Permutation Tests for Factorially Designed Neuroimaging Experiments

John Suckling* and Ed Bullmore

Brain Mapping Unit and Wolfson Brain Imaging Centre, University of Cambridge, Cambridge, United Kingdom

Abstract: Permutation methods for analysis of functional neuroimaging data acquired as factorially designed experiments are described and validated. The *F* ratio was estimated for main effects and interactions at each voxel in standard space. Critical values corresponding to probability thresholds were derived from a null distribution sampled by appropriate permutation of observations. Spatially informed, cluster-level test statistics were generated by applying a preliminary probability threshold to the voxel *F* maps and then computing the sum of voxel statistics in each of the resulting three-dimensional clusters, i.e., cluster "mass." Using simulations comprising two between- or within-subject factors each with two or three levels, contaminated by Gaussian and non-normal noise, the voxel-wise permutation test was compared to the standard parametric *F* test and to the performance of the spatially informed statistic using receiver operating characteristic (ROC) curves. Validity of the permutation-testing algorithm and software is endorsed by almost identical performance of parametric and permutation tests of the voxel-level *F* statistic. Permutation testing of suprathreshold voxel cluster mass, however, was found to provide consistently superior sensitivity to detect simulated signals than either of the voxel-level tests. The methods are also illustrated by application to an experimental dataset designed to investigate effects of antidepressant drug treatment on brain activation by implicit sad facial affect perception in patients with major depression. Antidepressant drug effects in left amygdala and ventral striatum were detected by this software for an interaction between time (within-subject factor) and group (between-subject factor) in a representative two-way factorial design. *Hum. Brain Mapp. 22:193–205, 2004.* © 2004 Wiley-Liss, Inc.

Key words: ANOVA; repeated measures; nonparametric; randomisation; pharmacological MRI

INTRODUCTION

The advantages of factorially designed experiments were first articulated clearly by Fisher [1935] and have since been endorsed widely by the biomedical scientific community. Factorial designs generally provide a more powerful basis for testing two or more experimental factors of interest (and their interactions) than alternative approaches, such as multiple tests each of a single factor. Factors may generally code fixed effects, random effects, or a mixture of both; in what follows, we are concerned with fixed effect designs. Typically, each factor will comprise two or more levels of an experimental or observational variable of interest and the design will be balanced, i.e., there will be an equal number of observations under all treatments or combinations of factors at different levels. The levels of a factor can be ordered or disjoint and the subjects may be measured re-

Contract grant sponsor: National Institute of Biomedical Imaging and Bioengineering; Contract grant sponsor: National Institute of Mental Health; Contract grant sponsor: GlaxoSmithKline plc; Contract grant sponsor: MRC.

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

Received for publication 29 August 2003; Accepted 9 February 2004
DOI 10.1002/hbm.20027
Published online in Wiley InterScience (www.interscience.wiley.com).

peatedly on all levels of a (within-subject) factor or each subject may be measured only on one level of a (between-subject) factor. Ideally, the subjects will have been randomly sampled from the population(s) of interest and randomly assigned to a particular set of treatments. The most usual method of analysing such data is by analysis of variance (ANOVA), which additionally assumes that the measurements are normally distributed and have equal variance between levels of each factor. Homogeneity of variance is particularly important for the condition of sphericity, which is necessary for the validity of repeated measures ANOVA [Mendoza et al., 1976]. The usual ANOVA statistic for hypothesis testing is the *F* ratio between mean sum of squares due to the factor in question and the mean sum of squares due to error.

Although factorial designs have been used already quite extensively in functional neuroimaging, it seems plausible that not all conditions for validity of their analysis by parametric *F* tests will always be upheld. For this reason alone, it may be useful to have tools for relatively "distribution-free," nonparametric analysis of factorially designed experiments. Additionally, nonparametric methods may provide the freedom to test potentially more sensitive but theoretically less tractable statistics than the *F* ratio estimated independently at each voxel of an image.

Nonparametric methods of hypothesis testing based on random resampling or permutation of the observed data have been described previously in the context of linear modelling of functional magnetic resonance imaging (fMRI) time series [Bullmore et al., 1996, 2001a; Forman et al., 1995; Locascio et al., 1997] as well as between-group comparisons based on positron emission tomography (PET) data [Arndt et al., 1996; Holmes et al., 1996], structural MRI [Bullmore et al., 1999a, 2001b; Thompson et al., 2001], and functional MRI data [Nichols and Holmes, 2002]. Moreover, there is prior literature on development of permutation tests for factorially designed experiments in biophysical and environmental areas of application [Edgington, 1995; Good, 2000; Still and White, 1981; Welch, 1990]. To the best of our knowledge, however, these versatile techniques have not been validated previously in relation to analysis of factorially designed neuroimaging experiments.

We describe algorithms for testing by permutation the statistical significance of main effects and interactions in any two-way factorial design. The methods are validated by comparison to parametric *F*-tests in voxel-level analysis of simulated images and extended to cope with spatially informed, cluster-level statistics. Exponential forms of the null distribution for cluster extent statistics have been theoretically derived previously for brain mapping [Cao, 1999; Cao and Worsley, 2001; Friston et al., 1994; Poline et al., 1997], but often may be over-conservative [Ashburner and Friston 2000; Bullmore et al., 1999a]. We have reported previously superior Type 1 error control by a permutation test (compared to a theoretical test) of a cluster-level statistic for analysis of a between-group difference in brain imaging data [Bullmore et al., 1999a]. We report here the generalisation of this approach to an arbitrary two-way factorial design and evaluate its sensitivity compared to voxel-level tests using receiver operating characteristic (ROC) curves constructed by analysis of simulated images. Finally, we illustrate application of the methods by analysis of a parallel group, repeated measures, placebo-controlled pharmacologic MRI study of antidepressant drug effects.

MATERIALS AND METHODS

Parametric and Permutation Tests for Main Effects and Interactions

Notation

In a factorially designed experiment, imaging data are acquired from subjects labeled $k = 1,2,3,\ldots,K$ within each treatment. After appropriate preprocessing and time-series modelling of each of these datasets, statistical maps of an estimated standardised parameter $\hat{\beta}$, describing some aspect of functional response in each individual, are coregistered into a standard stereotactic space. It is assumed that each intracerebral voxel, $v = 1,2,3,\ldots,V$, represents the same anatomic location within the parenchyma of the brain in every individual. These images are then treated as dependent variables in a univariate factorial analysis to assess at each voxel the effects of the treatments under which the data were collected. The treatments are specified by a combination of two fixed-effect factors *A* and *B* with levels indexed by $i = 1,2,3,\ldots,I$ and $j = 1,2,3,\ldots,J$, respectively, such that the estimate of the response of the *k*th individual (at voxel *v*) is denoted $\hat{\beta}_{ijk}$.

Estimation of F statistics

F statistics for main effects and interactions at each voxel are calculated via sum-of-squares. Account is taken in these calculations of whether the measures are independent or repeated measures or a mixed design with independent measures on one factor and repeated measures on the other [for details of the calculations, see Coolican, 1999].

In a one-way design, the *F* ratio is a comparison of two estimates of the sample variance, one from the variability within each group, the other from the variability of the means between each group. Under the null-hypothesis both estimates of sample variance, i.e., sum-of-squares (SS) divided by the appropriate degrees of freedom, are identical. In a two-way factorial design the total variability, SS_{total}, is partitioned into the variability of the treatments, SS_{treat}, and then partitioned further into the variability of each the main effects of *A* (ignoring levels of *B*), SS_A, and *B* (ignoring levels of *A*), SS_B, and the interaction between them, SS_{AB} (Fig. 1a). *F* ratios are then estimated for each main effect and interaction with error variance in the denominator.

