c_g \hat{\beta}_g) = c_g (X_g' V_g^{-1} X_g)^{-1} c_g'$$

Under the null hypothesis of no activation, H_0: $\boldsymbol{c}_g \hat{\beta}_g = 0$, the test statistic

$$T = \frac{c_g \hat{\beta}_g}{\sqrt{c_g (X_g' V_g^{-1} X_g)^{-1} c_g'}}$$

follows a T distribution with $N - p_g$ degrees of freedom (t_{n-p_g}). Under the alternative hypothesis of an activation of size Δ, H_A: $\boldsymbol{c}_g \hat{\beta}_g = \Delta$, T follows a noncentral T distribution, given by $T_{n-p_g,\mathrm{ncp}}$, where $n - p_g$ are the degrees of freedom, and the noncentrality parameter is given by

$$\mathrm{ncp} = \frac{\Delta}{\sqrt{c_g (X_g' V_g^{-1} X_g)^{-1} c_g'}} \tag{3}$$

Power for a test of size α is

$$P(T_{n-p_g,\mathrm{ncp}} > t_{1-\alpha,n-p_g}), \tag{4}$$

where $t_{1-\alpha,n-p_g}$ is defined as the value that satisfies $P(T_{n-p_g} > t_{1-\alpha,n-p_g}) = \alpha$. To calculate power for an F-test, allow $\boldsymbol{c}_g$ to be a contrast matrix, with each row corresponding to a single contrast and use a noncentral F distribution with $\mathrm{ncp} = \Delta' (\boldsymbol{c}_g (\mathbf{X}_g' \mathbf{V}_g^{-1} \mathbf{X}_g)^{-1} \boldsymbol{c}_g')^{-1} \Delta$, $r = \mathrm{rank}(\boldsymbol{c}_g)$ numerator and $n - p_g$ denominator degrees of freedom.

The power calculation requires specification of the parameters listed in Table 1. Although there are 10 pieces of information that need to be specified, all but the last four Δ, σ_k^2, $\mathbf{V}_k$ and σ_g^2, are known and must be estimated from previous analyses. We take a region of interest (ROI) approach where Δ, σ_k^2, $\mathbf{V}_k$ and σ_g^2 are averaged over all voxels within the ROI, producing a power calculation that applies to the average voxel in that ROI.

Estimation

Traditionally, the parameters in Table 1 are obtained from previous studies and we take the same approach here. The design matrices, contrasts, sample size, false-positive rate and effect size are assumed to be known by the investigator or readily available from a previous analysis. Since BOLD fMRI data do not have meaningful units, the design matrices and contrasts should be constructed so parameter and contrast estimates reflect % change from baseline fMRI signal. For block design studies, the regressors must be scaled so the baseline to activation distance is 1 and for an event-related design, the baseline to peak distance of an isolated event should be scaled to one. As long as the sum of the negative elements of the contrasts is −1 and the sum of the positive elements is 1, then all parameter and contrast values will represent % change from baseline.

The primary focus of this section will be on the estimation of parameters associated with the variance: σ_k^2, $\mathbf{V}_k$ and σ_g^2. We first focus on general estimation and then specific estimation based on prior analyses done with SPM2 or FSL are discussed in the Results section. All parameters involved will first be estimated in a voxelwise fashion and then averaged within ROI for use in the power calculation.

Table 1
Parameters necessary for carrying out a power calculation

Parameter	Parameter description
N	Number of subjects
α	False-positive rate
$\boldsymbol{c}$	First level contrast
$\mathbf{X}_k$	First level design matrix for subject k
$\boldsymbol{c}_g$	Group level contrast
$\mathbf{X}_g$	Group level design matrix
Δ	Size of the effect
σ_k^2	Within-subject variance for subject k
$\mathbf{V}_k$	Temporal autocorrelation matrix for subject k
σ_g^2	Between-subject variability

General: within-subject variance estimation

In order to capture the complicated structure of temporal autocorrelation of fMRI data, software models often calculate covariance estimates using many parameters. These covariance estimates, which typically differ over subjects and voxels, are appropriate when running a data analysis, but when carrying out a power analysis it is necessary to simplify the covariance structure. A parsimonious covariance model allows for specification of the covariance for a future study that will likely generalize over subjects, be applicable to study designs and sample sizes that differ from previous studies and is easy to communicate in print. The first step in developing a parsimonious model is to assume the covariance is the same across subjects ($\sigma_k^2\ \mathbf{V}_k=\sigma_w^2\ \mathbf{V}$, $k=1, \ldots, N$). Second, we choose a low parameter covariance structure to summarize the within-subject covariance, which not only simplifies the structure but allows comparison of variance across analyses that used different covariance structures. Although one would typically not assume $\mathbf{V}$ was the same across subjects, it is a standard assumption made when carrying out power analysis in order to calculate power for different sample sizes. A common assumption is that fMRI noise follows an AR(1)+WN structure (Zarahn et al., 1997; Purdon and Weisskoff, 1998; Marchini and Smith, 2003; Woolrich et al., 2001; Burock and Dale, 2000), so we recommend using this 3-parameter covariance model, which includes AR(1) correlation, ρ, the AR variance, σ^2_{AR}, and white noise variance, σ^2_{WN}. The structure and estimation of the AR(1)+WN covariance are discussed in Appendix A. Typically model residuals are used to estimate covariance, but since these data are often discarded by the software during model estimation, our procedure uses the covariance estimates, which are commonly included in the analysis results. Although this summary of the covariance is useful for the purposes of carrying out a power analysis, we recommend using the default covariance estimation procedure within the software package when carrying out fMRI analysis.

