Components of Variance in a Multicentre Functional MRI Study and Implications for Calculation of Statistical Power

John Suckling,^{1*} David Ohlssen,² Christopher Andrew,³ Glyn Johnson,⁴
Steven C.R. Williams,³ Martin Graves,⁴ Chi-Hua Chen,¹
David Spiegelhalter,² and Ed Bullmore^{1,5}

¹*Brain Mapping Unit, Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom*

²*MRC Biostatistics Unit, Institute of Public Health, University of Cambridge, Cambridge, United Kingdom*

³*Centre for Neuroimaging Sciences, Institute of Psychiatry, Kings College London, London, United Kingdom*

⁴*Department of Radiology, University of Cambridge, Cambridge, United Kingdom*

⁵*Clinical Unit Cambridge, Clinical Pharmacology Discovery Medicine, GlaxoSmithKline, Cambridge, United Kingdom*

Abstract: This article firstly presents a theoretical analysis of the statistical power of a parallel-group, repeated-measures (two-session) and two-centre design suitable for a placebo-controlled pharmacological MRI study. For arbitrary effect size, power is determined by the pooled between-session error, the pooled measurement error, the ratio of centre measurement errors, the total number of subjects and the proportion of subjects studied at the centre with greatest measurement error. Secondly, an experiment is described to obtain empirical estimates of variance components in task-related and resting state functional magnetic resonance imaging. Twelve healthy volunteers were scanned at two centres during performance of blocked and event-related versions of an affect processing task (each repeated twice per session) and rest. In activated regions, variance components were estimated: between-subject (23% of total), between-centre (2%), between-paradigm (4%), within-session occasion (paradigm repeat; 2%) and residual (measurement) error (69%). The between-centre ratio of measurement errors was 0.8. A similar analysis for the Hurst exponent estimated in resting data showed negligible contributions of between-subject and between-centre variability; measurement error accounted for 99% of total variance. Substituting these estimates in the theoretical expression for power, incorporation of two centres in the design necessitates a modest (10%) increase in the total number of subjects compared with a single-centre study. Furthermore, considerable improvements in power can be attained by repetition of the task within each scanning session. Thus, theoretical models of power and empirical data indicate that between-centre variability

Contract grant sponsors: Clinical Pharmacology and Discovery Medicine (GlaxoSmithKline R&D), National Institute of Mental Health (Human Brain Project), National Institute of Biomedical Imaging and Bioengineering.

*Correspondence to: John Suckling, Brain Mapping Unit, Department of Psychiatry, University of Cambridge, Addenbrooke's Hospital, Cambridge CB2 2QQ, United Kingdom.
E-mail: js369@cam.ac.uk

Received for publication 16 August 2006; Revised 9 May 2007; Accepted 22 June 2007

DOI: 10.1002/hbm.20451

Published online 6 August 2007 in Wiley InterScience (www.interscience.wiley.com).

can be small enough to encourage multicentre designs without major compensatory increases in sample size. *Hum Brain Mapp* 29:1111–1122, 2008. ©2007 Wiley-Liss, Inc.

Key words: ANOVA; Hurst exponent; factorial design; pharmacological fMRI

INTRODUCTION

Studies involving human subjects are both more generalisable and more efficient if the geographic catchment area for recruitment is enlarged through multicentre data collection. This practice is, of course, commonplace in pharmaceutical trials as it can improve the cost-effectiveness of clinical development programmes. Further, it may be the only pragmatic method of recruiting sufficient subjects within a reasonable time frame if the target patient group has a particularly rare condition or the study protocol enforces a stringent set of eligibility criteria. The negative impact of acquiring data from several distant sources mainly arises from inflation of the total variance by the introduction of a between-centre component of variability. This can be ameliorated by standardised protocols and careful quality control designed to minimise centre-related biases in measurement of the variable under investigation. Nevertheless, the effect of centre should be included, at least in the first instance, in the subsequent statistical analysis unless limited recruitment in one or more centres makes such an analysis impractical [ICH-E9, 1998].

Neural correlates of the action of a number of psychoactive drugs have been demonstrated with functional magnetic resonance imaging (fMRI) in the context of diverse cognitive paradigms [Honey and Bullmore, 2004] as well as during a no-task or 'resting' state [Wink et al., 2005]. These pharmacological MRI studies suggest that blood oxygenation level dependent (BOLD) responses to drug challenge could show measurable pharmacodynamic (PD) effects with comparable sensitivity to more common PD markers, such as symptom questionnaires or reaction times in cognitive testing, and with unprecedented neuro-anatomical specificity. Naturally, the question arises whether it is appropriate to use pharmacological MRI in the context of a multi-centre trial design.

MRI scanners from different manufacturers vary in the details of their construction and operation and it is clear that this will be reflected in differences in performance. Preliminary results from the Biomedical Informatics Research Network (BIRN; <http://www.nbirn.net/index.htm>) indicate large differences between centres in transient signals, sensitivity and image smoothness, shape of the haemodynamic response function and spectral properties of the time-series [Friedman and Glover, 2006a,b; Friedman et al., 2006; Zou et al., 2005]. In particular, the BIRN consortium has reported: (a) that intercentre variability

may be a persistent issue even when the centres in question are operating nominally identical scanners supplied by the same manufacturer and (b) that intercentre variability in BOLD signal can exceed intersubject variability [Turner et al., 2005].

In this article, we first report a theoretical analysis of the power of a hypothetical study design, entailing repeated scanning at two centres of two parallel groups at two points in time, of the kind that might be adopted for a placebo-controlled pharmacological MRI study of the effects on brain function of repeated dosing with a psychoactive drug. Our aim is to elucidate how the power of the study, for arbitrary effect size, depends on the measurement error, the between-session error, the ratio of measurement errors at the two centres, the proportion of subjects scanned at the centre with greatest measurement error and the total sample size.

