ELSEVIER

Journal of Neuroscience Methods 118 (2002) 115–128

JOURNAL OF NEUROSCIENCE METHODS

www.elsevier.com/locate/jneumeth

Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses

John E. Desmond [*], Gary H. Glover

Department of Radiology, Lucas MRS Center, MC: 5488, Stanford University, Stanford, CA 94305-5488, USA

Received 11 April 2001; received in revised form 20 December 2001; accepted 3 January 2002

Abstract

Estimation of statistical power in functional MRI (fMRI) requires knowledge of the expected percent signal change between two conditions as well as estimates of the variability in percent signal change. Variability can be divided into intra-subject variability, reflecting noise within the time series, and inter-subject variability, reflecting subject-to-subject differences in activation. The purpose of this study was to obtain estimates of percent signal change and the two sources of variability from fMRI data, and then use these parameter estimates in simulation experiments in order to generate power curves. Of interest from these simulations were conclusions concerning how many subjects are needed and how many time points within a scan are optimal in an fMRI study of cognitive function. Intra-subject variability was estimated from resting conditions, and inter-subject variability and percent signal change were estimated from verbal working memory data. Simulations derived from these parameters illustrate how percent signal change, intra- and inter-subject variability, and number of time points affect power. An empirical test experiment, using fMRI data acquired during somatosensory stimulation, showed good correspondence between the simulation-based power predictions and the power observed within somatosensory regions of interest. Our analyses suggested that for a liberal threshold of 0.05, about 12 subjects were required to achieve 80% power at the single voxel level for typical activations. At more realistic thresholds, that approach those used after correcting for multiple comparisons, the number of subjects doubled to maintain this level of power.

Keywords: Statistics; Neuroimaging; Power; Sample size; fMRI

1. Introduction

Statistical power is defined as the probability of rejecting the null hypothesis when it is false. In most experiments conducted in the behavioral sciences, the main factors that influence power are: (1) *the size of the effect*, determined by the difference of the means of the experimental and control conditions and the variability of this difference across subjects; (2) *the value of alpha* that is used, which is the probability of rejecting the null hypothesis when it is true; and (3) *the sample size*, i.e. the number of subjects tested.

Of these three factors, sample size is the most amenable to manipulation by the experimenter. Effect size can be influenced by the experimental design, but for a given contrast of interest is generally out of the experimenter's control. When the null hypothesis is false, increasing alpha increases power. However, the increased risk of falsely rejecting the null hypothesis is considered an unacceptable consequence if the null hypothesis is true. Increasing sample size increases power because the standard error of the mean decreases by the square root of N. As illustrated in Fig. 1, for a given alpha level and separation between the H_0 (null) and H_1 (alternative) distributions, there is a greater probability (larger area under the H_1 curve) of rejecting H_0 if it is false when the sample size is larger.

Power calculations for a within-subjects experiment depend on assessing the effect size (Kraemer and Thiemann, 1991), which is defined as follows:

$$\delta = (\mu_D - 0)/\sigma, \tag{1}$$

where δ is the effect size, μ_D is the difference in means between the experimental and control condition, 0 is the

[*] Corresponding author. Tel.: +1-650-498-5368; fax: +1-650-723-5795

E-mail address: jdesmond@stanford.edu (J.E. Desmond).

0165-0270/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0165-0270(02)00121-8

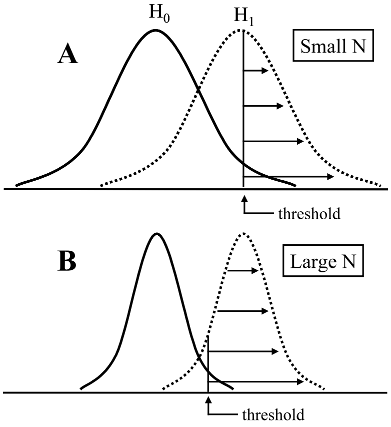

Fig. 1. Sample size increases power because the standard error of the mean decreases by the square root of N. For a given alpha level and separation between the H_0 and H_1 distributions, there is a smaller probability (smaller area under the H_1 curve) of rejecting H_0 if it is false when there are fewer subjects (A), than when there are more subjects (B).

difference in means under the null hypothesis, and σ is the variability in the difference in means. In functional MRI (fMRI) μ_D and σ are typically normalized as percent signal change (i.e. $100 \times (E-C)/C$, where E, experimental condition and C, control condition) because the raw signal values have no intrinsic physiological meaning and can vary considerably in magnitude in different regions of the brain.

fMRI signal-to-noise (SNR) is typically low, such that it is necessary to scan a subject for a period of 3–15 min during which time repeated presentations of the experimental and control conditions are presented to the subject while the scanner continually takes a snapshot of brain activation every 1–4 s. Thus a time series of approximately 50–400 time points is created for each of the approximately 120 000 voxel (i.e. smallest 3D element) locations in the brain. An inferentially valid statistical analysis of the data requires that the degrees of freedom at each voxel reflect the number of subjects. One way to achieve this is by creating a mean activation volume for the experimental and for the control condition by averaging all of the time points that were acquired during each condition. For each voxel, a paired t-test is then computed to determine if the experimental−control difference is significant, with the number of degrees of freedom equal to the number of subjects minus 1 (Holmes and Friston, 1998).

Because of the two-stage nature of fMRI group analysis, i.e. averaging time points within a scan for each subject followed by statistical tests on these averages across subjects, the variability, σ, represented in Eq. (1) consists of two parts, a within-scan (i.e. intra-subject) variability, σ_W, consisting of noise that occurs from one time point to another due to physiological fluctuations, thermal noise, and other random factors, and a between-subject (i.e. inter-subject) variability, σ_B, which is the subject to subject variability in the effectiveness of the experimental condition in producing a signal change. Estimation of effect size in Eq. (1) therefore requires estimating μ_D, σ_W, and σ_B.

Although analyses of power and sample size have been presented for positron emission tomography (PET) studies (Kapur et al., 1995; Andreasen et al., 1996; Grabowski et al., 1996; Van Horn et al., 1998; Wahl and Nahmias, 1998), similar analyses have not been performed for fMRI. The purpose of this report, which has appeared in abstract form (Desmond and Glover, 2000), was to estimate these parameters from real fMRI data, and then use the parameters in simulation experiments to generate power curves. Because increasing the number of time points within a scanning session tends to decrease the impact of σ_W on the ability to reject the null hypothesis, while increasing the number of subjects decreases the impact of σ_B, the effects of both time points and subjects on power were addressed in simulations. Finally, we sought to test the predictions of the simulation using a simple fMRI experiment involving somatosensory stimulation of the fingers.

2. Methods

The procedures for this report required first, an estimation of critical parameters from real fMRI data. Once this was accomplished, the estimated parameters were used in simulation experiments to generate power predictions and power curves. An empirical test using real fMRI data was then conducted to verify that the simulator, given accurate estimates of percent signal change and inter- and intra-subject variability, yields power predictions that accurately predict the results of standard random effects analysis methods. An overview of these procedures is illustrated in Fig. 2.

3. Estimation of σ_W

3.1. Subjects

Six subjects, four males and two females with a mean age of 36.7 (SD = 11.9) were scanned under resting conditions with eyes open for a total of 4 min. These rest segments were interspersed with blocks of finger stimulation (described below) that were designed for testing power predictions.