Special care must be taken if a high-pass filter was applied to reduce low frequency noise. If the covariance was estimated after filtering, the AR(1)+WN model will not fit well. Therefore, when a high-pass filter was used, extra steps may be necessary before fitting the AR(1)+WN model and the high-pass filter also must be incorporated into the power calculation.

General: between-subject variance estimation

Last, we need to obtain the between-subject variance, σ_g^2. This will require one group fMRI analysis and the procedure for estimating σ_g^2 depends on how the software estimated the group model variance. Some software packages may estimate a separate between-subject variance and so this can be averaged over voxels in the ROI to get a single estimate of σ_g^2.

Other software packages, for example SPM2, may assume that all within-subject variances are equal. In this case, the group model does not consider separate between- and within-subject variances, so

$$V_k = \sigma_{g*}^2 I_N \tag{5}$$

instead of

$$V_g = \mathrm{diag}(\sigma_k^2 c(X_k' V_k^{-1} X_k)^{-1} c') + \sigma_g^2 I_N. \tag{6}$$

To have the flexibility to calculate power for different design matrices, $\mathbf{X}_k$, separate between- and within-subject variances are needed. The goal is to equate Eqs. (5) and (6) and solve for σ_g^2. Since the underlying assumption of models that do not estimate a separate between-subject variance is that within-subject variance is homogeneous across subjects, it is reasonable to assume,

$$\sigma_k^2 c(X_k' V_k^{-1} X_k)^{-1} c' = 1/N \sum_{k=1}^{N} \hat{\sigma}_k^2 c(X_k' \hat{V}_k^{-1} X_k)^{-1} c' = \hat{\sigma}^2_{\mathrm{avg}},$$

where $\hat{\sigma}_k^2$ and $\hat{V}_k$ are the original software estimates. Since the within-subject variance was not part of the calculation of σ_{g*}^2, it is possible that the estimate of σ_{g*}^2 can be as large or larger than the contribution that would come from the within-subject covariance in Eq. (6). Therefore, simply setting $\hat{\sigma}_g^2 = \hat{\sigma}_{g*}^2 - \hat{\sigma}^2_{\mathrm{avg}}$ can lead to negative variances so we use

$$\hat{\sigma}_g^2 = \begin{cases} \hat{\sigma}_{g*}^2 - \hat{\sigma}^2_{\mathrm{avg}} & \text{if } \hat{\sigma}_{g*}^2 - \hat{\sigma}^2_{\mathrm{avg}} \geq 0, \\ 0 & \text{if } \hat{\sigma}_{g*}^2 - \hat{\sigma}^2_{\mathrm{avg}} < 0. \end{cases} \tag{7}$$

Constraining the between-subject variance to be positive prevents the awkward situation of reporting negative variance in a power analysis but may lead to a conservative power estimate.

This value is calculated within the voxels of the ROI's and then are averaged to get a representative ROI estimate of σ_g^2.

Data

We used the FIAC single subject block design data for subjects 0, 1, 2, 3, 4 and 6 (Dehaene-Lambertz et al., 2006). The stimulus consisted of French speakers reading a story "The Three Little Pigs". The data were collected on a 3-T whole body scanner. There are 30 slices of data where each slice has a thickness of 4 mm. There were 195 volumes collected, one volume of data was collected every 2.5 s (TR=2.5). For the block design study, there were four types of blocks: (1) same sentence-same speaker (SSt-SSp): a given sentence said by the same speaker was repeated six times; (2) same sentence-different speakers (SSt-DSp): a given sentence was repeated by six different speakers (3 males and 3 females); (3) different sentences-same speaker (DSt-SSp): a given speaker produced six different sentences; (4) different sentences-different speakers (DSt-DSp): six different speakers (3 males and 3 females) produces six different sentences. In all data analyses, all four blocks were modeled and the contrast of interest was SSt-SSp.

Results

Estimation within the FSL framework

To obtain the within-subject covariance and between-subject variance estimates, a group FSL analysis as well as the individual first level subject analyses are required.

FSL: within-subject variance estimation

In the first level FSL analysis, a voxelwise unstructured correlation is estimated from the OLS residuals, regularized by a Tukey taper, as well as a voxelwise variance (Woolrich et al., 2001). Since the residuals are discarded during model estimation, we propose the methods described in Appendix A to summarize the multi-parameter FSL covariance with the 3-parameter AR(1)+WN model.

