Published in final edited form as:
Stat Methodol. 2009 March ; 6(2): 133–146. doi:10.1016/j.stamet.2008.05.003.

Hypothesis testing, power and sample size determination for between group comparisons in fMRI experiments

Dulal K. Bhaumik[a,b,d], **Anindya Roy**[a,c], **Nicole A. Lazar**[a,f], **Kush Kapur**[a,e,g], **Subhash Aryal**[a,d,*], **John A. Sweeney**[b,e], **Dave Patterson**[h], and **Robert D. Gibbons**[a,b,d]

[a] Center for Health Statistics, University of Illinois at Chicago, 1601 W Taylor Street (MC 912), Chicago, IL 60612, United States

[b] Department of Psychiatry, University of Illinois at Chicago, United States

[c] Department of Mathematics and Statistics, University of Maryland Baltimore County, United States

[d] Division of Biostatistics, University of Illinois at Chicago, United States

[e] Center for Cognitive Medicine, University of Illinois at Chicago, United States

[f] Department of Statistics, University of Georgia, United States

[g] Department of BioEngineering, University of Illinois at Chicago, United States

[h] Discerning Systems Inc, Burnaby, BCV3N4S9, Canada

Abstract

Modern methods for imaging the human brain, such as functional magnetic resonance imaging (fMRI) present a range of challenging statistical problems. In this paper, we first develop a large sample based test for between group comparisons and use it to determine the necessary sample size in order to obtain a target power via simulation under various alternatives for a given pre-specified significance level. Both testing and sample size calculations are particularly critical for neuroscientists who use these new techniques, since each subject is expensive to image.

Keywords

Brain imaging; Equality of proportion; False discovery rate; Large sample test; Regions of interest

1. Introduction

Functional neuroimaging – the use of advanced imaging techniques such as functional magnetic resonance (fMRI) and positron emission tomography (PET) – to study the working human brain represents a major advance in our ability to understand brain function, and how that function differs among groups (men versus women, patients versus controls, children of different ages, stroke patients at different stages of recovery, and so forth). The data obtained from a typical fMRI study are large and complex, providing many statistical challenges. For a detailed discussion of the data collection process in fMRI, see Lange [17], Lazar et al. [18].

[*]Corresponding address: Center for Health Statistics, University of Illinois at Chicago, 1601WTaylor Street (MC 912), Room # 453, Chicago, IL 60612, United States. Tel.: +1 312 996 9522; fax: +1 312 996 2113., saryal1@uic.edu (S. Aryal).

Our goal in this article is to develop a large sample based multivariate proportion testing procedure to compare two groups, in order to examine issues of power and sample size for fMRI studies. Sample size considerations are important due to the cost (in time and money) of scanning a subject. Recently, researchers have been moving in the direction of larger studies, with dozens of subjects, and power calculations have become more relevant. While power considerations have been studied to some extent for PET (see for example [28]), relatively little has been done in fMRI. Friston et al. [11] compared the number of subjects needed in two different modeling approaches to fMRI data. More recently Desmond and Glover [9] performed an explicit power analysis for fMRI, by first estimating and obtaining the distributions of typical effect size, and between and within subject variances from real studies. They then looked at the effects of varying those components on the number of subjects needed to maintain power at a given level, for a given level of significance. In this paper, we take a more theoretical approach than Desmond and Glover [9]. As in any statistical analysis, in the analysis of fMRI data there are choices and assumptions that are made, and these will have implications for testing and sample size calculations that ensue. It is important to state these in advance. In particular, we focus on the following issues: nature of the outcome (dichotomous or continuous); extent of the analysis (particular regions only or the whole brain) and the role of independence.

To the first point, we consider binary outcomes. The data are by nature continuous, giving levels of activation or percent signal change at a voxel (volume element, the three-dimensional version of a pixel) over time (for the many thousands of voxels in the brain), so this choice might seem contrary to expectations. However, researchers are often interested in determining which voxels are active in response to a given task. A voxel is deemed either active or inactive; there are no intermediate levels. Hence, this decision is a binary one. Before proceeding further, let us provide a brief review of the literature in connection to classification of active voxels.

Some classification based procedures have been proposed in the literature to determine whether voxels are active or not. For example, in block design fMRI experiment, with two conditions (baseline and an experimental condition), the analysis involves calculating t tests on all of the individual voxels in the brain (or perhaps in a region of interest in the brain). Voxels with t-statistic values above a certain pre-specified threshold are then classified as being active. Though this procedure is simple to implement, the main concern regarding this technique is that it does not take into account correlations between voxels that are close to each other, or between voxels that may be separated physically, but assume similar (or related) functionality. It relies on adjustments for multiple testing: when thousands of tests are carried out on the same data, it is misleading to conduct each one at standard significance levels of $\alpha = 0.05$ or $\alpha = 0.01$. As a result, this method suffers from high false positive and false negative rates. Forman et al. [10], Holmes et al. [15], Worsley [30], Genovese et al. [12] have introduced various alternative techniques in order to address these concerns in connection to block design fMRI experiments. Several authors have used model-oriented approaches for block designs to refine the classification available from simple t-tests (see for example [29,33]). These authors use random fields of different types (Gaussian, t, χ^2) to model the observed patterns of activation in the brain incorporating some of the spatial correlations in the data and determining regions of activity areas in the random field where a given threshold value is exceeded. Bandettini et al.’s [1] “correlation method”, uses a reference function, $r(t)$, and correlates it with each voxel data time series v(t). At each voxel, the correlation coefficient is calculated and is used to select active voxels. Different types of reference functions such as simple square waves, sine waves and other parametric functions can be used resulting in several different analyses.

Another type of design that is being frequently used in fMRI experiments is known as "event-related" design. In event-related design a stimulus is presented for a short period of time (instead of for a relatively prolonged period of time, as in a block design). Event-related studies allow researchers to learn about the hemodynamic responses in addition to detecting areas of activation [4,24]. Buckner and colleagues [8,25,5] present a methodology based on *selective averaging* to get hemodynamic response curves, and *t*-tests to locate areas of activation for event-related studies with rapid presentation of stimuli and with mixed trials. Postle et al. [23] propose a general linear model (multiple regression) for their event-related study of working memory. The authors use least squares technique to estimate the coefficients of the covariates of interest, and generate statistical maps by computing *t*-statistics. This review is not comprehensive; we have provided an outline of some of the major techniques for classification of active voxels, for both block design and event-related design studies. Gibbons et al. [13] proposed a random effects approach for modeling fMRI data by combining hierarchical polynomial models, Bayes estimation, and clustering. The methodology requires fitting cubic polynomials to the voxel time courses of event-related design experiments. The coefficients of the polynomials are then estimated using empirical Bayes procedures in a two-level hierarchical model. The empirical Bayes cubic fit for each voxel is designed to borrow strength from all voxels. The voxel-specific Bayes polynomial coefficients are then transformed to the times and magnitudes of the minimum and maximum points on the hemodynamic response curve, which are in turn used to classify the voxels as being activated or not. Thus, the final output of the procedure is a binary number for each voxel indicating the activation level of the voxel.

