Global, Voxel, and Cluster Tests, by Theory and Permutation, for a Difference Between Two Groups of Structural MR Images of the Brain

Edward T. Bullmore,* John Suckling, Stephan Overmeyer,
Sophia Rabe-Hesketh, Eric Taylor, and Michael J. Brammer

***Abstract*— We describe almost entirely automated procedures for estimation of global, voxel, and cluster-level statistics to test the null hypothesis of zero neuroanatomical difference between two groups of structural magnetic resonance imaging (MRI) data. Theoretical distributions under the null hypothesis are available for 1) global tissue class volumes; 2) standardized linear model [analysis of variance (ANOVA and ANCOVA)] coefficients estimated at each voxel; and 3) an area of spatially connected clusters generated by applying an arbitrary threshold to a two-dimensional (2-D) map of normal statistics at voxel level. We describe novel methods for economically ascertaining probability distributions under the null hypothesis, with fewer assumptions, by permutation of the observed data. Nominal Type I error control by permutation testing is generally excellent; whereas theoretical distributions may be over conservative. Permutation has the additional advantage that it can be used to test any statistic of interest, such as the sum of suprathreshold voxel statistics in a cluster (or cluster mass), regardless of its theoretical tractability under the null hypothesis. These issues are illustrated by application to MRI data acquired from 18 adolescents with hyperkinetic disorder and 16 control subjects matched for age and gender.**

***Index Terms*— Brain, imaging/mapping, probability distributions, statistics.**

I. Introduction

HUMAN brain research using magnetic resonance imaging (MRI) is often motivated by an interest in one form or another of the alternative hypothesis that there is an anatomical difference between two groups of subjects. The distribution of the observed difference under this alternative hypothesis could be explicitly modeled (see [1] for an example of this approach to activation mapping in functional MRI) but it is more usual to resort to consideration of the null hypothesis that there is zero difference between groups. Two related questions then have to be addressed.

- What measure(s) of difference between two groups are likely to be most informative about departure from the null hypothesis?
- How can we ascertain the distribution of a potential test statistic under the null hypothesis?

In relation to the first of these questions, we here consider three types (or levels) of test statistic that can be estimated from two groups of structural MRI data. The first (global) type is simply a measure of the difference between groups in whole brain volume of any one of the three main tissue classes (grey matter, white matter, or cerebrospinal fluid). The second (voxel) type is a measure of the difference between groups in volume of a given tissue class at the level of a single voxel. The third (cluster) type is a measure of the difference between groups obtained by arbitrarily thresholding maps of a voxel test statistic and considering the properties of the spatial clusters of suprathreshold voxels that result. We will deal with two cluster statistics in particular: 1) the two-dimensional (2-D) area of a suprathreshold cluster and 2) the sum of suprathreshold voxel statistics, or mass, of a 2-D cluster. (Two other possible types of test statistic would be region of interest (ROI) volume measurement or a systems-level measure of multivariate difference between two groups [2]; however, these approaches are not discussed further here.)

Global statistics seem unlikely to be especially sensitive unless the anatomical difference between groups is diffuse. Voxel statistics are potentially more sensitive to focal differences between groups, but they will entail a much larger number of tests. In order to avoid an unacceptable number of false positive or Type I errors in assessment of a group difference at voxel level, it is therefore customary to adopt a very stringent criterion for statistical significance [3]. Typically, voxel statistics must have a probability under the null hypothesis, or P value, in the order of 10^{-3} or less to be regarded as significant. This inevitably increases the risk of false negative or Type II error compared, say, to a test at the conventional probability threshold of $P \leq 0.05$. Voxel statistics also have the disadvantage of neglecting the spatially coordinated nature of imaging data. Thus, each voxel is tested independently of its neighbors, despite prior knowledge that important group differences in regional anatomy may well be expected to extend over several spatially contiguous voxels.

Manuscript received May 26, 1998; revised December 9, 1998. This work was supported in part by the Wellcome Trust and by the European Union Programme for the Training and Mobility of Researchers. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was L. Clark. *Asterisk indicates corresponding author*.

*E. T. Bullmore, J. Suckling, S. Rabe-Hesketh, and M. J. Brammer are with the Department of Biostatistics and Computing, Institute of Psychiatry, King's College, University of London, De Crespigny Park, London SE5 8AF, U.K. (e-mail: e.bullmore@iop.kcl.ac.uk).

S. Overmeyer was with the Department of Child Psychiatry, Maudsley Hospital, Denmark Hill, London SE5 8AZ, U.K. He is now with the Psychiatrische Universitatsklinik, Abteilung Kinder und Jugendpsychiatrie, D-79104, Freiburg, Germany.

E. Taylor is with the Department of Child Psychiatry, Maudsley Hospital, Denmark Hill, London SE5 8AZ, U.K.

Publisher Item Identifier S 0278-0062(99)02023-6.

As Poline and Mazoyer first demonstrated in the context of functional brain-image analysis [4], these problems associated with voxel tests can be largely circumvented by testing at cluster level. The number of clusters to be tested is generally far fewer than the number of voxels, allowing more relaxed probability thresholds and reduced Type II error rates for a given number of false positive tests. Clusters can generally also be regarded as independent events under the null hypothesis, unlike voxels, which will be correlated under the null hypothesis at least to the extent that the image is smoothed by the point spread function of the scanning instrument. If strong Type I error control is required, therefore, a Bonferroni correction of the critical P value for a test at each cluster is both appropriate and straightforward to implement [4]. Finally, cluster area is generally a more sensitive measure of regional cerebral blood flow changes in simulated positron emission tomography (PET) data than a global test [4], and often more sensitive than a voxel test [5].

The only major problem associated with use of cluster size as a test statistic seems to relate to the (second) question of how best to ascertain its distribution under the null hypothesis. There are well-established theoretical approximations for the null distributions of many global and voxel statistics, but the issue is more controversial for cluster size. Several groups have adopted Monte Carlo procedures to sample the distribution of cluster size in images simulated under the null hypothesis [4], [6], [7]. Simulation necessitates making some assumptions, usually about the distribution of the voxel statistic and the spatial autocorrelation or smoothness of the voxel statistic image, under the null hypothesis. The null distribution of cluster area is conditional on both the smoothness W of the statistic image and the size of threshold u applied at voxel level. Therefore, cluster size distributions reported on the basis of Monte Carlo simulations are only appropriate for testing images which happen to have W and u identical to the arbitrary smoothness and voxel threshold of the simulated data [8], [9].

Friston *et al.* [5] derived an exponential form for the null distribution of cluster size which can be more generally used to estimate the probability of a cluster as a function of the smoothness, voxel threshold, and dimensionality D of any statistic image (see below for greater detail). However, as these and other authors have commented [10], [11], this distribution is derived from results in Gaussian field theory that are only exact in the limit of very high thresholds, i.e., $u \approx 6$ (assuming the voxel statistic has a standard normal distribution). This is unfortunate because a test of cluster size is likely to be most useful in detecting relatively large sets of spatially connected suprathreshold voxels that can only arise if the threshold applied at voxel level is fairly low, i.e., $u \approx 2$ [4].