High-pass filtering in FSL removes the fit of a Gaussian weighted running line smoother by premultiplying the data and design by a filtering matrix, reducing low frequency drift. Filtering is done prior

to covariance estimation hence the AR(1)+WN model will not be a good fit. Therefore, a non-high-pass filtered analysis must be carried out to supply a covariance estimate that the AR(1)+WN model will fit. Then the high-pass filtering matrix, which is a function of time series length and filter cutoff, is applied to the fitted AR(1)+WN estimate by pre- and post-multiplying the covariance estimate the by the filter matrix and its transpose, respectively. This filtered covariance is used in the power calculation.

To evaluate our AR(1)+WN covariance estimate, we compared *T*-statistics derived using the FSL covariance to *T*-statistics derived using the high-pass filtered AR(1)+WN covariance estimate in all voxels. Fig. 1 displays the results for a single subject and since the points on the plot are dense, the red line indicates the overall trend of the data from a fitted loess model (results were similar for other subjects). The fitted model runs close to the diagonal indicating the *T*-statistics are similar, although there is some bias for negative test statistic values that would result in conservative power estimates.

Using an ROI defined by voxels with a test statistic value larger than 2 for the SSt-SSp contrast of the block design FIAC study, average AR(1)+WN parameter estimates based on an FSL analysis for 6 subjects are given in columns 3–5 of Table 2. These values should be viewed when carrying out a power analysis to build intuition on the covariance structure of fMRI data.

FSL: between-subject variance estimation

Voxelwise estimates of the between-subject variance are created and stored as an image during group FSL analyses, so simply averaging this value over the ROI gives the value of σ_g^2 needed for the power calculation. For these data analysis, $\sigma_g^2=0.433$.

SPM2

As with FSL, a group analysis along with the corresponding single subject analyses are required to estimate the variance parameters necessary for calculating power.

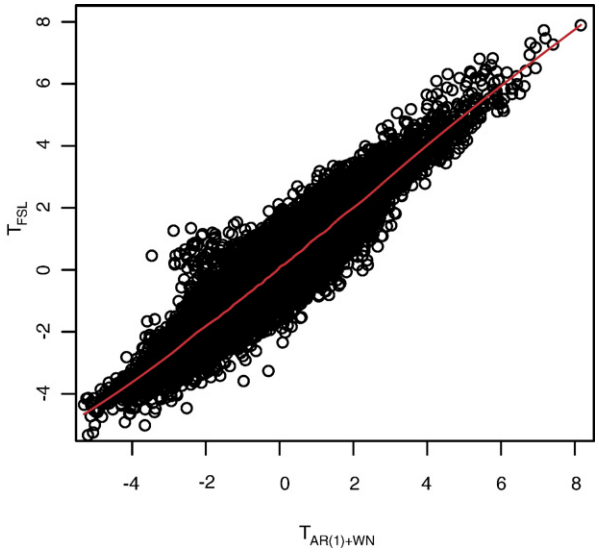

Fig. 1. Comparing *T*-statistics derived using the FSL unstructured covariance from an analysis using a high-pass filter (T_{FSL}) and the high-pass filtered AR(1)+WN covariance derived from the ASL covariance ($T_{\mathrm{AR(1)+WN}}$) across all voxels for a single subject. The red line shows the trend of the data using a loess fit.

Table 2
Estimated AR(1)+WN parameters from FSL and SPM2 for block design study, averaged over ROI defined by voxels with FSL group analysis *T*-statistics larger than 2

Subject	Study design	ρ_{FSL}	$\sigma^2_{\mathrm{AR_{tot}-FSL}}$	$\sigma^2_{\mathrm{WN_{FSL}}}$	ρ_{SPM2}	$\sigma^2_{\mathrm{AR_{tot}-SPM2}}$	$\sigma^2_{\mathrm{WN_{SPM2}}}$
1	Block	0.646	0.618	1.362	0.482	0.535	1.591
2	Block	0.700	0.655	0.949	0.536	0.852	1.086
3	Block	0.758	0.764	1.077	0.529	0.850	1.254
4	Block	0.740	0.535	0.910	0.515	0.535	0.309
5	Block	0.744	0.927	1.027	0.545	1.28	1.22
6	Block	0.809	2.38	2.551	0.550	3.14	2.461

Parameters are expressed in % change from baseline.

SPM: within-subject variance estimation

SPM2 uses a global estimate of the temporal autocorrelation that is based on a two-term Taylor series expansion of an AR(1) correlation about $\rho=0.2$ (Friston et al., 2002a,b). Specifically, the correlation is derived from $\mathrm{Cov}_{\mathrm{SPM2}}=\lambda_1 C1+\lambda_2 C2$, where $C1_{i,j}=0.2^{|i-j|}$ and $C2_{i,j}=|i-j|(0.2)^{|i-j|-1}$ and single global estimates of λ_1 and λ_2 are estimated using ReML. The variance, σ_k^2, is estimated at each voxel, and so $\sigma_k^2\widehat{\mathrm{Cov}}_{\mathrm{SPM2}}/\hat{\lambda}_1$ is the voxelwise covariance used to estimate the AR(1)+WN parameters as described in Appendix A.

