◆ Human Brain Mapping 31:1183–1195 (2010) ◆

Power Calculations for Multicenter Imaging Studies Controlled by the False Discovery Rate

John Suckling,[1]* Anna Barnes,[1] Dominic Job,[2] David Brenan,[3] Katherine Lymer,[4] Paola Dazzan,[5] Tiago Reis Marques,[5] Clare MacKay,[6] Shane McKie,[7] Steve R. Williams,[8] Steven C. R. Williams,[9] Stephen Lawrie,[2] and Bill Deakin[7]

[1]*Brain Mapping Unit, Department of Psychiatry and Behavioural and Clinical Neurosciences Institute, University of Cambridge, United Kingdom*
[2]*Division of Psychiatry, School of Molecular and Clinical Medicine, University of Edinburgh, United Kingdom*
[3]*Institute of Neurological Science, Southern General Hospital, Glasgow, United Kingdom*
[4]*SFC Brain Imaging Research Centre, Division of Clinical Neurosciences, University of Edinburgh, United Kingdom*
[5]*Division of Psychological Medicine, Institute of Psychiatry, Kings College London, United Kingdom*
[6]*FMRIB Centre, University of Oxford, Oxford, United Kingdom*
[7]*Neuroscience and Psychiatry Unit, University of Manchester, United Kingdom*
[8]*Imaging Science and Biomedical Engineering, University of Manchester, United Kingdom*
[9]*Centre for Neuroimaging Sciences, Institute of Psychiatry, Kings College London, United Kingdom*

Abstract: Magnetic resonance imaging (MRI) is widely used in brain imaging research (neuroimaging) to explore structural and functional changes across dispersed neural networks visible only via multi-subject experiments. Multicenter investigations are an effective way to increase recruitment rates. This article describes image-based power calculations for a two-group, cross-sectional design specified by the mean effect size and its standard error, sample size, false discovery rate (FDR), and size of the network (i.e., proportion of image locations) that truly demonstrates an effect. Minimum sample size (for fixed effect size) and the minimum effect size (for fixed sample size) are calculated by specifying the acceptable power threshold. Within-center variance was estimated in five participating centers by repeat MRI scanning of 12 healthy participants from whom distributions of gray matter were estimated. The effect on outcome measures when varying FDR and the proportion of true positives is presented. Their spatial patterns reflect within-center variance, which is consistent across centers. Sample sizes 3–6 times larger are needed when detecting effects in subcortical regions compared to the neocortex. Hypothesized multicenter studies of patients with first episode psychosis and control participants were simulated with varying proportions of the cohort recruited at each center. There is little penalty to sample size for recruitment at five centers compared to the center with the lowest variance alone. At 80% power 80 participants per group are required to observe differences in gray matter in high variance regions. *Hum Brain Mapp 31:1183–1195, 2010.*

Key words: power calculations; multicenter; false discovery rate; magnetic resonance imaging

Contract grant sponsor: Medical Research Council; Contract grant numbers: G0300610, G0300623, G0601652.

*Correspondence to: John Suckling, Brain Mapping Unit, Department of Psychiatry, University of Cambridge, Herchel Smith Building, Robinson Way, Cambridge CB2 0SZ, UK.
E-mail: js369@cam.ac.uk

Received for publication 28 May 2009; Revised 23 September 2009; Accepted 23 September 2009

DOI: 10.1002/hbm.20927
Published online 8 January 2010 in Wiley InterScience (www.interscience.wiley.com).

INTRODUCTION

Medical imaging has revolutionized the diagnosis and monitoring of several major diseases. The information that imaging techniques provides enables accurate assessment of the morphology of internal structures as well as biophysical and physiological data. Generally, the measurements that are derived from these images (volumes, flow rates, metabolic rates, and so on) have units that can be calibrated and compared directly and are increasingly found as endpoints in a wide range of clinical trials and experimental medicine studies.

Magnetic resonance imaging (MRI) uses magnetic fields and radio frequency irradiation to manipulate the magnetic moments of nuclei (usually the protons of the hydrogen atom in water molecules) to visualize their distribution and magnetic microenvironment and thus generate a high resolution image of anatomical structure. In fact, all kinds of MRI data can be acquired without exposing subjects to ionizing radiation or radioactive isotopes. In the early 1990s, MRI techniques were developed that are sensitive to the oxygenation status of hemoglobin. This has been exploited in functional MRI (fMRI) where changes in the local concentration of deoxygenated hemoglobin are associated with localized neural activity elicited by an experimentally manipulated stimulus [Logothetis, 2008]. Referred to collectively under the rubric of neuroimaging, structural MRI and fMRI have led to a renaissance in psychiatric neuroscience and experimental neuropsychology as a way of noninvasively probing normal and disordered brain structure and function.

What sets MRI apart from other hospital-based imaging modalities such as X-ray computed tomography (CT) or positron emission tomography (PET) is that the intensities recorded do not correspond to any quantitative unit such as opacity or rates of radioactive decay. Furthermore, the presence of an object with magnetic susceptibility (for example, the body of the patient) in the MRI scanner introduces local distortions in the applied magnetic fields causing spatial heterogeneity in signal intensity. Magnetic "shims" and optimized acquisition procedures are used to reduce this effect, but cannot be entirely effective leaving residual, difficult to predict fluctuations across the image. This is not of any particular concern for studies in which MRI data are used to derive secondary measures of the morphology or physiology of tumors or lesions, as these measures have properties measured against an absolute scale. However, psychiatric disorders such as schizophrenia, depression or autism do not present with a clear pattern of radiological signs that is recognizable by visual inspection. Instead, the acquired images are first processed via a pipeline of computational modules that estimate at each image location (voxel) a surrogate image property, commonly an estimate of the composition of tissue types (gray and white matters and cerebrospinal fluid) within a voxel is made from an analysis of the histogram of image intensities [Smith et al., 2004]. Image registration algorithms are also applied to align all processed images into the same stereotactic space within which statistical testing can take place. Observed structural differences may occupy diffuse "networks" including several spatially distant regions. Small effect sizes, resource constraints, and the desire to compress study timelines have led to a rise in the interest of multicenter imaging studies [Friedman et al., 2008; Keator et al., 2008; Schnack et al., 2004] to increase overall recruitment and thus power. However, these gains in precision must be reconciled against the negative impact of acquiring data from several centers and the corresponding inflation of variance arising from the introduction of a between-center component.

In general, between-center biases and increased variance can be mitigated by standardized acquisition protocols, careful quality control and calibration to a benchmark sample. An equivalent procedure for MRI is problematic in that a practical phantom (artificial calibration object) that mimics the inhomogeneity distortions produced by the human head is difficult to manufacture. Furthermore, whilst the internal environment of the living brain is homeostatically controlled, radio frequency energy deposition during scanning causes heating that alters the magnetic resonance properties of the materials from which the phantoms are constructed. In short, the most practical method for assessing within- and between-site variance as a precursor to power calculations is to scan a cohort of healthy normal individuals at each center participating in the study.

Power calculations for neuroimaging are under-represented in the literature and arguably under utilized by the community, although the relatively high cost of scanning and the ethical ambiguity of underpowered studies should compel investigators to ensure they have a sample size adequate to observe the hypothesized effect. The spatial inhomogeneities in MRI signal strength make such predictions more complex and calculations that limit themselves to one or a few regions-of-interest (ROIs) may suggest sample sizes sufficient to identify effects within these small areas, but not necessarily across dispersed neural networks. Moreover, assumed type I error rates are based on a single measurement that does not acknowledge the multiple comparisons issues associated with testing at a large number of brain voxels, which is almost certainly the method that will be deployed in the planned study.