Second, we build tests that work on the whole brain, as well as tests that focus on specified regions of interest. Typically, statistical analysis is done on all of the voxels in the brain (on a voxel-by-voxel basis). There are situations, on the other hand, where the experimental task is very robust and well understood (for example, some visual processing tasks), and it makes little sense to test all voxels for activation. Some are known *a priori*, on the basis of neurological and psychological theory and experience, to be irrelevant; for example, if a voxel in a language processing area of the brain were to be picked out as active in response to a visual stimulus, such as a flashing checkerboard, the researcher would probably deem that finding to be spurious and uninteresting. Furthermore, activation should only be observed in the gray matter of the brain, not in other tissue types, allowing a further reduction in the regions under test, should this be desired.

Finally, it is common in the analysis of fMRI data to assume that the observations at different voxels are independent. Thus, as mentioned above, analysis is on a voxel-by-voxel basis. While it is clear in an intuitive sense that voxels cannot be independent (since they are all in the brain of the same individual, who is reacting to a particular stimulus), it is a great simplification to act as if they are. Indeed, since the correlation among voxels is complicated and not entirely understood, it would be hard to proceed at all without some such simplifying assumption. In the procedures described below, some mild forms of dependence are possible.

We assume a binary ("on/off") response at each voxel, for pre-specified regions of interest. The research question of interest is whether or not two subject groups differ in the proportion of active voxels in the region. We address this issue under the assumption that the voxel within a region of interest are dependent but the regions are independent.

Section 2 discusses the general problem of working with the proportion of active voxels in regions of interest (ROI). In Section 3, we present large sample tests for equality of proportions. Section 4 introduces a common thresholding method and the control of the False Discovery Rate (FDR), which are used in this paper to compare ROIs of two groups.

In Section 5 we develop the methodology, which is illustrated by a real life data set in Section 6. The methodology is demonstrated by a simulated data set in Section 7. We conclude with a discussion of our results in Section 8.

2. Proportions in regions

In this section we discuss how to work with the proportion of active voxels in ROIs in a general way. The scenario we consider here is the following: the brain is divided into ROIs, which are areas of the brain defined by anatomy, function, or both. For a given task, such as processing a visual stimulus, certain ROIs will be deemed to be relevant, while others are irrelevant. Our goal is to determine whether the proportion of activation for voxels in a pre-specified region is the same for two groups, for instance, normal controls compared to autistic patients. To distinguish between the two groups, we test equality of the underlying population proportions. We assume that all subjects within a given group have the same underlying proportion of active voxels, and any observed differences are simply the result of random variation. However, practical experience indicates that some individuals are “high activators” (that is, they show extensive and intense activation in response to the stimulus), whereas others are “low activators” (that is, they show limited activation, both in terms of pervasiveness and level, to the same stimulus). In that situation there is no particular reason to assume the same underlying parameter for each individual in a group, and testing a hypothesis about the group parameters will only be feasible when multiple observations are available for each subject in each group. We do not address this case, as it is rare in fMRI studies to have repetitions of the type required for this more sophisticated model.

Working with proportions in ROIs raises two additional issues. First, there is the question of spatial normalization as brains differ in size, shape and configuration. In order to carry out any group analysis, including the definition of regions of interest, the brain maps of different subjects need to be warped onto a common reference. To date there is no ideal way of doing this, but one of the common solutions is to transform individual brains into *Talairach coordinates* [27]. This is accomplished using a small number of reference points and an interpolation scheme (often linear or cubic). By transforming all images into Talairach space, we guarantee that the ROIs for each individual contain the same number of voxels and are located in the same place (in Talairach coordinates). Neither of these statements would be true in original data space. For purposes of group statistical analysis, it is essential to do such a warping, although the true data become distorted and lose spatial resolution as a result.

The second issue concerns thresholding, that is, how to decide that a given voxel is active (or inactive). This is, in essence, a question of statistical hypothesis testing, with thousands, or even millions of tests, and lies at the heart of many neuroimaging research problems. Traditional approaches to multiple testing, such as the Bonferroni correction, are too conservative when the number of tests is as large as in the current context. Various other methods have been proposed in the literature, such as contiguity thresholding, which builds on the idea that active voxels should occur in clumps [10]; random field theory, which looks for “excursions” or peaks of activity that are above what would be expected for a field of a given type (Markov, Gaussian, etc.) [29–32]; the use of FDR procedures [2,12]; random-effect models with clustering [13].

We take as our starting point that brain space has been transformed into Talairach coordinates using the widely used software package AFNI, ROIs have been defined, and the p-threshold has been applied, so that voxels are classified as either active or inactive [27,6,7].

Our methodology is based on two technical concepts: large sample tests for equality of proportions and control of the FDR. Therefore, for the convenience of the reader first we include some details on these two topics before formally introducing our methodology.

3. Large sample tests for equality of proportions

In this section we present large sample tests for equality of proportions that we discussed in the previous section. Let $X_1, X_2, \ldots, X_{n_1}$ be a random sample from a K-dimensional binary distribution, where each component of $X_i = (X_{i,1}, X_{i,2}, \ldots X_{i,K})'$ is either 0 or 1. Here the parameters of the multivariate binary distribution are $E(X_1) = \mathit{\Pi}_1$, Var $(X_1) = \mathit{\Sigma}_1$. Let $Y_1, Y_2, \ldots, Y_{n_2}$ be another random sample from a K-dimensional binary distribution with $E(Y_1) = \mathit{\Pi}_2$, Var $(Y_1) = \mathit{\Sigma}_2$. The covariance matrices $\mathit{\Sigma}_1$ and $\mathit{\Sigma}_2$ are functions of the parameters $\mathit{\Pi}_1$ and $\mathit{\Pi}_2$, respectively, but also depend on unknown parameters that determine the correlation structure between the components of the X_is and the Y_is. Assume the two samples to be mutually independent. Let $n = n_1 + n_2$. In our large sample framework we assume the following conditions on the sample sizes.

Assumption SS: min $(n_1, n_2) \to \infty$ and $n^{-1}n_1 \to c$ for some $c \in [0, 1]$.