We then report empirical results of an fMRI experiment conducted at two centres in the UK (Institute of Psychiatry, London and University of Cambridge) that was designed to estimate the components of variance in two cognitive paradigms and in the resting state. Each of 12 subjects were scanned at both centres (in counterbalanced order), using very similar 1.5T MR systems supplied by the same manufacturer (General Electric, Milwaukee, MI). In each scanning session, subjects performed two versions of a facial affect processing paradigm: an event-related design and a blocked periodic design. Each version of the task was presented twice in each scanning session. Additionally subjects were scanned once in each session during a no-task or resting state. For the affect processing tasks, reaction times and accuracies of explicit decision making were recorded as well as estimates of task-related BOLD response in a network of brain regions known to be involved in sad facial affect processing [Fu et al., 2004].

Functional imaging of BOLD contrast generates time-series with long-memory or $1/f$ -like spectral properties. Fractional Gaussian noise (fGn) is an appropriate model for fMRI resting time-series [Maxim et al., 2005] and the Hurst exponent, H , that parametrises the fGn model, is sensitive to both acute pharmacological challenge [with scopolamine; Wink et al., 2005], early Alzheimer's disease [Maxim et al., 2005] and attention deficit and hyperactivity disorder [Anderson et al., 2006].

The total variance in the task-related variables was decomposed into the following components of variance:

subject, centre (London or Cambridge), paradigm (event-related or blocked periodic), within-session task repetition and residual measurement error. For the resting state data, total variance in the Hurst exponent was decomposed in the same way with the exception of paradigm and task repetition effects.

These empirical data on components of variance were then substituted into the theoretical power analysis for a pharmacological MRI design, in combination with a prior estimate of the effect size of antidepressant drug treatment (fluoxetine 20 mg/day) on BOLD response to the event-related affect processing paradigm [Fu et al., 2004]. We were then able to explore the dependency of the power of this design on the total number of subjects scanned, the proportion of subjects scanned at the centre with the greatest measurement error and the number of task repetitions within each scanning session. A key outcome of this analysis was calculation of the number of additional subjects that would need to be scanned in a two-centre design compared with the number required to achieve the same statistical power under the best-case scenario that all subjects were scanned at the single centre with the smallest measurement error.

MATERIALS AND METHODS

Power Calculations for a Hypothetical Pharmacological MRI Study

The hypothesised design is for two parallel groups of patients, randomised either to active treatment (group A) or placebo (group B), to be scanned twice (i.e., two scans per subject) with the proportions of the N subjects in each group scanned at the two centres, X and Y , denoted as Q and $(1 - Q)$, respectively.

Power is generally the probability of rejecting the null-hypothesis when it is false. In the hypothesised study, the factors that influence the power are: (a) the expected treatment effect size, δ ; (b) the expected uncertainty in estimating the treatment effect caused by between-session and between-centre variability and measurement error and (c) the type one error α (probability of rejecting the null model when it is true); typically $\alpha = 0.05$.

Assuming a standard normal distribution, the power is given by [Paintadosi, 1997]:

$$\text{power} = \int_{-\infty}^x \frac{1}{(2\pi)^{0.5}} \exp\left(-\frac{t^2}{2}\right) dt \quad (1)$$

where

$$x = \frac{\hat{\delta}}{\text{SE}(\hat{\delta})} - Z_{\alpha} \quad (2)$$

Here $\text{SE}(\hat{\delta})$ is the standard error of the estimated treatment effect $\hat{\delta}$ and Z_{α} is the Z-score at α ($Z_{\alpha} = 1.96$, for $\alpha = 0.05$).

Denoting the difference in the BOLD response at baseline (Session 0) and follow-up (Session 1) following treatment of the q th subject in group A as $\Delta A_q = A_{q1} - A_{q0}$ and similarly for the q th subject in group B as $\Delta B_q = B_{q1} - B_{q0}$, then the variance, $V(\hat{\delta})$, of the treatment effect, $\hat{\delta} = \overline{\Delta A} - \overline{\Delta B}$, at a specific cortical location is:

$$\begin{aligned} V(\hat{\delta}) &= V(\overline{\Delta A} - \overline{\Delta B}) = V(\overline{\Delta A}) + V(\overline{\Delta B}) \\ &= \frac{1}{N^2} \left(V\left(\sum_q \Delta A_q\right) + V\left(\sum_q \Delta B_q\right) \right) \\ &= \frac{1}{N^2} \left(V\left(\sum_q A_{q1} - A_{q0}\right) + V\left(\sum_q B_{q1} - B_{q0}\right) \right) \end{aligned} \quad (3)$$

with

$$\begin{aligned} V\left(\sum_q A_{q1} - A_{q0}\right) &= N(V(A_{q1} - A_{q0})) \\ &= N(V(A_{q1}) + V(A_{q0}) + 2\text{Cov}(A_{q1}, A_{q0})) = 2N(\sigma_e^2 + \sigma_r^2) \end{aligned} \quad (4)$$

and

$$V\left(\sum_q B_{q1} - B_{q0}\right) = 2N(\sigma_e^2 + \sigma_r^2) \quad (5)$$

Substituting Eqs. (4) and (5) into Eq. (3), we have:

$$V(\hat{\delta}) = \frac{4}{N}(\sigma_e^2 + \sigma_r^2) \quad (6)$$

Let us assume that the effect size $\hat{\delta}$ and the between-session variability σ_r^2 are the same at both centres. We can let $\sigma_e = \sqrt{\sigma_{eX}^2 \sigma_{eY}^2}$ denote the geometric mean of the error variances in centre X (σ_{eX}^2) and centre Y (σ_{eY}^2) and $r^2 = \frac{\sigma_{eX}^2}{\sigma_{eY}^2}$ denote the ratio of the error variances at the two centres. The total variance at each centre is then:

$$V_X(\hat{\delta}) = \frac{4}{NQ}(r\sigma_e^2 + \sigma_r^2) \quad (7)$$

$$V_Y(\hat{\delta}) = \frac{4}{N(1-Q)}\left(\frac{\sigma_e^2}{r} + \sigma_r^2\right) \quad (8)$$

and the overall standard error of the treatment effect for the two-centre study design is:

$$\text{SE}(\hat{\delta})^2 = V(\hat{\delta}) = \left(\frac{1}{V_X(\hat{\delta})} + \frac{1}{V_Y(\hat{\delta})} \right)^{-1} \quad (9)$$