When all subjects undergo all treatments in a two-way repeated measures design, calculations are made as if this were a three-way design with subjects as a factor and each individual a level of that factor (Fig. 1b). SS_{total} is partitioned initially into between-subject variability, SS_{bet}, and within-

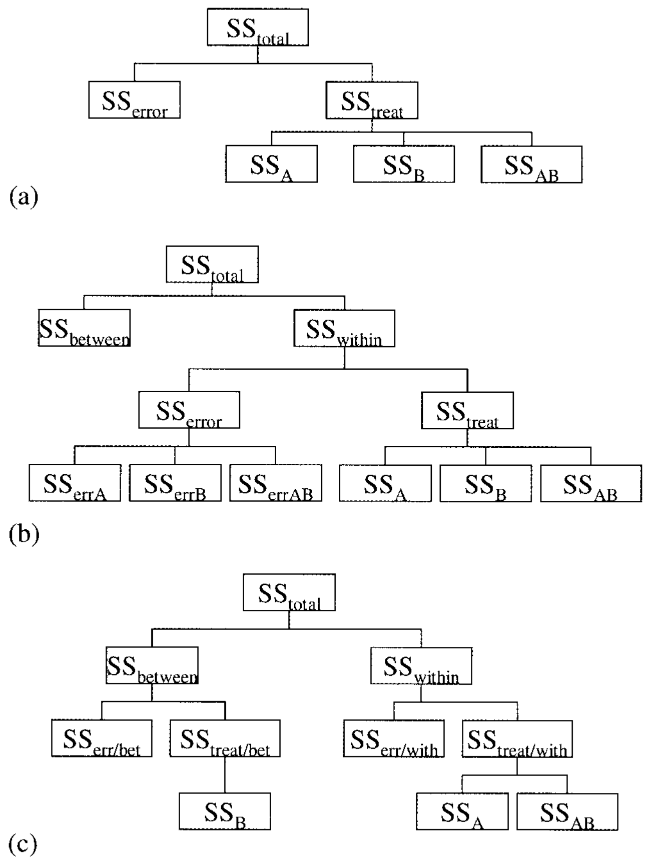

Figure 1.

Representation of the partitioning of variability (sum-of-squares) in a two-way ANOVA. **a:** Both factors are independent. **b:** Both factors are repeated measures. **c:** Factor *A* is a repeated measure and factor *B* is independent.

subject variability, SS_{with}, which is partitioned into the variability due to the treatments (SS_A, SS_B and SS_{AB}) and estimates of error SS for each main effect and interaction (SS_{errA}, SS_{errB} and SS_{errAB}). The error variances are then compared to the corresponding variance for each main effect and interaction to estimate the *F* ratios.

For a mixed design with repeated (factor *A*) and independent (factor *B*) observations within- and between-subject variability is partitioned further into variability for the appropriate effect and error terms (Fig. 1c). SS_{with} thus has a $SS_{err/with}$ term and terms for the factors involving repeated measures, SS_A and SS_{AB}, and the *F* ratios calculated for these effects. SS_{bet} is partitioned into an error term $SS_{err/bet}$ and term for the main effect of factor *B*, SS_B and the corresponding *F* value obtained.

Parametric tests for main effects and interactions

Tests against the parametric null-distribution were carried out with the DCDFLIB software (Department of Biomathematics, University of Texas). The validity of the parametric test is predicated on the assumptions of random sampling from the population, random assignment of subjects to treatments, normally distributed observations, and that these distributions should have homogeneous variance when observations are independent. For repeated measures, the assumption of homogeneity of variance is implicit in the assumption of sphericity [Mendoza et al., 1976].

Permutation tests for main effects and interactions

Informally, the null hypothesis to be tested by permutation of data acquired in a factorially designed experiment is that the magnitude of an observed test statistic is not determined by the treatments experimentally associated with each unit of observation $\hat{\beta}_{ijk}$, but would be reasonably likely to occur under any arbitrary reassignment of observations to treatments. To test this hypothesis operationally, observations are permuted randomly across levels of the factor of interest and the test statistic is re-estimated after each permutation. By repeating this procedure *m* number of times and ordering the set of permuted test statistics that results, the distribution of the test statistic under the null hypothesis is sampled and from it critical values for valid hypothesis testing are derived. In imaging, where test statistics may be estimated at several thousand voxels, the computational demands of sampling the permutation distribution may be mitigated substantially by permuting the data a small number of times, for example $m = 10$, at each voxel and then pooling permuted statistics over all voxels to construct a null distribution of mV observations.

This procedure must be refined according to whether the test is of a main effect or interaction, and whether the factor to be tested is a between-subject or a within-subject factor. In the latter case, the key issue is that repeated measurements on the same individual will be correlated and cannot be regarded as exchangeable.[1] The permutation must therefore be constrained so that each individual contributes only one observation to each level of the permuted within-subject factor.

For a test of significance of a main effect (in the absence of an interaction), the permutation distribution is sampled simply by permuting observations within levels of the corresponding factor. This supports an exact test of size α, i.e., the number of (false) positive tests (FP) under the null hypothesis is exactly as expected: $FP = E(FP) = \alpha V$. If a significant interaction has been detected, then subsequent testing for main effects is by an approximate test based on unrestricted permutation of the observations among all treatments [Anderson and ter Braak, 2003]. An approximate test is valid but possibly conservative in that the number of false positive tests may be less than or equal to expectation, i.e., $FP \leq E(FP) = \alpha V$.

[1] A set of *n* units of observation of the random variable X is termed exchangeable if the joint probability distribution $p(X_1, X_2, X_3, \ldots X_n)$ is invariant under permutation of the units; see Lindley and Novick [1981] for detail.

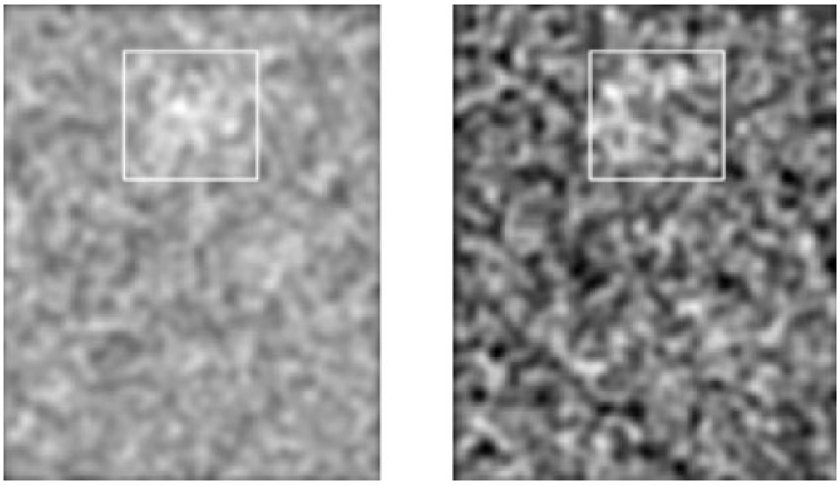

Figure 2.
Illustrations of the simulated images with effect size $\gamma = 1.0$. (left) Gaussian noise background and (right) non-Gaussian noise background. The effect region is indicated by the superimposed outline.

In general, there are no exact permutation tests with F statistics for interactions, although it is possible to construct approximate tests [Anderson and ter Braak, 2003; Still and White, 1981]. First the observations are replaced by residual values, $\tilde{\beta}_{\mathrm{ijk}}$, given as,

$$\tilde{\beta}_{\mathrm{ijk}} = \hat{\beta}_{\mathrm{ijk}} - \hat{\beta}_{\mathrm{i}\cdot\cdot} - \hat{\beta}_{\cdot\mathrm{j}\cdot} + \bar{\beta} \quad (1)$$

where · denotes the mean of the observations over the corresponding index and $\bar{\beta}$ denotes the overall mean. F statistics are then estimated for the residuals and tested against a permutation distribution sampled by random reassignment of residualised observations. If the interaction to be tested is between two between-subject factors, then permutation is unrestricted over all levels of both factors. If one or both of the interacting factors are within-subject factors, then permutation of the residuals is constrained within subject across all treatments.

Cluster-level tests

Permutation distributions of F statistics for main effects or an interaction, pooled across all voxels, were used to derive a preliminary, voxel-level threshold at $\alpha = 0.05$, which was then applied to observed and permuted F maps identically. The thresholding operation sets to zero any voxel with F less than the corresponding critical value (CV), i.e., $F_v < \mathrm{CV}_{0.05}$, and shrinks any suprathreshold voxel by subtraction of the threshold. This procedure results in a set of suprathreshold voxel clusters, each of which comprises the set of C voxels that are spatially contiguous in three dimensions, in both the observed F map and each of the permuted F maps. The sum or "mass" of suprathreshold voxel statistics M is computed for each cluster,

$$M = \sum_{v \in C} (F_v - \mathrm{CV}_{0.05}) \quad (2)$$

in both the observed and permuted maps. The values of M obtained from the permuted maps are then ordered to sample the permutation distribution from which critical values are derived to test the significance of clusters in the observed maps.