Suppose that we are interested in testing the equality of the marginal proportions in the two populations. More specifically, we want to test H_0: $\mathit{\Pi}_1 = \mathit{\Pi}_2 = \mathit{\Pi}$. Let $\widehat{\Pi}_1=n_1^{-1}\sum_{i=1}^{n_1}X_i$ and $\widehat{\Pi}_2=n_2^{-1}\sum_{i=1}^{n_2}Y_i$ be the respective sample means. If we assume that the correlation structure of the components of the X_is and the Y_is are the same, under the null hypothesis, then the two groups have a common covariance matrix, say $\mathit{\Sigma}$. Let $\widehat{\sum}=n^{-1}\{\sum_{i=1}^{n_1}(X_i-\widehat{\Pi})(X_i-\widehat{\Pi})'+\sum_{j=1}^{n_2}(Y_j-\widehat{\Pi})(Y_j-\widehat{\Pi})'\}$, where $\hat{\mathit{\Pi}} = n^{-1}\ (n_1\hat{\mathit{\Pi}}_1 + n_2\hat{\mathit{\Pi}}_2)$. Then a natural test statistic is a Wald type statistic given by

$$T_n=\frac{n_1n_2}{n}(\widehat{\Pi}_1-\widehat{\Pi}_2)'\widehat{\sum}^{-1}(\widehat{\Pi}_1-\widehat{\Pi}_2). \tag{3.1}$$

Note that, for sufficiently large n, $\hat{\mathit{\Sigma}}$ is nonsingular and hence in the large sample framework the test statistic is well defined. The following theorem gives the large sample distribution of T_n under the null hypothesis.

Theorem 1

Let $\mathrm{T_n}$ be the test statistic as defined in (3.1) and let Assumption SS hold. Then under the null hypothesis $\mathrm{H_0}$: $\mathit{\Pi}_1 = \mathit{\Pi}_2$ we have

$$T_n\xrightarrow{\mathcal{L}}\chi_K^2,$$

where χ_K^2 is a chi-square random variable with K degrees of freedom and the $\xrightarrow{\mathcal{L}}$ denotes convergence in distribution.

Proof

Under the null hypothesis, $E(\hat{\Pi}) = \Pi$ the common value of Π_1 and Π_2. By standard results, the sample variance matrices $\widehat{\sum}_1 = n_1^{-1}\sum_{i=1}^{n_1}(X_i - \widehat{\Pi})(X_i - \widehat{\Pi})'$ and $\widehat{\sum}_2 = n_2^{-1}\sum_{j=1}^{n_2}(Y_j - \widehat{\Pi})(Y_j - \widehat{\Pi})'$ converge in probability to Σ. Then

$$\widehat{\sum} = \frac{n_1}{n}\widehat{\sum}_1 + \frac{n_2}{n}\widehat{\sum}_2 \xrightarrow{p} \sum.$$

Also, by continuity of convergence in probability the unique symmetric square root $\hat{\Sigma}^{-1/2}$, of $\hat{\Sigma}^{-1}$, converges in probability to the unique symmetric square root $\Sigma^{-1/2}$, of Σ^{-1}. By the multivariate Central Limit Theorem,

$$W_i := \sqrt{n_i}(\widehat{\Pi}_i - \Pi) \xrightarrow{\mathcal{L}} N_K(0, \sum), \quad i = 1, 2,$$

where $N_K(0, \Sigma)$ denotes the K-dimensional normal distribution with mean zero and variance matrix Σ. Here W_1 and W_2 are independent. Let $Z_n = \sqrt{\frac{n_1 n_2}{n}}(\widehat{\Pi}_1 - \widehat{\Pi}_2)$. Then by Slutsky's theorem and the multivariate CLT,

$$\begin{aligned} Z_n = \widehat{\sum}^{-1/2}\left[\sqrt{\frac{n_2}{n}}\sqrt{n_1}(\widehat{\Pi}_1 - \Pi) - \sqrt{\frac{n_1}{n}}\sqrt{n_2}(\widehat{\Pi}_2 - \Pi)\right] &\xrightarrow{\mathcal{L}} \sum^{-1/2} N(0, (1-c)\sum + c\sum) \\ &= N(0, I). \end{aligned}$$

By continuity of convergence in distribution, we get the following result.

$$T_n = Z_n' Z_n \xrightarrow{\mathcal{L}} \chi^2_K.$$

The test statistic T_n can be used to define a large sample critical region as $\{T_n > \chi^2_{\alpha,k}\}$ where $\chi^2_{\alpha,K}$ is the upper α percentile for the χ^2_K distribution.

4. False discovery rate

Let us now introduce a common thresholding method, the control of the FDR, which we use in this article to compare ROIs of two groups. We assume that the voxels nested within a particular ROI are correlated and we have incorporated an unstructured correlation matrix while developing our test. On the other hand, our assumption is that the correlation among voxels across different ROIs is not significant. However, as multiple ROIs are involved in the study controlling the type I error rate is inevitable. It is true that the total number of ROIs in any study of this type will not be huge, most likely the number of ROIs will not exceed 25–30. A simple Bonferroni correction or an approach proposed by Meinshausen and Rice [21] can suffice in this case. When the number of significance tests to be performed in the course of a single study is large, it is crucial to correct for multiple testing error rates. If every test, out of possibly tens of thousands of tests, is evaluated at traditional levels of α =

0.05 or even $\alpha = 0.01$, the aggregate error rate will be considerably higher than the nominal level and many false positives will be detected. Let m denote the number of hypotheses being tested. The general technique is as follows:

1. Select a desired FDR bound q between 0 and 1. This is the maximum proportion of false discoveries that the researcher is willing to tolerate, on average. While q of 0.15–0.20 is reasonable in many cases (Benjamini, personal communication), for our illustration we have used values of q starting 0.1–0.5.

2. Order the p-values from smallest to largest:

   $$p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)}.$$

   $H_{(i)}$ is the hypothesis corresponding to p-value $p_{(i)}$.

3. Let r be the largest i for which

   $$p_{(i)} \le \frac{i}{m} q.$$

4. Reject the hypotheses $H_{(1)}, \ldots, H_{(r)}$.

Aside from ease of implementation, FDR has other advantages over alternative approaches. Benjamini and Hochberg [2,3] show that the FDR procedures have good power compared to the Bonferroni correction. The method is adaptive, in the sense that the thresholds that are chosen are automatically adjusted to the strength of the signal in the data. The parameter q has a definite and clear meaning that is comparable across studies. Finally, since the procedures work with p-values, and not test statistics, FDR methods can be applied with any valid statistical test, with equal ease.

5. Methodology

In this section we assume that our data (i.e. pre-processed and p-thresholded) are binary with a score of "one" indicating that the voxel has been classified as active and "zero" if it has been classified as inactive. Our response probability model is based on the following assumptions:

- Subjects under study are independent. They are nested within two groups, called "Treatment (or Patient)(1)" and "Control(2)".
- The K voxels nested within an anatomical ROI in a particular subject are possibly dependent.