3.2. Data acquisition and analysis

fMRI data were acquired on a 3 T GE Signa magnet using a T2*-weighted gradient echo spiral pulse se-

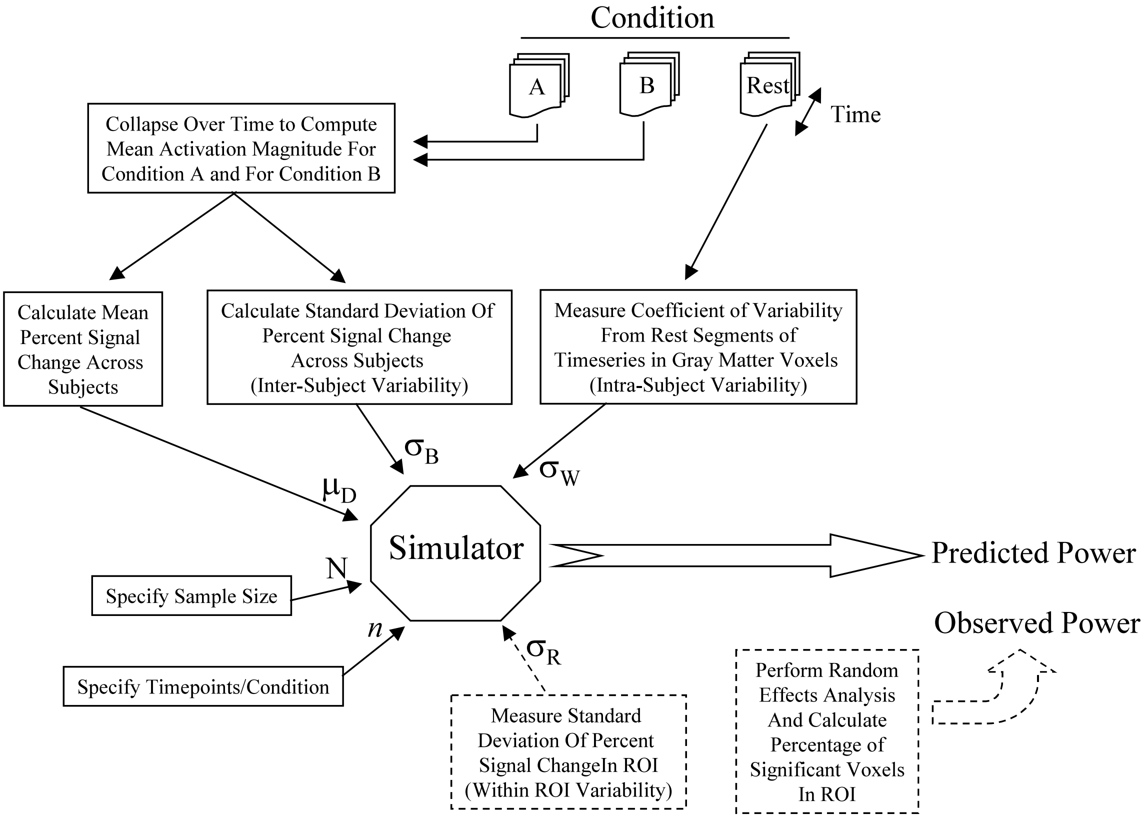
Condition
A
B
Rest
Time
Collapse Over Time to Compute Mean Activation Magnitude For Condition A and For Condition B
Calculate Mean Percent Signal Change Across Subjects
Calculate Standard Deviation Of Percent Signal Change Across Subjects (Inter-Subject Variability)
Measure Coefficient of Variability From Rest Segments of Timeseries in Gray Matter Voxels (Intra-Subject Variability)
σB
σW
μD
N
Simulator
Predicted Power
Specify Sample Size
n
Specify Timepoints/Condition
σR
Measure Standard Deviation Of Percent Signal ChangeIn ROI (Within ROI Variability)
Perform Random Effects Analysis And Calculate Percentage of Significant Voxels In ROI
Observed Power

Fig. 2. Overview of the methods used for estimating parameters, generating power predictions from simulations, and empirically testing the results. Components appearing in dashed lines were used only in the empirical test experiment. Note that in the test experiment, the intra-subject variability, σ_W, was not estimated from whole brain resting data, but instead was measured only from the region of interest during the tasks used for the analysis. Power curves were created by entering multiple values of sample size (N) and fixing all other parameters.

quence (Glover and Lai, 1998), using a custom-built quadrature 'top hat' elliptical bird cage head coil. Head movement was minimized using a bite-bar that was formed with the subject's dental impression, and further corrected using the SPM99 software package (Wellcome Department of Cognitive Neurology). fMRI scans were obtained from 25 axial slices using parameters of TR = 2000, TE = 30 ms, flip angle = 75°, single shot, inplane resolution = 3.75 mm, and thickness = 5 mm. A T2-weighted fast spin-echo was acquired in the same plane as the functional scans with parameters of TR = 3000, TE = 85 ms, echo train length = 8, and NEX = 1. These structural data were coregistered with the mean post-motion-corrected fMRI volume and spatially normalized to the Montreal Neurological Institute (MNI) brain template (2 × 2 × 2 mm voxels) using a nine-parameter affine transformation in SPM99 (Friston et al., 1995a). Both spatially smoothed (FWHM = 5 mm) and unsmoothed data were subjected to further analyses.

To estimate σ_W, segments of rest data were extracted from the time series of each voxel. The first 10 s of data from each segment were discarded to allow for hemodynamic changes from the previous block to settle. The coefficient of variability (CV), defined as 100 × (SD/mean), was calculated from a total of 3 min of rest data per subject as an estimate of σ_W. Spatially normalized data from all subjects were then averaged to make a σ_W volume. To insure that σ_W values were obtained from gray matter voxels, rather than regions of white matter or cerebro-spinal fluid, the mean structural brain volume, based on the T2-weighted fast spin-echo scans averaged over the six subjects, was segmented to isolate gray matter voxels using the method of Ashburner and Friston (1997) in SPM99. Undesired regions were thereby excluded from analysis.

To investigate whether autocorrelations in the fMRI time series affected measurements of σ_W, the CV on the last longest segment of resting data (40 time points) was calculated on subsampled time series. That is, the first calculation (SUBSAMP1) was calculated on the original 40 time points. The second calculation (SUBSAMP2) was performed on every other value of the time series (20 points per time series). Similar calculations were performed on every third (SUBSAMP3, 13 time points) and fourth (10 time points). Autocorrelations and CV values were calculated for each of the subsampled runs. To further investigate the distribution of noise in the time series, time series were converted into Z scores, and the distribution of these values was examined and compared to the normal distribution.

4. Estimation of μ_D and σ_B

4.1. Subjects

Twelve subjects, seven males and five females, gave their informed consent to participate in a study designed to measure brain activation during a verbal working memory task (to be published in a separate report). The mean age of the subjects was 32.9 years (SD = 10.3).

4.2. Tasks

Subjects were scanned under two conditions. The first condition utilized a verbal working memory task. For this task, six uppercase consonants were visually presented at a rate of one every 900 ms and for a duration of 750 ms each. After the letter presentation, the screen was kept blank for 2 s, and then a lowercase probe letter was presented for 2 s. The subject was instructed to press one button if the probe matched one of the original six letters that were being held in mind, and a second button if the probe did not match. The second condition was a resting condition in which the subject maintained fixation on a '+' symbol.

Stimuli were visually presented to the subject in the scanner by back-projecting the images via a magnet-compatible projector onto a screen located above the subject's neck. Visual images were viewed from a mirror mounted above the subject's head. Stimuli were presented from a Macintosh computer (Apple Computer, Inc., Cupertino, CA) using PsyScope software (Cohen et al., 1993).

4.3. Data acquisition and analysis

fMRI data were acquired on a 3 T Signa scanner as described above. For this experiment, 29 sections were acquired in the coronal plane and fMRI data were collected using parameters of TR = 3000, TE = 30 ms, flip angle = 83°, single shot, inplane resolution = 3.75 mm, slice thickness = 6 mm. A T2-weighted fast spin-echo scan was also acquired using parameters described above. Motion correction, spatial normalization (12 parameter, $2 \times 2 \times 2$ mm voxels), and Gaussian spatial smoothing (FWHM = 5 mm) were performed using SPM96.

To estimate μ_D and σ_B a two-step procedure was followed. For the first step, mean spatially normalized volumes were created for each subject in each of the conditions by averaging time series data for each condition (using the 'adjusted mean' routine in SPM96). The verbal working memory mean volume was averaged over 61 time points (183 s) collected in three blocks of 1 min duration each, and the rest condition over 42 time points (126 s) collected in three blocks of 42 s duration each. Using the general linear model approach available in SPM96 (Friston et al., 1995b), a random effects analysis was performed for the 12 subjects to identify regions in which activation from verbal working memory was greater than that of the resting condition. From this analysis, nine regions of interest were identified from the averaged activation map, using a threshold of $P < 0.01$, in Broca's area, left premotor cortex, left supramarginal gyrus, left middle frontal gyrus, left superior temporal gyrus, right superior temporal gyrus, right inferior and middle frontal gyri, right superior cerebellum, and right inferior cerebellum (3024 voxels total).

For the second step of this analysis, percent signal change, defined as $100(A-B)/B$, where A, activation magnitude under verbal working memory and B, activation during rest, was calculated for each of the voxels in the regions of interest (ROIs) for each subject. For each of the 3024 voxels, a mean percent signal change across the 12 subjects was computed. The distribution of these mean percent signal change values was taken as a measure of μ_D for this contrast. A similar approach was taken for estimating σ_B, except that instead of computing the mean of the percent signal change across subjects at each voxel, the standard deviation of the percent signal change was computed. Note that because σ_B is contaminated by the contribution of σ_W the standard deviation estimate of σ_B was corrected using the estimate of σ_W and the number of independent time points using the equation:

$$\sigma_B = \sqrt{\hat{\sigma}_B^2 - 2\sigma_W^2/n},$$

where $\hat{\sigma}_B^2$ is the measured standard deviation of the percent signal change, and n is the number of independent time points.