One of the objectives of the present study was to develop methods for ascertaining the null distributions of global, voxel, and cluster statistics by permutation procedures and to crossvalidate these permutation tests by comparison to the corresponding tests derived from normal theory. The basic principles of permutation testing are simple and well established [12]–[16]. Arguably, the most salient advantage of permutation testing in general is that it is applicable to any test statistic of interest, not just the subset of potentially interesting statistics with theoretically tractable distributions under the null hypothesis. Furthermore, since the permutation distribution is ascertained directly from the observed data, the critical values for testing will be appropriate for arbitrary properties of the observed data, such as image smoothness, which might prohibit appropriate use of critical values derived from prior Monte Carlo simulations. The advantages of permutation testing for functional brain-image analysis were first proposed by Blair *et al.* [17] and have since been widely recognized [2], [18]–[21]. However, to the best of our knowledge, this study represents the first comparative appraisal of permutation tests in structural brain-image analysis. It is also the first study to validate permutation tests at cluster level in any imaging modality.

II. Methods and Materials

A. Subjects

The patient group comprised 18 children (15 boys, 3 girls) with a mean age of 10.4 years and mean I.Q. of 99 (SD 14.9) who were attending the Child Psychiatry Department of the Maudsley Hospital, London, for treatment of hyperkinetic disorder (HD) diagnosed according to standard operationalized criteria [22]. HD is a child psychiatric syndrome characterized by inattentiveness, distractibility, and hyperactivity and is associated with delayed development of language and motor skills. Previous structural-imaging studies of HD, or the related syndrome of attention-deficit hyperactivity disorder, have identified grey matter volume deficits in frontal lobes and basal ganglia. For a review see [23].

The control group comprised 16 psychiatrically normal children (15 boys, 1 girl) with a mean age of 10.3 years who were the siblings of children receiving outpatient treatment at the same hospital. There was no significant difference between the groups in age, head circumference, height, or weight. Nine children in the patient group and one child in the control group were left handed.

Informed consent to participate in the study was given by the parents of each subject. The study was approved by the Ethics Committee of the Bethlem Royal and Maudsley NHS Trust.

B. Image Acquisition

Dual-echo fast-spin echo (FSE) MRI data were acquired at 1.5 T in the sagittal plane parallel to the interhemispheric fissure using a GE Signa system at the Maudsley Hospital: repetition time $(TR) = 4$ s, time to first echo $(TE_1) = 20$ ms, time to second echo $(TE_2) = 100$ ms, field of view $= 22$ cm, image matrix $= 256 \times 192$, in-plane resolution $= 0.86$ mm, number of interleaved slices $= 50$, slice thickness $= 3$ mm, number of signal averages $= 1$. Head movement was limited by foam padding within the head coil and a restraining band across the forehead.

This acquisition protocol represents each of 50 sagittally orientated brain volumes, or slices, by a pair of images: a proton density PD-weighted image acquired at TE_1 and a $T2$-weighted image acquired at TE_2. Since the time interval

between the two echos is only 80 ms, we neglect the possibility of misregistration of the image pair due to subject motion and assume that the two images differently represent an anatomically identical volume of the brain.

C. Segmentation of Extracerebral Voxels

The first step in image processing is segmentation of voxels representing extracerebral tissue, such as bone and skin. This is done by applying a computational algorithm, previously described and validated in detail [24], to the PD-weighted dataset for each subject. Briefly, the algorithm uses a linear scale-space set of features obtained from derivatives of the Gaussian kernel. Application of the second-order derivative (the Laplacian) at three different scales results in measures of the local grey-level curvature of the signal intensities. The first-order derivative feature represents the image edges and its inverse is applied as a binary mask to a combined image of second-order derivative features such that the extracortical cerebrospinal fluid is effectively circumscribed. A grey-level histogram analysis of the resulting image sets a lower threshold for the data volume and a binary mask is obtained in each slice of the image. After removal of small islands of data to improve computational speed, the largest three-dimensional (3-D)-connected object in the data volume is found and is assumed to be the brain. This binary image of the brain is then used to select voxels of the FSE dataset representing neural tissue (including cerebrospinal fluid) and to set to zero all voxels representing extracerebral tissue. The process is entirely automated.

D. Probabilistic Morphometry

The next step is estimation of the volumes of grey matter (G), white matter (W), and cerebrospinal fluid (C) represented at each voxel, and over all intracerebral voxels, in each individual FSE dataset.

Let V denote the number of intracerebral voxels in a single dual-echo image. At the ith voxel $i = 1, 2, 3, \ldots, V$ we have a pair of physical measurements denoted PD_i and $T2_i$ which we can refer to collectively as the observed feature vector $\mathbf{x}_i$. On the basis of these physical data we can derive the probabilities that the ith voxel would be correctly classified as representative of each of the three mutually exclusive tissue classes of interest as follows [25], [26]:

$$P(G \mid \mathbf{x}_i) = \frac{\exp(\beta_0 + \beta_1 PD_i + \beta_2 T2_i)}{1 + \exp(\beta_0 + \beta_1 PD_i + \beta_2 T2_i) + \exp(\beta_3 + \beta_4 PD_i + \beta_5 T2_i)}$$
$$P(W \mid \mathbf{x}_i) = \frac{\exp(\beta_3 + \beta_4 PD_i + \beta_5 T2_i)}{1 + \exp(\beta_0 + \beta_1 PD_i + \beta_2 T2_i) + \exp(\beta_3 + \beta_4 PD_i + \beta_5 T2_i)}$$
$$P(C \mid \mathbf{x}_i) = \frac{1}{1 + \exp(\beta_0 + \beta_1 PD_i + \beta_2 T2_i) + \exp(\beta_3 + \beta_4 PD_i + \beta_5 T2_i)} \quad (1)$$

The parameters of this polychotomous logistic discriminant function $\{\beta_j\}$, $j = 0, 1, 2, \ldots, 5$ are estimated by maximum likelihood from a relatively small subsample or training set of voxels (<1% of all intracerebral voxels) expertly selected from each image as representatives of each tissue class. The probabilities of class membership given the feature vector observed at the ith voxel $\{P(G \mid \mathbf{x}_i), P(W \mid \mathbf{x}_i), P(C \mid \mathbf{x}_i)\}$ are then simply computed by substituting the ith PD- and T2-weighted signal intensities into the trained or parameterized discriminant function. Repeating this substitution over all V voxels in the image generates a set of three maps, representing the probability of each voxel in the image belonging to each of the three tissue classes. Based on previous results, we equate these probabilities to the proportional volumes of each tissue class in the often heterogeneous volume of tissue represented by each voxel [25]. For example, if $P(G \mid \mathbf{x}_i) = 0.8$, we may say that 80% of the brain tissue represented by the ith voxel is grey matter. The absolute volume of grey matter represented is simply $P(G \mid \mathbf{x}_i) \times xyz$ where $x, y,$ and z denote the three dimensions of voxel size in millimeters. Absolute volume of grey matter estimated over the whole image is

$$Gv = \sum_{i=1}^{V} P(G \mid \mathbf{x}_i) \times xyz. \quad (2)$$

Substituting voxel probabilities $P(W \mid \mathbf{x}_i)$ or $P(C \mid \mathbf{x}_i)$ in (2) estimates total image volumes for white matter and cerebrospinal fluid, respectively.

E. Registration and Smoothing in Standard Space

The final step in image processing, prior to estimation and testing of group differences at voxel and cluster levels, is registration of the three probability maps (one for each tissue class) obtained from each individual image in the standard space of Talairach and Tournoux [27].