In SPM2, the high-pass filter is built into the design matrix as a set of discrete cosine transform functions. It does not interfere with the AR(1)+WN covariance estimation, although it should be included in the first level design matrix in power calculations and contrasts for hypothesis testing should be adjusted accordingly. Averages of the AR(1)+WN parameters based on the SPM covariance are displayed in columns 6–8 of Table 2. Note that since SPM is using a different covariance model than FSL, the AR (1) parameters will tend to differ. Although there are slight differences in the AR(1)+WN parameter estimates, there is little difference in calculated power between the two sets of estimates (see Supplemental materials).

SPM: between-subject variance estimation

The group model in SPM2 is estimated under the assumption of equal first level within-subject covariances. Therefore, separate within- and between-subject variances are not estimated during a group analysis, instead a single estimate, $\hat{\sigma}^2_{g*}$, is calculated. Therefore, we use the methods described in the General: within-subject variance estimation section. Using the same ROI described previously, the between-subject estimate based on the SPM analysis was found to be $\sigma_g^2=0.409$.

Calculating power

We first focus on consequences when the power model does not match the model used on the future data. Ignoring temporal autocorrelation when estimating power, but using a model that estimates temporal autocorrelation on the newly collected data, is an example of the models not matching. Does this impact power? Likewise, does the omission of hemodynamic response function (HRF) convolution with our first level regressors impact power?

The regressor used had 15 s activation blocks followed by 15 s rest blocks and we compared power for the 4 combinations of dependent/independent noise and with/without HRF convolution. The FSL high-pass filter was used whenever the covariance was modeled. We used the average subject parameter values from the FSL analysis in Table 2 $\rho=0.73$, $\sigma^2_{\mathrm{AR_{tot}}}=0.980$ and $\sigma^2_{\mathrm{WN}}=1.313$, as

well as the group effect size of $\delta=0.69\%$ and $\sigma_g^2=0.433$. For all power calculations, 20 subjects and a Type I error of $\alpha=0.005$ were used.

In Fig. 2, ρ, σ^2_{AR} and σ^2_{WN} are varied in the top, middle and lower panels, respectively, while in each panel the other 2 variance parameters are fixed at their mean values. In all cases the difference in power is largest between the correct model and the model with both incorrect noise and design, with a maximum difference of 14%. In the top panel, the difference in power between the model with the wrong noise and the correct model and can be as large as a 9%. Specifically, for the FIAC data, where $\rho=0.73$, a simplified power calculation omitting temporal autocorrelation overestimates a power of nearly 80% when the correct model yields a smaller power of 72%. Since the temporal autocorrelation can have such an impact on the power estimate, it is important to include it.

In the case where the temporal autocorrelation is modeled correctly, but the design is incorrect, the overestimation of power is as large as 8%. So it is also important not to simplify the power model by using a boxcar function that has not been convolved with an HRF as it will tend to overestimate power.

The main application of power calculations is for planning a study design. Power analyses can be used to prevent from spending time and money on an experiment that is under powered as well as preventing the collection of additional data that will have little impact on power. Using the mean AR(1)+WN parameters, we calculated power curves for different sample sizes for the block design described previously. The model included the boxcar regressor convolved with an HRF and an intercept. The bottom x-axis of Fig. 3 reflects how many on/off cycles of 30 s have occurred and the top axis shows the corresponding time in minutes. The power curves are restricted by a budget of \$7600 where each subject has a base cost of \$300, for subject preparation, and additional scanner time is \$10/minute.

Limitations from the budget and values of the mean and variance for this study imply only sample size between 18 and 22 subjects achieve at least 80% power. Focusing on the maximum power when the entire budget of \$7600 is used, the most powerful analysis would be with 21 subjects yielding 83% power.

Another feature shown with the power curves, is the point of diminishing returns where additional scanner time has very little impact on the power. For this study in particular, power increases less than 1% for each additional on/off cycle after approximately 14 cycles. For a sample size of 17 subjects, for example, higher levels of power are practically impossible to reach even if the subjects are scanned for very long periods of time. The within-subject variance decreases as the number of cycles increase and when the within-subject variance is