5. Simulations

Simulations were based on a block design experiment using a random effects model, and were written in the Interactive Data Language (IDL, RSI Systems, Inc., Boulder, CO). Because statistical tests in fMRI analyses are typically performed independently on each voxel, each simulated subject was represented as a time series in a single voxel. A population of 20 000 subjects (time series) was created, with each time series representing simulated brain activation for two conditions, an experimental and control condition. Values of μ_D, σ_W, and σ_B were based on parameter estimate studies described above, and the number of points in the time series was varied between 50 and 400.

To incorporate these parameters into realistic population time series, the following steps were performed: (1) a baseline value (10 000) was defined and a time series of n points at the baseline values was created; (2) a time series of n points of Gaussian noise with a mean of 0 and a standard deviation equal to σ_W was added to the baseline time series to create a noisy time series. (Note that these are n independent points without autocorrelations, and the value of n is therefore assumed to correspond to the effective degrees of freedom or independent observations in fMRI time series (Worsley and Friston, 1995).); (3) a square wave, representing alternating experimental and control conditions, was then added to the time series of step 2. The distribution of percent signal change between control and experimental conditions was defined by μ_D and σ_B. Fig. 3 illustrates two time series from the population after completion of these steps.

To create power curves, samples of 4, 6, 8, 10, 12, 16, 24, and 32 time series were drawn from the population. For each time series a mean value for the experimental and for the control condition was computed, and a paired t-test was performed with the degrees of freedom equal to the sample size minus 1. This procedure was repeated 1000 times for each sample size. The percentage of rejections of the null hypothesis (Ho) that the mean percent signal change was equal to zero was then computed to create power curves.

6. Empirical test experiment

6.1. Subjects

The six subjects described above for the estimation of σ_W were also scanned for the purpose of testing simulation predictions.

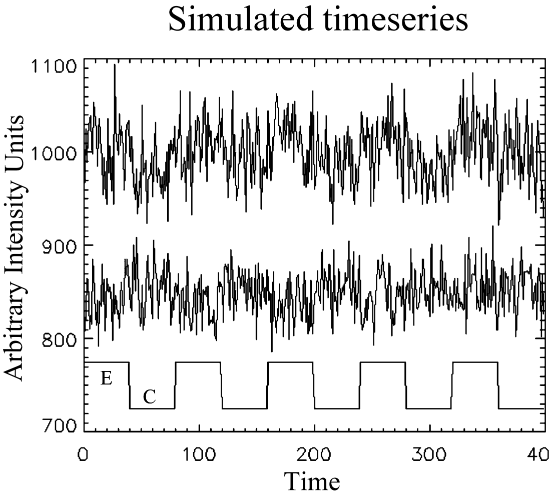

Fig. 3. Simulated fMRI waveforms used in power calculations. The top trace shows a time series in which the percent signal change in the experimental condition (E) with respect to the control condition (C) is positive. The middle trace illustrates a time series with a negative percent signal change. The bottom trace depicts the reference waveform indicating the times at which the experimental and control conditions occur.

6.2. Stimuli

Subjects received bilateral passive tactile stimulation of the fingers, using a custom-built MR-compatible device. This device consisted of five pneumatically-driven plungers that were imbedded in a foam mold of each hand. Computer-driven activation of the relay for any finger caused air pressure to translate the plunger upward and push up on the finger, thereby causing tactile sensation and passive movement. Subjects received movement of one finger at a time on each hand, and movement of each hand's finger occurred simultaneously. Stimulation was modulated by altering the number of finger movements per second, and subjects received alternating blocks at 4, 1, and 0 Hz (rest).

6.3. Data acquisition and analysis

Scan parameters were the same as those described for the estimation of σ_W. Subjects were given three scans in counterbalanced order. One scan, which was used to define primary somatosensory ROIs as well as to estimate σ_W, consisted of 12 alternating blocks of 1 Hz stimulation and rest (30 s/block, except for the last rest block, which was 90 s in duration). The other two scans each consisted of 16 alternating blocks of 4 and 1 Hz finger stimulation at 30 s/block. Data were motion corrected, coregistered with an in-plane structural scan, and then normalized to the MNI template using SPM99. Gaussian spatial smoothing was performed on the normalized volumes at FWHM = 5.0 mm. Average volumes for the 0, 1, and 4 Hz conditions were created using SPM99's adjusted mean function, for both spatially smoothed and unsmoothed data. In creating these mean volumes the data were high-pass filtered at a period of 120 s.

To define the primary somatosensory ROIs, each subject's smoothed and normalized volumes were statistically analysed using the general linear model implemented in SPM99. The contrast of 1 Hz vs. rest was performed for each subject, creating a t-value map for each subject's activations. The left and right ROIs were defined as the regions surviving a conjunction of each subject's activation at a P value threshold of 0.05 (one-tailed). Once the ROI was defined by the 1 Hz vs. rest contrast, the two 4 vs. 1 Hz scans were used to test the simulation's predictions.

The empirical test consisted of drawing sample sizes of 6, 8, 10, and 12 from the pool of 12 scans that were obtained from the six subjects (i.e. two runs of the 4 vs. 1 Hz condition per subject). Four samples of size 6 and 8, three samples of size 10, and one sample of size 12 were selected. For each sample size smoothed and unsmoothed versions of the data were analysed at two different levels of alpha (0.05 and 0.002) and for two separate ROIs (left and right somatosensory cortex), making a total of 96 pairs of predicted and observed power measurements. In generating predicted power, average σ_W, μ_D, and σ_B values were made from the voxels of the ROI. These parameters were then supplied to the simulator, and the prediction was tested by assessing the percentage of ROI voxels, out of the total number of voxels in the ROIs (225 voxels total), that were found to be statistically significant when a standard random effects analysis (using a paired t-test) was performed on the brain activation data. This method of testing was chosen because it was considerably easier to generate and test power from a large number of voxels than from a comparable number of subjects. However, a slight modification was required to use this method: because each subject is assumed to have a single percent signal change value representing the small somatosensory ROI, variability among the voxels for a given subject cannot be considered to be due to σ_B, nor can it be attributed entirely to σ_W. It was necessary to assume that this variability reflected a within-ROI component, σ_R. To the extent that σ_R is low, significance observed in one voxel of the ROI will likely predict the outcome of all the ROI voxels. Appendix A provides a detailed description of the equations used to measure these variance components. The simulation software was modified to incorporate this source of variability. Specifically, a population of 10 000 time series was created as described previously, and these time series reflected intra- and inter-subject variabilities. For each sample drawn from this population, a ROI of 100 voxels was created for each subject in the sample. The percent signal change values of the ROI voxels for each subject varied around that subject's mean value with variability defined by the σ_R value that was specified for the simulation. For each voxel a t-test was performed to test whether the mean percent signal change across subjects was significantly different from 0. The percentage of ROI voxels that were found to be significantly different from 0 was then calculated. A total of 500 repetitions of this procedure were performed to estimate the average power for rejecting Ho within the ROI.

7. Results

7.1. Parameter estimations

The distribution of σ_W values for smoothed and unsmoothed data is plotted in Fig. 4. It can be seen from the figure that spatial smoothing reduces the magnitude of σ_W, with a mean value of 0.74% (median = 0.70%) for the smoothed distribution, and a mean of 1.15% (median = 1.08%) for the unsmoothed distribution. Regional variability in the magnitude of σ_W was also apparent as illustrated in Fig. 5, which shows higher σ_W values predominately in visual areas and lower σ_W values mostly in anterior cingulate, basal ganglia, and insular cortex. The high and low values of σ_W illustrated in Fig. 5 were derived from the upper and lower tails of the smoothed distribution illustrated in Fig. 4 for segmented gray-matter voxels.