The validity of this test depends critically on the spatial covariance structure of the observed F maps being retained under permutation. This is achieved by ensuring that the set of permutations used to generate the permuted F maps are identical at each voxel. If different permutations are applied at each voxel, then the spatial covariance structure of the observed F maps will be destroyed or "whitened" and clusters generated by thresholding of the permuted F maps will tend to be smaller as a result. The null distribution of M will underestimate the true probability of a cluster of arbitrary size and there will consequently be uncontrolled Type 1 error on testing the observed cluster maps.

Operational details

Code was written in the C-language. Run-times were dependent on the number of images in the experiment, but were typically in the range of 3–5 min of processing time on a 2.6-GHz Pentium III with 1 Gb of memory.

Simulated Data

Simulated images were composed by adding an effect to a background of Gaussian or non-Gaussian noise. Gaussian noise was generated with mean = 1,024, standard deviation (SD) $\sigma_N = 1/3$ mean; non-Gaussian noise was generated by taking the cube of the exponential of a canonical pseudorandom variable on the interval [0,1]. Independent and repeated measures were simulated. In the latter case, a constant value was added to all the voxels of a set of images, one under each treatment, for a particular subject.

In a small region (1,024/136,800 voxels) an effect, i.e., an offset in grey levels, was added (Fig. 2). The maximum value ($\gamma = 1$) of effect was a signal-to-noise ratio (SNR) of 2.5 db, where

$$\mathrm{SNR} = 10.\log_{10}\left[\frac{S^2}{\sigma_N^2}\right] \quad (3)$$

TABLE I. Description of simulated experiments

Values of j	Values of i		
	$i = 1$	$i = 2$	$i = 3$
Study 1			
$j = 1$	0.00	0.25	0.50
$j = 2$	0.50	0.75	1.00
Study 2			
$j = 1$	0.00	0.50	1.00
$j = 2$	1.00	0.50	0.00
Study 3			
$j = 1$	0.50	0.75	1.00
$j = 2$	0.50	0.25	0.00

Study 1, main effects with no interaction; study 2, interaction with no main effects; study 3, one main effect and interaction. Table values are the multiplicative factors γ for a maximum effect signal-to-noise ratio = 2.5 db.

σ_N is the SD of the noise and S is the effect size. All simulations were smoothed via the Fast Fourier Transform with a 3-D Gaussian kernel of SD one voxel. The magnitudes of the effects under the various treatments (γ.SNR) were chosen to simulate results of three separate two-way factorial experiments, each comprising one two-level factor ($I = 2$) and one three-level factor ($J = 3$), with six observations under each treatment. As detailed in Table I, Study 1 simulated main effects with no interaction; Study 2 simulated an interaction with no main effects; and Study 3 simulated both a main effect and an interaction.

Receiver operating characteristic curves

Comparative evaluation of hypothesis-testing methods was based on construction of ROC curves. ROC curves are a well-established method for comparing alternative methods. Generally, ROC curves are monotonic and improved performance is indicated by increasing area under the curve. For a given size of test α, the number of true positives (TP) and false negatives (FN) identified among voxels from the region of simulated effect were recorded along with the numbers of true negatives (TN) and false positives (FP) identified among voxels comprising the background region of the image. The relationship between true positive ratio TPR = TP/(TP + FN) and false positive ratio FPR = FP/(FP + TN) was explored in the range of sizes of test: $0.00001 < \alpha < 0.05$.

Experimental (fMRI) Data

Experimental data were acquired as part of a pharmacologic MRI study of antidepressant drug effects in patients with major depression compared to healthy controls [see Mitterschiffthaler et al., 2003]. Two groups of 10 subjects were each scanned twice in two sessions 8 weeks apart. Functional MR images were acquired in a two-way factorial design comprising one between-subject factor *Group* with two levels (depressed patients satisfying DSM-IV criteria for major affective disorder and normal comparison subjects) and one within-subject factor *Time* with two levels (Session 1 and Session 2). Patients with depression were untreated at the time of the baseline scan but immediately afterwards they began treatment with the antidepressant drug fluoxetine (50 mg), which was continued for 8 weeks until the time of the second scanning session. The Group × Time interaction in this experiment is therefore an index of effects of antidepressant drug exposure on functional brain activation.

The study was approved by the Ethics (Research) Committee of the South London and Maudsley NHS Trust and all participants provided informed consent in writing.

At each scanning session, participants were shown a series of 60 facial stimuli expressing variable degrees of sadness randomly interspersed with 12 crosshair fixation trials. Each trial was presented for 3,000 msec in an event-related design with interstimulus interval randomly variable according to a Poisson distribution with mean = 5,000 msec; total duration of the experiment was therefore 9 min 36 sec. Participants were asked to decide on the gender of each face and indicate that decision by right-handed button press.

During stimulus presentation, gradient echo single-shot echoplanar imaging was used to acquire 180 T2*-weighted image volumes on a neuro-optimised 1.5-T IGE LX System (General Electric, Milwaukee, WI) at the Maudsley Hospital, South London, and Maudsley NHS Trust, London, UK. For each volume, 16 non-contiguous axial planes parallel to the intercommissural plane were collected with the following parameters: TR = 2,000 msec, TE = 40 msec, slice thickness = 7 mm, slice skip = 0.7 mm, in-plane resolution = 3 × 3 mm, and matrix size = 64 × 64.

After correction of slice timing differences and head movement-related effects in the fMRI time series at each voxel, linear regression was used to estimate experimentally induced signal changes, $\hat{\beta}$, for each individual [Bullmore et al., 1999b; Bullmore et al., 2001a]. Regression analysis modelled two mutually orthogonal aspects of brain activation at each voxel: a face-processing effect due to differential activation between baseline trials and all facial trials and an affective load-response effect due to differential activation between facial trials of variable affective intensity. Before model fitting, each contrast was convolved with two Poisson kernels (λ = 4 or 8 sec) to model locally variable hemodynamic response functions. The standardised linear model parameter $\hat{\beta}$ was calculated by dividing the estimated effect size, $\hat{b}$, from regression of the general linear model by its standard error, i.e., $\hat{\beta} = \hat{b}/\mathrm{SE}(\hat{b})$. This standardisation acts to suppress signals with large residual variances that often occur in regions of the image that are prone to artefact. The statistic maps representing face-processing effects for each individual in both scan sessions were registered into the standard space of Talairach and Tournoux [1988] by affine transformation to a template image [Brammer et al., 1997]. The analysis of affective load-response will not be discussed further here.