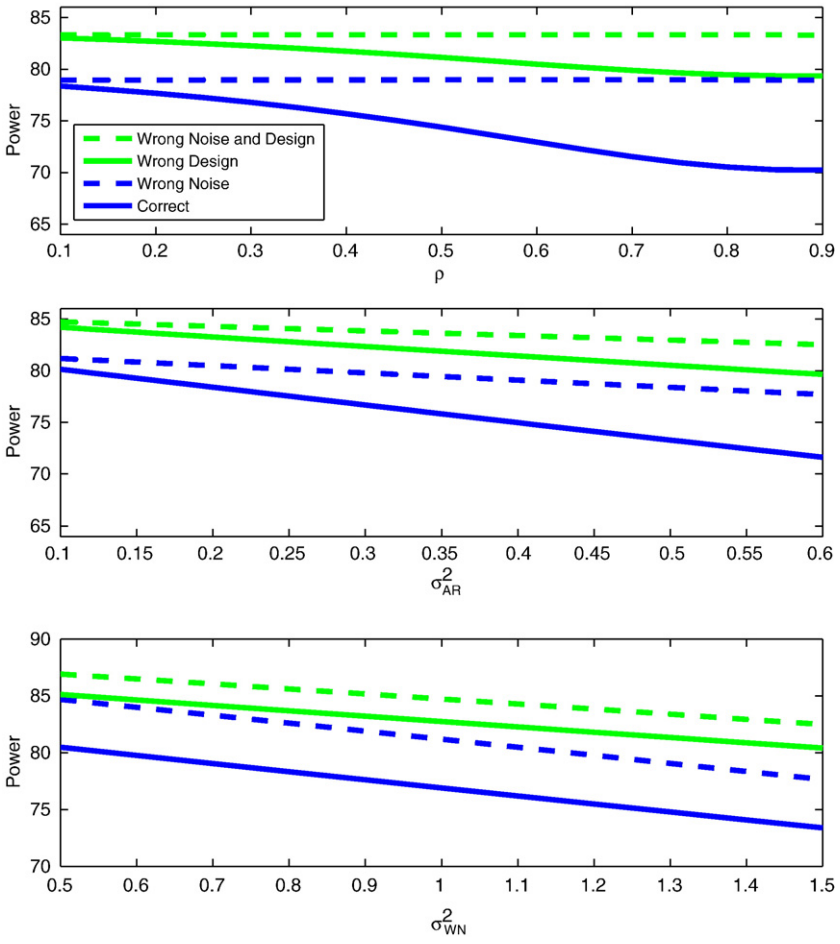

Fig. 2. An illustration of how power differs when the model is incorrectly specified either by not convolving a boxcar regressor with an HRF or assuming the noise is independent. This comparison is made over a range of ρ, σ^2_{AR} and σ^2_{WN} in the top, middle and bottom panels, respectively.

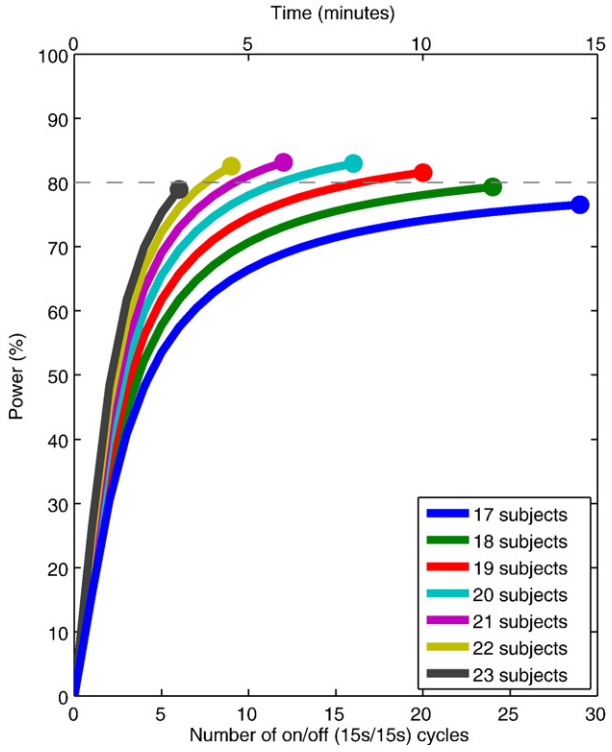

Fig. 3. Power estimates for a block design study where the total cost is limited to $7600. Each curve is for a different sample size and the grey dotted line indicates 80% power.

sufficiently smaller than the between-subject variance, the between-subject variance drives the power calculation. For this study in particular, the point of diminishing returns occurs when the within-subject variance is approximately 1/4 of the between-subject variance. This ratio varies with sample size as well as the size of δ, so the result does not generalize to other study designs or regions of the brain.

If we expect our future data to behave similarly to the data used in the power analysis, a sample size of at least 18 subjects should probably be used, even if the budget is increased to allow for more scanner time. Also, knowing the point of diminishing returns can prevent unnecessary spending. With 19 subjects, scanning each subject for 8 versus 10 min will not have much of an impact on power, but less time in the scanner saves money and prevents the subject from getting bored.

Another feature is different sample size/scanner time combinations result in the same amount of power. The left panel of Fig. 4 shows the number of cycles required to achieve 80% power for different sample sizes and the right panel shows the corresponding costs. Interestingly, in this case, the cost does not decrease linearly as the number of subjects increases; instead there is a minimum cost at 20 subjects. The cost of the study with the fewest subjects, 18, is the largest since it requires the subject to be scanned longer.

