NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 22 (2004) 95–108

An evaluation of thresholding techniques in fMRI analysis

Brent R. Logan[a] and Daniel B. Rowe[b,*]

[a] Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI 53226-0509, USA
[b] Department of Biophysics, Medical College of Wisconsin, Milwaukee, WI 53226-0509, USA

Received 12 May 2003; revised 16 December 2003; accepted 18 December 2003

This paper reviews and compares individual voxel-wise thresholding methods for identifying active voxels in single-subject fMRI datasets. Different error rates are described which may be used to calibrate activation thresholds. We discuss methods which control each of the error rates at a prespecified level α, including simple procedures which ignore spatial correlation among the test statistics as well as more elaborate ones which incorporate this correlation information. The operating characteristics of the methods are shown through a simulation study, indicating that the error rate used has an important impact on the sensitivity of the thresholding method, but that accounting for correlation has little impact. Therefore, the simple procedures described work well for thresholding most single-subject fMRI experiments and are recommended. The methods are illustrated with a real bilateral finger tapping experiment.

Keywords: Family-wise error rate; False discovery rate; Spatial correlation; Permutation; Resampling; Bonferroni procedure; Benjamini–Hochberg procedure; Random fields

Introduction

Many fMRI experiments have a common objective of identifying active voxels in a neuroimaging dataset. This is done in single-subject experiments, for example, by performing individual voxel-wise tests of the null hypothesis that the observed time course is not significantly related to an assigned reference function (Bandettini et al., 1993; Cox et al., 1995). A voxel activation map is then constructed by applying a thresholding rule to the resulting t statistics.

This paper describes three error rates that may be used to formally set activation thresholds based on individual voxel-wise test statistics, but not on cluster size. We review methods that control each of the error rates at a prespecified level α. These methods include simple procedures that ignore spatial correlation among the test statistics as well as more elaborate ones that incorporate this correlation information. The operating characteristics of the methods are shown through a simulation study, highlighting two results. First, as has been noted previously, the choice of error rate substantially impacts the power to detect true activations. Second, complicated procedures which explicitly account for the correlation structure do not improve the power in most practical situations, except when data are extremely strongly correlated. Therefore, for most single-subject analyzes, the simple procedures are recommended in practice. A real bilateral finger-tapping experiment is used to illustrate the methods and conclusions.

Problem and error rates

A common way of determining significance of a statistical hypothesis test is to specify the significance level or type I error rate of the test, usually denoted by α, and use this to determine a threshold. The type I error rate is the probability that, if the voxel were truly inactive, its test statistic would exceed the threshold, leading to the incorrect conclusion that it is active. This significance level determines the threshold, so that, for example, a 5% level voxel z test would have a threshold of 1.96 (two-sided) or 1.645 (one-sided). However, there is an important problem here. If we consider, for example, a $64 \times 64 \times 15$ volume image, as in the real experiment which follows, with no true activity attributable to treatment, we would expect $0.05 \times 61{,}440 = 3072$ voxels to exceed a 5% threshold by chance alone. Therefore, when we use this kind of thresholding rule, the result is a large number of false positives or voxels declared active when they are truly inactive. The reason for this problem is that there are multiple individual voxel hypotheses being tested. This is called the multiplicity problem; it occurs when multiple hypothesis tests are performed simultaneously and one must account for the possibility of errors occurring on each of these tests (Hochberg and Tamhane, 1987; Miller, 1981; Westfall and Young, 1993; Westfall et al., 1999).

As a result of this problem of excessive false positives, it is useful to consider other types of error rates which account for the multiplicity problem. Some notation needs to be laid out before proceeding. For voxels $i = 1,\ldots, m$, define T_i to be the random variable corresponding to the test statistic for treatment-related activation at voxel i, t_i to be its observed realization, $p_i = P\,(|T_i| > |t_i|)$ to be the (two-sided) p value for voxel i, μ_i to be the actual treatment-related activation for voxel i, and γ to be the fixed threshold set for determining whether a voxel is active. For example, $\mu_i = 0$ if voxel i is truly inactive and $\mu_i \neq 0$ if voxel i is truly active.

* Corresponding author. Department of Biophysics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226-0509.
E-mail address: dbrowe@mcw.edu (D.B. Rowe).

Available online on ScienceDirect (www.sciencedirect.com.)

1053-8119/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2003.12.047

When no account is made of the multiple testing, the error rate is the usual significance level or type I error rate. We will also call it the per comparison error rate (or per voxel error rate) and it refers to the probability of a false positive finding for an individual voxel *i*,

$$\mathrm{PCE} = P(\,|T_i|\, > \gamma \,|\, \mu_i = 0).$$

The most common way in the statistical literature to account for multiplicity is to consider the family-wise error rate (or image-wise error rate). The family-wise error rate (FWE) is the probability of at least one false positive on any voxel in the image,

$$\mathrm{FWE} = P\left(\mathop{\vee}_{i:\mu_i=0} \{\,|T_i|\, > \gamma\}\right).$$

Note that FWE ≥ PCE, so that any method which controls the FWE at level α will have a higher threshold γ than one which controls the PCE at level α. Sometimes a distinction is made between methods which only control the FWE under the overall null hypothesis (no voxels have treatment-related activity), called weak control of the FWE, and those which control the FWE under any null hypothesis (any subset of voxels have no treatment-related activity), called strong control of the FWE.

While several methods exist for controlling the FWE in fMRI data, it is important to consider whether the FWE is a relevant criterion for fMRI data. Is it relevant to focus on the probability of getting one or more false positives in a volume image of 61,440 voxels? The consequence of controlling the FWE at a set level α means that we will have relatively low power to detect truly active voxels. As in the Bonferroni procedure described below, we are adjusting the threshold for such a large number of voxel tests, and the signal or activity level will have to be very strong to be above this adjusted threshold.

One way of mitigating this problem is to consider a priori defined regions of interest (ROI). One can control the FWE in a region of the image which is of specific interest using Bonferroni adjustment or some other FWE method. This has the advantage that there is less multiplicity adjustment because of reduced family size (reduced number of voxels). Therefore, this method will have higher power to detect active voxels in that region. The disadvantage is that these ROIs must be identified a priori and must remain unchanged throughout the experiment and analysis. Otherwise, the FWE over that region will no longer be controlled. One example of an a priori identified ROI is to apply a mask to the image, so that only voxels inside the brain are considered in the multiplicity adjustment. This reduces the total number of voxel hypotheses and improves the power to detect activated voxels inside the brain and uncovered by the mask. Also, it is not difficult to specify a mask a priori, so this is a straightforward way to reduce the multiplicity adjustment. For simplicity, however, we do not consider masking further in the remainder of this paper.

The FWE criterion considers it unacceptable to have a false positive occurring anywhere in the thresholded volume image. Several alternatives exist for allowing a limited number of voxels to be falsely declared active to improve the power to detect truly active voxels. Sarkar (2002) proposed methods to control the probability of more than *k* falsely rejected voxels. However, *k* must be specified in advance, and this approach may be problematic when *k* is larger than the number of truly active voxels. An alternative is to allow some false positives in the thresholded image, but to relate the number of these acceptable false positives to the number of total positive findings. This is the basic concept of the false discovery rate (Benjamini and Hochberg, 1995), which is detailed next.

Table 1
True status vs. decision for all *m* voxels

True status	Decision		Total
	Declared inactive	Declared active	
Voxel inactive	N_T	D_F	m_0
Voxel active	N_F	D_T	m_1
	N	D	m

The false discovery rate (FDR) is the expected proportion of false positives to total positives, or the expected proportion of truly inactive voxels which are declared active to the total number of voxels declared active. This is illustrated in Table 1, where the entries in each cell refer to the counts of the number of individual hypotheses falling in the corresponding category. For example, false discoveries (D_F) are the number of inactive voxels which are declared to be active. Mnemonically, *N* refers to a nondiscovery (voxel declared inactive), *D* refers to a discovery (voxel declared active), *F* refers to a false conclusion, and *T* refers to a true conclusion.

Then the false discovery rate is

$$\mathrm{FDR} = E(D_F/D),$$

where the ratio D_F/D is defined to be 0 when $D = 0$. Note that using this notation,

$$\mathrm{FWE} = P(D_F > 0).$$

In the case where all null hypotheses are true (called the global null hypothesis), then the number of false positives D_F is equal to *D*, so that $D_F/D = 1$ if $D > 0$ and 0 otherwise. The FDR under this scenario simplifies to

$$\mathrm{FDR} = P(D > 0) = P(D_F > 0) = \mathrm{FWE}.$$

Therefore, any FDR-controlling procedure can be said to have weak control of the FWE.

Methods for controlling the FWE

The simplest way to control the FWE is through the Bonferroni method. To apply this, simply divide the individual threshold significance level α by the number of voxel hypotheses *m* to arrive at an adjusted threshold significance level $\alpha' = \alpha/m$ for each voxel test. This guarantees that FWE is no larger than α because

$$\mathrm{FWE} = P\left(\mathop{\vee}_{i:\mu_i=0} \{\,|T_i|\, > \gamma\}\right) \le P(\,|T_1|\, > \gamma \,|\, \mu_1 = 0) + \cdots + P(\,|T_m|\, > \gamma \,|\, \mu_m = 0) = m\alpha'.$$

One limitation of the Bonferroni method is that it results in conservative control of the FWE (i.e., fewer voxels declared active) in many situations. This conservativism is usually most severe when the test statistics are moderately to strongly correlated because there is a mismatch between the effective number of tests under correlation and the total number of tests *m* used in the

Bonferroni denominator. Functional MRI data are known to exhibit spatial autocorrelation, where closely spaced voxels are more strongly correlated with one another. The result of the conservative behavior of the Bonferroni method is potentially less power to detect truly active voxels.