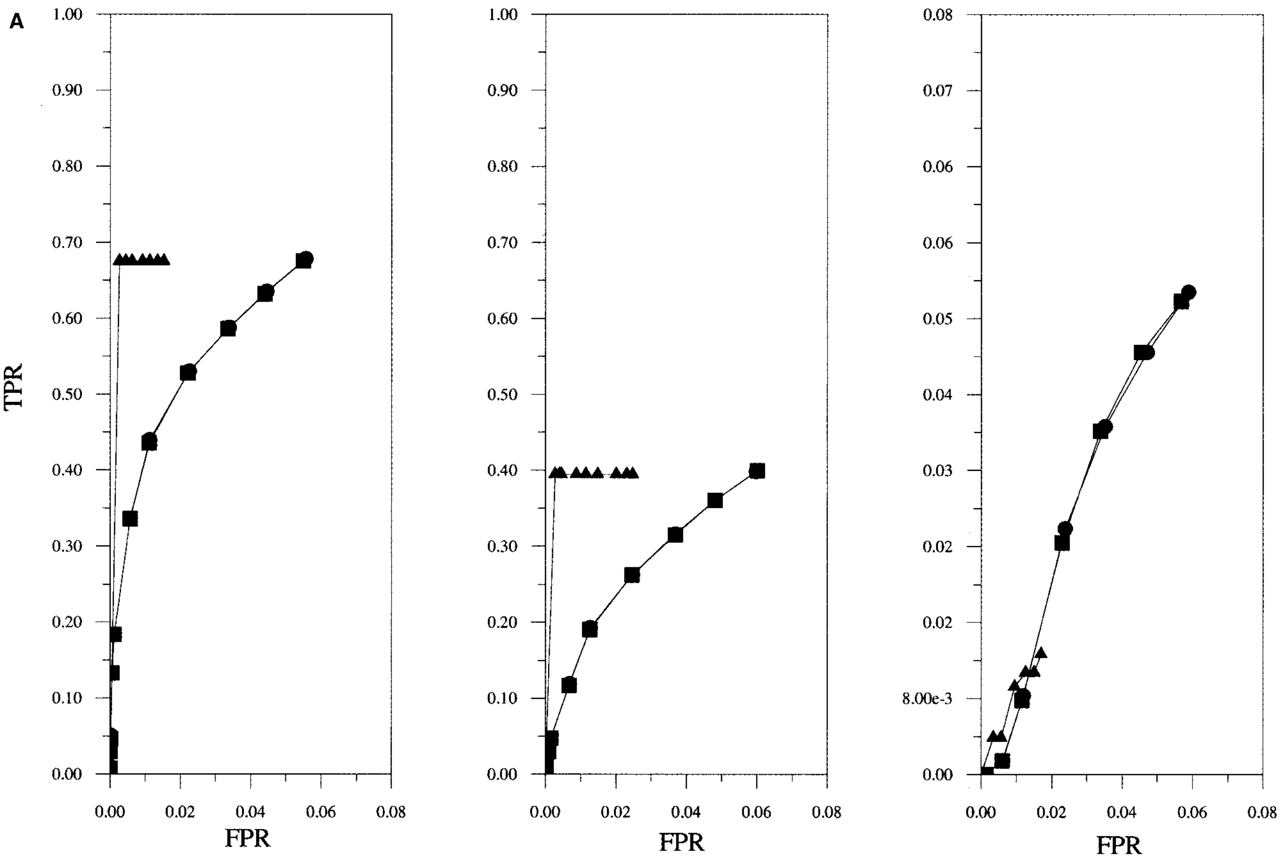

Figure 3.

Receiver operating characteristic (ROC) curves for simulated 2 × 3 factorial designs (see Table I) with independent measures on both factors and Gaussian random noise background. **a:** Left and centre, tests for factors *A* (two levels) and *B* (three levels), respectively; right, test for an interaction but without the presence of an effect. **b:** Left and centre, tests for factors *A* and *B*, respectively, but without the presence of an effect; right, test for interaction. **c:** Left, test for factor *A*; centre, test for factor B but without the presence of an effect; right, test for interaction. In each simulation, the ROC curves are shown for voxel-wise *F* statistics tested against the parametric *F* distribution (circles), a null-distribution sampled by permutation (squares) and for a spatial extent statistic, cluster mass, tested against its permutation distribution (triangles).

In short, preprocessing and time-series modeling of the individual images resulted in a set of 40 statistic maps, each representing the estimated face-processing effect at each voxel under all treatments of a balanced 2 × 2 factorial design. The *F* statistic for Group × Time interaction was tested for significance by the permutation test of cluster mass described above.

The rationale for using this experimental dataset to illustrate and validate these methods of factorial analysis is two-fold: (1) functional MRI is being used increasingly to investigate psychopharmacologic drug effects on brain function [see for example Bullmore et al., 2003; Honey et al., 1999, 2003, and references therein] and such pharmacologic MRI studies will invariably involve a factorial design of some degree; and (2) more specifically, there are comparable prior fMRI and PET studies of antidepressant drug effects on brain activation estimated by region-of-interest (ROI) analysis that provide a context for evaluation of the results of whole-brain analysis reported here.

RESULTS

Simulated Data

ROC curves for each simulated experiment with two between-subject factors are shown in Figure 3, and for each simulated experiment with two within-subject factors in Figure 4. In both cases the noise was Gaussian. In general, voxel-level tests for a main effect or interaction had the same area under the curve whether the corresponding *F* statistic was tested by permutation or against critical values of the *F* distribution. This means that permutation and parametric tests have virtually identical power to detect effects in these simulated data, which both validated the algorithms for data permutation and illustrated the robustness of the parametric

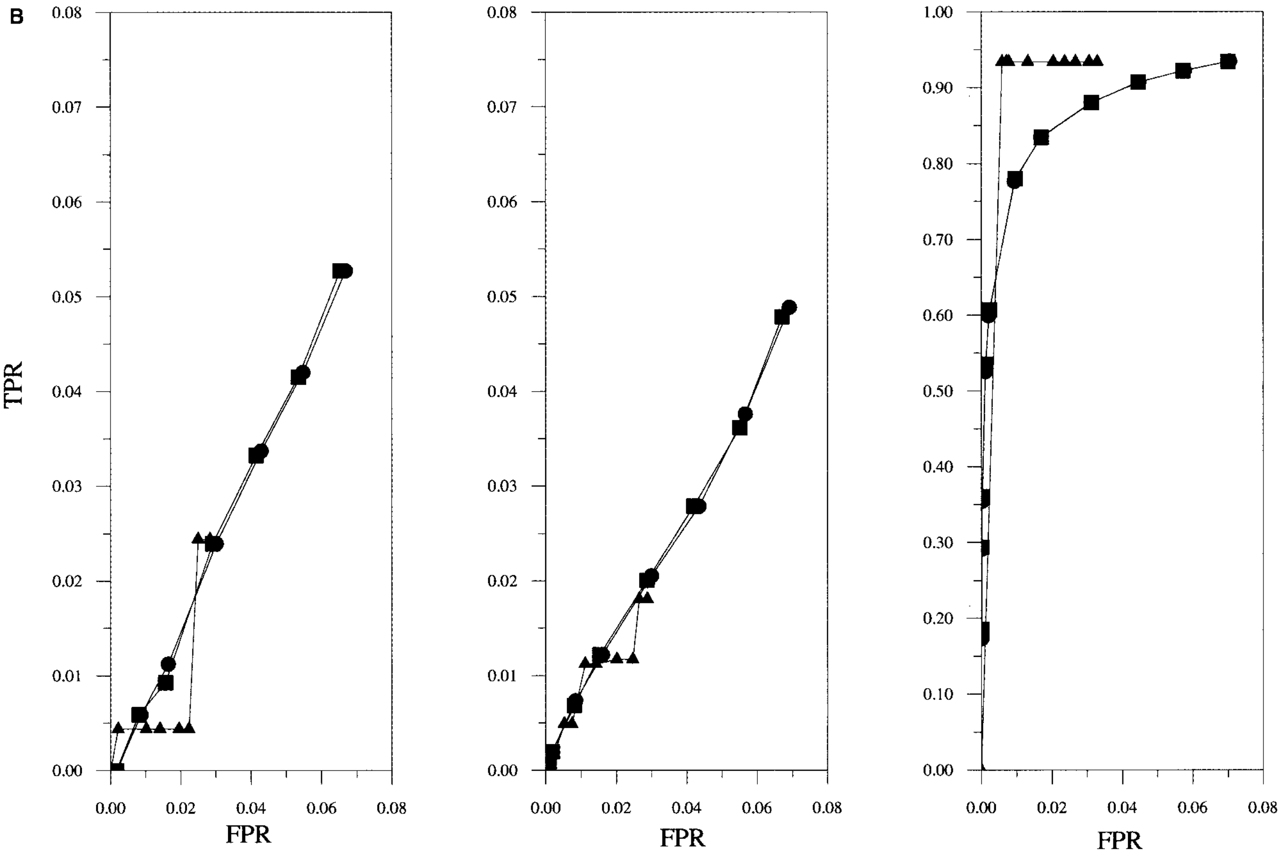

Figure 3.

test in the context of a realistically small sample size. These results were not changed substantially by substitution of non-Gaussian noise (data not shown); parametric and permutation tests at voxel level demonstrated near-identical performance under these conditions.