Discussion

Power calculations are useful tools for designing studies, especially in the case of fMRI where the cost per subject is high. We have introduced a new method of calculating power for group fMRI studies, overcoming many weaknesses of previous power calculation approaches by incorporating temporal autocorrelation, allowing for either block or event-related study designs, allowing multiple regressors in the first level model, having a flexible group model and using a general method can easily adapt to the models of a variety of fMRI software packages. Additionally, our model can easily be extended to calculate power for three level models, for example, where multiple runs per subject and multiple subjects are studied.

Specification of the variance is the most difficult task in power calculations, but especially in the case of group fMRI where there

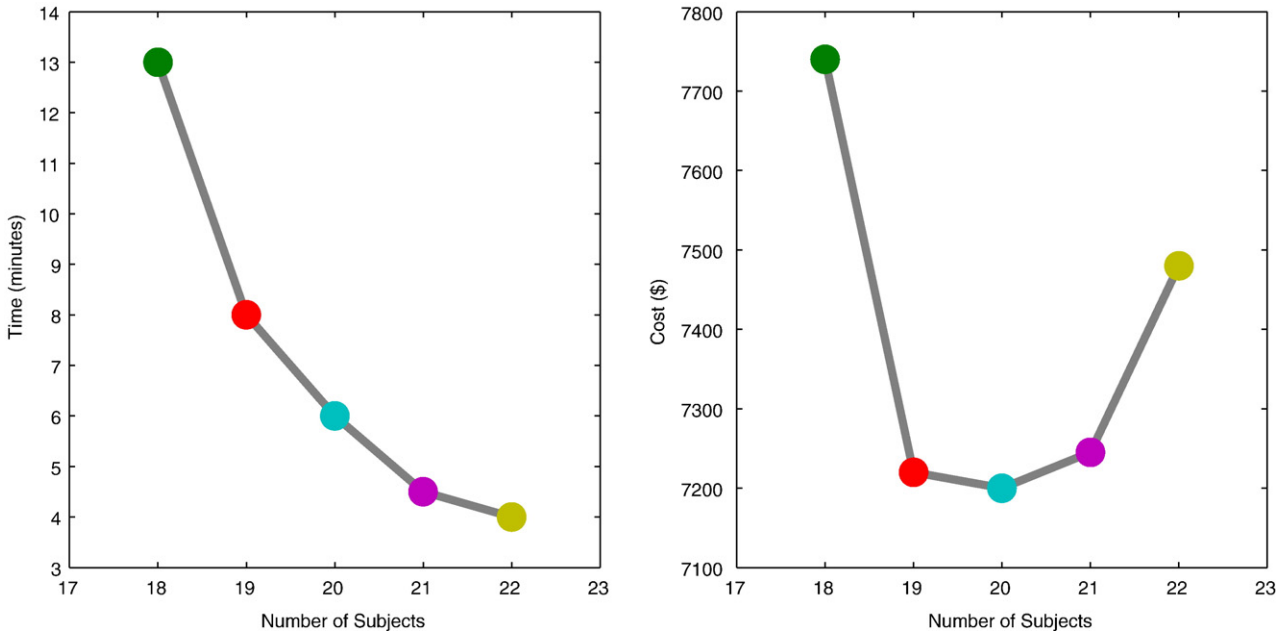

Fig. 4. Number of cycles and cost to achieve 80% power. The left panel shows how many cycles per subject are required and the right panel shows the total cost when there is a base cost of $300, for subject preparation, and additional scanner time is $10/minute.

is little intuition about variance since it is rarely reported. We break the task down into calculating the within-subject and between-subject variance separately. Although previous power methods have simplified the within-subject covariance by assuming the data are not temporally correlated, we illustrate how this assumption can lead to an overestimation of power. We propose estimating the covariance from a previous data analysis using a 3-parameter AR(1)+WN model. This model achieves two things: reduces the number of parameters to a manageable number that can easily be reported and compared and allows comparison of covariances from different software. It is common to assume the covariance is the same across subjects in a power analysis, in order to calculate power for a variety of sample sizes. Although this is not an appropriate assumption for an actual data analysis. Typically subjects do not exhibit the same amount of correlation and variance and Beckmann et al. (2003) show that under the assumption that variance are the same across subjects, test statistics tend to be conservative.

Power curves, such as the ones we constructed, supply useful information for study planning. For example, for each sample size, at some point additional scans cause very little improvement in power. This occurs when the between-subject variance is sufficiently smaller than the within-subject variance. The ratio of between- to within-subject variance at the point of diminishing returns varies with sample size and effect size, δ. Knowing the point of diminishing returns helps budget a study more efficiently. It is also possible that different scan/subject combinations result in the same amount of power. Depending on how scanner fees are charged, to reach a reasonable power level, it may be less expensive to scan more subjects for less time, as our results illustrated.