Substituting Eqs. (7) and (8) into Eq. (9), we have:

$$SE(\hat{\delta})^2 = \left(\left(\frac{4}{NQ} (r\sigma_e^2 + \sigma_r^2) \right)^{-1} + \left(\frac{4}{N(1-Q)} \left(\frac{\sigma_e^2}{r} + \sigma_r^2 \right) \right)^{-1} \right)^{-1} \quad (10)$$

Thus estimating the standard error of the effect size, for arbitrary sample size and proportion of subjects scanned at each centre, requires estimates of the between-session variance, σ_r^2 and the error variance at each centre, σ_{eX}^2 and σ_{eY}^2 (or equivalently r and σ_e^2). The standard error can then be substituted into Eqs. (1) and (2), together with a prior or arbitrary estimate of the effect size $\hat{\delta}$ to estimate the power of the design. Bearing in mind that the effect of repeating the measurement J times within each session will be to reduce the mean measurement error by a factor J , we can also use these results to explore the impact of repeating task presentation J times within a session on the power of the design.

Experimental Study Design

Data acquisition was conducted in two arms. In the first (Arm 1), data from six subjects were collected at Cambridge (University Department of Radiology, BUPA Lea Hospital, Impington, Cambridge) and again 2 days later in London (Centre for Neuroimaging Sciences, Maudsley Hospital, London). In the second (Arm 2), six different subjects were scanned first at London and again 3 days later Cambridge. In both arms all the experimental procedures were identical. The order and time-of-day subjects were scanned was maintained between centres. Subjects were asked to refrain from alcohol and caffeine 12 h prior to scanning.

Within each scanning session at a particular centre, blocked periodic and event-related versions of the facial affect processing paradigm were followed by a resting state scan and then the two affect processing paradigms were repeated on a second occasion in reverse order. The order of presentation of different versions of the affect processing paradigm was reversed during scanning at the second centre so that the overall design was counter-balanced for order of task presentation.

All participants provided written, informed consent. The project was approved by the Ethics Research Committees of the Institute of Psychiatry, London, UK and by the Addenbrooke's NHS Trust Local Research Ethics Committee, Cambridge, UK.

Subjects

Twelve, right-handed males were recruited locally by advertisement. Inclusion criteria were no personal history of psychiatric or neurological disorders, asthma or cigarette smoking or substance misuse syndromes, no currently prescribed medication for medical disorders and no contraindications to MRI. Six subjects were assigned to Arm 1 (mean age = 27.79 years, range = 23.90–38.66 years) and six to Arm 2 (mean age = 29.25 years, range =

25.25–33.33 years) with no between-arm difference in age ($t = 0.553$, $p = 0.31$, $df = 10$).

Data Acquisition

Identical imaging sequences were acquired on identical GE 1.5T Twinspeed Signa MRI scanners installed at both centres, running version 11.0M4 software. Using the whole body (23 mT/m) gradients, and a quadrature conventional birdcage coil for transmission and reception, sixteen noncontiguous axial planes (parallel to the AC-PC line) of gradient-echo echo-planar imaging data were collected at both centres with the following acquisition parameters: repetition time, TR = 2000 ms; echo time, TE = 40 ms; flip angle = 70 degrees; number of excitations, NEX = 1; section thickness = 7 mm; section skip = 0.7 mm; in-plane resolution = $3 \times 3 \text{ mm}^2$ (field-of-view = 240 mm); in-plane image dimensions = 64×64 . Auto-shimming and phase correction were both used.

Implicit Sad Facial Affect Processing Paradigm

Ten faces (4 male) from a standardized series of facial expressions of sadness [Fu et al., 2004] were morphed to represent low, medium and high intensities of sadness. For the block-designed paradigm, five faces of one level of intensity were presented sequentially. Each presentation was 5.2-s long resulting in blocks of 26 s in length. There were four repeats of each block such that each of the 10 faces was presented twice. There were also four blocks of the same duration of a baseline task (crosshair visual fixation). The length of the paradigm was 6 min 56 s.

For the event-related paradigm, the same facial stimuli and baseline trials were presented in random order. Each facial stimulus was presented twice at each intensity of sadness (60 faces in total, equivalent to the block design), along with 20 baseline trials (crosshair visual fixation), for a total of 80 trials. Trial duration was randomly varied according to a Poisson distribution with mean time of 5.2 s. The length of the paradigm was also 6 min 56 s.

Irrespective of paradigm design, for each facial trial, subjects were asked to indicate the gender of the face (male or female) by button press; no hand movement was required in response to a baseline trial. Reaction time (latency) and accuracy of the gender decision during imaging were recorded for each trial. 212 scans were continuously acquired during the paradigms, the initial four scans collected prior to the first trial were discarded to allow for T1-relaxation effects.

Resting State Data Acquisition

During acquisition of resting data, participants were instructed to lie quietly in the scanner with their eyes closed. 516 scans were continuously acquired over the

course of 17 min 12 s; the initial four scans were discarded to allow for T1-relaxation effects.