To do this, a template image was first constructed by proportional rescaling of a subset of five PD-weighted images from the control group. Using AFNI software [28] anatomical landmarks were identified, including the anterior and posterior commissures and lateral, superior, and inferior convexities of the cerebral surface. The distances between landmarks were then linearly rescaled to approximate each individual image to the size and shape of the reference brain depicted in a standard stereotactic atlas [27]. The five transformed images were then averaged to produce a single template image in standard space.

The affine transformation which minimizes the sum of absolute grey level differences between each (of $N = 34$) individual PD-weighted images and the template image was then identified by the Fletcher–Davidon–Powell algorithm [21], [29] and this individually estimated transformation matrix was applied in turn to each of that subject's three probability maps to register them in standard space.

To accommodate individual variability in anatomy and error in spatial normalization, all probability maps were then smoothed by convolution in the Fourier domain with a 2-D Gaussian filter with variance $= 5$ mm or full width at half maximum (FWHM) $= 5.67$ voxels.

We note that this procedure does not represent a unique solution to the problem of coregistering several MRI datasets in a standard space. It would be equally possible to adopt another

stereotactic system, to effect registration of each individual dataset in standard space by nonlinear transformation and/or to apply a different filtering regime to the data. But however the data are preprocessed, it is probably advisable to consider several alternative methods of testing the null hypothesis of zero between-group difference, as described in detail below.

F. Estimation of Voxel Statistics

Following registration of the three probability maps generated from each individual MRI dataset in the same space, we can estimate the difference between groups in proportional volume of a given tissue class by fitting a linear model at each voxel.

In matrix notation, linear models are generally of the form

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \tag{3}$$

where $\mathbf{y}$ is an $N \times 1$ vector of response or dependent variables, $\mathbf{X}$ is a $N \times p$ (design) matrix of constants, β is a $p \times 1$ vector of parameters, and $\mathbf{e}$ is a $N \times 1$ vector whose elements are independent identically and normally distributed $e_j \sim N(0, \sigma^2)$ for $i = 1, 2, 3, \ldots, N$. The standard errors of the parameters are the diagonal elements of the matrix $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, where ^{T} denotes transposition and $^{-1}$ denotes inverse of a matrix [30].

The simplest such model we will consider is a one-way analysis of variance (ANOVA) which can be written for a single observation of grey matter volume as follows:

$$P(G \mid \mathbf{x}_i)_{k,j} = \mu_i + a_{i,k} + e_{i,k,j} \tag{4}$$

where $P(G \mid \mathbf{x}_i)_{k,j}$ is the proportional volume of grey matter estimated at the ith voxel for the jth individual in the kth group with $j = 1, 2, 3, \ldots, N_1$ or N_2, $k = 1, 2$ and $i = 1, 2, 3, \ldots, V$, μ_i is overall mean at the ith voxel, $\mu_i + a_{i,k}$ is the mean of the kth group at the ith voxel, and $e_{i,k,j} \sim N(0, \sigma^2)$ denotes a residual quantity unique to the jth member of the kth group at the ith voxel. (Note that V now denotes the number of intracerebral voxels in standard space for which we have N proportional volume estimates, not the number of intracerebral voxels in an individual image as previously.)

To estimate the parameters $\{\mu, a\}$ we construct a $(N \times 2)$ design matrix $\mathbf{X}$, the first column of which is a vector of ones and the second column of which is a dummy variable coding membership of the control group by 1 and membership of the patient group by -1. Least squares estimates of the parameters are then given by $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

It can be shown [31] that the test statistic

$$A = \frac{\hat{a}}{SE(\hat{a})} \tag{5}$$

where $SE(\hat{a})$ is the standard error of $\hat{a}$ is equivalent to the familiar T statistic

$$T = \frac{\bar{\mu}_1 - \bar{\mu}_2}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}};$$

$$s^2 = \frac{1}{N_1 + N_2 - 2}\left\{\sum_{j=1}^{N_1}(y_{1,j} - \bar{\mu}_1)^2 + \sum_{j=1}^{N_2}(y_{2,j} - \bar{\mu}_2)^2\right\}. \tag{6}$$

The advantage of working with A rather than T is that it is straightforward to adjust A for the possibly confounding effects of other independent variables by adding the appropriate columns to the design matrix. For example, in this study, we wished to estimate the additive effect of group membership after controlling for age, gender, handedness, and absolute volume of a given tissue class over the whole image. Again taking grey matter to be the dependent variable, we can write this analysis of covariance (ANCOVA) model

$$P(G \mid \mathbf{x}_i)_{k,j} = \mu_i + a_{i,k} + b\,\mathrm{Age}_{k,j} + c\,\mathrm{Hand}_{k,j} + d\,\mathrm{Sex}_{k,j} + f\,Gv_{k,j} + e_{i,k,j}, \tag{7}$$

where $\mathrm{Age}_{k,j}$, $\mathrm{Hand}_{k,j}$, $\mathrm{Sex}_{k,j}$, and $Gv_{k,j}$ denote the age, handedness, gender, and image volume of grey matter in the jth individual from the kth group. Both ANOVA and ANCOVA models can be generalized to take proportional volumes of white matter or CSF as dependent variables [after substitution of the appropriate vector of image volumes in the design matrix for (7)]. Fitting either of these models at each intracerebral voxel in standard space yields a set of three effect maps of the test statistic A, one for each tissue class.

G. Estimation of Cluster Statistics

Applying an arbitrary threshold u to any one of these effect maps will yield a number M of voxel clusters, each comprising one or more spatially contiguous (eight-connected) voxels with values for $A \geq u$. We can apply the threshold such that if $A_i \geq u$ the value of the ith voxel in the thresholded map is one, otherwise the value of the ith voxel is zero. We can then measure the 2-D area (in voxels) of the mth such binary cluster and denote this ν_m with $m = 1, 2, 3, \ldots, M$.

Alternatively, we can apply the threshold such that if $A_i \geq u$ the value of the ith voxel in the thresholded map is $A_i - u = h_i$, otherwise the value of the ith voxel is zero. We can then measure the 2-D mass of the mth cluster and denote this τ_m

$$\tau_m = \sum_{i=1}^{\nu_m} h_i. \tag{8}$$

H. Computational Issues

All computations were performed on a Sun Ultra 1 workstation: 170 MHz with 128 MB of memory. Code was written in the `C` language, with the exception of the code for logistic discriminant analysis which was written in `S-PLUS`. Custom graphical user interfaces were written in the `X Window` system.

Typical processing times per subject were as follows: segmentation of extracerebral voxels, 30 minutes; selection of training data, discriminant function parameter estimation, and tissue classification, 30 minutes; registration in standard space and smoothing, 40 minutes. Estimation and testing of voxel and cluster statistics, using ten permutations at each voxel to sample the null distributions, required approximately five hours of processing time. If computational time costs were a major issue, it might be possible to reduce the number of permutations applied at each voxel. However, smaller permutation distributions will generally yield less stable critical values

TABLE I
GLOBAL BRAIN VOLUMES (IN MILLILITRES) OF GREY MATTER (G), WHITE MATTER (W), AND CEREBROSPINAL FLUID (CSF) AND TEST STATISTICS FOR A DIFFERENCE BETWEEN GROUPS AND PROBABILITIES OF THE OBSERVED DIFFERENCE ASCERTAINED BY THEORY AND PERMUTATION

	Controls Mean (SD)	Cases Mean (SD)	T	Theory P	Permutation P
G	870.49 (109.11)	873.92 (122.54)	-0.087	0.931	0.915
W	471.61 (73.98)	432.33 (94.55)	1.363	0.182	0.186
CSF	90.04 (36.16)	78.7 (24.04)	1.098	0.280	0.296

for conservative tests. If one wishes to apply a probability threshold of $P \leq 0.05$ then the size of the permutation distribution should be at least 5000 and if one wishes to set $P \leq 0.001$ then the size of the permutation distribution should be at least 10 000 [15]. These are approximate rules of thumb and it would be highly advisable to check the stability of critical values obtained from disproportionately small permutation distributions.