The experimental design of the data used for estimating power parameters should be similar to the design of the experiment for which power predictions are being made. Specifically, event-related designs should be used to predict power for event-related studies and block designs should be used to predict power for block design studies. This is due to nonlinearities in the BOLD response which are not accurately modeled by the usual HRF convolution of the stimulus (Vazquez and Noll, 1998), though see Wager et al. (2005).

Likewise, even if the study designs are the same, the approach to modeling the effect for the new study should be similar to the approach taken in the initial study used to estimate the parameters of the power calculation. As shown in Mechelli et al. (2003), for a block design study, a more sensitive data analysis was achieved by using an event-related analysis compared to a block design analysis of the data. If one were to use a simple block regressor convolved with an HRF for the power analysis and an event-related model for the new data analysis, power may be over- or underestimated. Similarly, if the model is not correct, in that it does not properly describe the BOLD response, the power estimate will not be correct since it assumes the correct model is being used.

Conclusion

We have presented a general method for estimating power for group level fMRI studies and have shown how this general method can be adapted to be used with models from different software packages. The flexibility of our method should make it appealing for investigators to use.

Acknowledgments

This work is supported by NIH grants R01 DA15410 and R01 EB004346-01A1.

Appendix A. Parameterization of the AR(1)+WN covariance

Under the assumption that the error, $\boldsymbol{\epsilon}=(\epsilon_1, \ldots, \epsilon_T)$, follows an AR(1)+WN covariance, the specific structure is given by

$$\mathrm{Cov}(\epsilon_i,\epsilon_j) = \begin{cases} (\sigma^2_{\mathrm{AR}}/(1-\rho^2))\rho^{|i-j|} & \text{if } i \neq j, \\ (\sigma^2_{\mathrm{AR}}/(1-\rho^2)) + \sigma^2_{\mathrm{WN}} & \text{if } i = j. \end{cases} \quad (8)$$

Note the total contribution to the variance from the AR(1) model can be denoted as $\sigma^2_{\mathrm{AR}_{\mathrm{tot}}}=\sigma^2_{\mathrm{AR}}/(1-\rho^2)$, so the model can also be expressed as

$$\mathrm{Cov}(\epsilon_i,\epsilon_j) = \begin{cases} \sigma^2_{\mathrm{AR}_{\mathrm{tot}}}\rho^{|i-j|} & \text{if } i \neq j, \\ \sigma^2_{\mathrm{AR}_{\mathrm{tot}}} + \sigma^2_{\mathrm{WN}} & \text{if } i = j. \end{cases} \quad (9)$$

Taking a Fourier transform of the covariance, the corresponding power spectrum is given by

$$F(\omega) = \sigma^2_{\mathrm{AR}}/(1 - 2\rho\cos(\omega) + \rho^2) + \sigma^2_{\mathrm{WN}}.$$

The estimation of the AR(1)+WN parameters will vary depending on what information the software saves when carrying out an analysis. If residual values are available, since the AR(1)+WN covariance is a special case of an autoregressive moving average with 1 autoregressive parameter and 1 moving average parameter (AR*MA*(1, 1)), any method such as nonlinear least squares or maximum likelihood-based approaches can be used to estimate the AR(1)+WN parameters. Most statistical software packages, such as R, Splus and Matlab, have functions that will carry out this estimation.

To save space, typically the residuals are discarded during an fMRI data analysis although the covariance estimates are saved. The covariance estimates can be used to obtain the AR(1)+WN parameters, as described below using a two step process. The first step estimates the WN parameter and the second step estimates the AR(1) parameters.

A property of the AR(1)+WN power spectrum is the height of the spectrum at high frequencies is a close approximation of σ^2_{WN}. Therefore, the white noise variance will be estimated by averaging the height, at high frequencies, of the power spectrum. The Wiener–Khinchin relation states that the Fourier transform of the autocovariance is the periodogram or estimated power spectrum for that time series. This is given 5 by

$$I(\omega) = \sum_{i-j=(T-1)}^{T-1} C_{i,j} e^{\omega(i-j)\sqrt{-1}}, \quad (10)$$

where $I(\omega)$ is the power at frequency ω, T is the number of time points, and $C_{i,j}$ is the (i, j)th element of the covariance matrix which would be the covariance estimate obtained from an FSL or SPM analysis. Note that $\omega=2\pi M/T$, for $M=1, \ldots m$ where $m=T/2$ for even values of T and $m=(T-1)/2$ for odd T. Then the estimate, $\hat{\sigma}^2_{\mathrm{WN}}$ is given by the average power of the top 1/3 frequencies of the periodogram. Smaller cutoffs produced similar covariance estimates, but are less efficient.