Several ways to sharpen the Bonferroni procedure are reviewed by Miller (1981), Hochberg and Tamhane (1987), Westfall and Young (1993), and Westfall et al. (1999), leading to less conservative control of the FWE and more voxels declared active. One common method is to set thresholds based on the distribution of the maximum $|T|$ statistic. This is because the FWE for threshold γ under the overall null hypothesis can be written as

$$\mathrm{FWE} = 1 - P(|T_1| \leq \gamma, \ldots, |T_m| \leq \gamma) = 1 - P(\max_i |T_i| \leq \gamma),$$

so that the exact threshold γ to obtain FWE of α is the $(1-\alpha)$ percentile of the maximum $|T|$ distribution. Equivalently, one can consider thresholds based on the minimum voxel P value. This distribution is dependent on the correlation structure of the t statistics and may be obtained in several ways.

Random field methods were first applied to functional neuroimaging data to approximate this max $|T|$ distribution by Friston et al. (1991) and Worsley et al. (1992). A unified approach was presented in Worsley et al. (1996) that improved upon the earlier approximations by accounting for a finite search volume; see Petersson et al. (1999) for a further review of the contributions in this area. They assume that the m t statistics can be viewed as a lattice representation of a continuous Gaussian random field. They derive an approximation to the probability $P(\max_i |T_i| \geq \gamma)$ inside a search region V using the expected Euler characteristic (EC) of the set of voxels exceeding γ. For high threshold γ, the expected Euler characteristic and the adjusted P value are approximately the same. To account for the finite search volume, Worsley et al. (1996) transform from voxel coordinates to unitless resel coordinates by dividing the voxel dimensions by the Full-Width Half Maximum (FWHM) of the random field. Then

$$P(\max_i |T_i| \geq \gamma) \approx \sum_{D=0}^{3} R_D(V) p_D(\gamma),$$

where $R_\mathrm{D}(V)$ is the resel count of the region V corresponding to dimension D and $p_D(\gamma)$ is the EC density function in D dimensions. Alternatively, these Gaussian random fields may be simulated, but this requires an assumption of stationarity which the random field methods for the peak height do not require.

Westfall and Young (1993) propose resampling techniques to estimate the distribution of the maximum $|T|$ statistic. These are applied to fMRI analysis by Holmes et al. (1996), Bullmore et al. (1996), and Brammer et al. (1997) (see also Nichols and Holmes, 2001 for a review), in which the exact γ values are simulated using permutation resampling of the multiple scans over time. If under the null hypothesis the data from these scans are exchangeable (have the same distribution), we can generate the exact empirical distribution of the max $|T|$ statistic by enumerating each permutation, recomputing each voxel t statistic, and determining the observed max $|t|$ statistic across voxels for each permutation. In practice, one takes a random sample of B possible permutations rather than enumerating each one, because the number of permutations becomes prohibitively large with the number of time points. For example, with $n = 128$ time points, there would be $128! \approx 3.8 \times 10^{215}$ possible permutations. A random sample of these permutations yields a max $|T|$ distribution which converges to the true distribution with increasing B.

To obtain exchangeability, two modifications may need to be made to the data. A time trend may need to be accounted for so that one instead would permute the residuals after fitting a model for the time trend. Also, the time courses may exhibit temporal autocorrelation. Locascio et al. (1997) estimated the temporal autocorrelation using a parametric model, whitened the data based on this estimate, and then applied the permutation procedure to the whitened data to estimate the distribution of the maximum $|T|$ statistic. However, it is important to note that fitting a time trend and whitening the data make the residuals only approximately exchangeable because the parameters computed from the data are only estimates of the true values and they may induce additional correlation in the residuals. While this is not likely to be a big problem if the model is correctly specified, the impact of this approximate exchangeability on the multiplicity correction is not fully known. Finally, note that other methods have been proposed to facilitate permutation-based inference in the presence of temporal autocorrelation, such as precoloring (Worsley and Friston, 1995) and wavelet resampling (Bullmore et al., 2001).

An algorithm using permutation resampling to estimate the FWE corresponding to threshold γ is given below:

1. Simulate a series of B images under the null hypothesis by permuting the suitably exchangeable residuals, and denote statistics computed on the simulated image b, $b = 1, \ldots, B$ with a superscript.
2. For simulated image b compute $\max_i |t_i^b|$.
3. Estimate the FWE for threshold γ using the simulated images as

$$\widehat{\mathrm{FWE}}(\gamma) = \frac{\{\#b \mid \max_i |t_i^b| > \gamma\}}{B}.$$

In conclusion, methods for controlling the FWE require a threshold to be set based on the distribution of the maximum $|T|$ statistic. Many possibilities for estimating this distribution exist, but the permutation resampling method is especially attractive because it can be applied to many situations, provided one can fit an appropriate model so that the residuals are suitably exchangeable.

Methods for controlling the FDR

Benjamini and Hochberg (1995) propose a simple step-up procedure for controlling the FDR at level q, which was applied to neuroimaging data by Genovese et al. (2002). This procedure is called step-up because it uses an adaptive threshold which depends on the ordered P values $P_{(1)} \leq P_{(2)} \leq, \ldots, \leq P_{(m)}$, where the subscript in parentheses denotes the order. Let $v_{(i)}$ denote the voxel corresponding to P value $P_{(i)}$, and let d be the largest i for which

$$P_{(i)} \leq \frac{i}{m} q.$$

The BH (Benjamini and Hochberg) procedure declares voxels $v_{(1)}, \ldots, v_{(d)}$ to be active. It is called a step-up procedure because of the sequential or stepping up method for finding d.

This procedure was originally shown to control the FDR at level q for independent test statistics or P values. This proof was extended to test statistics which are positive regression dependent on subsets (PRDS), a technical definition of positive dependency of which positively dependent voxel t statistics is a special case (Benjamini and Yekutieli, 2001). For a general correlation structure with potential negative correlations, Benjamini and Yekutieli (2001) show that the FDR is still controlled if you redefine d above to be the largest i for which

$$P_{(i)} \le \frac{i}{m}\frac{q}{\Sigma_{j=1}^{m} 1/j} \approx \frac{i}{m}\frac{q}{\log m + 1/2}.$$

However, this adjusted threshold is substantially smaller (approximately 1/9 of the original proposed BH threshold for m = 4096) and will result in smaller d and fewer voxels declared active. Also, simulations (not shown) indicate that the original BH procedure controls the FDR even when there is some moderate negative correlation. Therefore, it is preferable to use the first procedure unless the negative correlations are very strong.

Storey (2002, 2003) proposes a simplified version of the FDR which he calls the positive FDR (pFDR):

$$\mathrm{pFDR} = E(D_F/D \,|\, D > 0).$$

Under independence of the test statistics, Storey (2002, 2003) shows the pFDR has a natural Bayesian interpretation as

$$\mathrm{pFDR} = P(H_i \text{ is true} \,|\, |T_i| > \gamma),$$

or the posterior probability that the voxel is inactive, given that its test statistic is above the threshold. Storey uses the Bayesian interpretation to express the pFDR in the following way:

$$\mathrm{pFDR} = \frac{P(|T_i| \ge \gamma \,|\, \mu_i = 0)P(\mu_i = 0)}{P(|T_i| \ge \gamma)}$$

$$= \frac{F_0(\gamma)\pi_0}{F\gamma},$$

where F_0 is the complement of the CDF of $|T_i|$ when voxel i is inactive, F is the complement of the marginal CDF of $|T_i|$ regardless of activity (a mixture distribution of F_0 and an unknown alternative distribution F_1), and $\pi_0 = m_0/m$ is the proportion of null hypotheses (or inactive voxels).

Storey (2002) proposes estimates of the parameters of interest and establishes a heuristic connection between his procedure and BH in the following way. Suppose that the P value $P_{(i)}$ corresponded to the threshold $\gamma_{(i)}$, that is, $P_{(i)} = F_0(\gamma_{(i)})$. We can estimate $\hat{F}(\gamma_{(i)})$ by the proportion of rejected hypotheses, so that $\hat{F}(\gamma_{(i)}) = i/m$. A worst case or conservative estimate of pFDR can be obtained by using $\pi_0 = 1$. Then $\widehat{\mathrm{pFDR}} = mp_{(i)}/i$ so that

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

Therefore, this Bayesian formulation turns out to be equivalent to the BH method under a conservative estimate of $\pi_0 = 1$.

Several authors (Benjamini and Hochberg, 2000; Genovese and Wasserman, 2002; Storey, 2002; Storey and Tibshirani, 2001; Storey et al., 2004) have considered adaptive estimation of π_0 to further refine the FDR-controlling procedure and improve upon the conservative BH procedure. However, in most fMRI datasets, we expect that a relatively small proportion of the voxels in an image would actually be considered active as a result of treatment. In this setting, there may be limited utility in estimating π_0 because the estimate will typically be very close to one. In addition, estimation of π_0 may be sensitive to strong correlation structures. Storey et al. (2004) show that for weak dependence, where

$$\lim_{m\to\infty} \frac{D_F(\gamma)}{m_0} \overset{\text{a.s.}}{=} F_0(t) \text{ and } \lim_{m\to\infty} \frac{D_T(\gamma)}{m_1} \overset{\text{a.s.}}{=} F_1(t),$$

then the resulting adaptive estimates of pFDR will be asymptotically conservative (i.e., provide strong control over pFDR) as $m \to \infty$. This type of weak dependence includes dependence occurring in finite blocks, among others. However, for a finite m, less is known about how the adaptive estimation of π_0 in the presence of correlation affects the strong control of the FDR, so we do not consider it further.