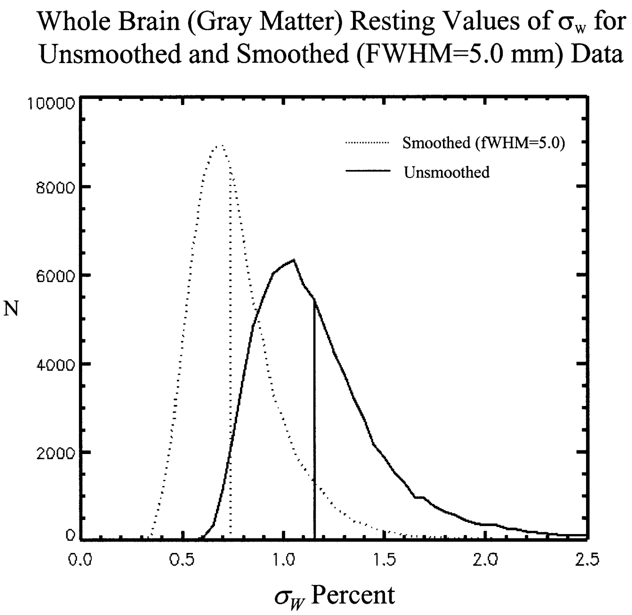

Fig. 4. Estimation of intra-subject variability, σ_W under spatially smoothed (FWHM = 5 mm) and unsmoothed conditions. Graphs depict the results obtained from six subjects who were scanned under resting conditions. Data were motion corrected, spatially normalized to an MNI template, and segmented to include only gray matter voxels.

Examination of the subsampled time series for the non-spatially-smoothed data revealed a significant effect of subsampling on the lag1 autocorrelation values ($F(3, 15) = 28.963$, $P < 0.0001$), with mean autocorrelation values of 0.314, 0.233, 0.108, and 0.057, respectively, at SUBSAMP values of 1–4. However, neither the mean ($F(3, 15) = 0.583$, $P = 0.64$) nor the median ($F(3, 15) = 0.556$, $P = 0.65$) values of σ_W showed any changes with subsampling. The distribution of time series noise within a short (80 s) segment showed good correspondence with the normal distribution, as illustrated in Fig. 6. The first four moments of the distribution were −0.0000464, 0.9397, 0.0405, and −0.0486. Inspection of each of the six subject's distribution showed similar near-normal distributions for all subjects.

The distribution of μ_D and σ_B for the working memory cognitive task is illustrated in Fig. 7. A mean percent signal change of 0.48% was observed for the working memory vs. rest comparison, and the mean value of σ_B was 0.77%. Note that these distributions are based on a $P < 0.01$ threshold used to define the voxels of interest, and that more stringent thresholds would likely bias the distribution of μ_D toward higher values.

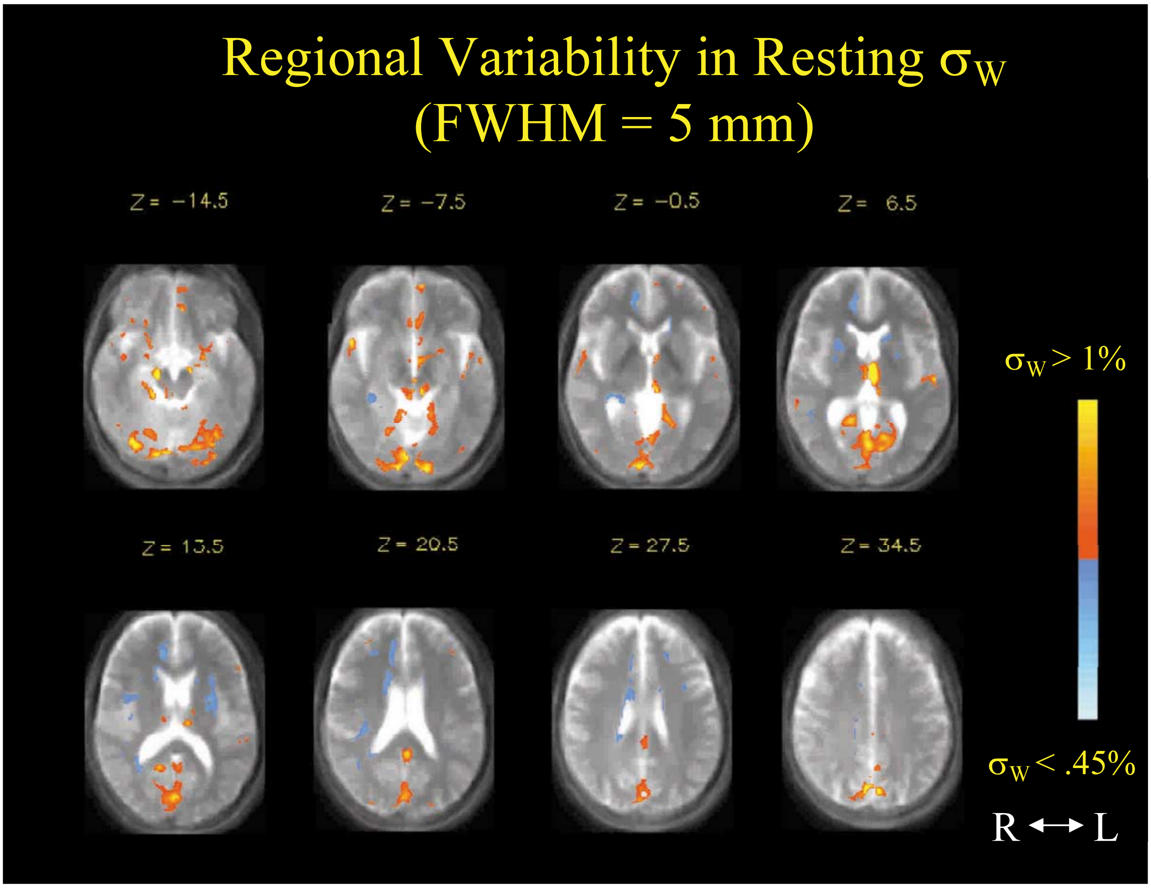

Fig. 5. Regional variability in intra-subject variability, σ_W, observed under resting conditions for six subjects with eyes open. Regions in the upper 10% of the smoothed distribution of Fig. 4 are illustrated in the red color scale and regions in the lower 10% of the distribution of σ_W values are shown in blue. These correspond to σ_W values of 1 and 0.45%, respectively.

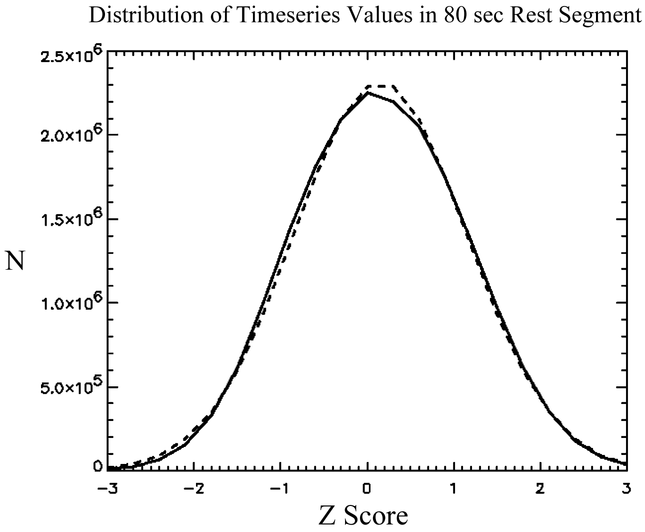

Fig. 6. Distribution of time series noise, converted into Z scores for each voxel using mean and standard deviation calculations for that voxel during an 80 s rest period, collected over six subjects for all gray matter voxels. Solid line denotes observed counts while dotted line represents expected values for normal distribution.

8. Simulations: power curves

For simulations, the following parameter values were chosen based on measurements described above:

$\sigma_W = 0.75 - 1.25\%$

$\sigma_B = 0.30 - 0.70\%$

$\mu_D = 0.25 - 0.75\%$

Power curves using a two-tailed alpha of 0.05 are depicted in Fig. 8, where the effects of different levels of either μ_D or σ_B can be seen. With a μ_D and σ_B of 0.5%, 11–12 subjects are needed to achieve 80% power at $\alpha = 0.05$, assuming a value of 0.75% for σ_W, typically observed in spatially smoothed data, and 100 time points per condition (n). Note that at $\mu_D = 0.75\%$, approximately six subjects are needed to achieve 80% power; a decrease of 0.25% in μ_D (from 0.75 to 0.5%) requires an additional 5–6 subjects to maintain 80% power, whereas an additional decrease in μ_D of 0.25% (from 0.5 to 0.25%), requires over 20 more subjects to maintain 80% power. In contrast, changes in σ_B seem to have a more linear effect on the number of subjects needed to maintain comparable power levels.