Voxel-based power calculations based on calibration data are a useful tool for planning neuroimaging studies. They have been reported previously for single-center designs based on conservative, but uncorrected type I error rates [Van Horn et al., 1998] as well as values corrected for the spatially correlated structure of parameter estimates derived from the original imaging datasets, described using random field theory [Hayasaka et al., 2007]. In this article, we present voxel-based power calculations for hypothesized multicenter, two-group cross-sectional imaging study designs using estimates of tissue composition from tissue segmentation of high-resolution

TABLE I. Summary of MRI scanner characteristics and T_1-weighted scan acquisition parameters at each center

Center	MRI scanner	Head-coil	Voxel size: *x*, *y*, *z* (mm)	TR (ms)	TE (ms)	TI (ms)	Flip angle (°)
V	Siemens 3T TimTrio	8-channel	1 × 1 × 1	9.0	2.98	900	9
W	GE 3T HD	8-channel	1.1 × 1.1 × 1	6.5	1.50	500	12
X	GE 3T HDx	8-channel	1.1 × 1.1 × 1	7.0	2.85	650	8
Y	Philips 3T Intera-Achieva	8-channel	1 × 1 × 1	8.2	3.80	885	8
Z	Siemens 3T TimTrio	12-channel	1 × 1 × 1	9.4	4.66	900	8

structural MRI. Five centers took part in a calibration study from which within-center variances were estimated. To estimate power, specification of the average effect size (difference in means), δ, the associated standard error and type II error rate, β, are required. However instead of the type I error rate, the model uses the false discovery rate, *q*, an increasingly prevalent tool in neuroimaging for controlling multiple comparisons [Genovese et al., 2002] in addition to the proportion of voxels that are true positives, θ, that estimates the size of neural network implicated. We use our model to explore the effect of δ, *q*, and θ on power for equal numbers of participants at each center. We then fix these values and consider the effects of altering the proportions of participants at each center and their impact on the sample sizes and detectable effect sizes of hypothesized studies in schizophrenia.

MATERIALS AND METHODS

Statistical Power Model

Power is the probability of rejecting the null-hypothesis when it is false. Assuming a standard normal distribution, the power is given by [Paintadosi, 1997]:

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

where

$$x = \frac{\delta}{\mathrm{SE}} - Z_\alpha \tag{2}$$

where SE is the standard error of the expected treatment effect δ (derived below), and Z_α is the *Z*-score at α (Z_α = 1.96, for α = 0.05). A cross-sectional study (i.e., two-sample *t*-test) is hypothesized in which the differences in means of two groups (for example, patient and control groups) of equal size, *N*, are tested with participants recruited at *C* centers with a proportion Q_c at each center; that is:

$$Q_c = \frac{N_c}{\sum_{c=1}^{C} N_c} = \frac{N_c}{N} \tag{3}$$

where N_c is the number of participants recruited in each group at each center (that is, each center recruits an equal number of participants from each group). Then, following a derivation based on a previous treatment [Suckling et al., 2008], the standard error is given by:

$$\mathrm{SE}^2 = \left[\sum_{c=1}^{C} \frac{NQ_c}{2\sigma_c^2}\right]^{-1} \tag{4}$$

where σ_c^2 is the within-center variance and includes both the between-subject and residual error variances at that center.

Estimation of Within-Center Variance

Voxelwise estimates of within-center variance were made from data acquired from healthy volunteers who were scanned at participating centers as part of the PsyGRID consortium (http://www.psygrid.org/) and the NeuroPsyGrid collaborative project (http://www.neuropsygrid.org/): The Wolfson Brain Imaging Center (University of Cambridge), Magnetic Resonance Imaging Facility (University of Manchester), the Institute of Psychiatry (Kings College, London), the Departments of Clinical Neurosciences at the Universities of Edinburgh and Glasgow, and the Center for Clinical Magnetic Resonance Research, (University of Oxford). The study was approved by the University of Manchester Ethics Committee and ratified by the appropriate committees at each of the participating centers.

Twelve male, right-handed healthy volunteers (mean age = 25 ± 6 years, range 20–35 years) with no previous history of head injury or psychiatric disorder gave informed consent to take part. Participants were scanned with a schedule that was not counterbalanced across centers, although, there was no prior expectation of order effects for structural MRI scanning. At each center a T_1-weighted high-resolution three-dimensional image was acquired with the sequence acquisition parameters given in Table I. Participating centers operated contemporary 3T systems from three of the major manufacturers. The variety of technology constrained the consistency of sequence parameter values, but was typical of those generally encountered.

Each image was processed with the same pipeline (http://www.fsl.ox.ac.uk/fsl; [Smith, 2002; Smith et al.,

2004]): The brain was initially identified (using the software: bet2) and extracerebral tissues excluded from further analysis. Partial volume estimates were calculated for gray (in addition to white matter and cerebral spinal fluid, although they are not further discussed in this article) from which each voxel was assumed to be composed (using the software: fast). Maps of the distribution of brain tissues from each participant were coregistered into the standard space of the Montreal Neurological Institute (MNI) using affine transforms (using the software: flirt) based on the matching of the original T_1-weighted image to the equivalent template image. Once completed, each map was smoothed with a Gaussian kernel of standard deviation 2 voxels (i.e., 2.2 mm), a value chosen to be sufficient to ameliorate noise introduced by the segmentation procedure and image registration.

At each intracerebral voxel in standard MNI space the partial volume estimates were regressed onto a random effects model:

$$\begin{aligned} y_{ic} &= \mu_0 + \varepsilon_c \\ \varepsilon_c &\approx N\left(0, \sigma_c^2\right) \end{aligned} \tag{5}$$

for subject i at center c, and the error variance, σ_c^2, separately estimated for each center. This model was fitted to each of the gray matter partial volume estimates using the mixed model software lme [Pinheiro and Bates, 2000] in the R library of statistical software (http://www.r-project.org/).

Type I Error Control by the False Discovery Rate

Controlling the number of type I errors for the multiple hypotheses testing as a result of the large number of voxels in an image has been the subject of considerable discussion [Nichols and Hayasaka, 2003]. An approach that has gained recent widespread use is the false discovery rate (FDR). This method [Benjamini and Hochberg, 1995; Genovese et al., 2002] controls the proportion of those voxels in which a difference is declared as significant that are false positives. A rate q, defined on [0,1], determines the average FDR if the experiment were to be repeated many times. The procedure is adaptive to the distribution of the observed parameter estimates with the equivalent type I error, α, dependent on δ (assumed equal at all voxels) and the proportion of voxels that are true positives, θ, by the following equation [Jung, 2005]:

$$q = \frac{\alpha}{\alpha + \frac{\theta}{1-\theta}\Phi^*(Z_\alpha - \delta/\mathrm{SE})} \tag{6}$$

where $\Phi^*(\ldots)$ is the survival function of the normal distribution (i.e., 1 − cumulative distribution function). It can be shown that q is increasing in α [Jung, 2005] and thus the value of α for any q was derived using the bisection method.