The region (or the set of K voxels) over which the testing of equality of proportion is performed deserves some discussion. The power of the test can be greatly enhanced by accurate identification of regions of activation (ROA) specific to a task. Such regions can be thought of as the collection of all voxels with reasonably high activation probability. As explained in Section 2, for a given task one generally concentrates on ROI that are obtained based on anatomical or functional considerations. However, such regions may still be quite conservative in the sense that they may be a large superset of the regions of activation. The voxels where the activation probability is uniformly low across groups and across subjects within groups will curtail the power of any test that is trying to detect a differential activation pattern between the groups. More specifically, let $\Pi_i'=(\pi_{i1},\pi_{i2},\ldots,\pi_{iK})'$ denote the overall activation probability of the ith group for the ROI. The activation probabilities depend on the task performed. Let the set of all possible tasks be $\mathcal{T}$ and let the voxel space for the ROI be denoted by $\mathcal{X}$. For a specific task $T \in \mathcal{T}$, let $A_T \subset \mathcal{X}$ denote the ROA. For notational convenience we suppress the task dependence of the activation probabilities. Then instead of testing

$$H_0{:}\pi_{1,k}=\pi_{2,k} \quad \forall k \in \mathcal{X}, \quad vs \quad H_A{:}\pi_{1,k} \neq \pi_{2,k} \quad \text{for some } k \in \mathcal{X},$$

considerable power can be gained by reducing the dimension of the test space to

$$H_0{:}\pi_{1,k}=\pi_{2,k} \quad \forall k \in A_T, \quad vs \quad H_A{:}\pi_{1,k} \neq \pi_{2,k} \quad \text{for some } k \in A_T.$$

However, one can potentially fail to identify group differences if the ROA is identified inaccurately and the test is restricted only to the ROA. The test of equality should be performed solely on the basis of the voxels in the ROA only if one strongly believes that such regions contain all possible active voxels for the given task. Thus, depending upon the situation, the number of voxels K will be either the total number of voxels in the ROI or that in the smaller ROA.

Let us denote the binary response variable from the kth voxel nested within the jth subject belonging to the ith group by x_{ijk}, where $i = 1, 2$, $j = 1, 2, \ldots, n_i$, and $k = 1, 2, \ldots, K$. Then for a single voxel in a single subject in one of the groups, $P(x_{ijk} = 1) = \pi_{ik}$, $E(x_{ijk}) = \pi_{ik}$, and $V(x_{ijk}) = \pi_{ik}(1 - \pi_{ik})$. We also define, for two distinct voxels belonging to the same subject $P(x_{ijk} = 1, x_{ijk'} = 1) = \pi_{ikk'}$, $\mathrm{Cov}(x_{ijk}, x_{ijk'}) = \pi_{ikk'} - \pi_{ik}\pi_{ik}'(= \delta_{ikk'}$, where $k \neq k'$). Write $\Sigma_i = ((\delta_{ikk'}))$. Finally, let $\widehat{\Pi}_{ik}=n_i^{-1}\sum_{j=1}^{n_i} x_{ijk}$ be an estimate of π_{ik}, and $\widehat{\pi}_{ikk'}=n_i^{-1}\sum_{j=1}^{n_i} x_{ijk}x_{ijk'}$ be an estimate of $\pi_{ikk'}$. Denote the corresponding estimate of Π_i by $\hat{\Pi}_i$. Hence $E(\hat{\Pi}_i) = \Pi_i$, and $\mathrm{Cov}(\widehat{\Pi}_i)=n_i^{-1}\sum_i$. We estimate Σ_i by $\hat{\Sigma}_i = ((\hat{\delta}_{ikk'}))$. The null hypothesis is

$$H_0{:}\Pi_1=\Pi_2. \tag{5.1}$$

Note that the dimension of the hypothesis for standard fMRI experiments is typically much larger than the sample size available, since even small ROIs have hundreds of voxels, and most studies comprise fewer than twenty subjects in total. This makes it infeasible to test the multivariate hypothesis without assuming some structure for the voxel-to-voxel dependence of the binary data. We did not assume any parametric structure for the correlation of the voxels as it is not well accepted by the experts in the field.

Note that H_0 defined in (5.1) can also be tested by using the following collection of univariate hypotheses.

$$H_{0,k}{:}\pi_{1k}=\pi_{2k}, \quad k=1, 2, \ldots, K. \tag{5.2}$$

In that case, the error rate that is controlled is the FDR. We will reject (5.1) if at least one of the hypotheses in (5.2) is rejected. The simplest FDR controlling methods assume a certain dependence structure among the hypotheses. Clearly, the p-values obtained in an fMRI example need not conform to this dependence structure. The number of hypotheses tested simultaneously is high, making the bias due to dependence possibly even higher. Thus, applying FDR to the collection of K hypotheses may lead to loss of power. We want to strike a balance between loss of efficiency in estimation of parameters due to small sample sizes relative to parameter dimension and lack of power due to application of FDR to a collection of possibly dependent p-values.

Suppose that we partition the region of interest into smaller regions, i.e., let $\mathcal{X} = B_1 \cup B_2 \cup \cdot \cdot \cdot \cup B_R$. Let $\Pi_i^{(r)}$ denote the probability response vector corresponding to the voxels in the subregion B_r. The partition is such that the number of voxels in each of the subregions is large enough to ensure efficient estimation of Σ_r, the variance–covariance matrix of the voxels in B_r. This makes it feasible to test each of the multivariate hypotheses

$$H_0^{(r)}:\Pi_1^{(r)}=\Pi_2^{(r)}, \qquad r=1, 2, \ldots, R. \tag{5.3}$$

Let $T_n^{(r)}$, $r = 1, 2, \ldots, R$ be the test statistic (3.1) computed for each of the subregions and let $p_1, p_2, \ldots, p_R$ be the corresponding p-values. The p-value for the rth region is obtained as

$$p_r=P(\chi^2_{k(r)}>T_n^{(r)}),$$

where $k(r)$ is the number of voxels in B_r. Let W be the number of subhypotheses (5.3) rejected when the Benjamini–Hochberg FDR procedure is applied to the sorted p-values, $p_{(1)} \le p_{(2)} \le \cdot \cdot \cdot \le p_{(R)}$. Then we will reject the hypothesis (5.1) if W is greater than zero, i.e., at least one of the subhypotheses is rejected based on the FDR procedure.