Functional MRI Data Preprocessing

Data were processed using the Camba software library (<http://www-bmu.psychiatry.cam.ac.uk/software/>). Three-dimensional image volumes collected during each paradigm were corrected geometrically for subject head movement by rigid body transform, and then temporally corrected for spin excitation effects [Bullmore et al., 1999; Suckling et al., 2006]. A two-dimensional Gaussian smoothing kernel of 1.5 mm standard deviation was applied to each volume prior to response estimation.

For the facial affect processing paradigm, differential response to facial trials (or blocks) was estimated at each voxel by a general linear model after convolving the design matrix with a canonical haemodynamic response function [Glover, 1999] and including an AR(3) autoregressive process to model correlated noise in the residuals. The model coefficient was divided by its standard error and squared to produce a map of the F statistic indicating magnitude of task-related response in each individual.

For the resting state paradigm, the Hurst exponent was estimated at each voxel by maximum likelihood in the wavelet domain [Fadili and Bullmore, 2002; Maxim et al., 2005].

The resulting whole brain maps of F and H were spatially transformed into the Montreal Neurological Institute (MNI)/Talairach [Talairach and Tournoux, 1988] standard coordinate system. Mappings were calculated from the affine registration of the mean time-series image onto the EPI template (<http://www.fil.ion.ucl.ac.uk/spm>). The standard space was regionally parcellated using a prior anatomical template image in the MNI coordinate system [Tzourio-Mazoyer et al., 2002], which resulted in a set of regional mean values for each of the F and H statistics.

Group activation maps demonstrating regions of significant median activation by the contrast of sad faces versus baseline for block and event related paradigms are shown in Figure 1. Estimates of BOLD responses under simulated conditions of the null-hypothesis were made following wavelet-based permutation of time-series. Maps of these estimates were registered into standard space and median values across subjects were obtained to sample the null-distribution against which the observed data were statistically thresholded using a cluster-based statistic [Bullmore et al., 1999, 2001; Suckling et al., 2006].

Twenty-eight bilateral regions of the template (Fig. 1) were selected based on prior evidence [Fu et al., 2004] implicating these regions in normal affect processing and/or the pathophysiology of depression. The selected homologous regions were: superior orbitofrontal cortex, middle frontal gyrus, middle and medial orbitofrontal cortices, rectus, insula, anterior cingulate cortex, hippocampus,

Figure 1.

Selected slices of MNI template (z slices: 30, 40, 50, 60 and 70) from the group median activation map pooled over occasions and repeats for (a) event related and (b) block paradigm designs. Overlaid are the outlines of the 28 selected parcellated regions. Both maps were statistically thresholded such that the expected number of false positive three-dimensional clusters < 1 .

parahippocampal gyrus, amygdala, caudate, putamen, pallidum and thalamus. Data were also summarised for the network consisting of all 28 regions combined.

Components of Variance Estimated by Fixed Effects Analysis of Variance

Analysis of variance (ANOVA) was used to investigate the main sources of variation in the 28 selected regions as well as the latency and accuracy of behavioural response recorded during the facial affect processing paradigms. The statistical analysis was based on the following fixed effects model:

$$y_{ijklm} = \mu_0 + \mu_i + \beta_k + \gamma_l + \omega_j + \varepsilon_{ijklm} \quad (11)$$

where y_{ijklm} is the outcome for subject i measured at occasion j at centre k in paradigm type l (block-designed or event-related) and at session m (for Arm 1, $m = 1$ for Cambridge and $m = 2$ for London and vice versa for Arm 2); μ_0 is the overall intercept; μ_i is a fixed effect for subject i ; β_k is the fixed effect for centre k ; ω_j is the fixed effect for occasion j ('habituation'); γ_l is the fixed effect for paradigm type l and ε_{ijklm} is the residual error. The residual error can be interpreted as the measurement error associated with scanning after accounting for subject, centre and paradigm type.

For facial affect processing paradigms, the model [Eq. (11)] was fitted for each region of interest. The sum-of-squares for each of the components of the model was calculated and expressed as a proportion of the total sum-of-squares to assess its relative contribution to the overall observed variance. This variance partitioning was also done for the mean reaction time and accuracy recorded during presentation of the stimuli. For the fixed effects of

centre, paradigm and occasion, t -values were obtained along with the associated P -values.

For the resting state paradigm, analysis of H followed a similar model [Eq. (11)], but without the effects for paradigm type, γ , and occasion, ω . The t -values and P -values were calculated for the remaining fixed effect of centre.

Components of Variance Estimated by Random Effects ANOVA

The data were also regressed onto a more complex random effects model¹ splitting the variation into a within-subject between-session component λ_{im} and the remaining measurement error:

$$\begin{aligned} y_{ijklm} &= \mu_0 + \mu_i + \beta_k + \gamma_l + \lambda_{im} + \varepsilon_{ijklm} \\ \lambda_{im} &\approx N(0, \sigma_r^2) \\ \varepsilon_{ijklm} &\approx N(0, \sigma_e^2) \end{aligned} \quad (12)$$

Estimates of the between-session variance, σ_r^2 , pooled across the centres and the error variance, σ_e^2 , for each centre, were obtained from Eq. (12). These were later substituted in Eq. (10) to estimate the standard error of the effect size for arbitrary N and Q .

Both models [Eqs. (11) and (12)] were fitted to each of the dependent variables independently using mixed model software lme [Pinheiro and Bates, 2000] in the R library of statistical software (<http://www.r-project.org/>).