The BH method is a simple, powerful procedure, but it has two main limitations. It has only been shown to control the FDR for independent or positively correlated voxel t statistics, and if the statistics are positively correlated, it may not be as powerful as another method which does incorporate correlation information. As discussed earlier, most fMRI datasets exhibit some spatial correlation. Therefore, it is useful to consider methods for controlling the FDR when there is correlation present.

Yekutieli and Benjamini (1999) propose a resampling-based estimate of the FDR for threshold γ which utilizes the correlation structure, and show that a step-up procedure using these estimates controls the FDR under an arbitrary correlation structure with the restriction that D_{F} and D_{T} are independent of one another. A sufficient but not necessary condition for this independence is that the P values corresponding to the true null hypotheses and the P values corresponding to the false null hypotheses are independent of one another. Specifically with regard to fMRI data, this means that the active voxels can be arbitrarily correlated with one another, and the inactive voxels can be arbitrarily correlated with one another, but the active and inactive voxels are independent of one another. This assumption may not be met in fMRI datasets, where the task-related signal changes are mean shifts within a general correlation structure. We investigate in subsequent simulations whether the YB procedure controls the FDR, when the active and inactive voxels are positively correlated with one another. This scenario of positive correlation between active and inactive voxels may fit the spatial correlation structure exhibited by many fMRI datasets. However, note that limited simulation studies such as the one presented later need not reveal problems with the YB method in terms of control of the FDR for correlation structures seen in fMRI studies. Further research is needed to establish mathematically whether this procedure still controls the FDR when D_{F} and D_{T} are positively correlated, or if not in general, then under what conditions does it control the FDR; note, however, that counter examples can be obtained to show that it does not necessarily control the FDR when D_{F} and D_{T} are negatively correlated.

The YB procedure works as follows. Define $D_{\mathrm{F}}(\gamma)$ to be the random variable representing the number of inactive voxels in the actual image (with some truly active and some truly inactive voxels), which are declared active using threshold γ. Similarly, define $D_0(\gamma)$ to be the random variable representing the number of truly inactive voxels which are declared active using threshold γ in a hypothetical image reflecting the overall null hypothesis where all voxels are inactive. Then $D_0(\gamma)$ is stochastically larger than $D_{\mathrm{F}}(\gamma)$ because it has more inactive voxels which may potentially be declared active. Yekutieli and Benjamini (1999) use this result and the independence of D_{F} and D_{T} to construct a conservative estimator of the FDR for threshold γ as

$$E\left[\frac{D_0(\gamma)}{D_0(\gamma)+D_T(\gamma)}\right] \text{ instead of } E\left[\frac{D_F(\gamma)}{D_F(\gamma)+D_T(\gamma)}\right].$$

In this expression, we can estimate $D_{\mathrm{T}}(\gamma)$ by

$$\hat{D}_T(\gamma) = D(\gamma) - mp_\gamma,$$

where $D(\gamma)$ is the number of t_i exceeding γ in the observed image and p_γ is the P value corresponding to the threshold γ. Then the final estimator is

$$\widehat{\mathrm{FDR}}_{\mathrm{YB}}(\gamma) = E\left[\frac{D_0(\gamma)}{D_0(\gamma)+\hat{D}_T(\gamma)}\right]. \tag{4.1}$$

This expectation is evaluated using the following steps:

1. For a given threshold (or P value) γ, estimate $\hat{D}_{\mathrm{T}}(\gamma)$ using the observed image data.
2. Simulate a series of B images under the null hypothesis, either through nonparametric resampling or Gaussian random field as detailed earlier, and denote statistics computed on simulated image b, $b = 1,\ldots, B$ with a superscript.
3. For simulated image b compute $D_0^b(\gamma)$, the number of $|t_i^b|$'s in the bth image exceeding γ.
4. Evaluate the expectation in Eq. (4.1) using the simulated images as

$$\widehat{\mathrm{FDR}}_{\mathrm{YB}}(\gamma) = \frac{1}{B}\sum_{b=1}^{B}\left[\frac{D_0^b(\gamma)}{D_0^b(\gamma)+\hat{D}_T(\gamma)}\right].$$

These FDR estimates can be used in a step-up procedure analogous to the BH procedure as follows. First order the observed P values so that $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$, and let $v_{(i)}$ denote the voxel corresponding to P value $P_{(i)}$. Compute the corresponding FDR estimates for each P value and denote the FDR estimate for $P_{(i)}$ as $\widehat{\mathrm{FDR}}_{\mathrm{YB}}(p_{(i)})$. Let d denote the largest i for which $\widehat{\mathrm{FDR}}_{\mathrm{YB}}(p_{(i)}) \leq q$. Conclude that the voxels $v_{(1)},\ldots, v_{(d)}$ are active, and the remaining ones are inactive.

The YB method can be used to control the FDR when there are potentially negative correlations, and it may be more powerful than the BH method because it explicitly incorporates the correlation structure.

fMRI simulation study

Basic design

Data are generated to simulate a bilateral finger-tapping fMRI block design experiment where the true motor activation structure is known so that each of the thresholding methods can be evaluated.

A 64 × 64 slice is selected for analysis within which two 7 × 7 ROIs as lightened in Fig. 1 are designated to have activation. For this slice, simulated fMRI data are constructed according to a regression model which consists of an intercept, a time trend for all voxels but also a reference function for voxels in each ROI which is related to a block experimental design.

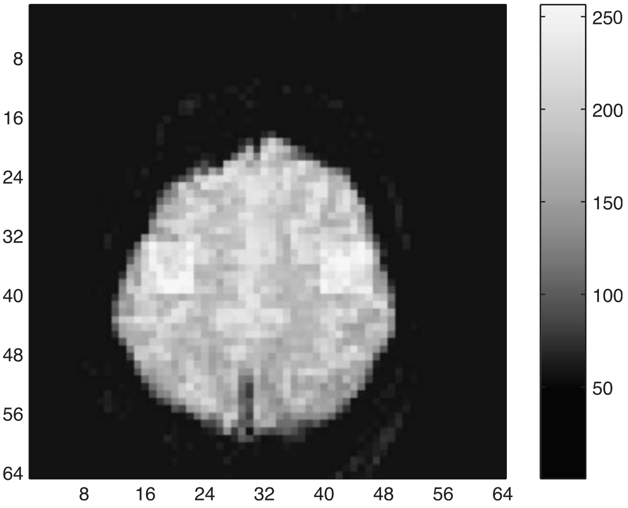


Fig. 1. Anatomical with ROI.

The multivariate regression model (Rowe, 2003) from which we generate the data for all m = 4096 voxels and all n = 128 time points is represented in terms of matrices as

$$\underset{n\times m}{Y} = \underset{n\times(q+1)}{X}\ \underset{(q+1)\times m}{B} + \underset{n\times m}{E} \tag{5.1}$$

where q is the number of independent variables, X is the design matrix, B is the matrix of true regression coefficients, and E is the error matrix.

The voxels in each ROI are numbered sequentially from left to right and top to bottom. The simulated data are generated according to Eq. (5.1), where the design matrix X is an n × 3 matrix whose first column is an n dimensional vector of ones, the second column is an n dimensional vector of the first n counting numbers, and the third column is an n dimensional vector consisting of eight replicates of eight ones then eight negative ones. The true regression coefficient matrix B outside each ROI consists of column vectors which are $(0.5,0.5,0)'$ plus random independent noise for the nonzero elements with zero mean and standard deviation 0.25. Inside each ROI, the regression coefficients associated with the reference function are given in terms of (i, j) coordinates by

$$B(i,j) = 2e^{-\frac{(i-i')^2+(j-j')^2}{2(2)}} + 1.5 \tag{5.2}$$

where (i', j') is the voxel number in the center of the ROI. These coefficients were chosen to have an activation region with the largest effect in the center and smaller effects towards the edge, but with reasonable power after multiplicity adjustment to detect the activations.

Correlation Structures Considered

The observation errors ϵ_i were randomly generated independently from a multivariate normal distribution with m dimensional zero mean vector and $m \times m$ positive definite covariance matrix $\Sigma = \sigma^2 R$, where $\sigma^2 = 64$. Several types of spatial correlation structures R were considered for the simulation study. First, the voxels were assumed to have a stationary, isotropic exponential function covariogram

(Cressie, 1993), which can be expressed in the simple form where each voxel is correlated with all other voxels according to ρ^d and d is the Euclidean distance between the voxels. For this exponential structure, values of ρ of 0.0, 0.7, and 0.95 were investigated and are denoted Exp(ρ).

Next, we assumed a population correlation matrix equal to the estimated sample correlation matrix from the example dataset in the Real fMRI example section to illustrate how the methods might perform on a "real" correlation structure.

Finally, we investigated how the methods might perform when the data are smoothed before analysis. Datasets were generated assuming spatial independence, and then we applied a 5-mm FWHM Gaussian kernel to smooth each dataset. The smoothed t statistics were rescaled to a $N(0,1)$ distribution so that appropriate thresholds could be determined. Note that the rescaling factor K requires knowledge of the spatial correlation among the residuals before smoothing, but under independence it just reduces to the square root of the sum of the squares of the kernel weights for voxels i and $j, K = \sqrt{\Sigma_{i,j}\ k_{ij}^2}$. When there is spatial correlation present, it may be necessary to either model this correlation or perform permutation resampling to determine an appropriate rescaling parameter. There are two main differences between the smoothed and unsmoothed data. First, smoothing tends to reinforce signals that are more spatially distributed and reduce in magnitude signals that are more isolated. Because our activation regions are fairly wide, we found it necessary to reduce the coefficients in the ROI by a multiplicative factor of K to facilitate comparisons among the thresholding procedures. Second, smoothing induces a correlation structure among the test statistics above and beyond the underlying spatial correlation present in the raw data before smoothing. We only consider the case where the residuals are uncorrelated before smoothing. The net effect of smoothing is a correlation matrix which is similar to the exponential correlation structure. Note that this correlation matrix induced by smoothing is also commonly used by fMRI researchers to model the residual spatial correlation without smoothing.