6. Illustration

As an illustration of our methodology, we consider the following experiment. The experiment used a visually guided saccade task to compare autistic and normal children. This is a simple visual task, in which the subject fixates on a crosshair in the middle of a screen, until a light appears in peripheral vision. The subject then needs to guide her eyes to the spot where the light appears. Saccade tasks are used in the study of diseases such as autism and schizophrenia since patients with these pathologies often exhibit disturbances in performance compared to healthy controls [20,22,19,14,26].

We chose the supplementary eye field of the left hemisphere of the brain as the ROI. This is one of the regions known to be involved in the performance of the visually guided saccade task. Data were collected using a standard block design, alternating periods of rest with periods of task. The study consisted of 24 subjects, 10 in the control group and 14 in the autistic group. The original dataset consisted of a section of the brain divided in a three-dimensional grid of voxels. There were in all 27 × 21 × 9 = 5103 voxels. Each of the 5103 voxels was classified as either 1 (active) or 0 (inactive), using a p-threshold of 0.001. The data obtained at the resolution of the 5103 voxels contained too many zero voxels and was mostly uninformative. Thus, the analysis was done on a coarser resolution and the three-dimensional grid was reduced by a factor of 3 in each direction. That is, 27 voxels in 3 × 3 × 3 grid of voxels were merged to construct larger voxels which we refer to as “supervoxels”. Suppose that the original voxels are indexed in a three-dimensional cube as V_{ijk}, $i = 1, \ldots, 27$, $j = 1, \ldots, 21$, $k = 1, \ldots, 9$, Then we define the supervoxels as $\tilde{V}_{lmn}=\cup_{i=3(l-1)+1}^{3l}\cup_{j=3(m-1)+1}^{3m}\cup_{k=3(n-1)+1}^{3n} V_{ijk}$ where $l = 1, \ldots, 9$, $m = 1, \ldots, 7$, $n = 1, \ldots, 3$. Combining voxels into supervoxels helps us to avoid the problem of excessive zero that we encounter in most of the original voxels and also simplifies computation. Thus, there were 5103/27 = 189(9 × 7 × 3) supervoxels. Each supervoxel was classified as 1 (active) if at least one of its 27 voxels had a score of 1. Otherwise the supervoxel was classified as 0 (inactive). Thus, if the value of the binary response at V_{ijk} was x_{ijk} then the value of the binary variable on the 'supervoxel' was defined as

$$\tilde{x}_{lmn}=\max\{x_{ijk}:3(l-1)+1\le i\le 3l;3(m-1)+1\le j\le 3m;3(n-1)+1\le k\le 3n\}.$$

The idea of supervoxel reduces the dimension of ROA that helps in implementing a chi-square test even for small samples. The resolution of 189 supervoxels is used to generate a framework for simulation in the next section. Before we present results from our simulation study on power properties of the proposed testing procedure, we end this section with an application of the proposed methodology to the given data. Following the description in Section 5, we subdivided the 189 supervoxels in a contiguous fashion into 63 subregions, each consisting of 3 supervoxels. Following the notation of Section 5, the subregions were $B_1 = \{\tilde{V}_{111}, \tilde{V}_{112}, \tilde{V}_{113}\}$, $B_2 = \{\tilde{V}_{121}, \tilde{V}_{122}, \tilde{V}_{123}\}$, …, $B_{63} = \{\tilde{V}_{971}, \tilde{V}_{972}, \tilde{V}_{973}\}$. The chi-square test, based on the binary values of the three supervoxels, was applied in each subregion and the values of the test statistic and the corresponding *p*-values for the 63 subregions are reported in Table 1. If the FDR methodology is applied to the 63 *p*-values, then the procedure fails to reject the hypothesis of equal activation probability in all of the 63 subregions for any reasonable value of the FDR control and hence fails to reject the overall hypothesis of equal activation profile. The values of the test statistic or *p*-values at these 63 regions are likely to be spatially correlated. The Benjamini–Hochberg FDR control procedure is known to be conservative for dependent data. This could be the reason for the procedure being not able to detect the significant difference. A more refined FDR control procedure which can take into account the dependence among the *p*-values is more likely to yield significance in some of the subregions. Table 1 shows the chi-square test statistic values, and their corresponding *p*-values for each of the 63 subregions. All of the *p*-values are quite large (>0.05), and so we fail to reject the null hypothesis in any location (with or without the FDR mechanism).

7. Statistical power

In order to compute the simulated power, we preserved this layout of 189 supervoxels for the simulation experiment and generated binary data for each of the supervoxels for 24 subjects. To generate realistic data, we needed to designate some of the 189 supervoxels as more likely to generate active response. First we estimated which supervoxels were more likely to be active for the control group. In order to do this, the average activation proportion over the ten subjects in the control group was calculated for each supervoxel. Thus, for the *k*th supervoxel, the population activation probability was estimated as $\hat{\pi}_k = s(k)/10$, where $s(k) \in \{0, 1, 2, \dots, 10\}$ denotes the number of subjects in the control group with a 1 for the *k*th supervoxel and $k = 1, 2, \dots, 189$. After computing $\hat{\pi}_k$ for each supervoxel, those with high (estimated) activation probability, *i.e.*, where $\hat{\pi}_k$ exceeded a given threshold, were considered active for the simulation experiment. For a threshold of 0.3, there were in all 34 supervoxels that were classified as active (which matches with expert experience).

For convenience, the numbers of subjects in the control group and in the patient group are taken to be the same, say *n*. The premise of our simulation model is that the nature of the voxels outside the region of activation is the same for both groups. By contrast, the activation probability for the voxels in the region of activation is not the same. For both groups, a model for the "on" voxels in the inactive region is

$$x_{i,j,k} \sim \text{Bernoulli}(\delta), \quad k \in \mathcal{X} \backslash A_T,$$

where δ is a small probability that captures the random noise in the inactive region and $\mathcal{X} \backslash A_T$ is the collection of all inactive voxels in the ROI. Thus, according to our model each voxel in the inactive region gets a small probability of activation. Therefore, the number of active voxels, U, in the complement of the region of activation is distributed as a Binomial (K_I, δ) where K_I is the number of voxels in $\mathcal{X} \backslash A_T$. Given U, the locations of the active voxels are distributed uniformly over $\mathcal{X} \backslash A_T$.

The number of active voxels in the region of activation for the control group, V_1, is distributed as Binomial (K_A, p_1) where K_A is the number of voxels in the region of activation. Given V_1 the locations of the active voxels are distributed uniformly over the region of activation. Thus, for the control group

$$x_{i,j,k} \sim \text{Bernoulli}(p_1), \quad k \in A_T.$$

Similarly, the number of active voxels in the region of activation for the patient group, V_2 is distributed as Binomial (K_A, p_2) and given V_2 the locations of the active voxels are distributed uniformly over A_T.