Estimation of Treatment Effect Size, $\hat{\delta}$

The effect size was estimated from fMRI data previously collected at the London centre on two groups of subjects viewing an event-related, facial affect processing paradigm identical to the event-related design in this study [Fu et al., 2004]. In this study, a group of $N = 19$ subjects with untreated depression was given fluoxetine (20 mg daily) over a period of 8 weeks. Functional MRI scans were acquired from all subjects at the beginning and end of this period. A group of $N = 19$, age-matched, healthy volunteers, who took no medication, were also scanned 8 weeks apart. These data were processed in an identical manner to that described earlier. A significant group $\times$ time interaction was observed in regions coincident with the 28 regions selected from the parcellated template. Mean F responses from these regions were used to provide a multi-regional estimate of the antidepressant treatment effect,

¹In the fMRI literature the use of terms ‘fixed’ and ‘random’ effects have taken the meaning of ‘subject-level’ and ‘group-level’ effects, respectively. Here the terms are used in the more general statistical sense. A fixed effect refers to a variable with specific levels, in this case including the subject effect, as we are interested only in longitudinal effects, within-subject. A random effect has levels drawn from a population; in this case it is not a specific repeat that is of interest, rather the additional variability induced when repeating.

$\hat{\delta} = 1.3$. However, we note that the calculations of statistical power described later could be generalised to any estimate of effect size at the discretion of the investigators planning a study.

RESULTS

Components of Variance: Fixed Effects Analysis

The proportional sum-of-squares for each of the components of the fixed effects ANOVA model for the facial affect processing paradigms are shown graphically in Figure 2 for the mean reaction time and accuracy and the fMRI response in each of the 28 parcellated regions.

By way of summary, means and standard deviation of the normalised sum-of-squares from the combined 28 parcellated regions were: subject, $23\% \pm 9\%$; centre, $2\% \pm 2\%$; occasion, $2\% \pm 2\%$; paradigm type, $4\% \pm 3\%$; residual, $69\% \pm 9\%$.

The component of variance in H attributable to subject differences accounted for 99% of total variance in all 28 regions; variance attributable to both centre and residual error were thus $<0.1\%$.

Significance of Fixed Effects of Centre, Occasion and Paradigm

Table I illustrates the P -values for the BOLD responses to facial affect processing paradigms in each of the 28 parcellated brain regions as well as the network of all regions.

Significant ($P < 0.05$) effects of paradigm were seen for accuracy (greater for the blocked paradigm) and significant centre and occasion effects were seen for reaction time (faster in London and on the second occasion). In brief, the BOLD response from the network as the dependent variable demonstrated a significant paradigm effect (greater activation for the event related design) and a trend ($P < 0.1$) centre effect (greater activation at Cambridge) was also observed. No significant effect of centre was observed from the combined regions with the Hurst exponent as the dependent variable ($t = 1.18$, $P = 0.26$).

Components of Variance: Random Effects ANOVA

Estimates for the parameters of the random effects ANOVA model [Eq. (12)] were obtained for each of the 28 parcellated regions in data acquired during facial affect processing. For the subsequent power calculations median values were used.

Pooling across centre and paradigm, the median between-session variance, $\sigma_r^2 = 0.0015$ and median error variance, $\sigma_e^2 = 0.78$. Median values of σ_e^2 were obtained for each paradigm and centre separately. The ratio between centre-specific error variances for each of the paradigms were: $r = 0.84$ (σ_e^2 greater for Cambridge than London) for the block paradigm design and $r = 0.79$ (σ_e^2

Figure 2.

Sum-of-squares for the four components (subject, centre, paradigm type and residual) of the fixed-effects ANOVA model, Eq. (11), normalized by the total sum-of-squares for the BOLD response to facial affect processing paradigms from the 28 parcellated regions, as well as reaction time and accuracy.

greater for London than Cambridge) for the event related design.

Power Calculations

Power [Eq. (1)] was plotted as a function of the number of subjects in each treatment group, N . A separate plot was generated representing different values of the error variance ratio, $r = 1.0, 0.8, 0.5$ and 0.25 . Within each plot a family of 10 curves was drawn corresponding to the deciles of patients scanned at a particular centre; that is, $Q = 1.0, 0.9, 0.8, \dots, 0.2, 0.1, 0.0$. Figure 3a assumes a single within session occasion of the facial affect processing paradigm per session, $J = 1$, Figure 3b assumes $J = 2$.

DISCUSSION

Power Calculations for Hypothetical fMRI Study Design

Power calculations are an essential part of developing study protocols. This article presents a theoretical treat-

ment to estimate the number of subjects in each group at each of two centres required to detect an effect in a parallel-group, repeated-measures study design. This analysis shows that the power of the experiment is inversely proportional to the standard error of the effect size, which is a function of the geometric mean of the measurement error over both centres, σ_e^2 ; the ratio of the measurement errors at the two centres, r ; the between-session variance pooled across centres, σ_r^2 ; the total number of subjects in each group N and the proportion of subjects in each group scanned at each centre Q . Since the mean measurement error is a function of the number of times measurement is repeated within each session J , this represents another potential factor influencing the power of the study.

The results of our experimental study showed that the between-centre variability was small (2% of total variance by fixed effects ANOVA) and the measurement error at the two centres were similar. Averaging over the two paradigm designs, the ratio of error variances at the two centres was approximately $r = 0.80$ (ignoring the direction of the ratio). If we assume initially that we will acquire

TABLE I. P-Values of fixed effects of ANOVA model [Eq. (11)] for dependent variables: accuracy, reaction, and BOLD response in each of the selected 28 regions of the template as well as over the network of the combined regions for the facial affect processing paradigm