Exploration of the Parameter Space

Power was calculated after providing the following parameters:

- Total number of participants in each group, N
- Proportion of the voxels representing the intracerebral space that truly demonstrate an effect, θ
- Average effect size estimate (difference in means) at a voxel, δ
- False discovery rate, q
- Number of centers, C (5 for the experiment reported).
- Proportion of participants (from each group equally) recruited at each center, Q_c

At each intracerebral voxel α was obtained from Eq. (6), the standard error of the effect was obtained from Eq. (4) using the estimates of within-center variance obtained from the calibration experiment, Eq. (5), and power assessed by substitution of α and SE into Eq. (1).

Additionally, by setting the acceptable level of type II errors (i.e., $\beta = 0.2$) and varying $\hat{\delta}$, the minimum effect size at each voxel that could be observed with fixed sample size was also calculated. In a similar manner, by varying N until the acceptable level of power was exceeded, the minimum sample size required to detect a fixed effect size was obtained.

Initially, maps of power, minimum effect size, and minimum sample size were calculated assuming equal proportions of participants at each center; that is $Q_c = 0.2$ for $c = 1, \ldots, 5$. Based on a meta-analysis of studies that investigated similar brain changes in first-episode psychosis [Ellison-Wright et al., 2008], the effect size was set to a typical value, $\hat{\delta} = 0.075$ and the sample size set to the mean sample size of the studies included in the meta-analysis, $N = 26$.

The range of q had an upper bound of 0.05, typical for neuroimaging studies that invoke FDR [Genovese et al., 2002]. The range of θ was estimated by considering the number of voxels required to represent an overall gray matter difference of 1–5% [Ellison-Wright et al., 2008] between patients with schizophrenia and control participants, as a proportion of the number of voxels representing intracerebral gray matter with mean partial volume occupancy of 0.5 and effect size of $\hat{\delta} = 0.075$.

Thus, the following parameters were set, varying q and θ

- $N = 26$; $\theta = 0.1$; $\hat{\delta} = 0.075$; $q = 0.005, 0.025, 0.05$
- $N = 26$; $q = 0.01$; $\hat{\delta} = 0.075$; $\theta = 0.05, 0.15, 0.25$

Comparison of Study Designs

Initially, one single-center and four multicenter cross-sectional studies were simulated to assess the potential improvements in power as the number of centers included in the study was varied. The purpose was to specify a prospective study for the identification of differences in gray matter distribution associated with the early stages of

psychosis in a population of adult males with an age range comparable to the cohort used for the calibration study (i.e., 20–35 years). We mandated conservative values for $q = 0.01$ and $\theta = 0.05$ and for each recruitment scenario the minimum sample size (with fixed $\hat{\delta} = 0.075$) and minimum effect size (with fixed $N = 26$) required to exceed 80% power ($\beta = 0.2$) were calculated.

The single-center scenario was defined as all participants recruited at the center with the lowest overall (averaged across all intracerebral voxels) within-center variance (i.e., best-case); that is, $Q_c = 1.0$ for the center with the lowest variance and $= 0.0$ for all other centers.

The two-center scenario ($C = 2$) was simulated by including the two centers with the lowest average within-center variance and the proportion of the overall cohort recruited at each was assumed equal (i.e., $Q_c = 0.5$ of the cohort recruited at each participating center). A three-center scenario included the center with the third lowest within-center variance and similarly for the four-center and five-center scenarios, in each case adjusting the recruitment in corresponding equal proportions (i.e., for $C = 3$, $Q_c = 0.33$ at each center; $C = 4$, $Q_c = 0.25$ at each center; $C = 5$, $Q_c = 0.20$ at each center).

Values of both the minimum sample size and minimum effect size were extracted in amygdala, caudate and inferior frontal gyrus bilaterally (Fig. 1) defined by the anatomical parcellation of standard MNI space provided by the automated anatomical labeling template [Tzourio-Mazoyer et al., 2002]. These regions are both consistently associated with changes in gray matter in first episode psychosis and schizophrenia [Ellison-Wright et al., 2008] and sample the range of within-center variances giving lower, intermediate and upper values to subsequent predictions.

A second comparison was made between: best- (all participants scanned at the center with the lowest average within-center variance, as described earlier) and worst-case (all participants scanned at the center with the greatest average within-center variance) single-center scenarios; a five center scenario with equal proportions of the cohort recruited at each center (as described earlier); five center scenarios with a distribution of participants recruited at each center in the proportions: 0.067, 0.133, 0.2, 0.267, 0.333 associated with both the lowest-to-highest ordering (worst-case), and highest-to-lowest ordering (best case) of average within-center variance. Regional average values of minimum effect size and sample size were again extracted from bilateral amygdala, caudate and inferior frontal gyrus to sample across the range of within-center variances.

RESULTS

Within-center variances are shown for selected axial slices of the MNI template in Figure 1. The spatial distribution of within-center variance is broadly comparable between centers, with increased variance in subcortical structures, cerebellum and regions centered on the retrosplenial cortex. The relative average whole-brain within-center variances were: σ_c^2 0.22, 0.18, 0.25, 0.16, 0.19 for centers V to Z (Table I), respectively.

Maps of power, minimum effect size (fixed N) and minimum sample size (fixed δ) are shown as a function of q (Fig. 2a) and θ (Fig. 3a). Notably, there is sufficient power (>80%) to detect an effect size of $\hat{\delta} = 0.075$ with $N = 26$ per group with appreciable coverage of cortical gray matter at the smallest values of q and θ. However, the minimum effect sizes observable (for fixed $N = 26$) and corresponding minimum sample sizes (for fixed $\hat{\delta} = 0.075$) are spatially inhomogeneous and reflect the patterns of within-center variances (Fig. 1). From the ROIs: subcortical regions, where within-center variance is greatest, at $q = 0.025$ and $\theta = 0.1$ typically the minimum effect sizes is $\hat{\delta} \sim 0.13$ (for fixed $N = 26$; Fig. 2b) and minimum sample size is $N \sim 80$ (for fixed $\hat{\delta} = 0.075$; Fig. 2c), whilst in the inferior frontal cortex the respective values are $\hat{\delta} = 0.06$ and $N = 15$.

As expected, as power increases with q (Fig. 2a), the minimum effect sizes that are detected are reduced (Fig. 2b) as are the minimum sample sizes (Fig. 2c) required to achieve 80% power. The proportion of true positive voxels, equivalent to the extent of the hypothesized network of brain differences, has a similar effect on power, increasing with θ. Over the ranges of q and θ considered, changes are observed of around 15% in minimum effect (Figs. 2b and 3b) and samples sizes (Figs. 2c and 3c) with a relative change amplified by increases in the within-center variance between regions. As θ increases the corresponding type I error rate [α; Eq. (6)] becomes more lenient. The within-center variance determines the width of the alternative distribution and overlap (i.e., β) with the null distribution. For changes in α (or equivalently, q and θ) applied to the null distribution, the corresponding changes in β will vary by greater degree when the within-center variance is large and thus the alternative distribution wider. Alternatively, with small within-center variance the critical value lies in the tails of the alternative distribution where small changes in α do not change β to a great extent.

For equal recruitment rates at each center projected minimum sample sizes and minimum effect sizes for the single-center and multicenter simulations are shown in Figure 4 and averages for the ROIs in Figure 5a,b. None of the outcome is strongly dependent on the number of centers included in the design. In the amygdala and caudate, where there is relatively large within-center variance (Fig. 1), there is a small decline in both the minimum number of participants and minimum sample size as the number of centers is increased to two, followed by an increase and subsequently unchanging values with three, four or five centers. This is a reflection of changes in regional variance dissimilar to average changes across other brain regions. Specifically, when $C = 2$ the newly included (second) center has increased variance in amygdala and caudate, whilst variance is reduced in the neocortex leading to smaller minimum effect and samples sizes in these regions.