To illustrate the effect of the spatial autocorrelation structure on how the t statistic image appears, the first subfigure of Figs. 3–7 contains a single sample image (realization) of observed t statistics generated from the corresponding model and correlation structure. The Exp(0.0) image has little to no clustering of colors, the Exp(0.7) image has moderate clustering of the t statistic values, while the Exp(0.95) image has large areas of clustering. Note that the sample image for the smoothed simulated dataset appears to fall between the exponential covariance structures with ρ = 0.0 and ρ = 0.7 in terms of the degree of clustering, while the covariance structure generated from the example dataset results in very slight clustering located only in certain parts of the image. The spatial correlation structure of most single-subject fMRI data is expected to resemble the ρ = 0.0 or ρ = 0.7 scenarios, and is not likely to be as strong as the ρ = 0.95 scenario.

Another way to illustrate the spatial correlation structures considered is to show the correlation matrices corresponding to each structure. Fig. 2 illustrates the population correlation matrices for the five correlation structures considered by ordering the 4096 voxels from left to right and top to bottom, computing the correlation between each pair of voxels, and mapping the correlations to a color map. Fig. 2a shows a diagonal line of nonzero values when ρ = 0.0, while for ρ = 0.7, the correlation image in Fig. 2b has

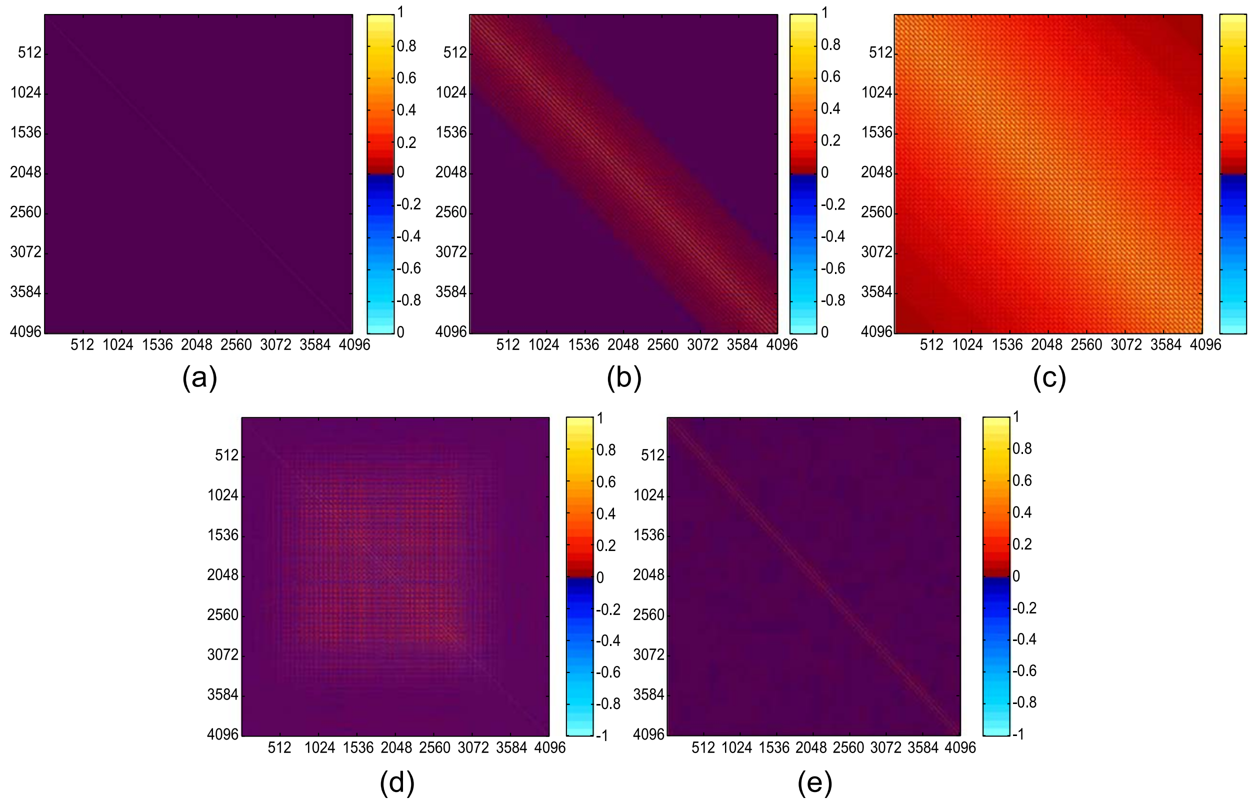


Fig. 2. Population correlation images. (a) ρ = 0.00, (b) ρ = 0.70, (c) ρ = 0.95, (d) real ρ's, (e) FWHM ρ's.

a narrow band of moderate correlation, indicating that the correlation is strong locally but not very dispersed. We refer to a correlation matrix which does not have a lot of moderate to strong correlations throughout the image as a sparse correlation structure (both $\rho = 0.0$ and 0.7 would qualify). When $\rho = 0.95$, this band of moderate correlation widens substantially, indicating that voxels even very far away are still moderately associated with one another. Although the correlation matrix from the real dataset shown in Fig. 2d does not have the same spatial correlation structure as the exponential structure, it appears that the magnitude and sparsity of the correlation structure based on this real dataset is more similar to the situation where $\rho = 0.7$, and not as strong as when $\rho = 0.95$. Note also that there is a central box-like shape to the larger correlation values; this region corresponds to the intracerebral voxels. This suggests long-range correlations in the original dataset which may be due to unmodeled signal or physiological noise. Finally, an empirical correlation matrix is shown in Fig. 2e to illustrate the correlation structure induced by smoothing the data with a 5-mm FWHM Gaussian kernel. This empirical correlation matrix is constructed from 5000 simulated datasets with no true activation that were smoothed. Note that the correlation matrix is quite sparse (it has moderate correlation for voxels within two of one another) and falls somewhere between the Exp(0.0) and Exp(0.7) models.

In addition to the positive spatial correlation structures above, we also studied a model where several blocks or regions of voxels were moderately negatively cross-correlated with one another, while voxels within each region followed a positive spatial correlation from a Exp(ρ) covariance function. The results of this structure were similar to the Exp(ρ) covariance function which did

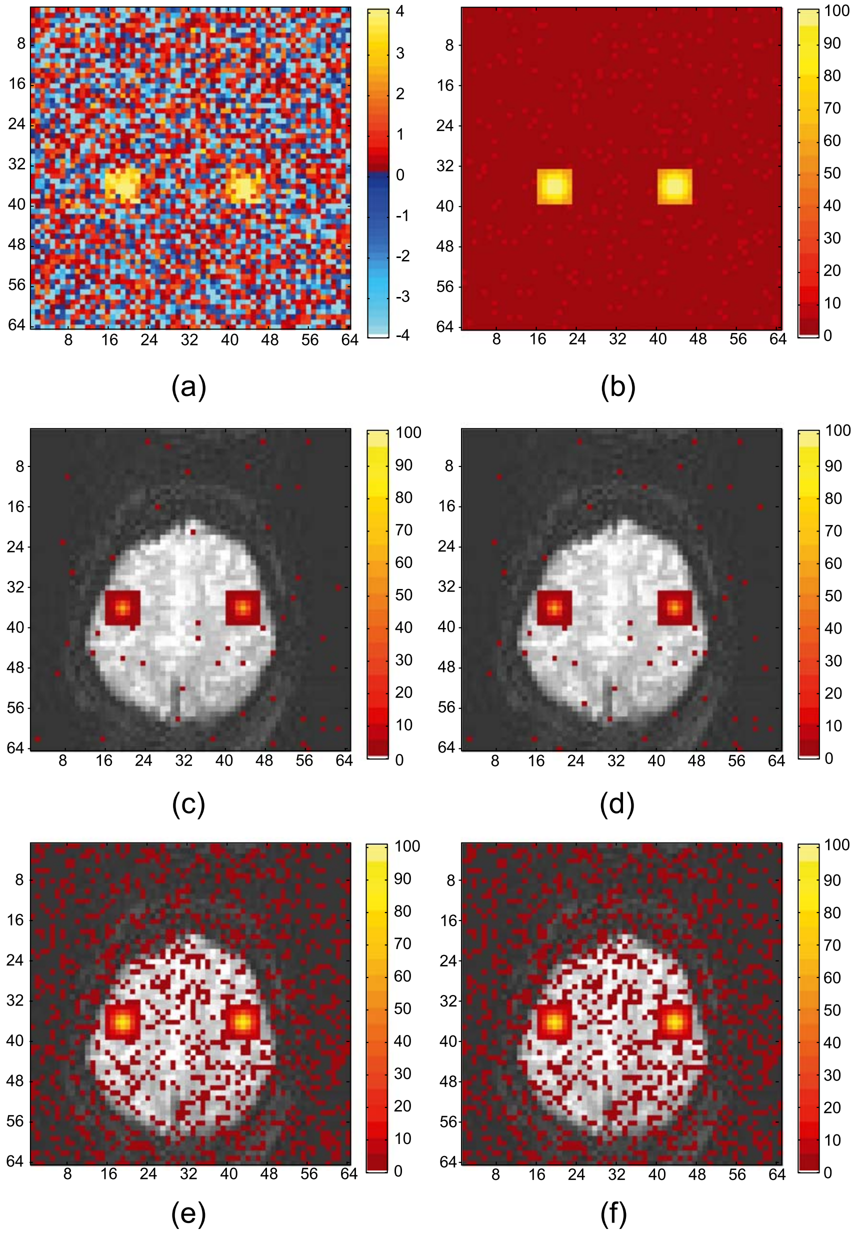


Fig. 3. Sample t statistic image and $\alpha = 0.05$ power images for $\rho = 0.00$. (a) Sample t statistic image, (b) unadjusted threshold, (c) FWE Bonferroni method, (d) FWE permutation method, (e) FDR BH method, (f) FDR YB method.

not have negatively correlated blocks and are therefore omitted for brevity. In particular, it is worth noting that the BH procedure controlled the FDR despite the negative correlations present, indicating further that the conservative approximation discussed by Benjamini and Yekutieli (2001) is unnecessary for most fMRI applications.