Region	Paradigm type	Centre (F)	Occasion	Centre (H)
Front sup. orb. L	5.15E-04*	0.313	2.76E-03	0.293
Front sup. orb. R	0.006	0.460	0.008	0.330
Front middle L	0.046	0.608	0.413	0.330
Front middle R	0.084	0.334	0.400	0.092
Front middle orb. L	5.44E-04*	0.791	0.097	3.97E-03
Front middle orb. R	0.041	0.935	0.288	0.085
Front medial orb. L	0.011	0.878	0.013	0.721
Front medial orb. R	0.085	0.105	0.046	0.350
Rectus L	4.75E-03	0.120	0.651	0.915
Rectus R	0.070	0.595	0.843	0.409
Insula L	0.987	0.249	0.339	0.412
Insula R	0.126	0.160	0.912	0.512
Cingulum ant. L	0.021	0.073	0.198	0.172
Cingulum ant. R	0.073	0.209	0.156	0.167
Hippocampus L	4.55E-03	0.041	0.168	0.239
Hippocampus R	4.73E-03	0.009	0.973	0.361
Parahippocampus L	0.073	0.158	0.410	0.590
Parahippocampus R	0.301	0.319	0.212	0.785
Amygdala L	0.966	0.008	0.261	0.700
Amygdala R	0.924	0.015	0.746	0.400
Caudate L	0.520	0.758	0.252	0.718
Caudate R	0.362	0.090	0.295	0.374
Putamen L	0.016	0.259	0.379	0.162
Putamen R	1.58E-03*	0.144	0.565	0.137
Pallidum L	0.505	0.969	0.980	0.469
Pallidum R	0.129	0.044	0.011	0.017
Thalamus L	0.592	0.219	0.012	0.731
Thalamus R	0.373	0.808	0.142	0.246
Accuracy	2.75E-25**	0.216	0.891	N/A
Reaction time	0.643	0.034**	0.001**	N/A
Network	0.018**	0.081***	0.203	0.26

P-Values are also shown for the centre fixed-effect with H as the dependant variable.

*Significant effects at $P < 0.002$ (Bonferroni corrected for 28 regions).

**Significant effects at $P < 0.05$.

***Trend effects at $P < 0.1$.

data only on one occasion per session ($J = 1$), the corresponding power curve of Figure 3a indicates that, compared with an equal split of subjects across the two-centres ($Q = 0.5$), scanning all subjects at the centre with the lowest measurement error ($Q = 1.0$) would require 4 less subjects ($N = 24$ reduced to 20) in each group to achieve the same power (arbitrarily, 0.8). Conversely, scanning all subjects at the centre with the highest variance ($Q = 0.0$) would require 6 more subjects ($N = 24$ increased to 30) to be scanned in each group.

Substituting the same estimated parameters in the power calculation, we can also consider the situation when the activation paradigm is repeated twice in each session (i.e., $J = 2$), so that the error variance σ_e^2 is reduced by 50%.

The corresponding power curve of Figure 3b predicts smaller N overall and a smaller change in the number of subjects relative to equal numbers at each centre: for $Q = 1.0$, N is reduced from 11 to 10; and for $Q = 0.0$, N is increased from 11 to 13. The considerable improvements in power and robustness to multicentre data acquisition achieved by task repetition within session are potentially valuable given the modest marginal cost associated with repeating a short experimental task.

Whilst these changes in subject size as a result of scanning at two centres rather than one appear reassuringly small, it is worth bearing in mind that between-centre differences in the measurement error can profoundly alter these results. Figure 3a,b demonstrates that for, say, $r = 0.50$, the worst-case scenario ($Q = 0.0$) can lead to increases in N by around 25 subjects for $J = 1$ and 10 subjects for $J = 2$. This emphasises the need for any investigator contemplating a multicentre study to verify experimentally the local measurement errors and their ratio between centres.

Previous power calculations for fMRI [Desmond and Glover, 2002] have accurately estimated sample sizes in somatosensory paradigms from estimates given by working memory paradigms. The effect size they considered was the difference between active and baseline tasks of the paradigm and is clearly different to the between-group effect size that is the focus of this work. These authors included in their calculations the intrasubject variance component that we have omitted from our statistical treatment. They noted that the benefits of increasing the number of scans (and thus reducing the intrasubject variance) diminished after acquisition of more than 100 scans. The facial affect processing paradigms were collected over twice this number of scans, but this component of variance may become an important parameter for studies entailing shorter periods of data collection.

The network of regions included in the analysis overlap with the areas of regional activation generated by the paradigm (Fig. 1). Within these regions the largest components of variance for the BOLD responses—between-subject variability and measurement error—together accounted for 92% of the overall variance. Subjects recruited for this study were all male, right-handed and of similar ages. This is a common demographic seen in many studies involving fMRI, and indeed other biomedical or psychological studies. Nevertheless, from the evidence provided by these data, as much effort as possible should be invested into reducing subject variance as it is by far the largest factor over which experimenters have some degree of control.

The changes in the precise proportions of variance components across the regions (Fig. 2) suggest an uneven spatial relationship between measurement error and between-subject variability. However, on closer inspection, homologous regions have similar values. The ratio of the measurement errors in left and right regions was compared with an $F(82,82)$ distribution. A single positive test was

Figure 3.

Power curves for facial affect processing paradigms as a function of the number of subjects, N , in each treatment group. A plot is given for square-root ratios of error variances, $r = 0.25, 0.5, 0.8$ and 1.0 . Within each plot, curves are drawn representing the proportions of subjects attending the centre with lowest error variance; that is, $Q = 0.0, 0.1, 0.2 \dots 1.0$. Two circumstances are considered: (a) one ($J = 1$) and (b) two ($J = 2$) acquisitions per scanning session.

observed with $P < 0.004$ (Bonferroni corrected for 14 tests), consistent with the number expected under the null-hypothesis. Variance components due to centre, occasion and paradigm were small for the entire set of regions considered on average (2, 2 and 4%, respectively) as well as for each region individually. In particular, the proportional variance attributable to centre effects was much less than for subjects (23%). Significant effects of paradigm (Table I) were observed in 3 regions at $P < 0.002$ (Bonferroni corrected for 28 regions). At this level of significance there were no regional effects of centre or occasion (Table I). Nevertheless, across the entire network of regions there was a trend effect of centre (Table I).