A comparison of the single- and multicenter best- and worst-case scenarios is shown in Figure 6a,b for the

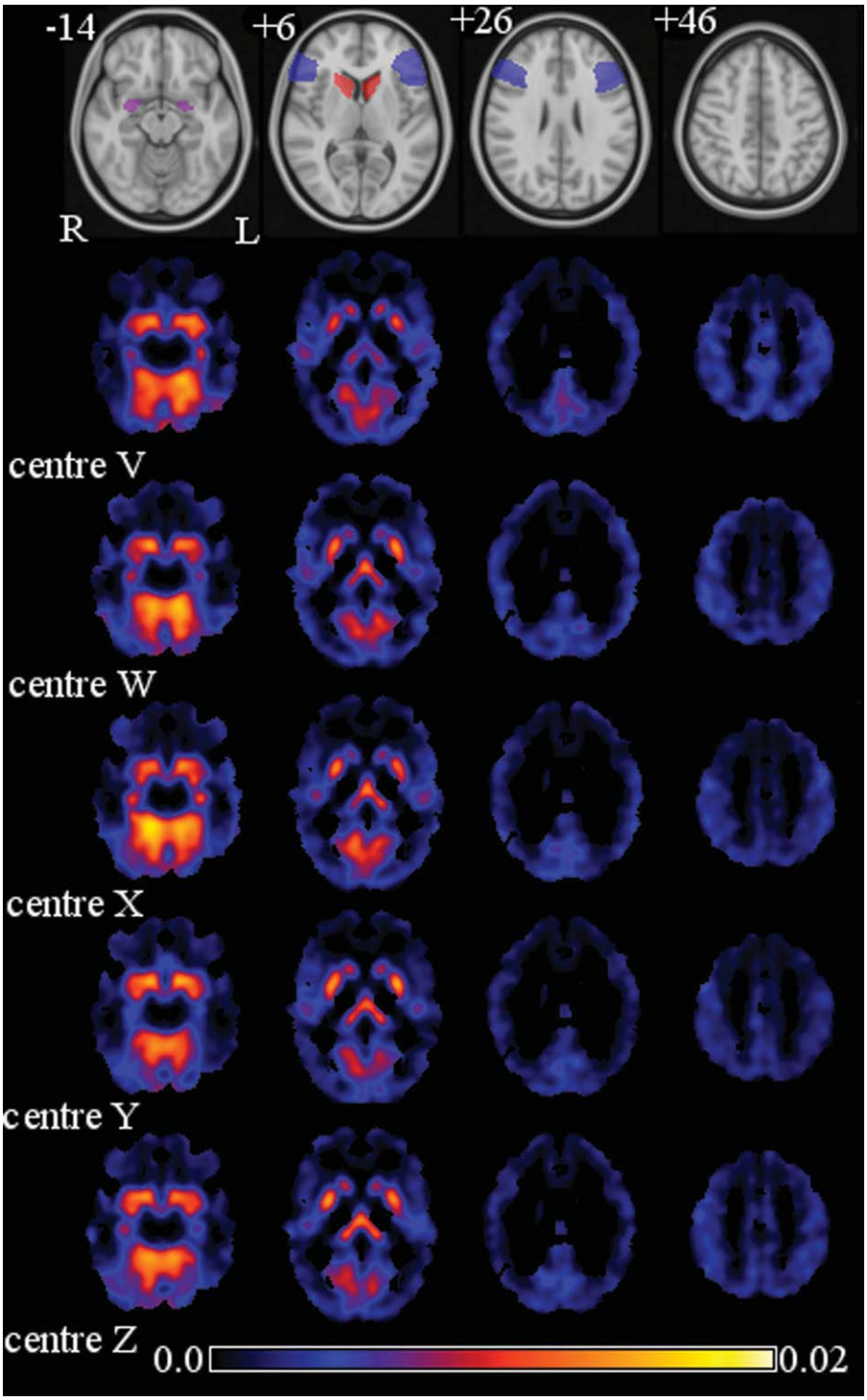

Figure 1.

Voxelwise estimates of within-center variances. The top row shows axial slices of the MNI template (axial slice positions, in mm, are given in the column headings) with overlays for the regions-of-interest: amygdala (purple), caudate (red), and inferior frontal gyrus (blue). The right-hand of the image is the left-hand (L) side of the brain.

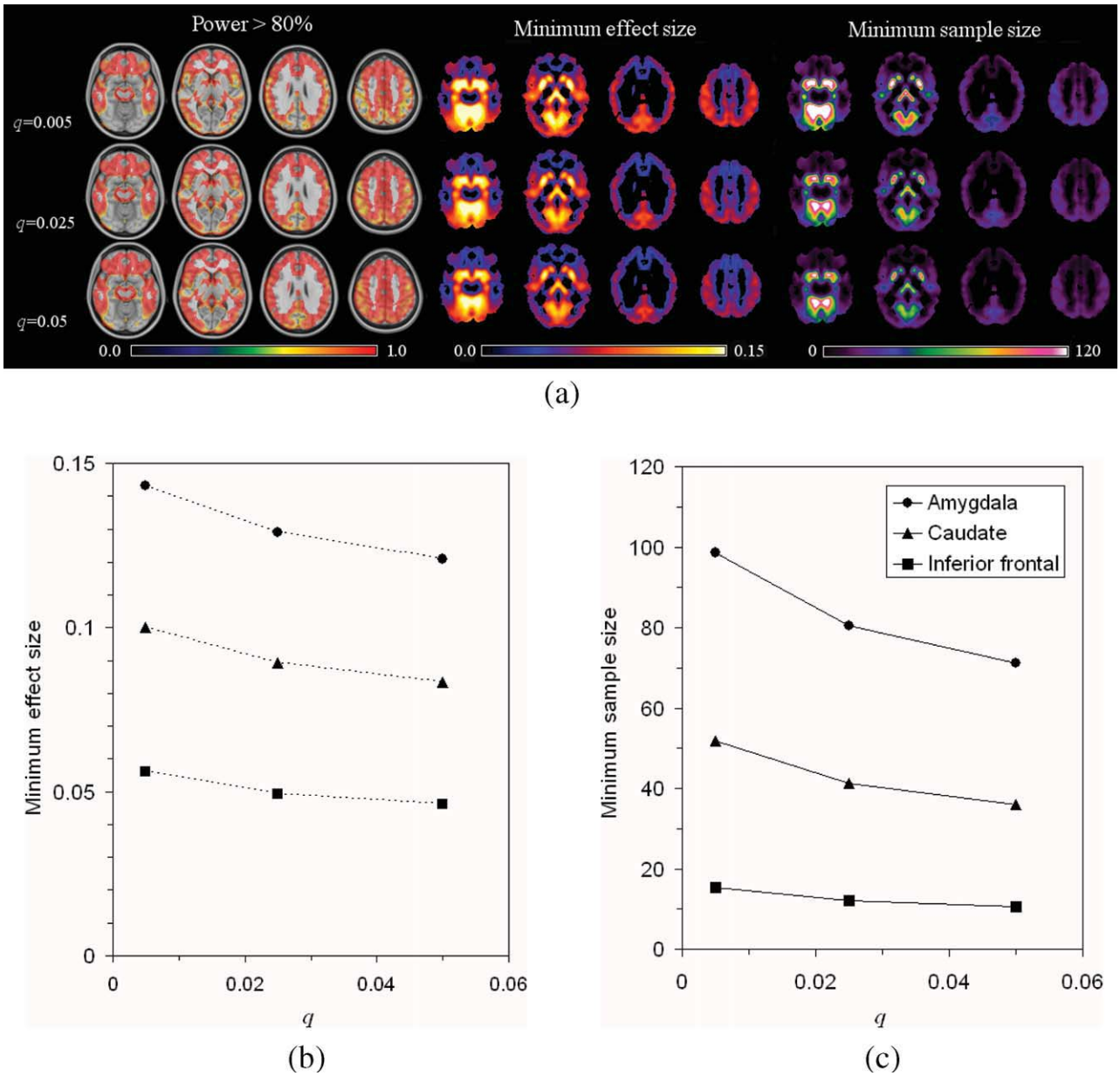

Figure 2.