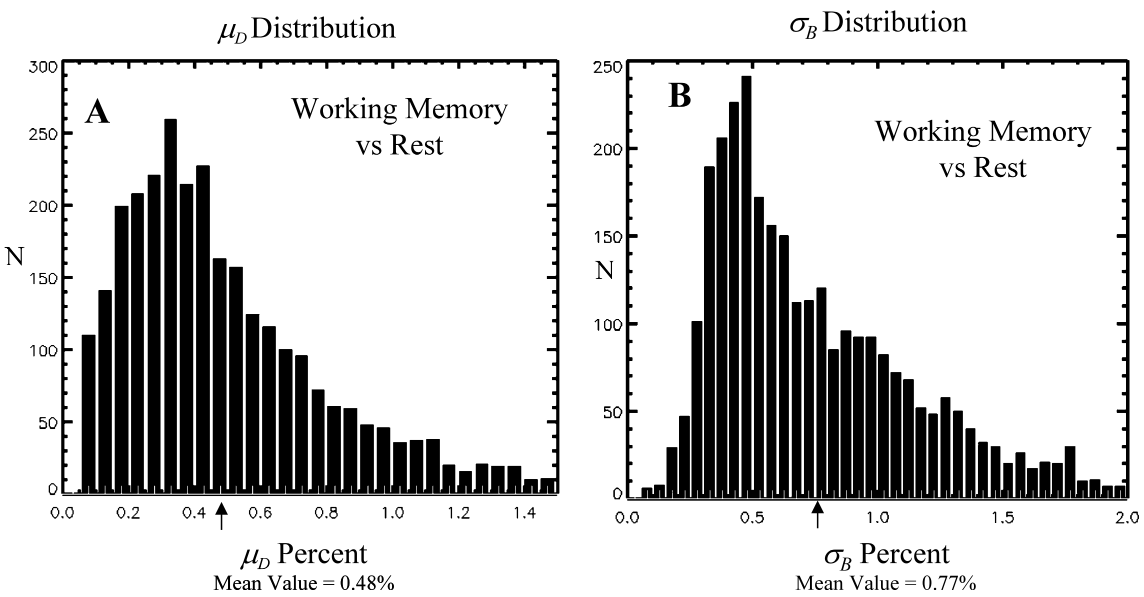

Fig. 7. (A) Distribution of percent signal change value, μ_D, during a working memory task relative to rest for 3024 voxels representing nine regions of interest, with each voxel averaged over 12 subjects. (B) Distribution of inter-subject variability, σ_B, values during the working memory task for the same voxels and subjects.

Power curves using a more conservative level of $\alpha = 0.002$ (two-tailed) are illustrated in Fig. 9. For the same values of $\mu_D = 0.5$ and $\sigma_B = 0.5\%$ it can be seen that approximately twice the number of subjects are needed to maintain 80% power for this level of alpha than at $\alpha = 0.05$. For $\alpha = 0.000002$ (two-tailed), higher signal (i.e. $\mu_D \geq 0.75$ at $\sigma_B = 0.5\%$) or lower inter-subject variability (i.e. $\sigma_B = 0.3$ at $\mu_D = 0.5\%$) are needed to maintain 80% power with approximately 25 subjects (Fig. 10).

The effects of the number of time points per condition (n) is illustrated in Fig. 11 at $\alpha = 0.05$, and in Fig. 12 at $\alpha = 0.002$. In both figures note that at lower levels of σ_W n has less impact than when σ_W is higher. Fig. 13 illustrates that the effect of n may be greatest when σ_W is high and σ_B is low. There also appears to be diminishing

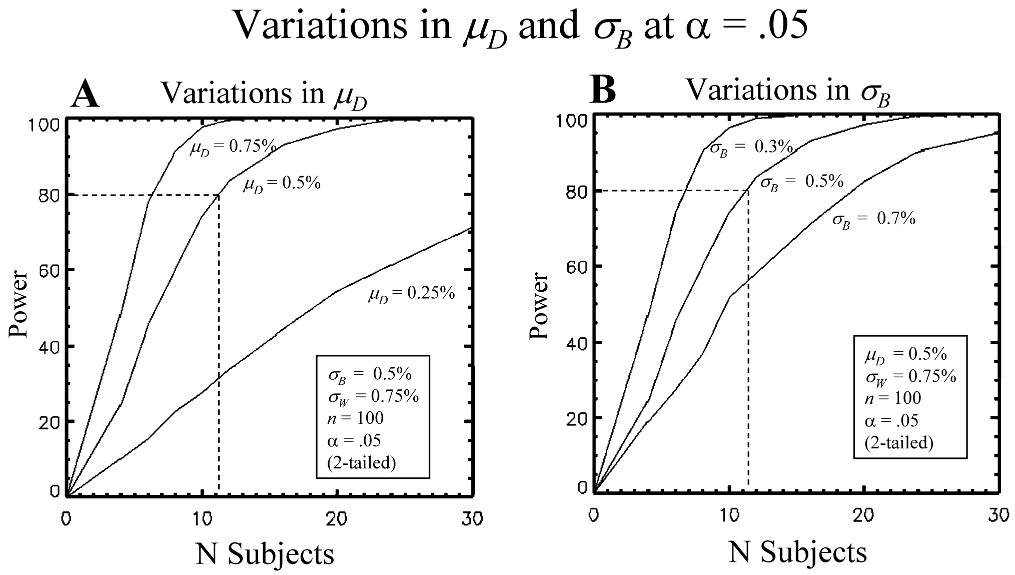

Fig. 8. Power curves generated at an alpha of 0.05 (two-tailed). The effects of different levels of μ_D (percent signal change) with fixed inter-subject variability (σ_B), intra-subject variability (σ_W), and time point per condition (n) are depicted in A, whereas the effects of different levels of σ_B with fixed μ_D, σ_W, and n are illustrated in B.

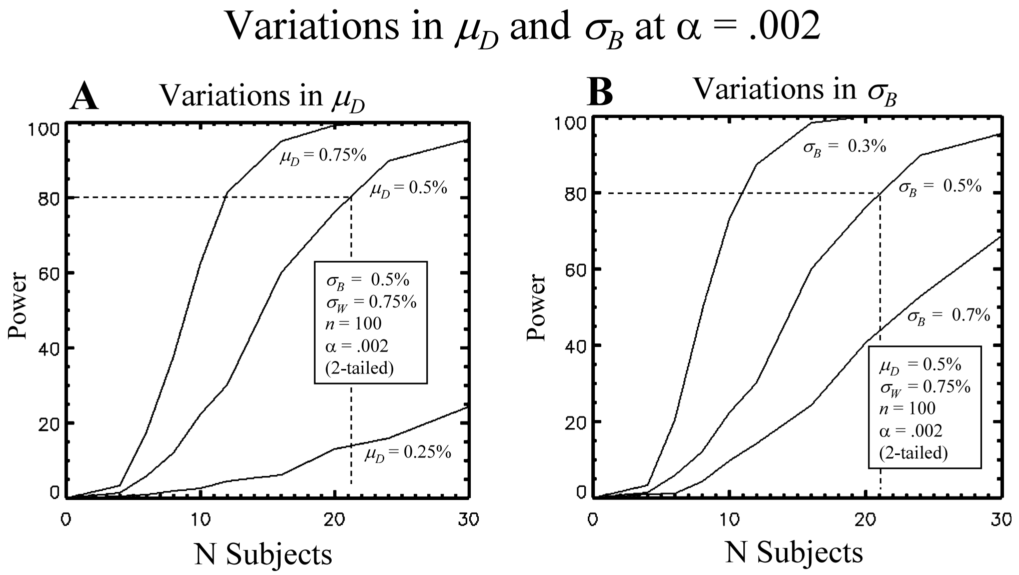

Fig. 9. Power curves generated at an alpha of 0.002 (two-tailed). The effects of different levels of percent signal change (μ_D) with fixed inter- (σ_B), intra-subject variability (σ_W), and time points per condition (n) are depicted in A, whereas the effects of different levels of σ_B with fixed μ_D, σ_W, and n are illustrated in B.

returns for increasing n beyond 100 independent samples.