III. RESULTS

A. Global Tests

The estimates of absolute image volume of grey matter, white matter, and cerebrospinal fluid obtained by (2) for each group are summarized in Table I. We used T (6) as our global test statistic and assessed the probability of the observed value of T under the null hypothesis by theory and by permutation.

The theoretical test is based on the familiar assumption that, under the null hypothesis, T has a t distribution on $N_1 + N_2 - 2 = 32$ degrees of freedom (df), i.e., $T \sim t_{32}$. The two-tailed P value for the observed T statistic is then $2P\{t_{32} > |T|\}$ where $|T|$ is the absolute value of the observed test statistic [31].

The corresponding permutation test was based on the repeated and random reassignment of the individual estimates of global tissue class volume to two groups of sizes N_1 and N_2. T was estimated by (6) after each random reassignment. This process was repeated 999 times, resulting in 999 estimates of T under the null hypothesis. Taken together with the single estimate of T observed in the original data, these were ordered in size to sample the permutation distribution of T. The two-tailed P value was then simply the number of entries in the permutation distribution with absolute value greater than observed $|T|$, divided by the total number of permutations plus one $=$ 1,000.

As shown in Table I, the P values obtained by these two methods for the observed difference between groups are very similar for all three global statistics. By neither test would any of the observed differences be conventionally considered significant. This suggests that the permutation and theoretical distributions are reasonably similar, which is also supported by direct comparison of the distributions in Fig. 1.

B. Voxel Tests

The test statistics at voxel level were the standardized regression coefficients A estimated by fitting either of the two linear models (4) or (7). A theoretical test of these statistics can be conducted by assuming that A has a t distribution on $N_1 + N_2 - p$ df. The corresponding permutation test is conducted along basically similar lines to the test described above. The data are repeatedly and randomly reassigned to two groups, and the test statistic is estimated after each permutation to sample its distribution under the null hypothesis. However, since the number of voxels to be tested was large, i.e., $V = 342\,733$, we preferred to permute the data only ten times at each voxel, then pool the resulting estimates of A over all voxels in the search volume to form a permutation distribution comprising 3 427 330 estimates of A under the null hypothesis. A P value was assigned to each voxel by referring its observed value of A to this pooled permutation distribution.

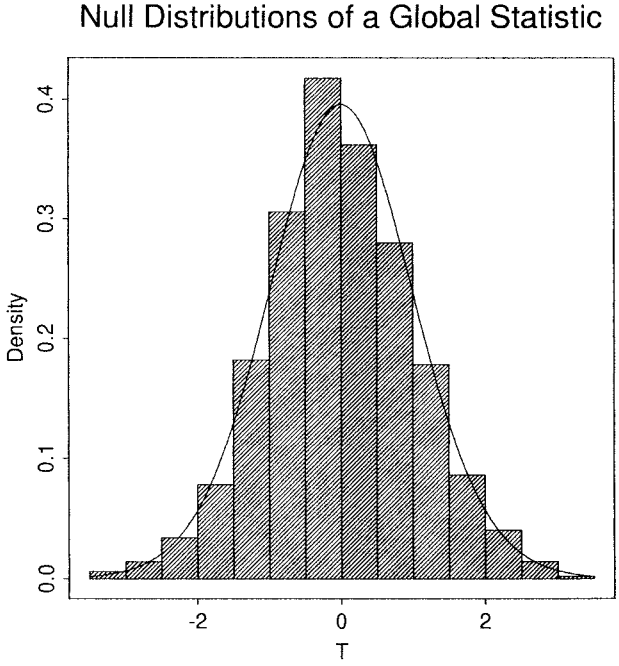
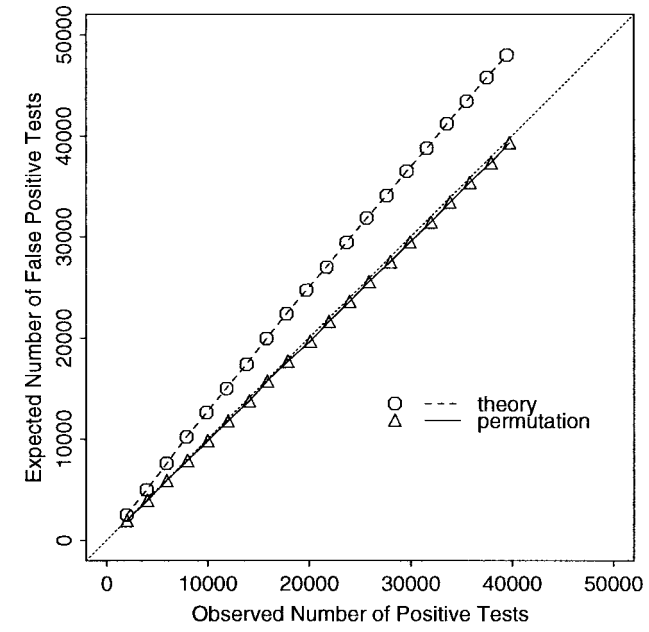
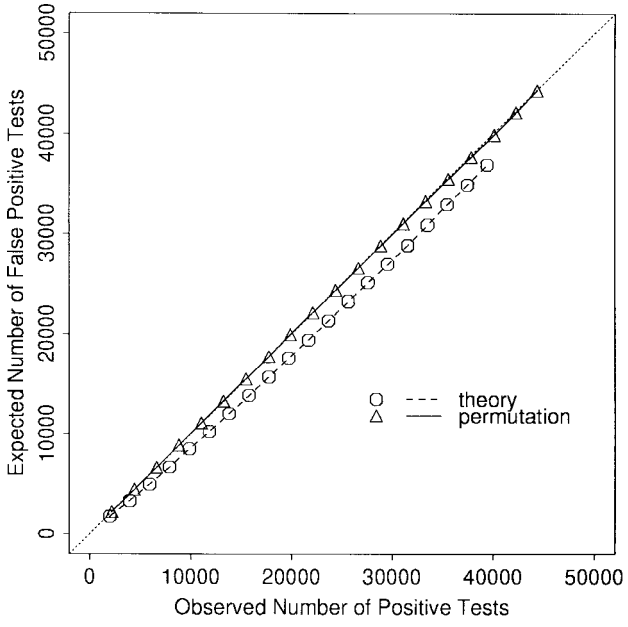


Fig. 1. Theoretical and permutation distributions for the difference between groups in global grey matter volume under the null hypothesis. The histogram shows the null distribution sampled by 999 random permutations of the observed data; the solid line shows the t distribution on 32 df.