(**a**) Maps of power (left), minimum effect size (center) and minimum sample size (right) for simulated studies with $N = 26$; $\theta = 0.1$; $\delta = 0.075$; $q = 0.005$ (top rows), 0.025 (middle rows), and 0.05 (bottom rows). (**b**) Mean minimum effect size and (**c**) mean minimum sample size within regions of interest.

minimum sample size (with fixed $\hat{\delta} = 0.075$) and minimum effect size (with fixed $N = 26$). Whilst the recruitment distribution for the three multicenter simulations is deliberately accentuated, the effect this has on minimum sample and effect sizes is small; typically <1 participant per group irrespective of the region. The smallest minimum sample and effect sizes are observed when all participants are recruited at the single-center with lowest within-center variance results. In contrast, a single-center study at the center with the largest within-center variance is associated with the largest minimum sample and effects sizes.

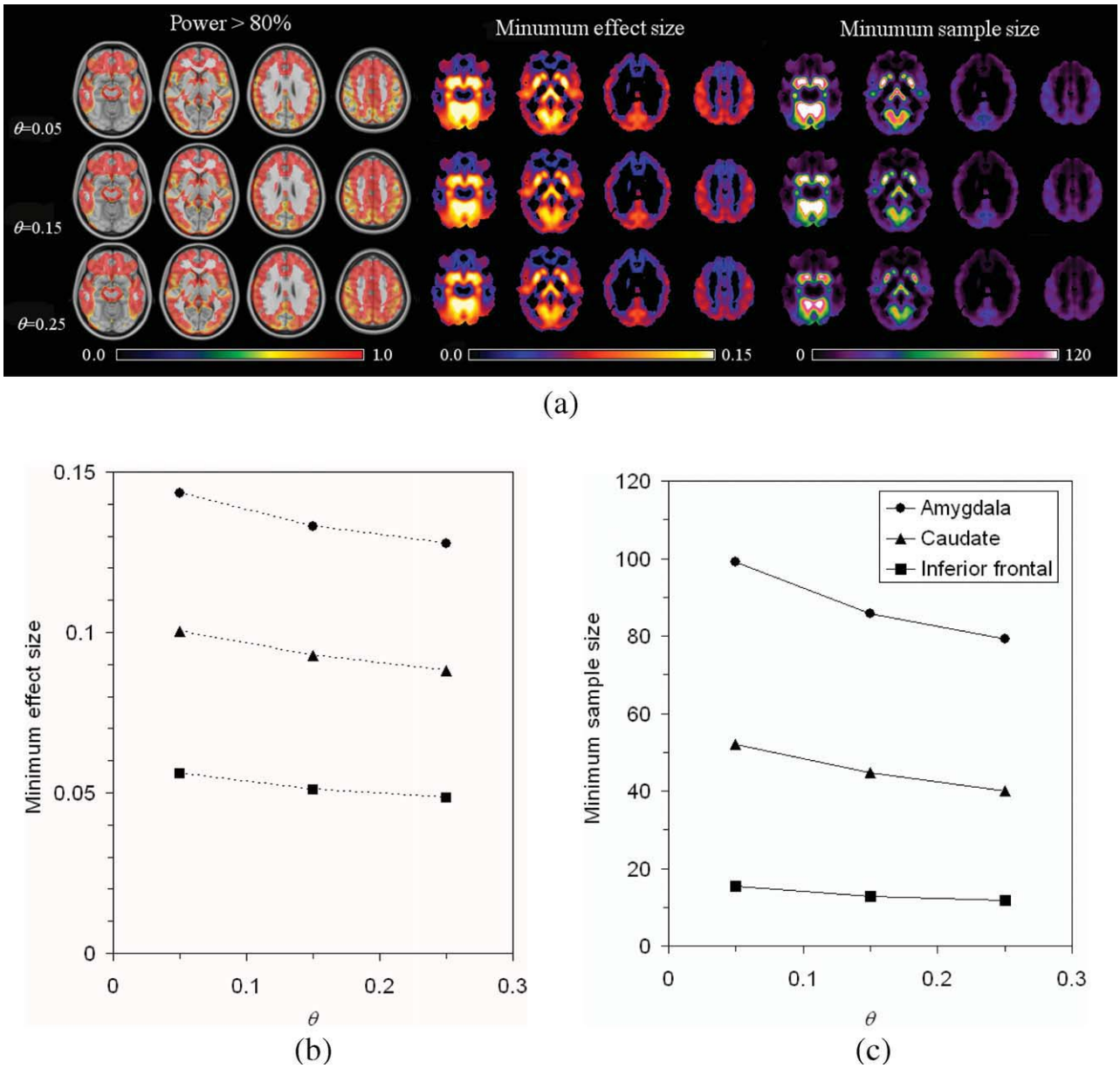

Figure 3.

(**a**) Maps of power (left), minimum effect size (center) and minimum sample size (right) for simulated studies with for $N = 26$; $q = 0.01$; $\delta = 0.075$; $\theta = 0.05$ (top rows), 0.15 (middle rows) and 0.25 (bottom rows). (**b**) Mean minimum effect size and (**c**) mean minimum sample size within regions of interest.

DISCUSSION

A voxel-based method of calculating statistical power for multicenter imaging studies has been presented for given effect size, sample size, FDR, and the proportion of true positives (or network extent). This method has been applied to distributions of cortical gray matter derived from T1 weighted structural magnetic resonance images of the brain. Maps of within-center variance for each of five centers calculated from data acquired in a calibration study were used to extend the method to multiple acquisition centers, specifying the proportion of the cohort recruited at each center. An illustration of how this procedure might be utilized in planning studies has been given

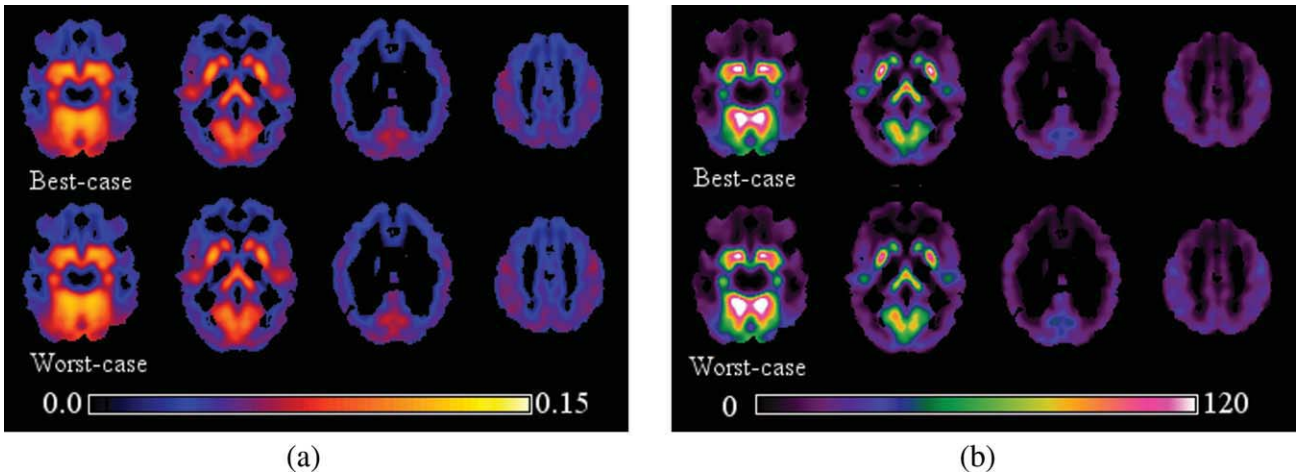

Figure 4.