9. Empirical test experiment

The results of the conjunction analysis for the 1 Hz vs. rest contrast that were used to define the somatosensory ROIs are illustrated in Fig. 14. For each of the 96 combinations of sample size, alpha, ROI, and spatial smoothness, estimates of σ_W, μ_D, σ_B and σ_R were computed from the samples and used as parameters for the simulation. An n of 28 was used in the simulations, based on the effective (i.e. independent) degrees of freedom estimated by SPM99. For each of the 96 tests, a simulation using the estimated parameters generated a predicted power value. A random effects analysis on the adjusted mean volumes for the 4 vs. 1 Hz contrast was then performed to calculate the observed power (i.e. percent of the voxels in the ROI found to be significant). The plot of predicted vs. observed measurements is illustrated in Fig. 15, which shows a high

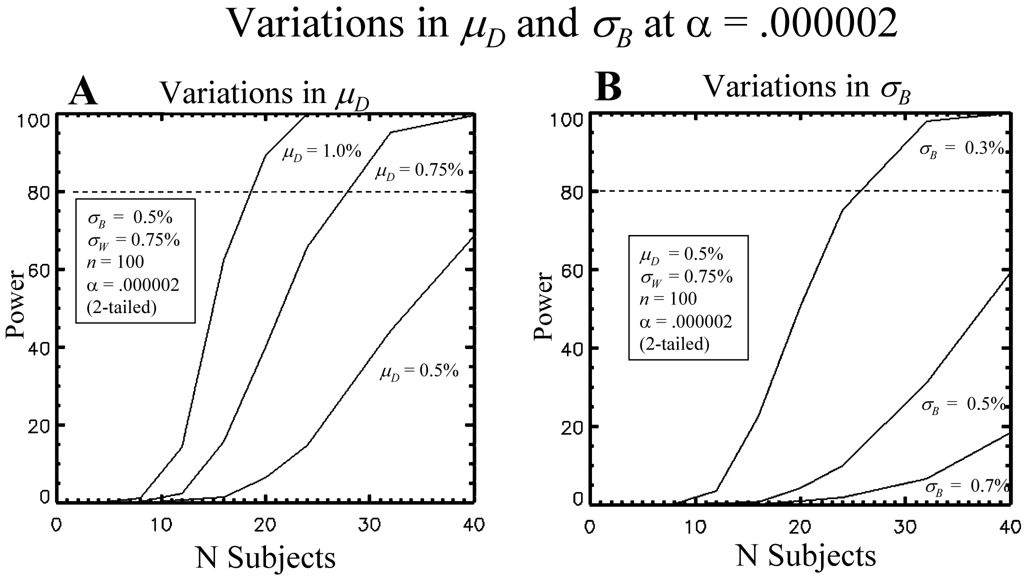

Fig. 10. Power curves generated at an alpha of 0.000002 (two-tailed). The effects of different levels of percent signal change (μ_D) with fixed inter- (σ_B), intra-subject variability (σ_W), and time points per condition (n) are depicted in A, whereas the effects of different levels of σ_B with fixed μ_D, σ_W, and n are illustrated in B.

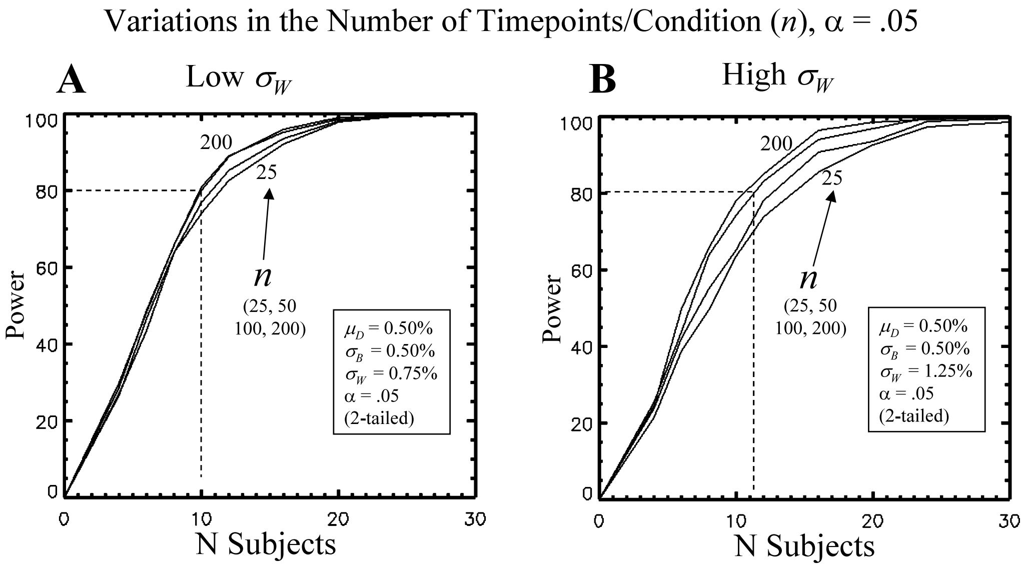

Fig. 11. Effects of varying time points per condition (n) at a low value of intra-subject variability (σ_W) (A) and at a higher value (B) using an alpha of 0.05 (two-tailed).

correlation between these values ($r = 0.98$). Note that the 96 tests cannot be regarded as independent as they are permuted from the same set of data. The different permutations were intended to show the entire range (0–100%) of predicted and observed values to see if, given accurate parameter estimates, the simulator, with its simplifying assumptions, would reasonably predict the results of a random effects analysis.

10. Discussion

This report has presented a method for estimating the number of subjects needed in fMRI research, and the results of an empirical test experiment revealed a good correspondence between simulation-based power predictions and power observed in somatosensory regions. The power curves were based on a random effects analysis, which generally requires more subjects than a fixed effects analysis, but has greater inferential validity to the population from which the subjects were drawn. Alternatives to the random effects approach that retain some of the sensitivity of fixed-effects models and are capable of making restricted inferences to the population have also been proposed (Friston et al., 1999a,b), but were not addressed in this report.

The most difficult aspect of deciding how many subjects to use in an fMRI experiment is estimating

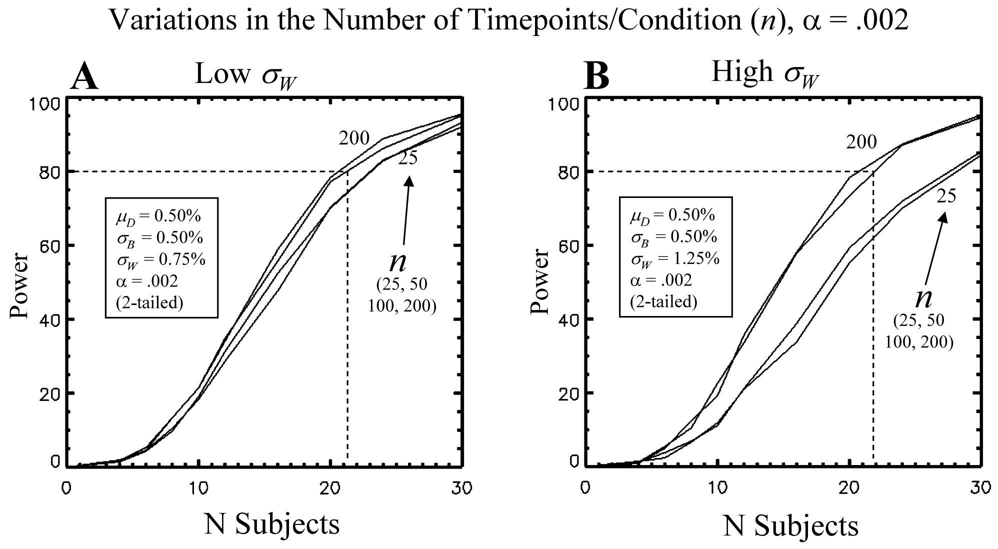

Fig. 12. Effects of varying time points per condition (n) at a low value of intra-subject variability (σ_W) (A) and at a higher value (B) using an alpha of 0.002 (two-tailed).