Once the estimate $\hat{\sigma}^2_{\mathrm{WN}}$ is known, we can simply remove it from the autocovariance estimate using subtraction, $C_{\mathrm{AR}_{i,i}}=C_{i,i}-\hat{\sigma}^2_{\mathrm{WN}}$ and $C_{\mathrm{AR}_{i\neq j}}=C_{i\neq j}$. The resulting covariance then follows an AR(1) structure and the estimation of the AR(1) correlation parameter, ρ, can be found using the Yule–Walker equations

$$\begin{pmatrix} \hat{\phi}_0 \\ \hat{\phi}_1 \\ \vdots \\ \hat{\phi}_{T-1} \end{pmatrix} (\rho) = \begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \vdots \\ \hat{\phi}_T \end{pmatrix} \tag{11}$$

where $\hat{\phi}_{i-j}$ are the correlation estimates from C_{AR}, $\hat{\phi}_{i-j}=C_{\mathrm{AR}_{i,j}}/C_{\mathrm{AR}_{0,0}}$. Now that we have an estimate, $\hat{\rho}$ from the Yule–Walker equations, since the AR(1) variance is given by $C_{\mathrm{AR}_{0,0}}=\sigma^2_{\mathrm{AR}}/(1-\rho^2)$, we can obtain the estimate by $\hat{\sigma}^2_{\mathrm{AR}}=C_{\mathrm{AR}_{0,0}}(1-\hat{\rho}^2)$.

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2007.07.061.

References

Beckmann, C.F., Jenkinson, M., Smith, S.M., 2003. General multilevel linear modeling for group analysis in FMRI. NeuroImage 20 (2), 1052–1063 (Oct.).

Burock, M.A., Dale, A.M., 2000. Estimation and detection of event-related fMRI signals with temporally correlated noise: a statistically efficient and unbiased approach. Hum. Brain Mapp. 11 (4), 249–260 (Dec.).

Dale, A.M., 1999. Optimal experimental design for event-related fMRI. Hum. Brain Mapp. 8, 109–114.

Dehaene-Lambertz, G., Dehaene, S., Anton, J.-L., Campagne, A., Ciuciu, P., Dehaene, G.P., Denghien, I., Jobert, A., Lebihan, D., Sigman, M., Pallier, C., Poline, J.-B., 2006. Functional segregation of cortical language areas by sentence repetition. Hum. Brain Mapp. 27 (5), 360–371 (May).

Desmond, J., Glover, G., 2002. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. J. Neurosci. Methods 118, 115–128.

Friston, K.J., Zarahn, E., Josephs, O., Henson, R.N., Dale, A.M., 1999. Stochastic designs in event-related fMRI. NeuroImage 10 (5), 607–619 (Nov.).

Friston, K.J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002a. Classical and Bayesian inference in neuroimaging: theory. NeuroImage 16 (2), 465–483 (Jun).

Friston, K.J., Glaser, D.E., Henson, R.N.A., Kiebel, S., Phillips, C., Ashburner, J., 2002b. Classical and Bayesian inference in neuroimaging: applications. NeuroImage 16 (2), 484–512 (Jun).

Josephs, O., Henson, R.N., 1999. Event-related functional magnetic resonance imaging: modelling inference and optimization. Philos. Trans. R. Soc. Lond., B Biol. Sci. 354, 1215–1228.

Liu, T.T., Frank, L.R., Wong, E.C., Buxton, R.B., 2001. Detection power, estimation efficiency, and predictability in event-related fMRI. NeuroImage 13 (4), 759–773 (Apr).

Marchini, J.L., Smith, S.M., 2003. On bias in the estimation of autocorrelations for fMRI voxel time-series analysis. NeuroImage 18 (1), 83–90 (Jan).

Mechelli, A., Henson, R.N.A., Price, C.J., Friston, K.J., 2003. Comparing event-related and epoch analysis in blocked design fMRI. NeuroImage 18 (3), 806–810 (Mar).

Mumford, J., Nichols, T., 2006. Modeling and inference of multisubject fMRI data: using mixed-effects analysis for joint analysis. IEEE Eng. Med. Biol. Mag. 25 (2), 42–51.

Purdon, P.L., Weisskoff, R.M., 1998. Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI. Hum. Brain Mapp. 6 (4), 239–249.

Smith, S., Jenkinson, M., Beckmann, C., Miller, K., Woolrich, M., 2007. Meaningful design and contrast estimability in FMRI. NeuroImage 34 (1), 127–136 (Jan).

Vazquez, A., Noll, D., 1998. Nonlinear aspects of the BOLD response in functional MRI. NeuroImage 7 (2), 108–118.

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

Wager, T., Vasquez, A., Hernandez, L., Noll, D., 2005. Accounting for nonlinear BOLD effects in fMRI: parameter estimates and a model for prediction in rapid event-related studies. NeuroImage 25 (1), 206–218.

Woolrich, M.W., Ripley, B.D., Brady, M., Smith, S.M., 2001. Temporal autocorrelation in univariate linear modeling of FMRI data. NeuroImage 14 (6), 1370–1386 (Dec).

Zarahn, E., Aguirre, G.K., D'Esposito, M., 1997. Empirical analyses of BOLD fMRI statistics. I. Spatially unsmoothed data collected under null-hypothesis conditions. NeuroImage 5 (3), 179–197 (Apr).