Therefore, for the treatment group

$$x_{i,j,k} \sim \text{Bernoulli}(p_2), \quad k \in A_T.$$

Finally, we partition the region of interest, $\mathcal{X}$ into $B_1 \cup B_2 \cup \cdots \cup B_R$. In our simulation study, the total number of voxels in $\mathcal{X}$, K, is 189 and that in the ROA, K_A, is 34. We choose a regular grid of 21 disjoint three-dimensional cubes as our partition where each cube contains 9 voxels. Of course in practice if natural partitions are known (for example, groups of voxels that are anatomically close and not necessarily contiguous in space) then one can choose such partitions to potentially improve the performance of the tests. We compute R(=21) test statistics for testing equality of the nine-dimensional probability vectors for each cube across the two groups. The large sample χ^2 null distributions are used for computing the p-values and then FDR is applied to the sorted p-values. Thus, in our case, the number of hypotheses, m, is 21. One may argue that for such a small number of tests, controlling the family wise error rate (FWER) with Bonferroni or similar procedures is more reasonable. However, in practice, the number of tests can be considerably larger, making FDR the more appropriate measure for control. For our simulation, we set $0 < p_1, p_2 \leq 0.5$ at an increment of 0.02, $\delta \in \{0.05, 0.1\}$, $n \in \{20, 25, 30\}$ and $q \in \{0.10, 0.15,\ldots, 0.60\}$ where q is the control level in the FDR procedure.

Figs. 1 and 2 both show the power surfaces as a function of p_1 and p_2 for different sample sizes and different values of δ. The general findings are that the power is an increasing function of the sample size and a decreasing function of the noise parameter δ. Note that power is also an increasing function of the critical difference $|p_1 - p_2|$. This can be used for sample size determination when a certain level of power is desired for a given parameter combination. To this end, we tabulated the critical difference $|p_1 - p_2|$ required for attaining a desired power γ for different values of n, q and for $\delta = 0.05$.

Tables 2, 4 and 6 show that the critical difference $|p_1 - p_2|$ required for a fixed power level γ = (0.60, 0.80, 0.90) is a decreasing function of both n and α for all the regions (denoted by R in equation 5.3) 9, 21 and 63. The critical difference values are averages over the range $0 < p_1 < 1$. The values have a slight increasing trend for $0 < p_1 < 1/2$ and a slight decreasing trend for $1/2 < p_1 < 1$. However, as evident in the standard errors reported at the bottom of the table, the variability of the critical difference is not severe. Thus, the power for a fixed n, α and δ can be taken as a function of the difference $|p_1 - p_2|$ rather than a bivariate function of p_1 and p_2. Tables 3, 5 and 7 report the control level q required for attaining the pre-fixed size of the test for number of ROIs 9, 21and 63 respectively. Tables 2, 4 and 6 reveal that as the number of regions increase the corresponding critical differences $|p_1 - p_2|$ decreases. However, we do not see this kind of monotonicity property of q in Tables 3, 5 and Specifically, we looked at the control level q needed to maintain the size of the tests at 1%, 5% or 10%. For smaller sample sizes, the control level cannot be too stringent, otherwise the tests will be undersized.

Discussion

We have presented a methodology for performing between group comparisons in fMRI, based on the proportions of active voxels in a region of interest for a group of control subjects and a group of patients. The test is based on certain simplifying assumptions, such as being able to partition the region of interest into a set of subregions, and having a constant proportion of active voxels across subjects in a group. These constraints are due to characteristics of the data; for instance, having many more voxels than subjects makes it impossible to get good estimates of the correlation structure across voxels, hence some “parcelling” of voxels and subdividing of regions are necessary. Similarly, without having multiple repetitions of the same experiment for each subject, something that is not currently routine in fMRI studies, it is impossible to get individual estimates of the activation parameter, and we need to assume constancy of the probability of activation.

Using this test we have investigated its power depending on various sample sizes. Not surprisingly, the power of the test increases as sample size increases, and decreases as the probability of activation in supposedly “inactive” parts of the studied region goes up. The latter indicates that the test loses power in the presence of noise, a sensible finding. Finally, even for small sample size (e.g. n = 10 per group) we can achieve quite good power, of the order of 80% or 90%, if the probabilities of activation in the two groups are sufficiently separated (e.g. 15%–20% difference between p_1 and p_2). Lower thresholds, as determined by the FDR procedure (but presumably similar findings would hold for other thresholding methods), require more separation of the base probabilities, as do smaller sample sizes. For the sample sizes common in fMRI (where it is usually not feasible to have even as few as 20 subjects per group), and the generally low q-thresholds that are used by fMRI researchers, the degree of separation would have to be that much clearer for high levels of power to be attained. Our test therefore behaves as we would expect a “reasonable” test of the equality of proportions to behave.

The study of power and sample size is relatively new in this application area, and the current work is just one of the first steps necessary for providing fMRI researchers with guidance on this aspect of their experimental design. Our basic strategy can, and should be, expanded to explore some of the other comparisons of interest [16].

Acknowledgments

The authors thank an associate editor and two referees for their insightful comments that significantly improved the quality of this article. This work was supported in part by a grant from the National Institute of Mental Health (1P50HD055751-01).