Results of simulation study

Five thresholding methods were considered: unadjusted method with type I error of 5%, a Bonferroni procedure with FWE of 5%, a permutation resampling procedure to control the FWE at 5%, the BH procedure with FDR of 5%, and the YB permutation resampling procedure with a FDR of 5%. Both resampling methods used 500 randomly generated permutations. Because of the computational burden, 1000 simulated images were created, on which each of these methods was applied. For each procedure, a power image was constructed which summarized the frequency over the 1000 simulated images with which each voxel was detected as active (above the respective threshold). For clarity, all voxels which are never detected as active are kept as image grayscale, while those which are detected active are given a color expressing the power or frequency with which they are declared active. These images are given in Figs. 3–7 for Exp(0.0), Exp(0.7), Exp(0.95), real data, and 5 mm FWHM correlation structures, respectively. As described above, the first subfigure of each figure is a single sample set of observed t statistics generated from the corresponding model and covariance matrix.

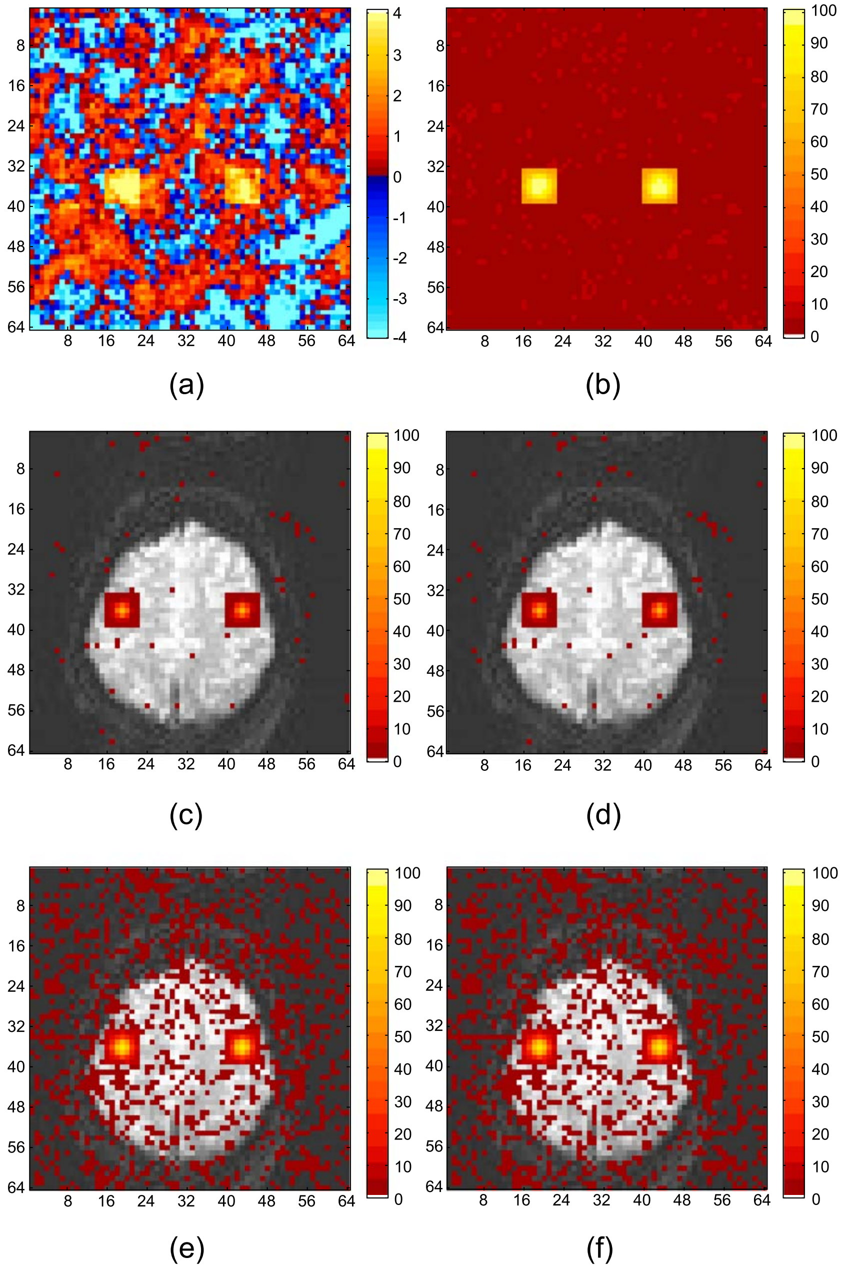


Fig. 4. Sample t statistic image and $\alpha = 0.05$ power images for $\rho = 0.70$. (a) Sample t statistic image, (b) unadjusted threshold, (c) FWE Bonferroni method, (d) FWE permutation method, (e) FDR BH method, (f) FDR YB method.

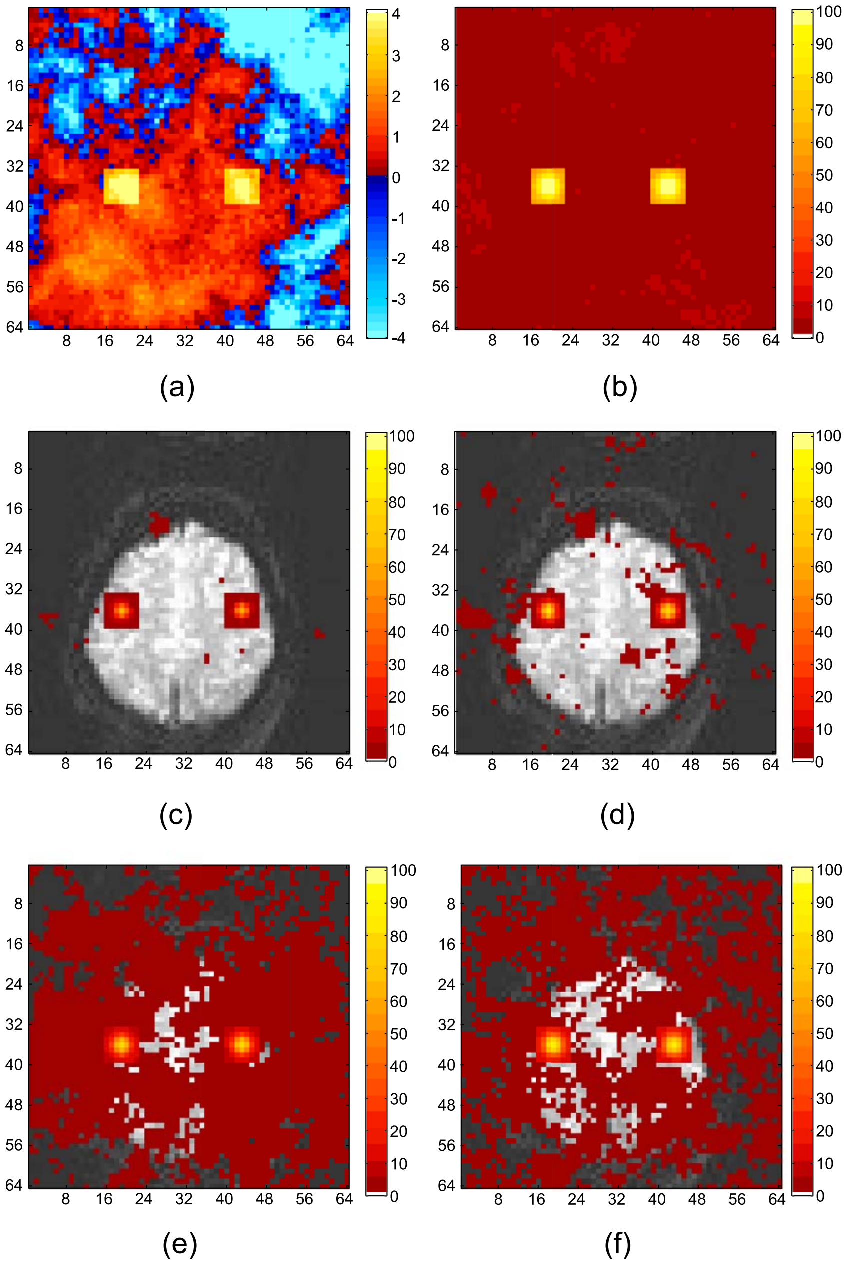


Fig. 5. Sample *t* statistic image and $\alpha = 0.05$ power images for $\rho = 0.95$. (a) Sample *t* statistic image, (b) unadjusted threshold, (c) FWE Bonferroni method, (d) FWE permutation method, (e) FDR BH method, (f) FDR YB method.

In all cases considered, the unadjusted method detects the active region with high power, but at the cost of a substantial number of false positives. In fact, every voxel in the entire 64 × 64 image is declared active at least once and often multiple times in the 1000 simulated images.