(**a**) Maps minimum effect sizes ($N = 26$) predicted for a cross-sectional study with: $\theta = 0.1$; $q = 0.01$, for (top-row) the worst-case multicenter design and (bottom-row) the best-case multicenter design. See text for details of the simulation parameters. (**b**) Similarly, maps of minimum sample sizes ($\hat{\delta} = 0.075$) for (top-row) the best- and (bottom-row) the worst-case multicenter designs.

in the context of a cross-sectional study using distributions of gray matter obtained via segmentation of T_1-weighted magnetic resonance images.

The spatial distribution of the within-center variance is broadly similar at each center participating in the calibration study and dominates the corresponding patterns of the minimum effect size required to observe a given sample size and the minimum sample size required to observe a given effect size, both at fixed parameters of statistical testing (i.e., power, FDR and proportion of true positives).

Power Calculations Controlled by the False Discovery Rate

The application of type I error control on voxel statistics appears in several guises within the neuroimaging literature. The Bonferroni correction adjusts the false positive rate by dividing by the number of tests (i.e., voxels) conducted. This manner of specifying α was explored previously [Van Horn et al., 1998] for single-center voxel-based power calculations where it was noted that the reproducibility of studies may well be adversely affected by strict

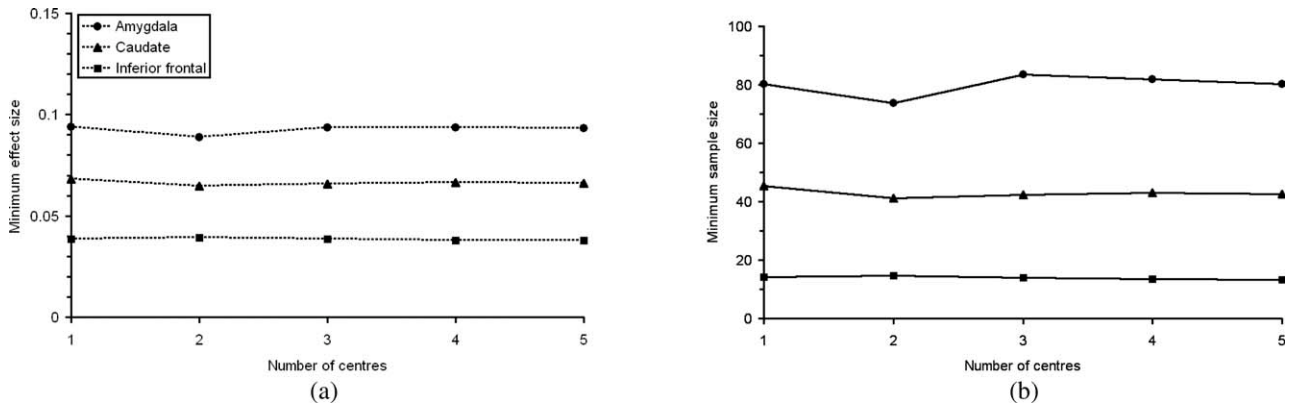

Figure 5.

Graphs of minimum (**a**) effect size and (**b**) sample size for each ROI as the number of centers in a simulated multicenter study is increased. The order of adding centers is from lowest-to-highest within-center variance. See text for details of the simulation parameters.

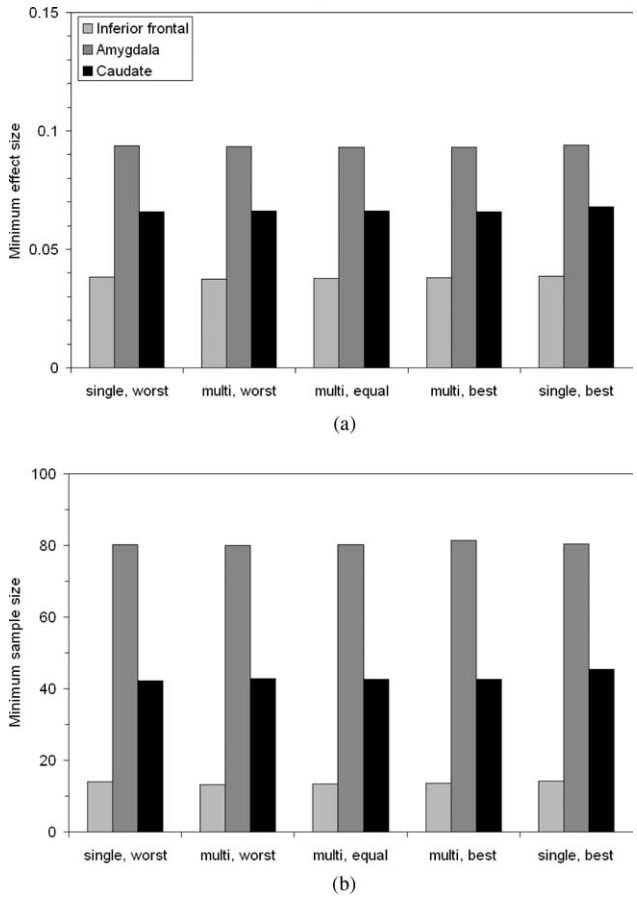

Figure 6.

Charts comparing minimum (**a**) effect size and (**b**) sample size for each ROI for the single-center and multicenter design scenarios. See text for details of the simulation parameters.

type I error control due to the exclusion of a large proportion of the true positive voxels. In general, there was a strong dependence of power on α.

The determination of an appropriate factor to correct type I error rates for multiple comparisons is compounded by the spatial coherence of voxel statistics, which reduces the number of independent tests conducted on the data. Theoretical approaches to correct for multiple comparisons in power calculations have been made within the framework of random field theory [Hayasaka et al., 2007]; a popular method that has been adopted extensively by the neuroimaging community, but is valid only for suitably smoothed data where the assumptions hold. An alternative approach to determining objective thresholds for statistical testing is to control for the proportion of false positives among the tests for which the null hypothesis was rejected, rather than amongst all the hypotheses tested. Controlling the FDR is a less conservative approach than controlling α that understandably has gained a good degree of acceptance in neuroimaging as well as other fields in which multiple comparisons present difficulties, such as ecology or gene expression microarrays. Nevertheless, there has been criticism of the technique [Chumbley and Friston, 2009] for use with fMRI applications: that FDR does not control the false discovery rate regionally and is under-conservative in spatial fields that lack compact support. Consequently, signal is present throughout the image. The validity of FDR for large spatial smoothing kernels is thus questionable and investigators should bear this in the mind when designing studies with power calculations. Permutation methods to estimate FDR are available with few assumptions on the exact form of the null distribution and reducing the need for large smoothing kernels (>6 mm). Nevertheless, these methods have their own constraints [Xie et al., 2005].