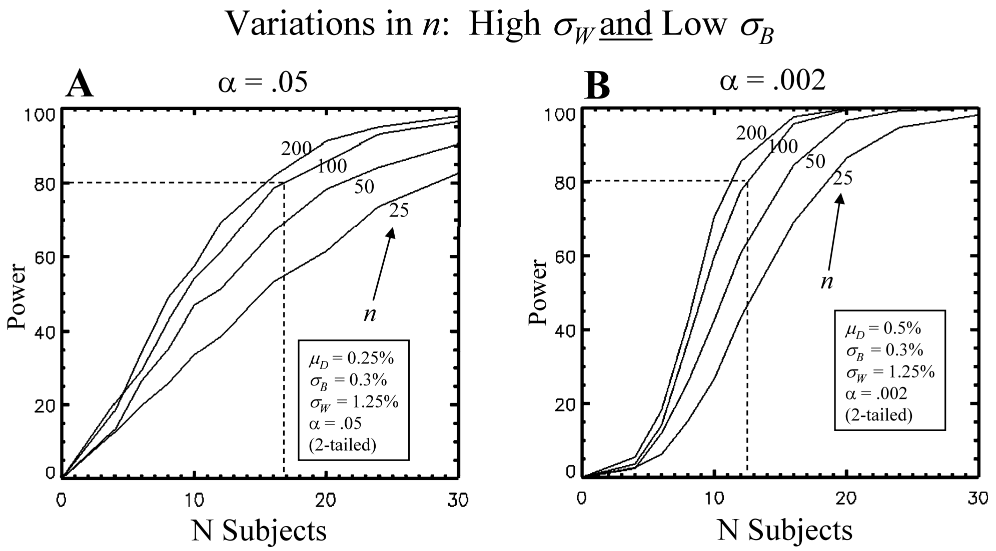

Fig. 13. The number of time points/condition (n) has a greater effect on power curves when intra-subject variability (σ_W) is high and inter-subject variability (σ_B) is low. This is observed at $\alpha = 0.05$ and percent signal change (μ_D) of 0.25% in (A), and at $\alpha = 0.002$ and $\mu_D = 0.5\%$ in (B).

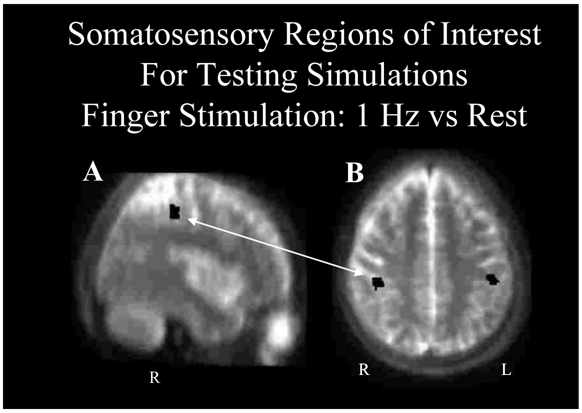
Somatosensory Regions of Interest For Testing Simulations Finger Stimulation: 1 Hz vs Rest

Fig. 14. Somatosensory ROIs used in the empirical test experiment. Voxels surviving the conjunction across six subjects at $P < 0.05$ for each subject are depicted in black on the sagittal (A) and axial (B) sections. ROIs are depicted on a T2-weighted fast spin echo volume that was normalized and averaged across six subjects.

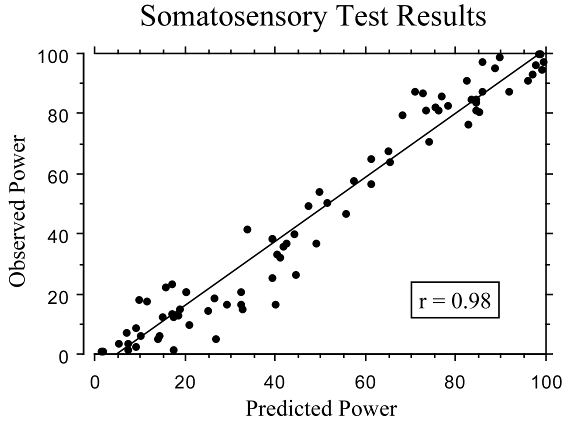

Fig. 15. Results of somatosensory test experiment. The graph depicts predicted (from the simulation) vs. observed (from t-tests on ROI voxels, using subjects as a random factor) power for the 96 combinations of subject sample size, alpha, spatial smoothness, and ROI that were analysed.

the critical parameters μ_D and σ_B. Because this study focused on only one type of cognitive task, verbal working memory, for estimating μ_D and σ_B, our simulations were generated using a range of parameter values rather than a single value, with the hope that this range will be relevant for many types of tasks.

It is evident from the power curves that selecting a sample size imposes limits on the range of μ_D, σ_B, and σ_W values that are likely to be represented in an fMRI activation map. Thus, before accepting the null hypothesis that a region of the brain does not show activation, and therefore does not contribute to a cognitive task, it is useful to consider the alternative interpretation that the sample size has effectively filtered out voxels whose μ_D, σ_B, or σ_W, or some combination of these parameters, are not in an optimal range. It is difficult to define a priori what the optimal range should be or to assign a level of meaningfulness to any specific percent signal change value, because this value is highly dependent on the particular contrast used in the analysis. For example, a well-designed control condition is likely to yield low percent signal changes, which if reliable, may have a high degree of theoretical or interpretive meaning.

The results of the simulations indicate that with percent signal changes of approximately 0.5% and spatial smoothing at FWHM of 5 mm, a minimum of 12 subjects are needed to insure 80% power at $\alpha = 0.05$ at the single voxel level. For a more conservative alpha, twice as many subjects are recommended to maintain this level of power. With μ_D values as high as 0.75%, or with low σ_B values of approximately 0.3%, 10 subjects may be adequate for 80% power for an alpha of 0.002, but approximately 25 subjects are needed for a stricter alpha of 0.000002.

The benefits of increasing n, the number of time points per condition, appeared to diminish after 100. It should be noted that in simulations n represented independent time points; however, in actual fMRI experiments attempts to increase n by decreasing TR will not result in a one-to-one increase in n. In the empirical somatosensory experiment, the number of effective degrees of freedom obtained from SPM99 was used to estimate the number of independent time points. The beneficial effects on power of increasing n appeared to be greatest when σ_W was high and σ_B was low.

Measurements made from subsampled time series revealed that σ_W estimates were likely not distorted by autocorrelations at shorter lags, as σ_W remained unchanged when autocorrelations were reduced by subsampling. The distribution of σ_W measured from 80-s segments of rest showed close correspondence to a normal distribution. However, the measurement of noise from relatively short segments likely precludes measurement of the lower frequency components of fMRI noise that are known to be prevalent (Zarahn et al., 1997; Friston et al., 2000). Lower frequency components of the noise may affect estimates of σ_B and this effect may be different at different task frequencies (Skudlarski et al., 1999). Many variables inherent to fMRI that occur within a scanning session, including the effects of task frequency, low frequency noise, heart rate, respiration, head motion, voxel size, and TR were not explicitly modeled in the simulations, for the sake of simplicity. The effects of these variables on power in random effects analyses will be addressed in future refinements of the simulator. For these reasons, simulations were performed using a range of different parameter values rather than trying to define a single value that is appropriate for all tasks.

Spatial smoothing reduced considerably the distribution of σ_W values, and because the effects of n were less pronounced at lower values of σ_W, smoothing would appear to be a reasonable pre-processing strategy to compensate for shorter scan times, especially if the resulting loss of spatial resolution is not a concern. Of course, an alternative to spatial smoothing is to use acquisitions with lower intrinsic resolution. For the same scan a SNR benefit will result. Other researchers have found spatial smoothing to be beneficial (Skudlarski et al., 1999), but the benefits of enhanced signal due to averaging can depend on the size and shape of the desired signal relative to the size and shape of the convolution kernel (Petersson et al., 1999). We also observed regional differences in resting σ_W values, raising the possibility that some areas of the brain may have statistical power advantages or disadvantages relative to other regions simply by virtue of the intra-subject variability inherent to the region. The properties of σ_W may change regionally depending on the type of condition under which it is measured. For example, the map depicted in Fig. 5 was obtained under an eyes-open resting condition, and this may account for the higher σ_W values observed in the occipital regions.

This study focused on power analysis for typical within-group fMRI experiments, in which inferences concerning the difference in activation between two or more conditions (with each condition measured from each subject) are intended to be made to a single population. Different parameter values for σ_W, σ_B, and μ_D may be observed in different populations, e.g. younger vs. older populations, or in different clinical populations. Future studies will address these possible parameter differences and the estimation of sample size when inferences regarding group differences are desired. In this regard, routine reporting of μ_D and σ_B by researchers would be beneficial for building a database that could be used for assessing statistical power in fMRI studies.

Acknowledgements

Supported by NIA (AG12995), NIMH (MH60234), NIAAA (AA10723), and NCRR (RR09784).

Appendix A

The following equations were used to calculate parameters for the empirical test experiment.