It is also evident from inspection of the ROC curves that the permutation test for main effects and interactions at the cluster level had consistently superior sensitivity compared to both voxel-level tests. True-positive and false-positive rates in the ROC curves were based on the number of voxels detected in the effect and non-effect regions of the simulation, respectively. As the *P*-value was increased, clusters of multiple voxels with a cluster-level statistic above the critical value were included and abrupt changes in the ROC curve resulted.

Significance testing for main effects (Studies 2 and 3) and interaction (Study 1) in data for which no effect was simulated were used to calibrate Type 1 error control. All tests demonstrated exact or slightly conservative Type 1 error control by this standard (Figs. 3 and 4).

Experimental Data

A cluster-level permutation test for Group × Time interaction identified a significant effect in the following left brain regions: amygdala, ventral striatum, pregenual anterior cingulate cortex (Brodmann's area [BA] 24, 32), and thalamus (Fig. 5). This interaction was characterised by increased effect size in the depressed patients at baseline compared to controls ($t = -4.29$, $df = 36$, $P = 0.0001$), which "normalised" over the course of the experiment so that there was no significant difference between groups at the second (8-week) session ($t = 3.157$, $df = 36$, $P = 0.0033$; see Fig. 5 for boxplot).

DISCUSSION

An algorithm for the analysis of any two-way factorially designed neuroimaging experiment has been described, validated by analysis of simulated data, and illustrated by application to mapping of a pharmacologic MRI experiment.

Permutation tests for voxel-level analysis of main effects and interactions have been shown to have virtually identical performance compared to equivalent parametric *F*-tests of the same data. This result validates the methods prescribed for data permutation, including the operational refinements required to test main effects of within-subject factors and interactions between factors. It also indicates that our largely pragmatic decision to sample the permutation distribution by pooling permuted *F* statistics over all voxels in the image does not seriously bias estimation of the null distribution for a voxel-wise test.

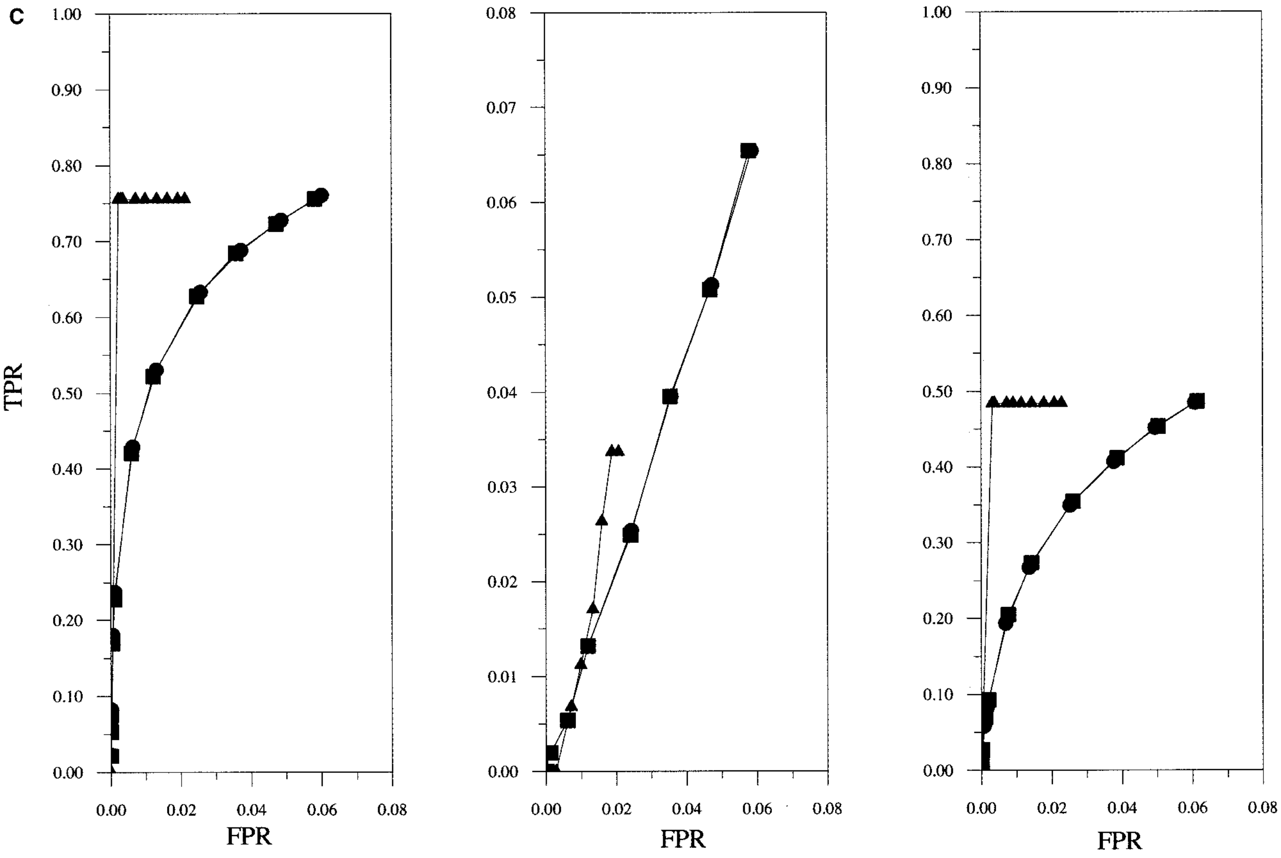

Figure 3.

We have also confirmed previous reports of radically enhanced sensitivity by cluster-level analysis [Poline and Mazoyer, 1993] and validated Type 1 error control for a cluster-level permutation test of main effects and interactions by analysis of factors in Gaussian and non-Gaussian noise images with zero simulated effect. We have not compared the cluster level permutation test for factorial effects to a parametric alternative because, to the best of our knowledge, no well-validated parametric test for cluster "mass" statistics in brain mapping has yet been defined. A validation of a cluster size statistic [Hayasaka and Nichols, 2003] has shown that parametric methods for this metric are valid, although conservative, above certain image smoothness. In the same study, permutation methods performed well for all degrees of image smoothness and were robust when assumptions underpinning the parametric test were violated.

Finally, we have illustrated the methods by analysis of a representative pharmacologic MRI study. We chose this dataset for illustrative purposes because factorial designs are an inevitable aspect of pharmacologic MRI studies and, more specifically, because there is prior data on antidepressant drug effects in amygdala. Sheline et al. [2001] reported a two-way factorial experiment in which two groups of participants (patients with major depression and healthy controls) were each scanned twice, at baseline and 8 weeks later, during visual presentation of masked emotional faces. The patients received treatment with sertraline (100 mg daily), a selective serotonin reuptake inhibitor, for 8 weeks beginning immediately after the baseline scan. By ROI analysis of voxel statistics, focused on the amygdala, Sheline et al. [2001] demonstrated a treatment-related change in amygdala activation by emotional face processing that was very similar to the effect reported here, i.e., there was initially increased left amygdala activation in the patient group that normalised over the course of 8-week antidepressant treatment. Our replication of this result by a whole-brain analysis, without prior specification of anatomic ROIs, provides some informal validation of our methods applied to experimental data analysis. Moreover, the capacity of our analysis to demonstrate a plausible additional locus of antidepressant treatment effect in the left ventral striatum indicates the potential benefits of using more sensitive cluster-level statistics to define regions of significant drug effect in a factorially designed pharmacologic MRI experiment.