In order to calibrate Type I error control by this procedure, group differences between two randomly decided subsets of the control group data were tested. Each control subject was randomly assigned to one of two groups, each of size $N_1/2 = 8$, and the difference between these groups was estimated by linear modeling. Both theoretical and permutation tests were then used to assess the probability of each observed voxel statistic under the null hypothesis, and the observed effect maps were thresholded over a range of sizes of two-tailed test $0.005 < P \leq 0.1$. Since no true difference is expected to exist between these two subsets of the control group, all significant voxels identified by a given size of test should be false positive tests. In other words, the observed number of positive tests should equal the predicted number of false positive tests $= PV$. As shown in Fig. 2, the number of positive tests observed by permutation testing closely corresponded



(a)



(b)

Fig. 2. Type I error-calibration curves for voxel-level tests by theory and permutation. The expected number of false positive voxels over a range of sizes of test $0.005 < P \leq 0.1$ can be compared directly to the observed number of significant voxels following tests by theory (dashed line and circles) and permutation (solid line and triangles). The two groups tested were randomly decided subsets of the control subjects, therefore the number of observed and expected tests should be equal (dotted line). (a) An ANOVA model was fitted at each of 342 733 voxels to estimate group difference in grey matter volume (4). (b) An ANCOVA model was fitted at each voxel (7).

to the expected number of false positive tests, regardless of the linear model used to estimate the voxel statistics. The number of positive tests observed by the theoretical test of the ANOVA coefficient was generally less than the expected number, implying that this test is somewhat over conservative. Correspondence between observed and expected positive tests was closer for the theoretical test of the ANCOVA coefficient.

The results of testing by permutation for a difference at voxel level between the two original groups are shown in Fig. 3.

C. Cluster Area

Effect maps estimated by fitting both ANOVA and ANCOVA models were thresholded such that if $|A| < 2$ the value of that voxel was set to zero, else if $|A| > 2$ the value of that voxel was set to $h = A - 2$ if $A > 0$ and set to $h = A + 2$ if $A < 0$. This size of threshold is approximately equivalent to applying a threshold of size $u = 1.85$ to a standard normal voxel statistic. The area and mass of each cluster of nonzero voxels sharing the same sign of h were estimated as described above.

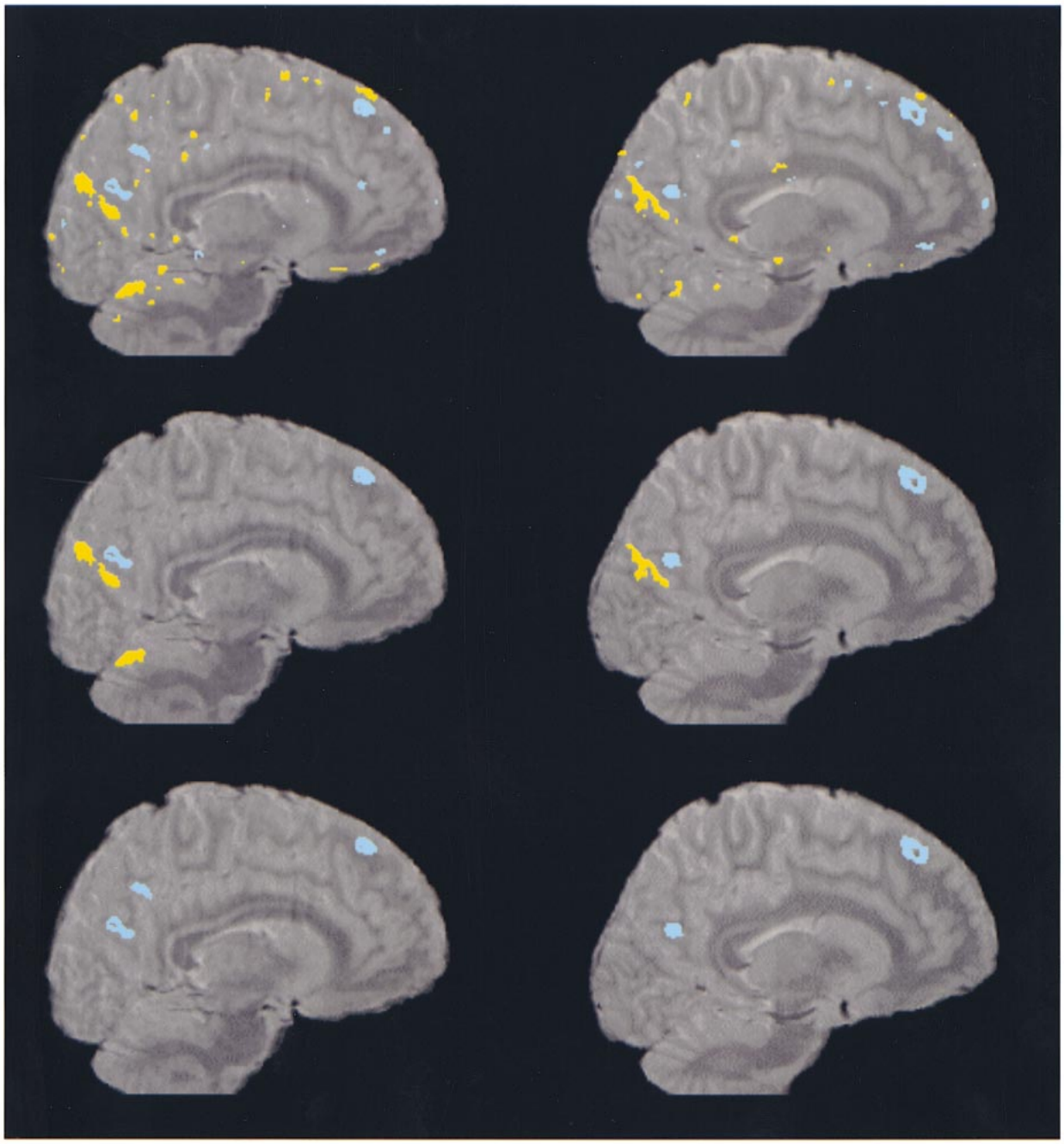
The probability of a given cluster area under the null hypothesis was estimated using an exponential distribution derived from Gaussian field theory [5]. For cluster size measured in two spatial dimensions, i.e., cluster area ν, the probability that ν has some particular value x is

$$P(\nu = x) = \psi e^{-\psi x} \tag{9}$$

where ψ denotes the expected number of suprathreshold clusters $E(M) = V(2\pi)^{-3/2} W^{-2} u e^{-u^2/2}$ divided by the expected number of suprathreshold voxels $E(V+) = V\Phi(-u)$. It can be seen that the theoretical probability of a given cluster area is a strong function of the smoothness of the voxel statistic image W, which can be estimated in 2-D by $\mathrm{FWHM}/\sqrt{4\log 2}$, and the threshold u, which has a P value under the standard normal distribution denoted by $\Phi(-u)$. This distribution is plotted in Fig. 4 with $\mathrm{FWHM} = 5.67$ voxels and $u = 1.85$.