References

1. Bandettini PA, Jesmanowicz A, Wong EC, Hyde JS. Processing strategies for time-course data sets in functional MRI of the human brain. Magnetic Resonance in Medicine. 1993; 30:161–173. [PubMed: 8366797]
2. Benjamini Y, Hochberg Y. Controlling the FDR: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B. 1995; 57:289–300.
3. Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. Journal of Education and Behavioural Statistics. 2000; 25:60–83.
4. Buckner RL. Event-related fMRI and the hemodynamic response. Human Brain Mapping. 1998; 6:373–377. [PubMed: 9788075]
5. Buckner RL, Goodman J, Burock M, Rotte M, Koutstaal W, Schacter D, Rosen B, Dale AM. Functional-anatomic correlates of object priming in humans revealed by rapid presentation event-related fMRI. Human Brain Mapping. 1998; 6:373–377. [PubMed: 9788075]
6. Cox RW. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Computers and Biomedical Research. 1996; 29:162–173. [PubMed: 8812068]
7. Cox RW, Hyde JS. Software tools for analysis and visualization of fMRI data. NMR in Biomedicine. 1997; 10:171–178. [PubMed: 9430344]
8. Dale AM, Buckner RL. Selective averaging of rapidly presented individual trials using fMRI. Human Brain Mapping. 1997; 5:329–340. [PubMed: 20408237]
9. Desmond JE, Glover GH. Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses. Journal of Neuroscience Methods. 2002; 118:115–128. [PubMed: 12204303]
10. Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC. Improved assessment of significant change in functional magnetic resonance imaging (fMRI): Use of a cluster size threshold. Magnetic Resonance in Medicine. 1995; 33:636–647. [PubMed: 7596267]
11. Friston KJ, Holmes AP, Worsley KJ. How many subjects constitute a study? NeuroImage. 1999; 10:1–5. [PubMed: 10385576]
12. Genovese CR, Lazar NA, Nichols TE. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage. 2002; 15:870–878. [PubMed: 11906227]
13. Gibbons RD, Lazar NA, Bhaumik DK, Sclove SL, Chen HY, Thulborn KR, Sweeney JA, Hur K, Patterson D. Estimation and classification of fMRI hemodynamic response patterns. NeuroImage. 2004; 22:804–814. [PubMed: 15193609]
14. Goldberg MC, Lasker AG, Zee DS, Garth E, Tien A, Landa RJ. Deficits in the initiation of eye movements in the absence of a visual target in adolescents with high functioning autism. Neuropsychologia. 2002; 40:2039–2049. [PubMed: 12208001]
15. Holmes AP, Blair RC, Watson JDG, Ford I. Nonparametric analysis of statistic images from functional mapping experiments. Journal of Cerebral Blood Flow Metabolism. 1996; 16:7–22. [PubMed: 8530558]
16. Keles S. Mixture modeling for genome-wide localization of transcription factors. Biometrics. 2007; 63:10–21. [PubMed: 17447925]
17. Lange N. Statistical approaches to human brain mapping by functional magnetic resonance imaging. Statistics in Medicine. 1996; 15:389–428. [PubMed: 8668868]
18. Lazar NA, Eddy WF, Genovese CR, Welling J. Statistical issues in fMRI for brain imaging. International Statistical Review. 2001; 69:105–127.
19. McDowell JE, Brenner CA, Myles-Worsley M, Coon H, Byerley W, Clementz BA. Ocular motor delayed-response task performance among patients with schizophrenia and their biological relatives. Psychophysiology. 2001; 38:153–156. [PubMed: 11321616]
20. McDowell JE, Clementz BA. Ocular motor delayed response task performance among schizophrenia patients. Neuropsychobiology. 1996; 34:67–71. [PubMed: 8904734]
21. Meinshausen N, Rice J. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. Annals of Statistics. 2006; 34:373–393.
22. Minshew NJ, Luna B, Sweeney J. Oculomotor evidence for neocortical systems but not cerebellar dysfunction in autism. Neurology. 1999; 52:917–922. [PubMed: 10102406]

23. Postle BR, Zarhan E, D’Espisito M. Using event-related fMRI to assess a delay-period activity during performance of spatial and nonspatial working memory tasks. Brain Resonance Protocol. 2000; 5:57–66.

24. Rosen BR, Buckner RL, Dale AM. Event-related functional MRI: Past, present, future. Proceedings of National Academy of Science. 1988:773–780.

25. Schacter DL, Buckner RL, Koutstaal W, Dale AM, Rosen BR. Late onset of anterior prefrontal activity during true and false recognition: An event-related fMRI study. NeuroImage. 1997; 6:259–269. [PubMed: 9417969]

26. Takarae Y, Minshew N, Luna B, Krisky C, Sweeney J. Pursuit eye movement deficits in autism. Brain. 2004; 127:2584–2594. [PubMed: 15509622]

27. Talairach, J.; Tournoux, P. Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System: An Approach to Cerebral Imaging. Georg Thieme Verlag; Stuttgart: 1988.

28. Wahl LM, Nahmias C. Statistical power analysis for PET studies in humans. Journal of Nuclear Medicine. 1998; 39:1826–1829. [PubMed: 9776297]

29. Worsley KJ. Local maxima and the expected euler characteristic of excursion sets of χ^2, f, t fields. Advances in Applied Probability. 1994; 26:13–42.

30. Worsley KJ. The geometry of random images. Chance. 1996; 9:27–40.

31. Worsley KJ. Testing for signals with unknown location and scale in a χ^2 random field, with an application to fMRI. Advances in Applied Probability. 2001; 33:773–793.

32. Worsley KJ. Detecting activation in fMRI data. Statistical Methods in Medical Research. 2003; 12:401–418. [PubMed: 14599003]

33. Worsley KJ, Evans AC, Marrett AC, Neelin S. A three-dimensional statistical analysis of rCBF activation studies in human brain. Journal of Cerebral Blood Flow Metabolism. 1992; 12:900–918. [PubMed: 1400644]

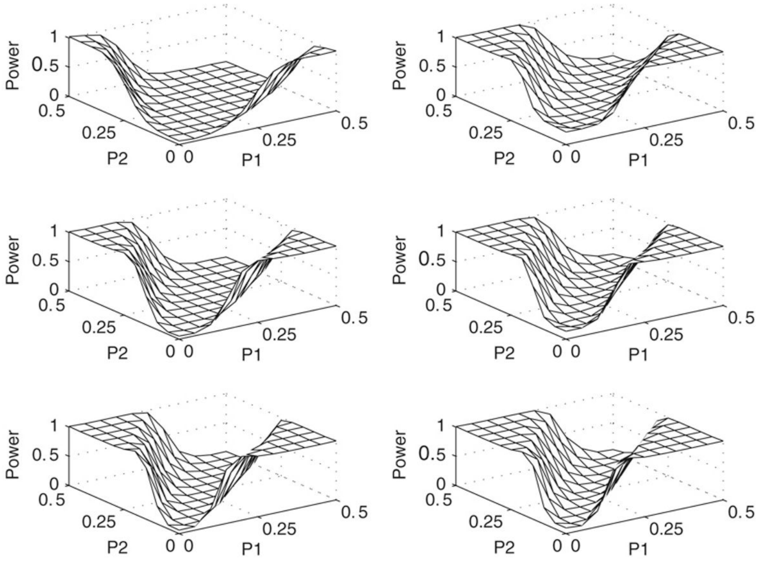

Fig. 1.

Power surface plots for $\delta = 0.05$: Left column: $R = 9$; Right column: $R = 63$: First row: $n = 20$; Second row: $n = 25$; Third row: $n = 30$.