For the zero correlation or independent voxels scenario, there appears to be no benefit to using a permutation sampling method to account for correlation. The FWE Bonferroni method and the FWE permutation method give virtually identical power images as do the FDR BH method and the FDR YB method. However, there is a substantial difference between the FWE-controlling procedure and the FDR-controlling procedure. The FDR-controlling procedures maintain higher power in the ROI than the FWE-controlling procedures, albeit at the cost of more falsely detected voxels. However, the proportion of falsely detected voxels is still maintained at a rate of 5% or fewer on average of the total number of voxels declared active.

Similar results can be seen for the Exp(0.7) structure. Even though the spatial correlation is stronger, there is still little to no effect of incorporating the correlation information through a permutation method, either for the FWE- or FDR-controlling procedures. This is probably because, even though $\rho = 0.7$, the correlation between any voxel and a neighbor 6 voxels away is only 0.12, which is very low and indicative of the sparse correlation matrix seen in Fig. 2b. Because the correlation matrix is sparse, there is little advantage to incorporating such correlation information into

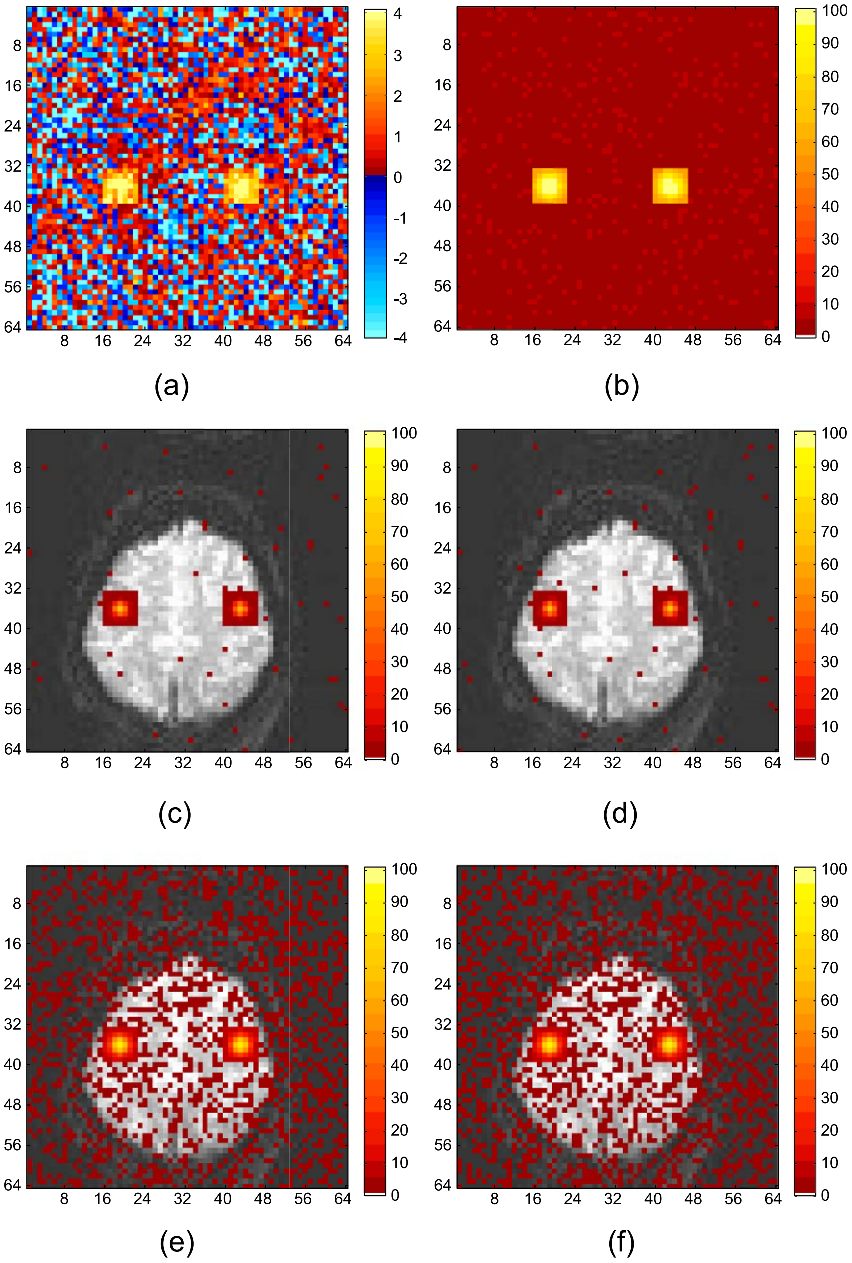
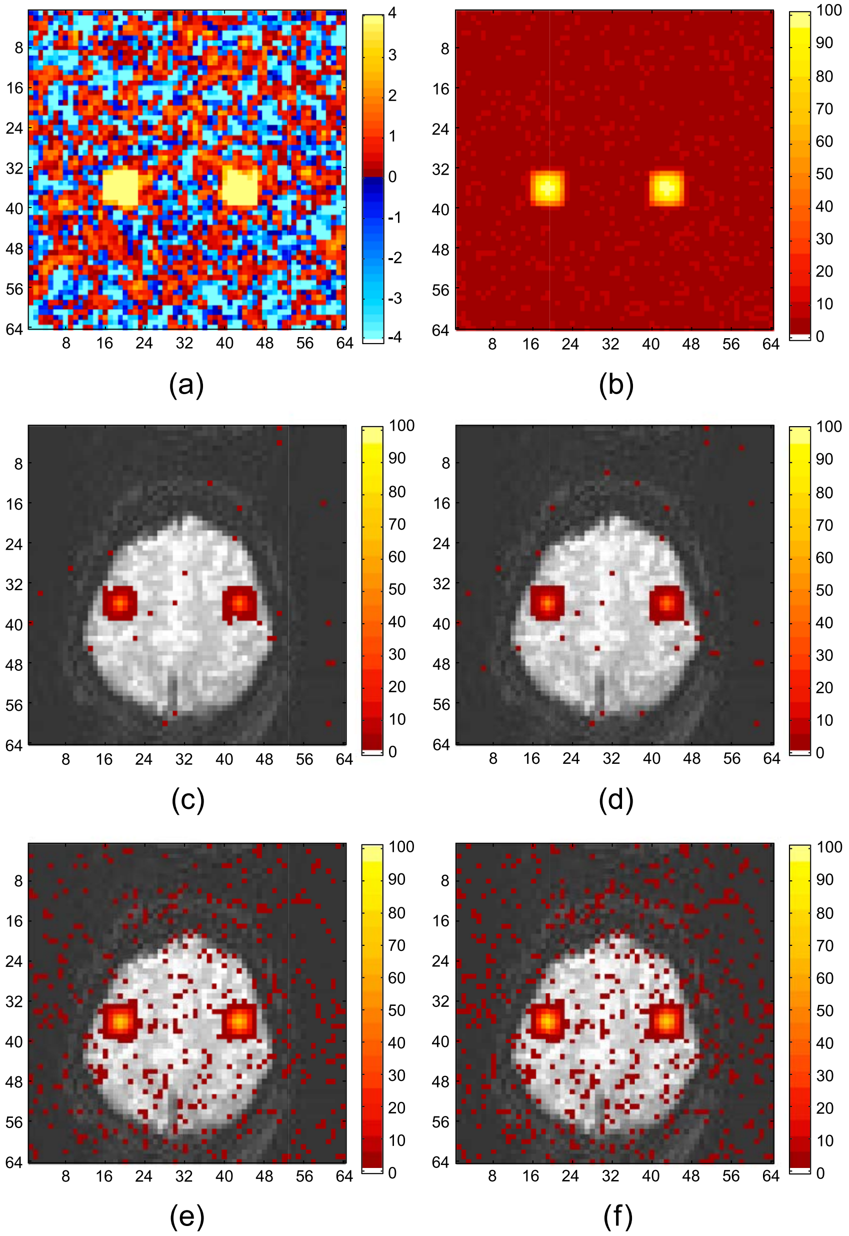


Fig. 6. Sample t statistic image and $\alpha = 0.05$ power images for $\rho = 0.00$. (a) Sample t statistic image, (b) unadjusted threshold, (c) FWE Bonferroni method, (d) FWE permutation method, (e) FDR BH method, (f) FDR YB method.

the multiplicity adjustment. However, there is still an important advantage in terms of power to use a FDR-controlling procedure rather than a FWE-controlling procedure again at the cost of more false positives, similar to when the correlation was 0.0.

When the spatial correlation is very strong (Exp(0.95)), then we can see evidence that permutation-resampling methods improve the power to detect voxel activations. The FWE permutation-sampling procedure detects a larger portion of the activation region with higher power than the FWE Bonferroni procedure. Similarly, the FDR YB resampling method also has higher power than the FDR BH method to detect voxel activations. For the spatial correlation of 0.95, the correlation between any voxel and a neighbor 6 and 12 voxels away is 0.74 and 0.54, respectively, and this is illustrated in Fig. 2c by a less sparse correlation map with a larger frequency of moderate to high correlation values. This substantially stronger spatial correlation is utilized by the resampling methods to improve the power relative to their non-resampling counterparts. As above, the FDR-controlling methods are more powerful than the FWE-controlling methods at the cost of more false negatives. However, it is important to recall that the FDR procedures control the expected rate of false discoveries relative to total discoveries. The actual false discovery rate may be higher than the targeted rate (e.g., 5%), and this can be especially problematic when the correlation is strong enough to induce substantial clustering in the t statistics. In this case, it may be more appropriate to consider methods which control the actual false discovery rate $\leq q$ with probability β, as discussed in Pacifico et al. (2003). However, as indicated by the sample t



(a) (b) (c) (d) (e) (f)

Fig. 7. Sample t statistic image and $\alpha = 0.05$ power images, 5 mm FWHM smoothed. (a) Sample t statistic image, (b) unadjusted threshold, (c) FWE Bonferroni method, (d) FWE permutation method, (e) FDR BH method, (f) FDR YB method.

statistic image, a spatial correlation of this extent and severity is unlikely to be encountered in single-subject fMRI data.