The corresponding permutation distribution of ν was ascertained as follows. We randomly reassigned the N observations at each voxel to two groups of sizes N_1 and N_2, estimated A by ANOVA or ANCOVA at each voxel, applied a threshold to the resulting effect maps, and then measured the area of each of the M clusters in the thresholded maps after each permutation. This process was repeated ten times and the resulting estimates of cluster area under the null hypothesis were pooled over the search volume to sample the permutation distribution. Note that in order to preserve the spatial correlations between adjacent voxels in the observed maps, the same set of reassignments must be identically applied to all voxels in the image at each permutation. If the voxels are permuted independently, this will lead to systematic underestimation of the probability of a given cluster area under the null hypothesis [11]. The resulting distribution of ν is plotted in Fig. 4. It can be seen that the permutation and theoretical distributions for cluster area are less closely approximate than the corresponding pair of distributions for global grey matter volume presented in Fig. 1. In particular, the theoretical distribution relatively underestimates the probability of the smallest clusters and overestimates the probability of medium sized clusters.

To calibrate Type I error control by both procedures, we again subdivided the control group into two randomly decided subsets and compared the observed number of positive tests



(a)

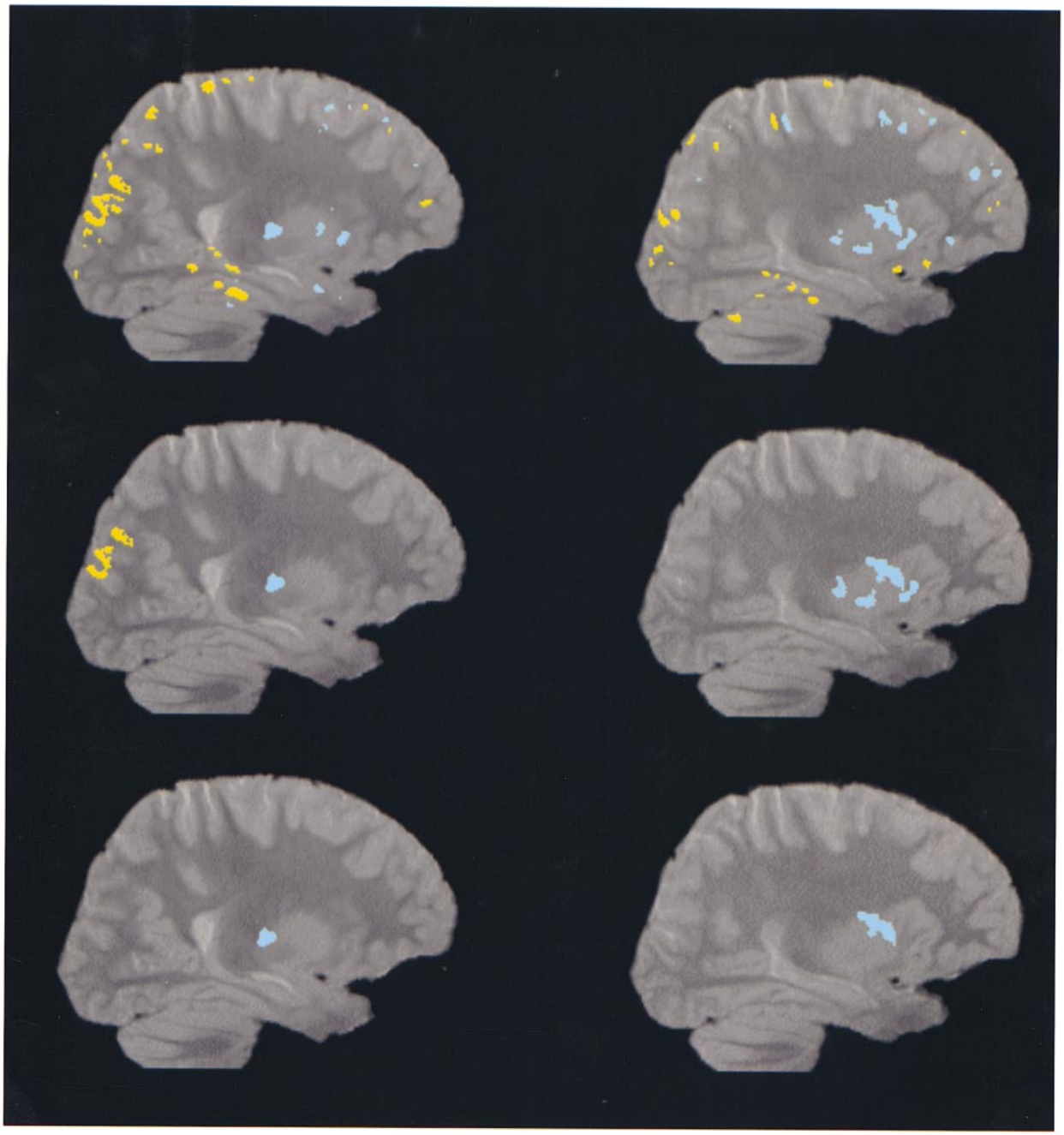
Fig. 3. The results of testing for a neuroanatomical difference between two groups of MRI data, acquired from 18 adolescents with hyperkinetic disorder and 16 matched control subjects. (a) A pair of sagittal slices through the right hemisphere are shown +9 mm (left column) and +12 mm (right column) lateral to the cerebral midline in standard space.

(at cluster level) to the expected number of false positive tests. As in the case of calibrating voxel testing procedures, the observed number of positive tests should be equal to the expected number of false positive tests. It can be seen from Fig. 5 that the permutation distribution consistently yields almost exactly the number of positive tests expected under the null hypothesis, whereas the theoretical distribution generally yields fewer positive tests than expected. This disparity is less obvious for smaller test sizes. However, in general it seems that the theoretical distribution is somewhat over conservative.

The results of using cluster area to test (by permutation) for a difference between the two original groups are shown in Fig. 3.

D. Cluster Mass

The permutation distribution for cluster mass was ascertained in exactly the same way as described for cluster area, except for the obvious difference that the mass τ rather than area ν of each suprathreshold cluster was estimated after each permutation and pooled over the search volume (see Fig. 6).



(b)

Fig. 3. (*Continued.*) The results of testing for a neuroanatomical difference between two groups of MRI data, acquired from 18 adolescents with hyperkinetic disorder and 16 matched control subjects. (b) Another pair of sagittal slices are shown +24 mm (left column) and +27 mm (right column) lateral to the cerebral midline. In both displays, the top row shows the results of thresholding a voxel test statistic (ANCOVA model coefficient) with $P \leq 0.05$. Voxels demonstrating significant grey matter volume deficit in the hyperactive group are colored blue and voxels demonstrating significant grey matter volume excess in the hyperactive group are colored yellow. The middle row shows the results of thresholding cluster area with $P \leq 0.05$. Clusters demonstrating significant grey matter deficit in the hyperactive group are colored blue and clusters demonstrating significant grey matter excess in the hyperactive group are colored yellow. The bottom row shows the results of thresholding cluster mass with $P \leq 0.05$. The color table is as for cluster area. The expected deficits in grey matter volume of right prefrontal cortex and basal ganglia are demonstrated most clearly in the hyperactive group by the test based on cluster mass.

We could find no theoretical distribution for τ in the existing literature.

Type I error-calibration curves for the permutation distribution of τ are presented in Fig. 7. It can be seen that the observed number of positive tests is almost exactly equal to the number of false positive tests expected under the null hypothesis.

The results of using cluster mass to test (by permutation) for a difference between the two original groups are shown in Fig. 3.