The use of permutation methods brings with it the opportunity to test statistics for factorial designs other than *F*. These include metrics that are rank equivalent to *F* [Edgington, 1995] but rather more rapid to calculate. The large computational requirement of permutation methods has been a criticism of permutation methods in general and algorithms to speed up calculations have been sought. Contemporary processor clock speeds continue to increase un-

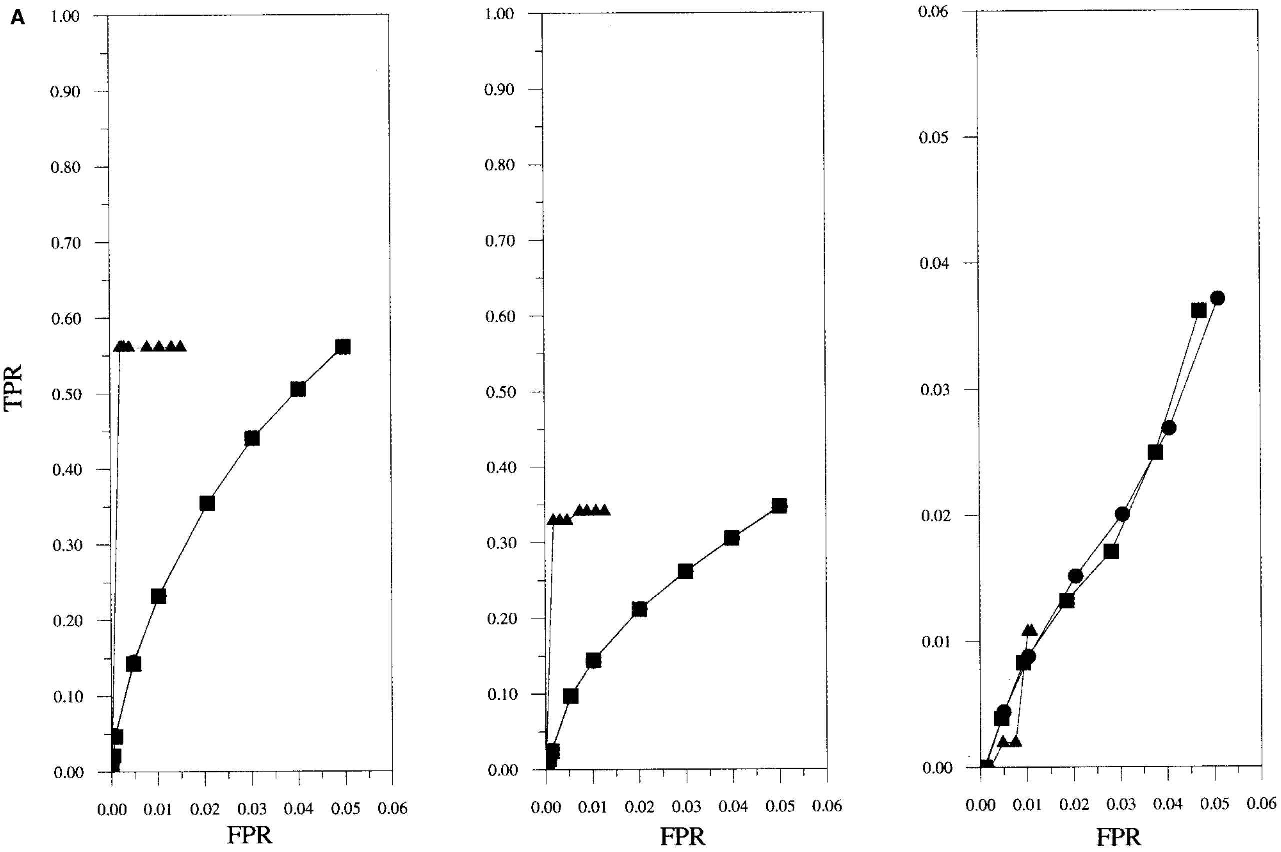

Figure 4.

Receiver operating characteristic (ROC) curves for simulated 2 × 3 factorial designs (see Table I) with repeated measures on both factors and Gaussian random noise background. **a:** Left and centre, tests for factors *A* (two levels) and *B* (three levels), respectively; right, test for an interaction but without the presence of an effect. **b:** Left and centre, tests for factors *A* and *B*, respectively, but without the presence of an effect; right, test for interaction. **c:** Left, test for factor *A*, centre, test for factor *B* but without the presence of an effect; right, test for interaction. In each simulation, the ROC curves are shown for voxel-wise *F* statistics tested against the parametric *F* distribution (circles), a null-distribution sampled by permutation (squares), and for a spatial extent statistic, cluster mass, tested against its permutation distribution (triangles).

abated making this debate somewhat specious. With compute power no longer at a premium there are many other statistics that could be devised, accounting for unbalanced designs [Good, 2000] or for exact tests of interactions [Pesarin, 2001].

The key point with any statistic assessing main effects and interactions in the context of the algorithm presented herein is that it should be pivotal, that is, it should not depend on the parameters of the distribution of the original observations from which it was calculated, especially the grey-level mean of the images at that voxel. This is a necessary condition to pool *F* ratios from all permutations and all voxels, allowing for a more accurate sampling of the null-distribution from which the initial probabilistic threshold $CV_{0.05}$ is found, and to apply this threshold uniformly to all voxels to generate the 3-D clusters subsequently tested by *M*.

Implicit in the uniform voxel threshold is that the statistic *M* is also pivotal with regard to its spatial distribution, i.e., it is assumed that the spatial covariance (smoothness) of the statistic image is homogeneous. Regions of increased smoothness are more likely to generate clusters of large *M* (or indeed any other similar statistic) and thus detection of significant effects in these regions is enhanced. Nevertheless, the test remains valid overall, although specificity may be nonuniform. Changes in the spatial smoothness in functional images may be modality specific, for example, susceptibility artefacts in MRI data. Image edges are a further source of inhomogeneous smoothness, especially if the data have been masked previously to process parenchymal regions. Generally, however, such artefacts serve to reduce smoothness and thus specificity in these regions.

Notwithstanding these issues, the broader class of statistics that can be tested by permutation methods illustrates the fundamental difference in the null hypothesis tested in comparison to parametric methods. In general, a parametric test tests a specific and quantitative null hypothesis concerning,

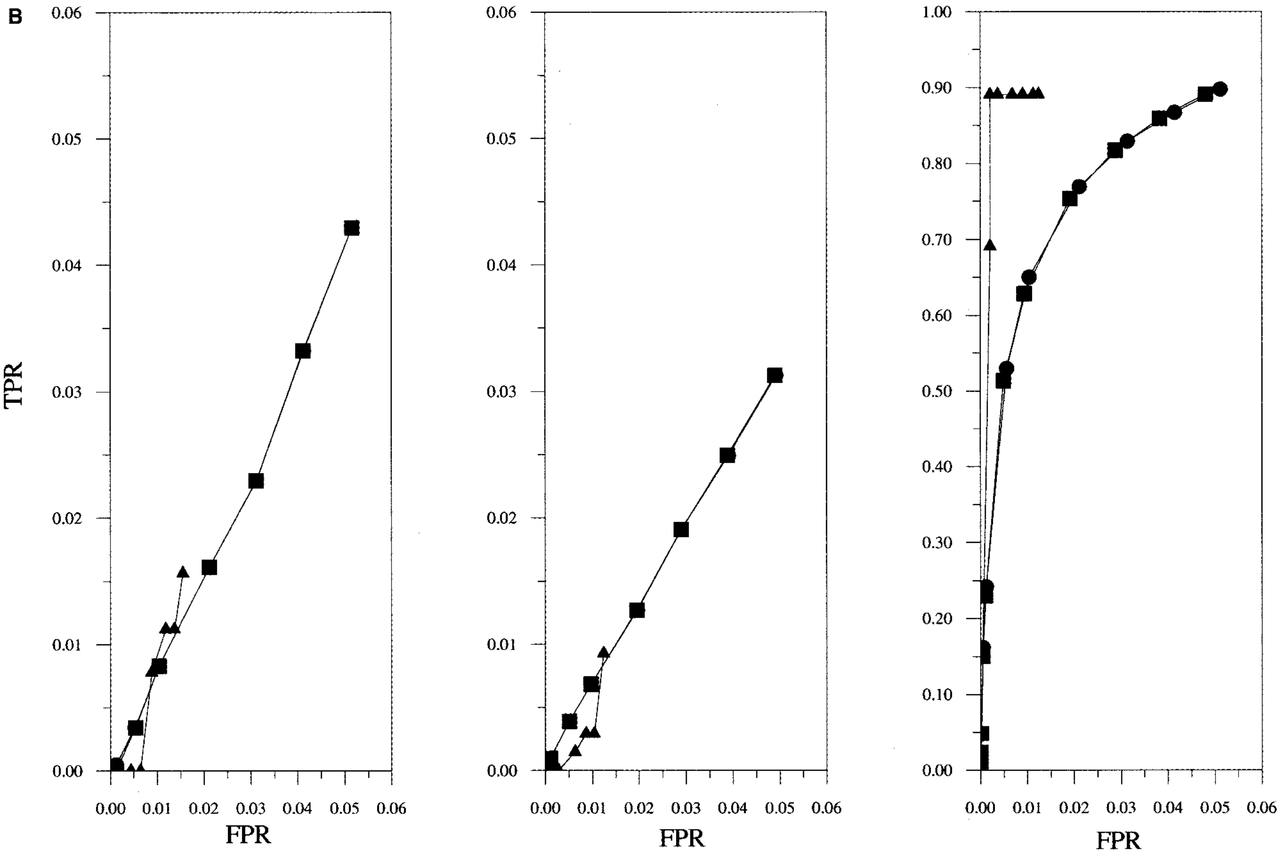

Figure 4.