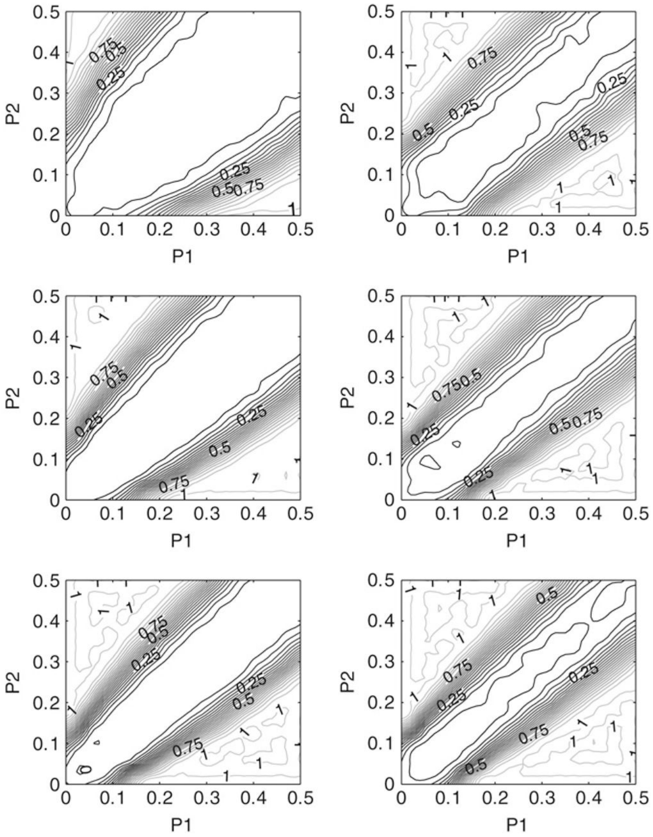

Fig. 2.

Power contour plots for $\delta = 0.05$; Left column: $R = 9$; Right column: $R = 63$: First row: $n = 20$; Second row: $n = 25$; Third row: $n = 30$.

Table 1

Chi-square test statistic and the corresponding *p*-values for testing control and autistic group for 63 subregions

Subregion χ^2-test statistic	*p*-value	Subregion χ^2-test statistic	*p*-value
1.4400	0.6962	2.4686	0.4810
1.3714	0.7122	2.5019	0.4750
2.4332	0.4875	0.9375	0.8164
3.7034	0.2953	4.2514	0.2356
4.6945	0.1956	4.3337	0.2276
2.9727	0.3958	4.1646	0.2442
0.8687	0.8330	4.5426	0.2085
0.9301	0.8181	0.0000	1.0000
2.8726	0.4117	2.3808	0.4972
4.0962	0.2513	4.3680	0.2244
3.1543	0.3684	3.0677	0.3813
6.8571	0.0766	5.9468	0.1142
1.4338	0.6976	3.2262	0.3580
1.4338	0.6976	1.6000	0.6594
2.2664	0.5190	4.1546	0.2452
2.7664	0.4291	2.6057	0.4565
1.6124	0.6566	1.9830	0.5759
3.4412	0.3285	2.5412	0.4679
0.7945	0.8508	3.2229	0.3585
2.6416	0.4502	6.3552	0.0956
3.1657	0.3668	3.8741	0.2754
1.9228	0.5886	2.7318	0.4349
3.4433	0.3282	1.5889	0.6619
5.0449	0.1685	1.9714	0.5784
7.5624	0.0560	0.8874	0.8285
6.2707	0.0992	6.1714	0.1036
6.3059	0.0976	3.2229	0.3585
6.5049	0.0895	3.2816	0.3502
1.5095	0.6801	2.2015	0.5317
1.0955	0.7782	1.6000	0.6594
2.0875	0.5545	1.6000	0.6594
6.2948	0.0981		

Table 2

Critical difference, $|p_1 - p_2|$ forgiven power level, γ, at different sample sizes, n, type I error, α, δ = 0.05 and number of ROIs R = 9

α	***γ*= 0.60**			***γ* = 0.80**			***γ*= 0.90**		
	***n* = 20**	***n* = 25**	***n* = 30**	***n* = 20**	***n* = 25**	***n* = 30**	***n* = 20**	***n* = 25**	***n* = 30**
0.01	0.352	0.205	0.163	0.416	0.255	0.186	0.437	0.266	0.203
0.05	0.203	0.148	0.125	0.264	0.174	0.155	0.231	0.203	0.175
0.10	0.151	0.120	0.117	0.175	0.151	0.143	0.220	0.178	0.151

Standard errors of entries in the table range between 5% and 50% of the entries.

Table 3

Control level *q* required for attaining prescribed type I error rates for different sample sizes (*n*), $\delta = 0.05$ and number of ROIs $R = 9$

α	$n = 20$	$n = 25$	$n = 30$
0.01	0.15	0.15	0.15
0.05	0.35	0.30	0.30
0.10	0.50	0.50	0.40

Table 4

Critical difference, $|p_1 - p_2|$ for given power level, γ, at different sample sizes, n, type I error, α, $\delta = 0.05$ and number of ROIs $R = 21$

α	$\gamma = 0.60$			$\gamma = 0.80$			$\gamma = 0.90$		
	$n = 20$	$n = 25$	$n = 30$	$n = 20$	$n = 25$	$n = 30$	$n = 20$	$n = 25$	$n = 30$
0.01	0.214	0.172	0.142	0.240	0.191	0.166	0.262	0.206	0.180
0.05	0.137	0.120	0.111	0.174	0.148	0.136	0.188	0.162	0.151
0.10	0.119	0.110	0.097	0.149	0.132	0.120	0.176	0.150	0.142

Standard errors of entries in the table range between 5% and 50% of the entries.

Table 5

Control level q required for attaining prescribed type I error rates for different sample sizes (n), $\delta = 0.05$ and number of ROIs $R = 21$

α	$n = 20$	$n = 25$	$n = 30$
0.01	0.15	0.10	0.05
0.05	0.45	0.40	0.30
0.10	0.55	0.50	0.40

Table 6

Critical difference, $|p_1 - p_2|$ for given power level, γ, at different sample sizes, n, type I error, α, $\delta = 0.05$ and number of ROIs $R = 63$

α	$\gamma = 0.60$			$\gamma = 0.80$			$\gamma = 0.90$		
	$n = 20$	$n = 25$	$n = 30$	$n = 20$	$n = 25$	$n = 30$	$n = 20$	$n = 25$	$n = 30$
0.01	0.199	0.162	0.138	0.227	0.184	0.163	0.242	0.191	0.172
0.05	0.136	0.131	0.130	0.168	0.152	0.146	0.177	0.165	0.150
0.10	0.118	0.116	0.103	0.139	0.126	0.126	0.150	0.146	0.139

Standard errors of entries in the table range between 5% and 50% of the entries.

Table 7

Control level q required for attaining prescribed type I error rates for different sample sizes (n), $\alpha = 0.05$ and number of ROIs $R = 63$

α	$n = 20$	$n = 25$	$n = 30$
0.01	0.05	0.05	0.05
0.05	0.40	0.20	0.15
0.10	0.55	0.50	0.45