Behavioural measures also exhibited a number of effects. Paradigm type accounted for a large component of variance in accuracy (Fig. 2) and a significant difference between the blocked and event related versions was also observed with this measurement (Table I). Presumably, this was because of the unanticipated nature of the presentation of faces in the event related design with pseudo-random interstimulus-intervals, rather than the paced presentation with the block paradigm. Significant occasion effects were seen on reaction time presumably due to beneficial effects of practice on speed of task performance. Significant centre effects were also observed on reaction time. This is an unexpected result as the stimulus delivery equipment; button-boxes for responses and associated hardware were designed, installed and operated by the same team at both centres. Moreover, the study was counterbalanced specifically to avoid order effects. The causes of this difference are thus subtler, but nevertheless observed in reaction times for both paradigms and in parallel to the trend to a centre effect on BOLD response (Table I). In contrast to the trend centre effects seen with responses to the cognitive paradigm, analyses based on H did not reach significance (Table I), with >99% of the total variance explained by measurement error.

Relationship to Previous Studies

A previous study of multicentre fMRI reliability [Zou et al., 2005] has used the number of activated voxels as a primary metric. Whilst this is indeed an important measure ensuing from fMRI data processing, it is threshold-dependent, subject to assumptions on the parametric form of null hypothesis and the assumption that this distribution is invariant across centres. Furthermore, the set of detected voxels contains a number of type I errors and the exact value of the probabilistic threshold determines the well-known compromise between false-positives and sensitivity. In place of this inference procedure, we chose to extract mean values of activation from predefined anatomical regions known to be involved in the task. In doing so, some sensitivity was inevitably lost, as the selected regions of the anatomical template do not conform exactly to the activation pattern elicited by the task. However, the sam-

ples of H and F are unbiased in the sense that they are extracted from regions defined by independent criteria.

The facial affect processing paradigm was selected for this study because it had demonstrable success in detecting differences in studies of antidepressant treatment [Fu et al., 2004] and has also been used in the patient groups that may benefit from such an intervention [Mayberg et al., 1999]. The regions selected from the automated anatomical template included ventrolateral and medial prefrontal cortex that are reciprocally connected to limbic and subcortical regions (including anterior cingulate, amygdala-hippocampus complex, basal ganglia, thalamus, insula and brainstem), which modulate emotional processing, bodily sensations and response [Ongur and Price, 2000]. It has been suggested that over-activation of limbic and subcortical structures important for emotional responses, and dysfunctional frontal regions important for effortful emotion regulation, may be associated with the biological substrates of major depressive disorder [Phillips et al., 2003]. With this pattern of activation, facial affect processing is a candidate paradigm for future pharmacological trials of antidepressant drugs and the power calculations ensuing from these data could be extended to predict appropriate sample sizes. Furthermore, it is a straightforward paradigm for subjects to understand and perform and should thus be accurate and reliable. Nevertheless, paradigms containing emotional stimuli have been associated with habituation effects over repeated presentations, resulting in poor reproducibility [Phan et al., 2003; Stark et al., 2004]. However, no occasion effects were observed on the mean BOLD activation (Table I) produced by the facial affect processing paradigm.

The Hurst exponent characterises the spectral noise properties of the fMRI signal and is estimated via a maximum likelihood procedure, implemented in the wavelet domain. This algorithm has previously been described and validated in detail, and evaluated in terms of its bias and efficiency relative to several other possible estimators of H in time and wavelet domains [Fadili and Bullmore, 2002; Maxim et al., 2005]. Moreover, it has been shown to be substantively informative as a sign of pathophysiological change in early Alzheimer's disease [Maxim et al., 2005] as well as indicative of changes in normal ageing and the effect of pharmacological manipulation with scopolamine [Wink et al., 2005].

From these data, it appears that large between-subject and measurement errors dominate fMRI measurements. This has consequences for the ability of fMRI to detect between-group differences in subject groups where strong control of the participants' demographics is not possible. Comprehensive efforts were made to harmonise stimulus delivery at both centres and counterbalancing was used in the study design. Nevertheless, although appearing as small components of variance, centre and paradigm effects on task-related activation were seen, reflecting behavioural differences in reaction time and accuracy. However, a centre difference was not observed with H . Resting state

acquisitions are simple to acquire and are less influenced by local conditions than cognitive paradigms. This simplicity and reliability makes resting state data acquisition a good candidate paradigm for multicentre calibration studies.

CONCLUSIONS

We have shown theoretically how the power of a multicentre, parallel group, repeated measures fMRI design can be predicted; and we have shown experimentally how key components of variance can be estimated to inform these power calculations. The theoretical results are generally applicable (regardless of the test statistics used to detect activation) and our experimental data indicate that it is possible to control between-centre variability and differences in measurement error to such an extent that multicentre data acquisition is not compromised by the need to scan a much larger total number of subjects (especially if the task-related activation is measured more than once in each scanning session).

As we have shown, for arbitrary power and effect size and based on local estimates of components of variance, the predicted subject numbers for a two group study, assuming equal recruitment at the London and Cambridge centres and using a facial affect processing paradigm as the PD measure, would increase by a maximum of 12 subjects in total over the sample size required if all scanning was conducted at the centre with lowest measurement error. By including a second within-session scan ($J = 2$), the worse case increase in subjects is just four in total. In both cases, the additional subjects required to compensate for data acquisition over both centres amount approximately to a 10% increase in total sample size, which could be economically attractive if recruitment rate is substantially increased by sampling from the larger geographical catchment area of south-east England subtended by these two centres.