IV. Discussion

The main aim of this paper has been to compare two ways of testing three different kinds of statistics that might be of

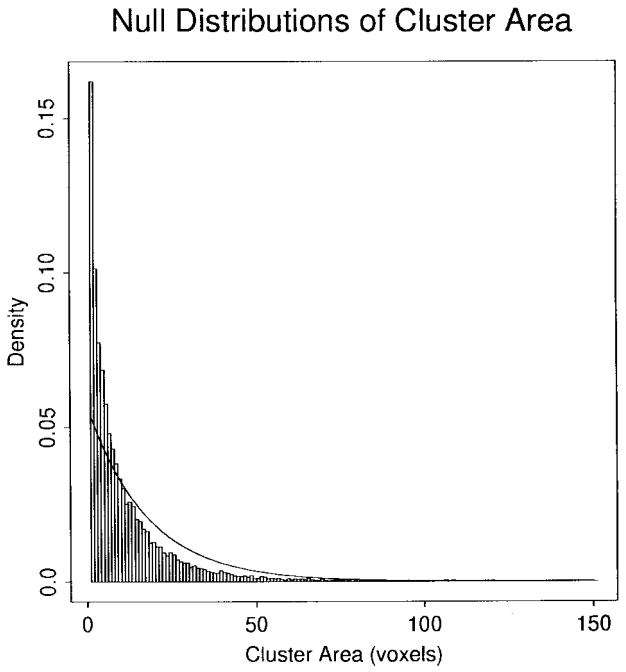
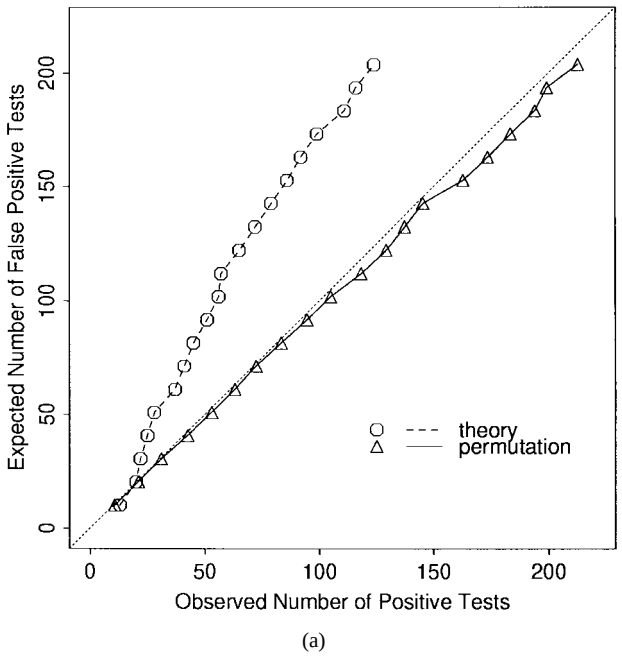


Fig. 4. Theoretical and permutation distributions for the area of suprathreshold clusters under the null hypothesis. The histogram shows the permutation distribution and the solid line shows the theoretical distribution (9) with $\mathrm{FWHM} = 5.67$ voxels and $u = 1.85$.

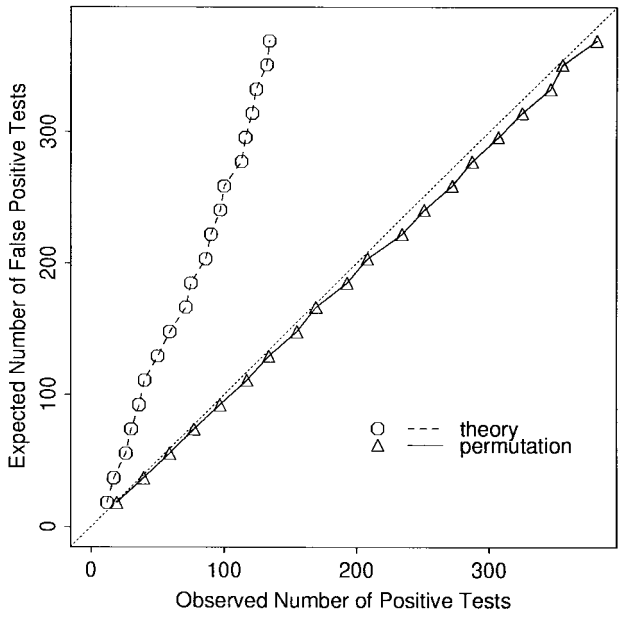
value in identifying a neuroanatomical difference represented in structural MRI data acquired from two groups of subjects. Specifically, we have estimated global, voxel, and cluster statistics and tested each of these (wherever possible) against both a null distribution derived from normal theory and a null distribution derived from repeated permutation of the observed data.

One of our findings has been that cluster statistics appear to be more specifically informative about the neuroanatomical differences between these two quite small groups of subjects than either global or voxel statistics. This is perhaps not a particularly surprising observation given the extensive literature advocating cluster statistics in the context of functional neuroimaging [4]–[6], [9]. Even without the benefit of this literature, one might expect cluster statistics in structural neuroimaging to be more informative than voxel tests simply because measurements on suprathreshold clusters are more informed by the full dimensionality of the data. A voxel statistic is based on only one dimension: the feature or effect estimated at that voxel. Cluster area is additionally informed by the x and y spatial dimensions of the data, but some information in the effect dimension is lost by setting all suprathreshold voxels to the same value. Cluster mass is also informed by two spatial dimensions and has the further advantage of preserving information in the effect dimension. Extrapolating, we might expect to find that estimating cluster mass in all three spatial dimensions would provide even more informative measures of the difference between groups.

Another main finding concerns the development and (cross) validation of permutation procedures for testing global, voxel, and cluster statistics. We have shown that permutation distributions can be ascertained for all the statistics considered



(a)



(b)

Fig. 5. Type I error-calibration curves for tests of cluster area by theory and permutation. The expected number of false positive clusters over a range of sizes of test $0.005 < P \leq 0.1$ can be compared directly to the observed number of significant clusters following tests by theory (dashed line and circles) and permutation (solid line and triangles). The two groups tested were randomly decided subsets of the control subjects. Therefore, the number of observed and expected tests should be equal (dotted line). (a) Cluster area was measured in the thresholded effect maps obtained by fitting an ANOVA model (4) at each voxel. (b) Cluster area was measured in the thresholded effect maps obtained by fitting an ANCOVA model (7) at each voxel.

here without imposing unrealistic demands on computational resources. Nominal Type I error control for all voxel and cluster permutation tests has been demonstrated by analysis of two randomly decided subsets of the control group data. For some test statistics, we found the theoretically ascertained null distribution was closely approximate to the permutation distribution and provided approximately equivalent quality