The correlation image for the “real” correlation structure in Fig. 2d exhibits similar characteristics of sparsity as the correlation image for Exp(0.7). Therefore, as expected, the relative performance of the thresholding methods is similar to the exponential covariance model structure with $\rho = 0.0$ or 0.7.

Smoothing induces a correlation matrix whose sparsity falls between that of the Exp(0.0) and Exp(0.7) covariance structures. As a result, there is little advantage to using permutation resampling to

Mult values by 2.

Table 2
Average power for various thresholding methods

Method	Exp(0.0)	Exp(0.7)	Exp(0.95)	“Real”	5 mm FWHM
Unadjusted	0.7459 (0.0014)	0.7481 (0.0036)	0.7530 (0.0071)	0.7557 (0.0028)	0.3329 (0.0151)
FWE Bonferroni	0.0924 (0.0008)	0.0915 (0.0020)	0.0937 (0.0034)	0.0937 (0.0015)	0.0340 (0.0019)
FWE permutation	0.0926 (0.0009)	0.0957 (0.0021)	0.1721 (0.0049)	0.0981 (0.0015)	0.0386 (0.0022)
FDR BH	0.2326 (0.0016)	0.2307 (0.0044)	0.2406 (0.0075)	0.2331 (0.0033)	0.0892 (0.0034)
FDR YB	0.2359 (0.0016)	0.2369 (0.0043)	0.3067 (0.0078)	0.2446 (0.0033)	0.0963 (0.0049)

Table 3
FDR for various thresholding methods

Method	Exp(0.0)	Exp(0.7)	Exp(0.95)	"Real"	5 mm FWHM
Unadjusted	0.7325 (0.0005)	0.7286 (0.0015)	0.6005 (0.0077)	0.7105 (0.0029)	0.7538 (0.0011)
FWE Bonferroni	0.0057 (0.0008)	0.0077 (0.0016)	0.0042 (0.0020)	0.0049 (0.0010)	0.0077 (0.0017)
FWE permutation	0.0060 (0.0008)	0.0083 (0.0016)	0.0159 (0.0033)	0.0056 (0.0010)	0.0108 (0.0022)
FDR BH	0.0502 (0.0014)	0.0490 (0.0022)	0.0303 (0.0040)	0.0512 (0.0026)	0.0446 (0.0024)
FDR YB	0.0527 (0.0015)	0.0557 (0.0027)	0.0538 (0.0050)	0.0571 (0.0027)	0.0532 (0.0026)

adjust the threshold for the correlation structure induced by smoothing, as illustrated by the power images in Fig. 7. Note, however, that permutation resampling may be useful in determining the appropriate scaling factor or marginal distribution after smoothing from which the thresholds are set, as discussed above. Finally, as in other scenarios, there are substantial advantages to using the FDR rather than the FWE to set the thresholds.

Tables 2–4 provide additional information on the magnitude of the power differences between the various methods as well as the error rates obtained. Table 2 presents the average power in the true activation regions for the various methods and correlation structures, while Tables 3 and 4 contain the estimates of the false discovery rates and family-wise error rates based on the simulated data. Standard errors are given next to each estimate in parentheses.

The FDR-controlling procedures appear to improve the average power in the active region by approximately 14%. The resampling adjustments to incorporate correlation improve the power by about 7%, but only when $\rho = 0.95$. Also of particular interest is whether the YB resampling method controls the FDR for positively correlated active and inactive voxels. The observed FDR for the YB method is slightly more than 1.96 SE above the nominal 5% value in some cases (Exp(0.7), "real"). There are several possible reasons for this slight elevation of the FDR rate. First, in a simulation study of this magnitude, we can only use a limited number of permutation samples (500). As a result, the permutation distribution may not be adequately filled in, and the error rate may be elevated as a result. Second, estimation of the correlation among the residuals by resampling the residual vectors may be inadequate because of the limited sample size ($n = 128$) and high dimension (${}_{4096}C_2$ correlation parameters) involved. The uncertainty associated with such estimation may inflate the error rate above the nominal value. Finally, the YB procedure has not been mathematically proven to control the FDR when D_F and D_T are correlated, and both cases qualify. However, the magnitude of the inflation of the FDR error rate is not large and should not be considered a major barrier to using this procedure. It will slightly inflate the power estimates, however, so that we should not give too much credence to small differences in power between the BH and YB procedures. Finally, keep in mind that a limited simulation study such as the one presented here may not reveal problems with the YB method in terms of control of the FDR for correlation structures encountered in fMRI studies.

As mentioned above, additional correlation structures were also considered, including some with negatively correlated regions of the brain. In all others, similar results were observed, where the correlation matrix was too sparse for the correlation-based multiplicity adjustments to improve the power to detect active voxels over methods which do not account for correlation.

Real fMRI Example

To illustrate the thresholding methods described in this paper, a bilateral finger-tapping experiment was performed with the same design as the previous simulation study. To generate the functional data, bilateral finger tapping was performed in a block design with eight epochs of 16 s on and 16 s off. Scanning was performed using a 3-T Bruker Biospec in which 15 axial slices of size 64 × 64 were acquired. Each voxel has dimensions in mm of 3.125 × 3.125 × 5, with TE = 27.2 ms. Observations were taken every TR = 2000 ms so that there are 128 in each voxel. Data from a single axial slice through the motor cortex were selected for analysis. A multiple regression model was fit to the data with an intercept, a time trend, and a reference function. The reference function was eight replicates of eight ones then eight negative ones, which mimics the experimental design in the simulation study.

After fitting the regression model, the correlation matrix was computed from the residuals and a correlation image was constructed. This sample correlation image shown in Fig. 2d does not have the same spatial correlation structure as the others considered in the simulation study. It appears that the magnitude and the sparsity of this real dataset are more similar to the situation where $\rho = 0.7$, and not as strong as when $\rho = 0.95$. Therefore, we would expect there to be little difference between using a thresholding method which does not account for spatial correlation and one which does account for spatial correlation. The real data t statistic image is given in Fig. 8, along with thresholded images using each of the methods discussed with a 5% error rate.

As expected, there are little to no differences between the Bonferroni and the permutation resampling FWE adjustment, or between the BH and the YB FDR adjustment. Also as expected,

Table 4
FWE for various thresholding methods

Method	Exp(0.0)	Exp(0.7)	Exp(0.95)	"Real"	5 mm FWHM
Unadjusted	1.0000 (0.0000)	1.0000 (0.0000)	0.9980 (0.0014)	1.0000 (0.0000)	1.0000 (0.0000)
FWE Bonferroni	0.0500 (0.0069)	0.0440 (0.0065)	0.0070 (0.0026)	0.0400 (0.0062)	0.0480 (0.0096)
FWE permutation	0.0540 (0.0072)	0.0480 (0.0068)	0.0510 (0.0070)	0.0460 (0.0066)	0.0640 (0.0110)
FDR BH	0.6910 (0.0146)	0.5300 (0.0158)	0.1140 (0.0101)	0.5080 (0.0158)	0.5400 (0.0223)
FDR YB	0.7090 (0.0144)	0.5640 (0.0157)	0.2150 (0.0130)	0.5570 (0.0157)	0.6040 (0.0219)