ACKNOWLEDGMENTS

EB is employed 50% by GlaxoSmithKline and 50% University of Cambridge.

REFERENCES

- Anderson CM, Lowen SB, Renshaw PF (2006): Emotional task-dependent low-frequency fluctuations and methylphenidate: Wavelet scaling analysis of 1/f-type fluctuations in fMRI of the cerebellar vermis. *J Neurosci Methods* 151:52–61.
- Bullmore E, Long C, Suckling J, Fadili J, Calvert G, Zelaya F, Carpenter TA, Brammer M (2001): Colored noise and computational inference in neurophysiological (fMRI) time series analysis: Resampling methods in time and wavelet domains. *Hum Brain Mapp* 12:61–78.
- Bullmore ET, Suckling J, Overmeyer S, Rabe-Hesketh S, Taylor E, Brammer MJ (1999): Global, voxel and cluster tests, by theory and permutation for a difference between two groups of structural MR images of the brain. *IEEE Trans Medical Imaging* 18:32–42.
- Desmond JE, Glover GH (2002): Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses. *J Neurosci Methods* 118:115–118.
- Fadili J, Bullmore ET (2002): Wavelet-generalized least squares: A new BLU estimator of linear regression models with $1/f$ errors. *Neuroimage* 15:217–232.
- Friedman L, Glover GH (2006a): Report on a multicenter fMRI quality assurance protocol. *J Mag Reson Imaging* 23:827–839.
- Friedman L, Glover GH, for the FIRST BIRN (2006b): Reducing interscanner variability of activation in a multicenter fMRI study: Controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. *Neuroimage* 33:471–481.
- Friedman L, Glover GH, Krenz D, Magnotta V, for the FIRST BIRN (2006): Reducing inter-scanner variability of activation in a multicenter fMRI study: Role of smoothness equalization. *Neuroimage* 32:1656–1668.
- Fu CHY, Williams SCR, Cleare AJ, Brammer MJ, Walsh ND, Kim J, Andrew CM, Pich EM, Williams PM, Reed LJ, Mitterschiffthaler MT, Suckling J, Bullmore ET (2004): Attenuation of the neural response to sad faces in major depression by antidepressant treatment. *Arch Gen Psychiatry* 61:877–889.
- Glover GH (1999): Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage* 9:416–429.
- Honey G, Bullmore E (2004): Human pharmacological MRI. *Trends Pharmacol Sci* 25:366–374.
- Maxim V, Sendur L, Fadili J, Suckling J, Gould R, Howard R, Bullmore E (2005): Fractional Gaussian noise, functional MRI and Alzheimer's disease. *Neuroimage* 25:141–158.
- Mayberg HS, Liotti M, Brannan SK, McGinnis S, Mahurin RK, Jerabek PA, Silva JA, Tekell JL, Martin CC, Lancaster JL, Fox PT (1999): Reciprocal limbic-cortical function and negative mood: Converging PET findings in depression and normal sadness. *Am J Psychiatry* 156:675–682.
- Ongur D, Price JL (2000): The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys and humans. *Cereb Cortex* 10:206–219.
- Paintadosi S (1997): *Clinical Trials: A Methodological Perspective*. Hoboken, NJ: Wiley.
- Phan KL, Liberzon I, Welsh RC, Britton JC, Taylor SF (2003): Habituation of rostral anterior cingulate cortex to repeated emotionally salient pictures. *Neuropsychopharmacology* 28:1344–1350.
- Phillips ML, Drevets WC, Rauch SL, Lane R (2003): Neurobiology of emotion perception. II. Implications for major psychiatric disorders. *Biol Psychiatry* 54:515–528.
- Pinheiro JC, Bates DM (2000): *Mixed-Effects Models in S and S-Plus*. Berlin: Springer-Verlag.
- Stark R, Schienle A, Walter B, Kirsch P, Blecker C, Ott U, Schäfer A, Sammer G, Zimmermann M, Vaitl D (2004): Hemodynamic effects of negative emotional pictures: A test-retest analysis. *Neuropsychobiology* 50:108–118.
- Statistical principles for clinical trials (1998): International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use (ICH). ICH-E9 (Step 4).
- Suckling J, Davis MH, Ooi C, Wink AM, Fadili J, Salvador R, Welchew D, Sendur L, Maxim V, Bullmore ET (2006): Permutation testing of orthogonal factorial effects in a language-processing experiment using fMRI. *Hum Brain Mapp* 27:425–433.

- Talairach J, Tournoux P (1988): A coplanar stereotaxic atlas of the human brain. Stuttgart: Thieme Verlag.
- Turner J, Potkin SG, Brown GG, Glover GH, Fallon JH, McCarthy G, Lim KO, Gollub RL, Rosen BR, Toga AW, Kikinis R, Pieper S, Dale AM, Lauriello J, O'Leary DS, Lieberman JA, Belger A, Greve DN, Friedman L, Cannon TD, Keator DB, Grethe JS, for the Function BIRN (2005): Multi-center fMRI methods and design: Function BIRN. International Congress of Schizophrenia Research, Savannah, GA.
- Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M (2002): Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15:273–289.
- Wink AM, Bernard F, Salvador R, Bullmore E, Suckling J (2005). Age and cholinergic effects on hemodynamics and functional coherence of human hippocampus. *Neurobiol Aging* 10:1395–1404.
- Zou KH, Greve DN, Wang M, Pieper SD, Warfield SK, White NS, Manandhar S, Brown GG, Vangel MG, Kikinis R, Wells WM, for the FIRST BIRN Research Group (2005): Eproducibility of functional MR imaging: Preliminary results of prospective multi-institutional study performed by Biomedical Informatics Research Network. *Radiology* 237:781–789.