A.1

Calculation of intra-subject variability, σ_W

Let $q_{ij}(t)$ represent the signal intensity for subject i ($i = 1, \ldots, N$), in ROI voxel j ($j = 1, \ldots, R$) during a resting condition at time t ($t = 1, \ldots, n$). σ_W for subject i and voxel j is the coefficient of variability that is computed using the equation:

$$\sigma_{W_{ij}} = 100 \times \frac{\sqrt{\sum_{t=1}^{n} [q_{ij}(t) - \bar{q}_{ij}]^2/(n-1)}}{\bar{q}_{ij}},$$

where $\bar{q}_{ij}$ is the mean value averaged over time for subject i and voxel j. The overall value of σ_W is then computed as the average of the individual estimates:

$$\sigma_W = \frac{\sum_{j=1}^{R} \sum_{i=1}^{N} \sigma_{W_{ij}}}{NR}.$$

A.2

Calculation of percent signal change, μ_D

Let $q_{ijk}(t)$ represent the signal for subject i ($i = 1, \ldots, N$), in ROI voxel j ($j = 1, \ldots, R$) during condition k ($k = 1, \ldots, 2$, i.e. experimental and control condition) at time t ($t = 1, \ldots, n$ time points per condition). The mean signal for conditions 1 and 2 is:

$$\bar{q}_{ij1} = \frac{\sum_{t=1}^{n} q_{ij1}(t)}{n}, \quad \bar{q}_{ij2} = \frac{\sum_{t=1}^{n} q_{ij2}(t)}{n},$$

and the percent signal change for subject i at voxel j, p_{ij} (illustrated in Fig. 16) is then given by:

$$p_{ij} = 100 \times (\bar{q}_{ij1} - \bar{q}_{ij2})/\bar{q}_{ij2},$$

μ_{D} was then defined by averaging the p_{ij} values:

$$\mu_{\mathrm{D}} = \frac{\sum_{j=1}^{R} \sum_{i=1}^{N} p_{ij}}{NR}.$$

A.3

Calculation of within-ROI variability, σ_R

If we assume that each subject has a single percent signal change value for the small somatosensory ROI chosen for the experiment, then voxel to voxel variability within the ROI is defined as within-ROI variability, or σ_{R}. σ_{R}^2 for any given subject i was estimated from the equation:

$$\hat{\sigma}_{\mathrm{R}_i}^2 = \sum_{j=1}^{R} \frac{(p_{ij} - \bar{p}_i)^2}{R - 1},$$

where $\bar{p}_i$ is mean percent signal change value for that subject averaged over all the voxels in the ROI. Averaging this value across subjects gave an overall estimate of $\hat{\sigma}_{\mathrm{R}}^2$ of:

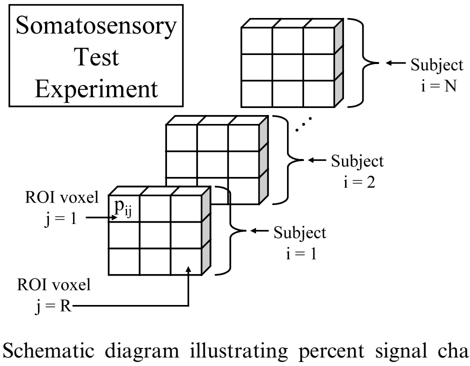

Fig. 16. Schematic diagram illustrating percent signal change measurements (p_{ij}) for each of the $j = 1, \ldots, R$ voxels and each of the $i = 1, \ldots, N$ subjects described in Appendix A.

$$\hat{\sigma}_{\mathrm{R}}^2 = \frac{\sum_{i=1}^{N} \hat{\sigma}_{\mathrm{R}_i}^2}{N}.$$

However, because each p_{ij} measure is derived from an average over time points in the time series (i.e. $q_{ijk}(t)$), $\hat{\sigma}_{\mathrm{R}}^2$ reflects the contribution of both σ_{R}^2 and σ_{W}^2. The value of σ_{R} corrected for intra-subject variability was computed from the equation:

$$\sigma_{\mathrm{R}} = \sqrt{\hat{\sigma}_{\mathrm{R}}^2 - 2\sigma_{\mathrm{W}}^2/n},$$

where n is the number of time points per condition. Note that the simulation assumed independent time points, and thus, the effective degrees of freedom obtained from SPM99 were used to estimate independent time samples. The effective degrees of freedom were found to be 56, so an n of 28 was used.

A.4

Calculation of inter-subject variability, σ_B

Inter-subject variability at any given voxel j was estimated using the equation:

$$\hat{\sigma}_{\mathrm{B}_j}^2 = \sum_{i=1}^{N} \frac{(p_{ij} - \bar{p}_j)^2}{N - 1},$$

where $\bar{p}_j$ is the mean percent signal change value for that voxel computed across subjects. The overall value for $\hat{\sigma}_{\mathrm{B}}^2$ was obtained by averaging over all the ROI voxels, i.e.:

$$\hat{\sigma}_{\mathrm{B}}^2 = \frac{\sum_{j=1}^{R} \hat{\sigma}_{\mathrm{B}_j}^2}{R}.$$

Because $\hat{\sigma}_{\mathrm{B}}^2$ contains the contributions of σ_{R}^2 and σ_{W}^2, as well as that of σ_{B}^2, the corrected estimate of σ_{B} was obtained using the equation:

$$\sigma_{\mathrm{B}} = \sqrt{\hat{\sigma}_{\mathrm{B}}^2 - \sigma_{\mathrm{R}}^2 - 2\sigma_{\mathrm{W}}^2/n}.$$

References

Andreasen NC, Arndt S, Cizadlo T, O'Leary DS, Watkins GL, Ponto LL, Hichwa RD. Sample size and statistical power in [15O]H$_2$O studies of human cognition. Journal of Cerebral Blood Flow and Metabolism 1996;16:804–16.

Ashburner J, Friston K. Multimodal image coregistration and partitioning—a unified framework. Neuroimage 1997;6:209–17.

Cohen JD, MacWhinney B, Flatt M, Provost J. PsyScope: a new graphic interactive environment for designing psychology experiments. Behavioral Research Methods, Instruments, and Computers 1993;25:257–71.

Desmond JE, Glover GH. Estimating sample size in random effects analyses of fMRI data. Society for Neuroscience Abstracts 2000;26:2235.

Friston KJ, Holmes AP, Worsley KJ. How many subjects constitute a study? Neuroimage 1999;10:1–5.

Friston KJ, Holmes AP, Price CJ, Buchel C, Worsley KJ. Multisubject fMRI studies and conjunction analyses. Neuroimage 1999;10:385–96.

Friston KJ, Ashburner J, Frith CD, Poline JB, Heather JD, Frackowiak RSJ. Spatial registration and normalization of images. Human Brain Mapping 1995;3:165–89.

Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RSJ. Statistical parametric maps in functional imaging: a general linear approach. Human Brain Mapping 1995;2:189–210.

Friston KJ, Josephs O, Zarahn E, Holmes AP, Rouquette S, Poline J. To smooth or not to smooth? Bias and efficiency in fMRI time-series analysis. Neuroimage 2000;12:196–208.

Glover GH, Lai S. Self-navigated spiral fMRI: interleaved versus single-shot. Magnetic Resonance in Medicine 1998;39:361–8.

Grabowski TJ, Frank RJ, Brown CK, Damasio H, Ponto LLB, Watkins GL, Hichwa RD. Reliability of PET activation across statistical methods, subject groups, and sample sizes. Human Brain Mapping 1996;4:23–46.

Holmes AP, Friston KJ. Generalisability, random effects and population inference. Neuroimage: Abstracts of the 4th International Conference on Functional Mapping of the Human Brain, 1998. Vol. 7. p. S754.

Kapur S, Hussey D, Wilson D, Houle S. The statistical power of [15O]-water PET activation studies of cognitive processes. Nuclear Medicine Communications 1995;16:779–84.

Kraemer HC, Thiemann S. How Many Subjects? Statistical Power Analysis in Research. Newbury Park, CA: Sage Publications, Inc, 1991:38–52.

Petersson KM, Nichols TE, Poline JB, Holmes AP. Statistical limitations in functional neuroimaging. II. Signal detection and statistical inference. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 1999;354:1261–81.

Skudlarski P, Constable RT, Gore JC. ROC analysis of statistical methods used in functional MRI: individual subjects. Neuroimage 1999;9:311–29.

Van Horn JD, Ellmore TM, Esposito G, Berman KF. Mapping voxel-based statistical power on parametric images. Neuroimage 1998;7:97–107.

Wahl LM, Nahmias C. Statistical power analysis for PET studies in humans. Journal of Nuclear Medicine 1998;39:1826–9.

Worsley KJ, Friston KJ. Analysis of fMRI time-series revisited again. Neuroimage 1995;2:173–81.

Zarahn E, Aguirre GK, D'Esposito M. Empirical analyses of BOLD fMRI statistics. I. Spatially unsmoothed data collected under null-hypothesis conditions. Neuroimage 1997;5:179–97.