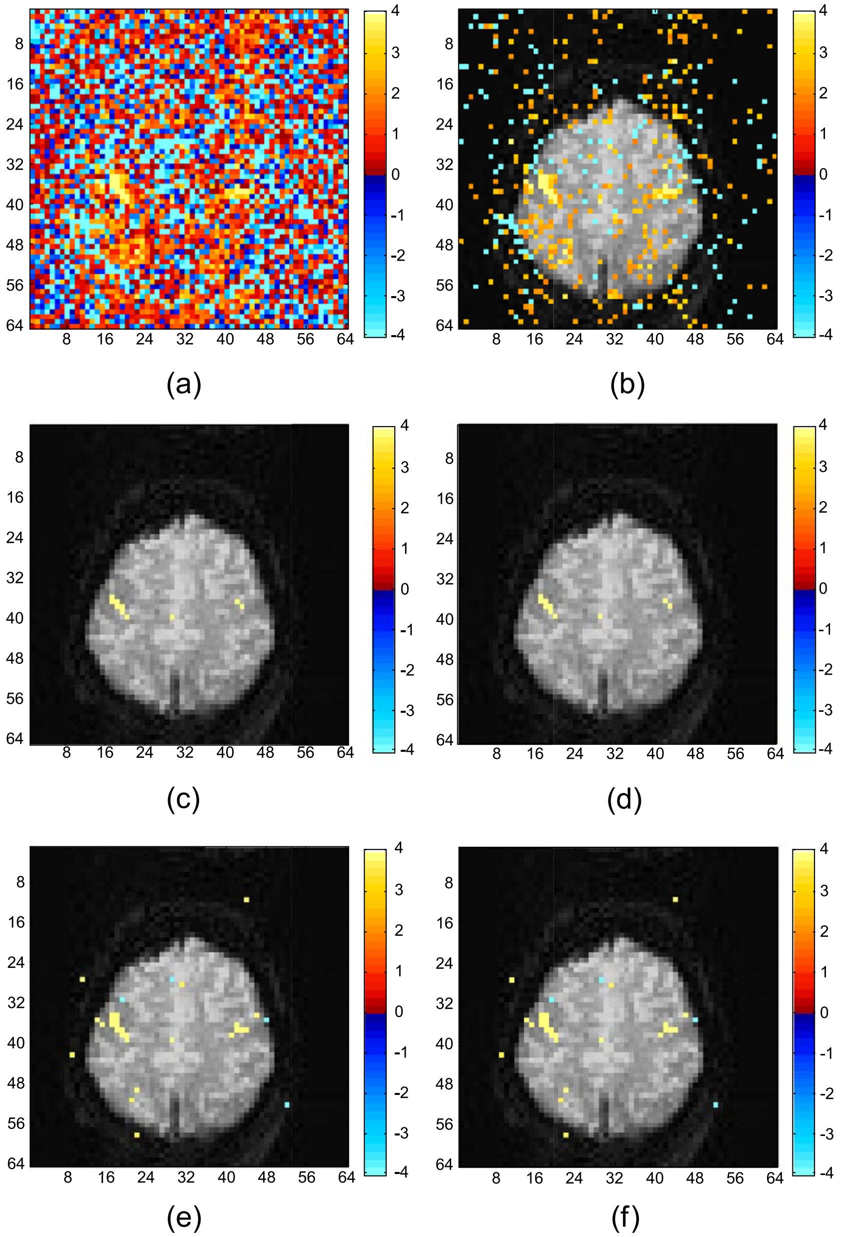


Fig. 8. Real data thresholded t statistic images for α = 0.05. (a) Sample t statistic image, (b) unadjusted threshold, (c) FWE Bonferroni method, (d) FWE permutation method, (e) FDR BH method, (f) FDR YB method.

there are more voxels above the threshold using the FDR adjustment than using a FWE adjustment because the error rate is less stringent.

Conclusion

This simulation study highlights two important findings. First, as has been indicated previously by other authors, the FDR-controlling methods generally have higher power than FWE-controlling methods to detect active voxels. The average magnitude of this power improvement was approximately 14% in the simulations considered, but this is likely to be sensitive to the underlying parameters involved and the size of the image considered. In general, the FDR criterion is more robust to the size of the image being considered than the FWE criterion (Holland and Cheung, 2002). However, the procedures are controlling two different error rates, so this higher power comes at the cost of a greater rate of false positives. For most fMRI applications, because of the large number of voxels considered, controlling the FWE is less appealing than controlling the FDR at a fixed rate α because the number of allowable voxels which are falsely declared active is then linked to the total number of voxels declared active.

Second, except when the spatial correlation is extremely strong (Exp(0.95)), voxel-wise thresholding methods which use resampling to account for correlation in the multiplicity adjustment do not have much impact on the power. This is probably due to the sparseness of the overall covariance matrix in many fMRI applications, in which most of the entries in the entire covariance matrix are close to 0, resulting in an ‘‘average’’ correlation close to 0. This phenomenon was found even for moderate local spatial correlation such as the Exp(0.7) or the 5 mm FWHM structures. Therefore, because of the extreme computational burden of doing such resampling and the difficult assumptions they require, the simple procedures, such as Bonferroni for controlling the FWE or Benjamini–Hochberg for controlling the FDR, are recommended in practice, unless there is a strong indication of a high correlation between a large number of the voxels in the image. Note that while we have focused on single-subject fMRI, multiple-subject fMRI studies often utilize more extensive smoothing which induces stronger spatial correlation. This may indicate that adjustment with one of the more computationally intensive methods is more appropriate for multiple subject studies; however, more research needs to be done on the effect of smoothing on multiplicity adjustment for multiple subject fMRI studies.

While incorporating correlation information does not appear to be important to voxel-wise thresholding rules, note that these findings do not apply to cluster thresholding methods, as in Friston et al. (1994), where an a priori cluster size is also used to set the threshold. In this case, the spatial correlation information may have a significant impact on the expected cluster size, as indicated by our sample *t* statistic images in Figs. 3–7, and the resulting *t* statistic–cluster size thresholding rule needs to be sensitive to this.

Finally, our simulation studies have focused on the effect of spatial correlation on the voxel-wise thresholding methods. fMRI datasets may include temporal autocorrelation, and while one may whiten the data as in Locascio et al. (1997) before applying a permutation resampling method, further work needs to be done to investigate whether our conclusions hold for prewhitened data as well.

References

Bandettini, P.A., Jesmanowicz, A., Wong, E.C., Hyde, J.S., 1993. Processing strategies for time-course data sets in functional MRI of the human brain. Magn. Reson. Med. 30, 161–173.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J.R. Stat. Soc., 289–300.

Benjamini, Y., Hochberg, Y., 2000. On adaptive control of the false discovery rate in multiple testing with independent statistics. J. Educ. Behav. Stat. 25, 60–83.

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29, 1165–1188.

Brammer, M., Bullmore, E., Simmons, A., Williams, S.C.R., Grasby, P.M., Howard, R.J., Woodruff, P.W.R., Rabe-Hesketh, S., 1997. Generic brain activation mapping in functional magnetic resonance imaging: a nonparametric approach. Magn. Reson. Imaging 15, 763–770.

Bullmore, E., Brammer, M., Williams, S.C.R., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., Sham, P., 1996. Statistical methods of estimation and inference for functional MR image analysis. Magn. Reson. Med. 35, 261–277.

Bullmore, E., Long, C., Suckling, J., Fadili, J., Calvert, G., Zelaya, F., Carpenter, T.A., Brammer, M., 2001. Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains. Hum. Brain Mapp. 12, 61–78.

Cox, R.W., Jesmanowicz, A., Hyde, J.S., 1995. Real-time functional magnetic resonance imaging. Magn. Reson. Med. 33, 230–236.

Cressie, N.A.C., 1993. Statistics for Spatial Data. Wiley, New York.

Friston, K.J., Frith, C.D., Liddle, P.F., Frackowiak, R.S.J., 1991. Comparing functional (PET) images: the assessment of significant change. J. Cereb. Blood Flow Metab. 11, 690–699.

Friston, K.J., Worsley, K.J., Frackowiak, R.S.J., Mazziotta, J.C., Evans, A.C., 1994. Assessing the significance of focal activations using their spatial extent. Hum. Brain Mapp. 1, 214–220.

Genovese, C.R., Wasserman, L., 2002. Operating characteristics and extensions of the false discovery rate procedure. J. R. Stat. Soc., Ser. B 64, 499–517.

Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15, 772–786.

Hochberg, Y., Tamhane, A.C., 1987. Multiple Comparison Procedures. Wiley, New York.

Holland, B., Cheung, S.H., 2002. Familywise robustness criteria for multiple comparison procedures. J. R. Stat. Soc., Ser. B 64, 63–77.

Holmes, A.P., Blair, R.C., Watson, J.D.G., Ford, I., 1996. Nonparametric analysis of statistic images from functional mapping experiments. J. Cereb. Blood Flow Metab. 16, 7–22.

Locascio, J.J., Jennings, P.J., Moore, C.I., Corkin, S., 1997. Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging. Hum. Brain Mapp. 5, 168–193.

Miller Jr., R.G., 1981. Simultaneous Statistical Inference, Second ed., Springer, New York.

Nichols, T.E., Holmes, A.P., 2001. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15, 1–25.

Pacifico, M.P., Genovese, C., Verdinelli, I., Wasserman, L., 2003. False Discovery Rates for Random Fields, Technical Report No. 771, Department of Statistics, Carnegie Mellon University.

Petersson, K.M., Nichols, T.E., Poline, J.-B., Holmes, A.P., 1999. Statistical limitations in functional neuroimaging II. Signal detection and statistical inference. Philos. Trans. R. Soc. Lond. B Biol. Sci. 354 (1387), 1261–1281.

Rowe, D.B., 2003. Multivariate Bayesian Statistics. CRC Press, Boca Raton, FL, USA.

Sarkar, S.K., 2002. Some results on false discovery rate in stepwise multiple testing procedures. Ann. Stat. 30, 239–257.

Storey, J.D., 2002. A direct approach to false discovery rates. J. R. Stat. Soc., Ser. B 64, 479–498.

Storey, J.D., 2003. The positive false discovery rate: a Bayesian interpretation and the Q-value. Ann. Stat. 31, 2013–2035.

Storey, J.D., Tibshirani, 2001. Estimating the positive False Discovery Rate Under Dependence, with Applications to DNA Microarrays, Technical Report No. 2001-28, Department of Statistics, Stanford University.

Storey, J.D., Taylor, J.E., Siegmund, D., 2004. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. J. R. Stat. Soc., Ser. B 66, 187–205.

Westfall, P.H., Young, S.S., 1993. Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment. Wiley, New York, USA.

Westfall, P.H., Tobias, R.D., Rom, D., Wolfinger, R.D., Hochberg, Y., 1999. Multiple Comparisons and Multiple Tests Using the SAS System. SAS Institute, Cary, NC, USA.

Worsley, K.J., Friston, K.J., 1995. Analysis of fMRI time series revisited-again. NeuroImage 2, 173–181.

Worsley, K.J., Evans, A.C., Marrett, S., Neelin, P., 1992. A three-dimensional statistical analysis for CBF activation studies in human brain. J. Cereb. Blood Flow Metab. 12, 900–918.

Worsley, K.J., Marrett, S., Neelin, P., Vandal, A.C., Friston, K.J., Evans, A.C., 1996. A unified statistical approach for determining significant signals in images of cerebral activation. Hum. Brain Mapp. 4, 58–73.

Yekutieli, D., Benjamini, Y., 1999. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J. Stat. Plan. Inference 82, 171–196.