for example, the difference in means or variance of the observations. In contrast, permutation methods, irrespective of the test statistic, have a more informal null hypothesis: namely, that there is no differential effect of any of the treatments for any of the subjects [Edgington, 1995]. In this case, therefore, the alternative hypothesis is that measurements from at least one subject depend on the treatment. This is just a re-expression of the statement that nonparametric tests are assumption free, but it also raises an important point about the scope of any subsequent interpretation of results.

A frequent criticism of permutation methods is that the results only apply to the dataset under scrutiny and cannot strictly be generalised to the population from which the sample of subjects was drawn. In contrast, using a parametric test, inference about the population can be made based on the sample provided a number of specific assumptions about the execution of the experiment are fulfilled: random sampling, random assignment, normally distributed observations with equal variance. It is generally not too difficult to violate one or more of these assumptions in practice and it might be thought that permutation methods have the advantage in the analysis of such data. In fact, parametric methods are robust to violations of these assumptions and indeed perform slightly better than permutation methods with small samples unless the distribution is highly non-normal [Gonzalez and Manly, 1998; Routledge, 1997]. The results produced by this work corroborate this finding even with data with highly non-normal noise, although differences between inference techniques become less apparent with larger sample sizes. Permutation tests, if properly constructed, will always provide good Type I error control and reliable results, whereas parametric tests do so only under restricted conditions.

Permutation tests are of course not entirely without assumptions themselves. To construct a permutation test, the appropriate exchangeable (independent) units are identified and their labels randomly rearranged to construct the null distribution. For this, we require random assignment of subjects to treatments. If this cannot be assured then a permutation test remains valid if the subjects are drawn randomly from the population and the null hypothesis can be formulated as: distributions from populations are the same. For one observation per subject across the experiment, permuting values unrestrictedly within the levels of the factor assessed main effects. With a repeated measures design, permutations were restricted additionally within subject. In either case, an exact test was possible. For interactions, no exact test is possible with the F statistic as there are no permutations other than the observed ordering that meets the restriction of permuting within levels of the factors [Anderson and ter Braak, 2003; Edgington, 1995]. In these

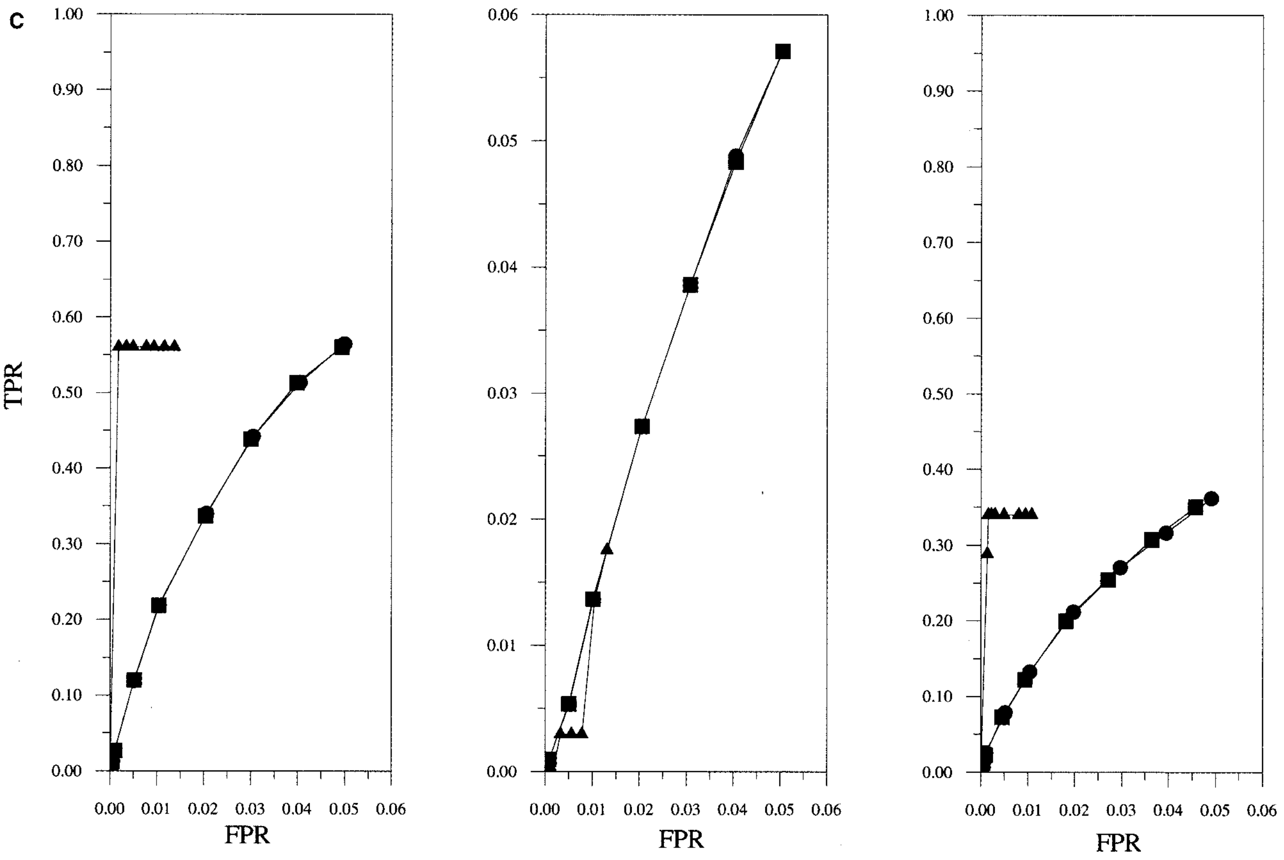

Figure 4.

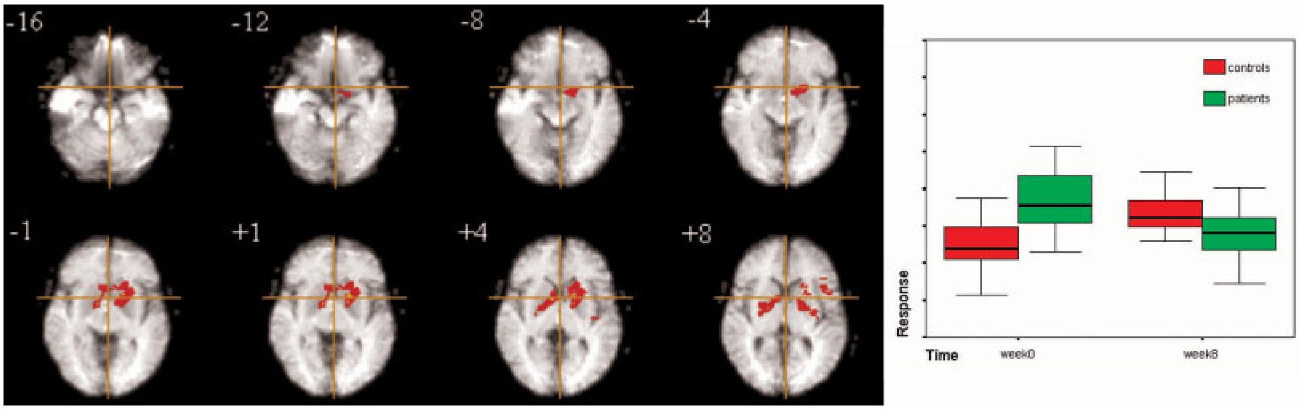

Figure 5.