Introducing this method of type I error control into power calculations requires that the FDR, q, and the proportion of true false positives, θ, be specified. In absolute terms, the effect of q or θ on the power calculation depends strongly on the within-center variance. Thus, increasing q from 0.025 to 0.05 (Fig. 2c) results in a predicted fall in sample size of around 10 per group for high variance regions, but only 1–2 for low variance regions. Doubling θ, from 0.1 to 0.2, results in comparable changes in sample size in similar regions (Fig. 3c).

In the calculation of power we have assumed that each voxel is an independent test and that the average effect size is equal at all voxels, following the formalism of Jung [Jung, 2005]. Extensions of this treatment [Shao and Tseng, 2007] allow for differential effect sizes and correlated tests, but require additional parameters to be set, such as the average correlation between tests or subsets of tests. Power calculations are, of course, predicated on random sampling and the normal distribution and do not account for deviations from these assumptions that might occur especially in studies with small numbers of participants or patient groups. Improving estimates of within-center variance needs to be balanced against the additional resources needed to expand the cohort of the calibration study or the need to ethically justify the recruitment of patients in these circumstances. Enhancement of accuracy could also be achieved by modification of the calculations themselves. For small sample sizes, the t-distribution is an accurate approximation for the null and alternative distributions. Under these conditions Jung et al. [2005] derived an equivalent formula to Eq. (6) relating q, θ, and α, although it does not have a closed form and thus requires the use of numerical methods for its solution making it less practical for exploratory predictions. The power calculation [Eq. (1)] can also be expressed under the assumption of a t-distribution. In comparison, estimates of variance with a normal assumption (as used here) lead to slight increases for small sample sizes, although the effect is only 1 or 2 participants for sample sizes around 20 [Snedecor et al., 1989].

In summary, the guidance these calculations provide on sample sizes and power is a lower bound to the experimental parameters they predict and appropriate

allowances should be made during recruitment. Investigators that use these models should be mindful of Box's often quoted aphorism that "all models are wrong, but some are useful" [Box and Draper, 1987].

Empirical Measurement of Within-Center Variance

The effect size and its standard error are commonly estimated from the literature or prior experimentation. In a multicenter context, the relative contributions from each center must also be estimated. For the purposes of this study, we conducted a calibration experiment in which a group of healthy participants were repeatedly scanned at the centers. Not only does this permit an in situ estimation of the within-center error, but also gives a map of its distribution that is, in this case, spatially inhomogeneous (Fig. 1). Increased variance in subcortical and retrosplenial structures is a strong feature and leads to marked differences in power, minimum sample, and effect sizes at different intracerebral voxels. Broadly the pattern is consistent across centers and might therefore be attributable to the processing pipeline, which was identical for all images. In fact, power calculations could be an instructive method of assessing or comparing the processing modules that are chained together as pipelines for analysis.

Calibration studies carry with them significant operational challenges including scheduling, transportation, and cost. Counterbalancing data acquisition across participants is a key design feature of calibration studies, although this may be difficult to achieve in practice and randomized orders of attendance at each center may have to suffice. Combining within-center errors in the calculation of the standard error [Eq. (4)] makes the assumption that between-subject variance remains constant at each center. However, extended phases for data acquisition may lead to changes in brain structure due to, for example, brain maturation in adolescents or young adults. For power calculations of proposed cross-sectional studies in these populations this may well result in biased estimates of variance. For structural MRI calibration datasets acquired over a short period this is a reasonable position, although hydration [Kempton et al., 2009] and the menstrual cycle [Protopopescu et al., 2008] are reported to have acute effects on brain morphology. Furthermore, when considering measures of brain function rather than morphology as part of a multicenter trial it should be noted that functional MRI and positron emission tomography can also be affected by diurnal cycles and patterns of sleep [Cunningham-Bussel et al., 2009; Germain et al., 2007; Habeck et al., 2004] and subject x center interactions have been observed in previous fMRI calibration studies [Suckling et al., 2008]. Thus, additional care is required in scheduling scanning at similar times of day and ensuring that the participant reception, scanning environment, presentation of stimuli, and so on is as uniform as possible at each center.

The treatment taken in this article is not specific to segmented structural MRI datasets. Previous power calculations specifically for fMRI have considered effects between activation elicited by different event types, taking into account the properties of the time-series [Desmond and Glover, 2002; Mumford and Nichols, 2008]. We have previously reported between-group power calculations in regions of interest in a two-center calibration study of fMRI data acquired during a working memory paradigm and at rest [Suckling et al., 2008]. In general, the only requirement is that the within-center (between-subject) variance be normally distributed, a condition that can be violated in up to 30% of intracerebral voxels [Thirion et al., 2007]. Between-center variance can also be inflated in task-activation paradigms due to the difficulty in ensuring consistent stimulus presentation and scanning environment and the related significant effects of center on behavior and network activation. Conversely, estimates of inherent signal characteristics such as fractal properties from signals acquired during rest demonstrates no between-center effects [Suckling et al., 2008] and are consequently ideal candidate markers for calibration studies of this type.

Implications for Multicenter Imaging Studies

Multicenter studies offer advantages over single-center designs other than reductions in completion times. Studies involving human subjects are both more generalizable and more efficient if the geographic catchment area for recruitment is large enough to sample the population that is to be the target of future studies. The impact of additional variance from between-center components can be moderated by working practices. Nevertheless, the effect of center should be included, at least in the first instance, in the subsequent statistical analysis unless limited recruitment in a center makes such an analysis impractical (ICH, 1998; Vierron and Giraudeau, 2007]. This might be more effectively achieved by using covariates to adjust the dependent variable for predictable, but unwanted variance. A range of measures derived from images, have been tendered for fMRI equalizing effect sizes across sites [Friedman and Glover, 2006; Friedman et al., 2006] reducing type I errors introduced by center biases. Similar measures for structural MRI would be a useful avenue of investigation.

Strictly, the results presented in this article apply only to the centers involved and for projected studies involving a cohort with similar demographics. Within these boundaries we are able to say that increasing the number of participating centers does not unduly reduce the minimum observable effect size or, equivalently, increase the sample size needed (Fig. 5a). Thus, increasing the rate of recruitment by increasing the number of centers will typically reduce overall study completion time in inverse proportion to the expanding population they subtend. Furthermore, skewing the simulated recruitment distribution across the centers does not have a significant impact on

power (Fig. 5b) and confirms that studies with unequal numbers of participants at each center can be efficient [Senn, 1998]. Some differences are apparent in comparison of multicenter and single-center designs. However, even the best-case (lowest within-center variance) single-center only reduces the study sample size by 1–2 participants compared to the other designs.