Antidepressant drug effects on left amygdala and ventral striatum identified by factorial analysis at cluster level of a pharmacologic MRI experiment. Selected slices of data in standard space showing loci of significant Group × Time interaction in an experiment analysed as a 2 × 2 factorial design. Total activation elicited by the sad facial affect paradigm in the significant region was extracted and plotted as the figure inset. The software used a mixed repeated measures (Time) and an independent measures (Group) design with permutation tests of the spatial extent statistic. The probability threshold was set such that there was less than one estimated false positive cluster in each map. There was one significant 3-D cluster.

cases, approximate tests are the alternative. As demonstrated by the simulations presented here, there is both good nominal Type 1 error control (ROC curves on tested factors where there was no simulated effect) and close correspondence at voxel level between permutation and parametric methods.

There is a clear need for software for the comprehensive analysis of factorially designed neuroimaging experiments. Permutation methods build in robustness to skewed or extreme distribution of response data especially as the number of subjects scanned increases. Further, they permit the use of statistics such as the cluster-level metric described here, which impart additional sensitivity to often low-power experiments. Data from any imaging instrument may be used. The variance of these measures is assumed to be homogeneous, although this may not be case from subject-to-subject or even voxel-to-voxel. Future refinements of this method will adapt to the nonhomogeneous case by adopting a weighted least-squares approach to the calculation of *F* (or other statistics).

This software is only designed for two factors with fixed effects. It is, however, entirely possible to extended to additional factors with random effects [Anderson and ter Braak, 2003] and more complex versions of factorial designs such as the Latin square or nested factors, or indeed, a whole range of other experimental designs.

ACKNOWLEDGMENTS

We thank colleagues at the MRI Unit, Maudsley Hospital, London, UK for their help with acquisition of the experimental fMRI data and the journal reviewer for their comments. This work was presented at the Human Brain Mapping Conference, New York, 2003.

REFERENCES

Anderson MJ, ter Braak CJF (2003): Permutation tests for multifactorial analysis of variance J Stat Comput Simul 73:85–113.

Arndt S, Cizadlo T, Asndreasen NC, Heckel D, Gold S, O'Leary DS (1996): Tests for comparing images based on randomisation and permutation methods. J Cereb Blood Flow Metab 16:1271–1279.

Ashburner J, Friston KJ (2000): Voxel-based morphometry—the methods. Neuroimage 11:805–821.

Brammer MJ, Bullmore ET, Simmons A, Williams SC, Grasby PM, Howard RJ, Woodruff PW, Rabe-Hesketh S (1997): Generic brain activation mapping in functional magnetic resonance imaging: a non-parametric approach. Magn Reson Imaging 15:763–770.

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

Bullmore ET, Suckling J, Overmeyer S, Rabe-Hesketh S, Taylor E, Brammer MJ (1999a): Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. IEEE Trans Med Imaging 18:32–42.

Bullmore ET, Brammer MJ, Rabe-Hesketh S, Curtis VA, Morris RG, Williams SC, Sharma T, McGuire PK (1999b): Methods for diagnosis and treatment of stimulus correlated motion in generic brain activation studies using fMRI. Hum Brain Mapp 7:38–48.

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

Bullmore E, Suckling J, Brammer MJ (2001b): In praise of tedious permutation. In: Moore M, editor. Spatial statistics: methodological aspects and some applications. Lecture notes in statistics, Vol 159. New York: Springer. p 183–200.

Bullmore E, Suckling J, Zelaya F, Long C, Honey G, Reed L, Routledge C, Ng V, Fletcher P, Brown J, Williams SC (2003): Practice and difficulty evoke anatomically and pharmacologically dissociable brain activation dynamics. Cereb Cortex 13:144–154.

Cao J (1999): The size of the connected components of excursion sets of χ^2, *t*, and *f* fields. Adv Appl Probab 31:579–595.

Cao J, Worsley KJ (2001): Applications of random fields in human brain mapping. In: Moore M, editor. Spatial statistics: methodological aspects and some applications. Lecture notes in statistics, Vol 159. New York: Springer. p 169–182.

Coolican H (1999): Research methods and statistics in psychology (3rd ed). London: Hodder and Stoughton Educational.

Edgington ES (1995): Randomization tests (3rd ed). New York: Marcel Dekker.

Fisher RA (1935): The design of experiments. Edinburgh: Oliver and Boyd.

Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC (1995): Improved assessment of significant activation in functional magnetic resonance imaging: use of a cluster size threshold. Magn Reson Med 33:636–647.

Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, Evans AC (1994): Assessing the significance of focal activations using their spatial extent. Hum Brain Mapp 1:214–220.

Gonzalez L, Manly BF (1998): Analysis of variance by randomization with small data sets. Environmetrics 9:53–65.

Good PI (2000): Permutation test: a practical guide to resampling methods for testing hypotheses (2nd ed). Berlin: Springer-Verlag.

Hayasaka S, Nichols TE (2003): Validating cluster size inference: random field and permutation methods. Neuroimage 20:2343–2356.

Holmes AP, Blair R, Watson J, Ford I (1996): Nonparametric analysis of statistical images from functional mapping experiments. J Cereb Blood Flow Metab 16:7–22.

Honey GD, Bullmore ET, Soni W, Varatheesan M, Williams SC, Sharma T (1999): Differences in frontal cortical activation by a working memory task after substitution of risperidone for typical antipsychotic drugs in patients with schizophrenia. Proc Natl Acad Sci USA 96:13432–13437.

Honey GD, Suckling J, Zelaya F, Long C, Routledge C, Jackson S, Ng V, Fletcher PC, Willimas SC, Brown J, Bullmore ET (2003): Dopaminergic drug effects on physiological connectivity in a human cortico-striato-thalamic system. Brain 126:1801–1813.

Lindley DV, Novick MR (1981): The role of exchangeability in inference. Ann Stat 9:45–58.

Locascio JL, Jennings PJ, Moore CI, 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.

Mendoza JL, Toothaker LE, Crain BR (1976): Necessary and sufficient conditions for F ratios in the L × J × K factorial design with two repeated factors. J Am Stat Assoc 71:992–993.

Mitterschiffthaler MT, Fu CH, Kim J, Williams SC, Walsh ND, Cleare AJ, Brammer MJ, Pich EM, Andrew CM, Suckling J,Bullmore ET (2003): Neural substrates of happy and sad facial emotion processing in depression. Neuroimage 19:122.

Nichols TE, Holmes AP (2002): Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15:1–25.

Pesarin F (2001): Multivariate permutation tests: with applications in biostatistics. Chichester, UK: John Wiley and Sons.

Poline JB, Mazoyer BM (1993): Analysis of individual tomography activation maps by clusters. J Cereb Blood Flow Metab 13:425–437.

Poline JB, Worsley KJ, Evans AC, Friston KJ (1997): Combining spatial extent and peak intensity to test for activations in functional imaging. Neuroimage 5:83–96.

Routledge RD (1997): *P*-values from permutation and *F*-tests. Comput Stat Data Anal 24:376–386.

Sheline YI, Barch DM, Donnelly JM, Ollinger JM, Snyder AZ, Mintun MA (2001): Increased amygdala response to masked emotional faces in depressed subjects resolves with antidepressant treatment: an fMRI study. Biol Psychiatry 50:651–658.

Still AW, White AP (1981): The approximate randomization test as an alternative to the F test in analysis of variance. Br J Math Stat Psychol 34:243–252.

Talairach J, Tournoux P (1988): A coplanar stereotactic atlas of the human brain. Stuttgart, Germany: Thieme.

Thompson PM, Cannon TD, Narr KL, van Erp T, Poutanen VP, Huttunen M, Lönnquist J, Standertskjöld-Nordenstam CG, Kaprio J, Khaledy M, Dail R, Zoumalan CI, Toga AW (2001): Genetic influences on brain structure. Nat Neurosci 4:1253–1258.

Welch WJ (1990): Construction of permutation tests. J Am Stat Assoc 85:693–698.