Dominating statistical power is within-center variance; sample sizes ~3–6 times greater would be required should the hypothesized effect reside in subcortical and other midline structures (for example, amygdala or caudate) compared to regions of the neocortex (Figs. 4 and 6b). Increased volumes of the basal ganglia are associated with treatment with typical antipsychotic medication [Ellison-Wright et al., 2008] and switching to atypical antipsychotics appears to normalize these changes [Ellison-Wright et al., 2008; Scherk and Falkai, 2006]. Additionally, neuropathological brain changes in amygdala are associated with early episodes of psychosis [Ellison-Wright et al., 2008] leading to speculation that it may mediate social and cognitive decline [Hulshoff Pol et al., 2001]. Thus, predictions for adequately powered sample sizes will depend upon the focus of the study (e.g., treatment sensitivity or neurocognitive decline) and in turn on the implicated target region or regions. In general, the overall recommendation should be based on the most conservative prediction. To take these cases as examples: for a structural MRI, five-site, two-group, cross-sectional study with a FDR of 0.01, around 80 participants in both patient and control groups (total of 160) would be necessary to observe a difference in gray matter of $\hat{\delta} = 0.075$ in amygdala, but approximately half that amount $N = 42$ in each group (84 in total) in caudate. Alternatively, for a fixed sample size, $N = 26$ (the average sample size previously reported [Ellison-Wright et al., 2008]), the minimum effect size (difference in group means) that could be expected to be detected in amygdala is 0.095 and 0.065 in caudate, which conditions expectations from studies of similar size. As we have shown that these parameters are robust to the effect of the distribution of participants recruited at each of the five centers involved, multicenter recruitment strategies have no deleterious effects on sample sizes and should serve to reduce study recruitment times.

ACKNOWLEDGMENTS

Katherine Lymer is a member of the SINAPSE collaboration (www.sinapse.ac.uk). The authors thank the participants and the radiographic and administrative staff in each of the centers for their concerted and sustained support throughout this project.

REFERENCES

Benjamini Y, Hochberg Y (1995): Controlling the false discovery rate—A practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 57:289–300.

Box GEP, Draper NR (1987): Empirical Model-Building and Response Surfaces. New York, NY: Johh Wiley and Sons.

Chumbley JR, Friston KJ (2009): False discovery rate revisited: FDR and topological inference using Gaussian random fields. Neuroimage 44:62–70.

Cunningham-Bussel AC, Root JC, Butler T, Tuescher O, Pan H, Epstein J, Weisholtz DS, Pavony M, Silverman ME, Goldstein MS, Altemus M, Cloitre M, Ledoux J, McEwen B, Stern E, Silbersweig D (2009): Diurnal cortisol amplitude and frontolimbic activity in response to stressful stimuli. Psychoneuroendocrinology 34:694–704.

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

Ellison-Wright I, Glahn DC, Laird AR, Thelen SM, Bullmore E (2008): The anatomy of first-episode and chronic schizophrenia: An anatomical likelihood estimation meta-analysis. Am J Psychiatry 165:1015–1023.

Friedman L, Glover GH (2006): 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 (2006): Reducing inter-scanner variability of activation in a multicenter fMRI study: role of smoothness equalization. Neuroimage 321656–1668.

Friedman L, Stern H, Brown GG, Mathalon DH, Turner J, Glover GH, Gollub RL, Lauriello J, Lim KO, Cannon T, Greve DN, Bockholt HJ, Belger A, Mueller B, Doty MJ, He J, Wells W, Smyth P, Pieper S, Kim S, Kubicki M, Vangel M, Potkin SG (2008): Test-retest and between-site reliability in a multicenter fMRI study. Hum Brain Mapp 29:958–972.

Genovese CR, Lazar NA, Nichols T (2002): Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15:870–878.

Germain A, Nofzinger EA, Meltzer CC, Wood A, Kupfer DJ, Moore RY, Buysse DJ (2007): Diurnal variation in regional brain glucose metabolism in depression. Biol Psychiatry 62:438–445.

Habeck C, Rakitin BC, Moeller J, Scarmeas N, Zarahn E, Brown T, Stern Y (2004): An event-related fMRI study of the neurobehavioral impact of sleep deprivation on performance of a delayed-match-to-sample task. Brain Res Cogn Brain Res 18:306–321.

Hayasaka S, Peiffer AM, Hugenschmidt CE, Laurienti PJ (2007): Power and sample size calculation for neuroimaging studies by non-central random field theory. Neuroimage 37:721–730.

Hulshoff Pol HE, Schnack HG, Mandl RC, van Haren NE, Koning H, Collins DL, Evans AC, Kahn RS (2001): Focal gray matter density changes in schizophrenia. Arch Gen Psychiatry 58:1118–1125.

ICH (1998): Statistical principles for clinical trials. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). CPMP/ICH/363/96. The European Agency for the Evaluation of Medicines, London, 1998.

Jung SH (2005): Sample size for FDR-control in microarray data analysis. Bioinformatics 21:3097–3104.

Keator DB, Grethe JS, Marcus D, Ozyurt B, Gadde S, Murphy S, Pieper S, Greve D, Notestine R, Bockholt HJ, Papadopoulos P; BIRN Function; BIRN Morphometry; BIRN-Coordinating (2008): A national human neuroimaging collaboratory enabled by the Biomedical Informatics Research Network (BIRN). IEEE Trans Inf Technol Biomed 12:162–172.

Kempton MJ, Ettinger U, Schmechtig A, Winter EM, Smith L, McMorris T, Wilkinson ID, Williams SC, Smith MS (2009):

Effects of acute dehydration on brain morphology in healthy humans. Hum Brain Mapp 30:291–298.

Logothetis NK (2008): What we can do and what we cannot do with fMRI. Nature 453:869–878.

Mumford JA, Nichols TE (2008): Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. Neuroimage 39:261–268.

Nichols T, Hayasaka S (2003): Controlling the familywise error rate in functional neuroimaging: A comparative review. Stat Methods Med Res 12:419–446.

Paintadosi S (1997): Clinical Trials: A Methodological Perspective. Hoboken, NJ: John Wiley and Sons.

Pinheiro JC, Bates DM (2000): Mixed-Effects Models in S and S-Plus. Berlin: Springer-Verlag.

Protopopescu X, Butler T, Pan H, Root J, Altemus M, Polanecsky M, McEwen B, Silbersweig D, Stern E (2008): Hippocampal structural changes across the menstrual cycle. Hippocampus 18:985–988.

Scherk H, Falkai P (2006): Effects of antipsychotics on brain structure. Curr Opin Psychiatry 19:145–150.

Schnack HG, van Haren NE, Hulshoff Pol HE, Picchioni M, Weisbrod M, Sauer H, Cannon T, Huttunen M, Murray R, Kahn RS (2004): Reliability of brain volumes from multicenter MRI acquisition: a calibration study. Hum Brain Mapp 22:312–320.

Senn S (1998): Some controversies in planning and analysing multicentre trials. Stat Med 17:1753–1765; discussion 1799–1800.

Shao Y, Tseng CH (2007): Sample size calculation with dependence adjustment for FDR-control in microarray studies. Stat Med 26:4219–4237.

Smith SM (2002): Fast robust automated brain extraction. Hum Brain Mapp 17:143–155.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM (2004): Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 (Suppl 1):S208–S219.

Snedecor G, Cochran W, Cox D (1989): Statistical Methods, 8th ed. The Iowa State University Press. Ames, 1989.

Suckling J, Ohlssen D, Andrew C, Johnson G, Williams SC, Graves M, Chen CH, Spiegelhalter D, Bullmore E (2008): Components of variance in a multicentre functional MRI study and implications for calculation of statistical power. Hum Brain Mapp 29:1111–1122.

Thirion B, Pinel P, Mériaux S, Roche A, Dehaene S, Poline JB (2007): Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. Neuroimage 35:105–120.

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.

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

Vierron E, Giraudeau B (2007): Sample size calculation for multicenter randomized trial: Taking the center effect into account. Contemp Clin Trials 28:451–458.

Xie Y, Pan W, Khodursky AB (2005): A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. Bioinformatics 21:4280–4288.