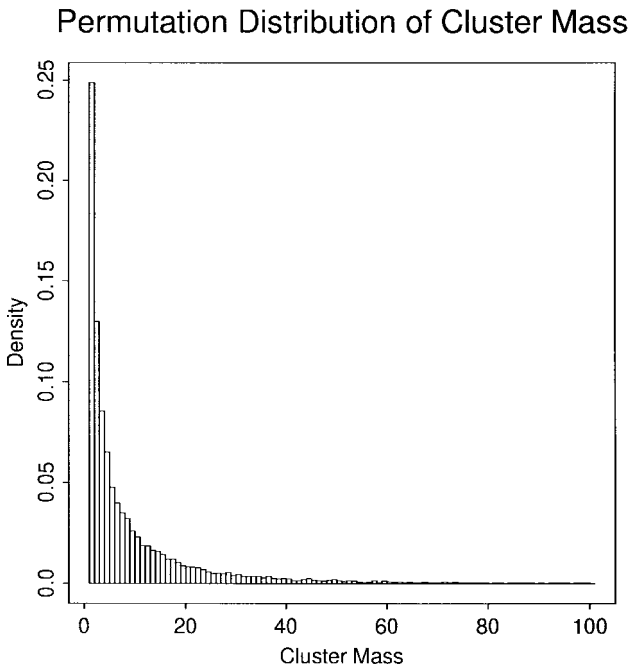


Fig. 6. Permutation distribution for the mass of suprathreshold voxel clusters under the null hypothesis.

of Type I error control. However, in the case of the voxel statistic estimated by fitting a simple ANOVA model, and in the case of cluster area, the theoretical distributions seemed somewhat over conservative for tests of size $\alpha \geq 0.01$. For the voxel test this may not matter much since the large number of tests conducted is likely to enforce a more stringent probability threshold for significance than 0.01. However, one of the potential advantages of analysis at a cluster level is that the smaller number of tests allows more relaxed probability thresholds, in the order of 0.01, so the behavior of the theoretical distribution in this case is more likely to be problematic in practice. The most likely explanation for any discrepancy between theory and permutation tests is that the assumptions entailed by the theoretical test were not entirely justified in the context of the analysis. In our first example, the residuals of the ANOVA model may not, in fact, have had a standard normal distribution, due to the unmodeled effects of other factors such as handedness. Likewise, the exponential distribution adopted for cluster area is based on theoretical results that are only exact in the case of much higher values for the voxel threshold u than we have applied here.

On the basis of these findings, it would clearly be wrong to advance any general conclusions about the validity of theoretical tests for structural brain-image analysis. If the assumptions they entail are justified by the data, it seems likely that they will yield very similar results to the corresponding permutation test [20]. However, the validity of permutation tests is generally conditional on far fewer assumptions and permutation tests can be readily devised for any statistic of interest. For example, here we were interested in cluster mass, which could be easily tested by permutation (and only by permutation).

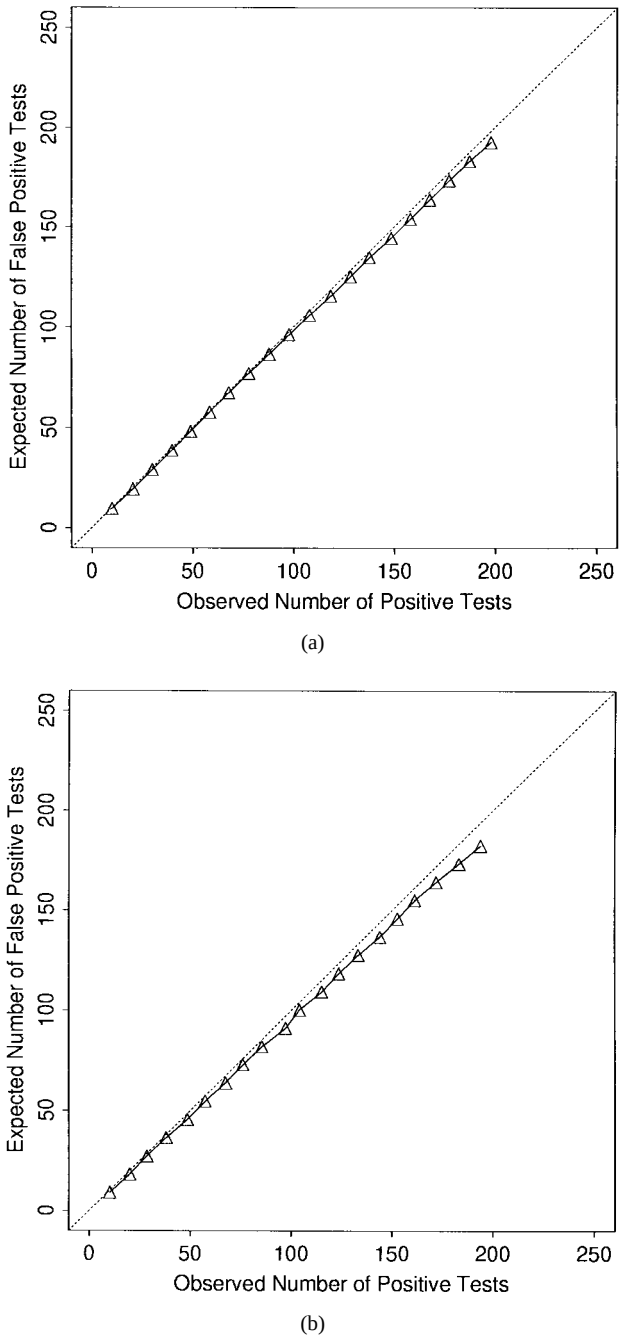


Fig. 7. Type I error-calibration curves for tests of cluster mass by permutation. The expected number of false positive clusters over a range of sizes of test $0.005 < P \leq 0.1$ can be compared directly to the observed number of significant clusters following tests by permutation (solid line and triangles). The two groups tested were randomly decided subsets of the control subjects, therefore, the number of observed and expected tests should be equal (dotted line). (a) Cluster mass was measured in the thresholded effect maps obtained by fitting an ANOVA model (4) at each voxel. (b) Cluster mass was measured in the thresholded effect maps obtained by fitting an ANCOVA model (7) at each voxel.

Historically, the advantages of permutation testing have been well recognized, but mitigated by the computational cost entailed. To paraphrase a remark made by R. A. Fisher [12], results obtained by theory are valid only insofar as they are corroborated by permutation, but this elementary method is tedious. With the increasing accessibility of powerful micro-

processors, the tedium of permutation testing is much reduced, and there seems no important argument remaining against preferred use of this elementary but exact and flexible method.

REFERENCES

[1] B. S. Everitt and E. T. Bullmore, "Mixture model mapping of brain activation in functional magnetic resonance images," *Human Brain Mapping*, vol. 7, pp. 1–14, 1999.

[2] A. R. McIntosh, F. L. Bookstein, J. V. Haxby, and C. L. Grady, "Spatial pattern analysis of functional brain images using partial least squares," *NeuroImage*, vol. 3, pp. 143–157, 1996.

[3] I. C. Wright, P. K. McGuire, J. B. Poline, J. M. Travere, R. M. Murray, C. D. Frith, R. S. J. Frackowiak, and K. J. Friston, "A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia," *NeuroImage*, vol. 2, pp. 244–252, 1995.

[4] J. B. Poline and B. M. Mazoyer, "Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise pixel clusters," *J. Cerebral Blood Flow Metabolism*, vol. 13, pp. 425–437, 1993.

[5] K. J. Friston, K. J. Worsley, R. S. J. Frackowiak, J. C. Mazziotta, and A. C. Evans, "Assessing the significance of focal activations using their spatial extent," *Human Brain Mapping*, vol. 1, pp. 214–220, 1994.

[6] P. E. Roland, B. Levin, R. Kawashima, and S. Akerman, "Three-dimensional analysis of clustered voxels in 15 O-butanol brain activation maps," *Human Brain Mapping*, vol. 1, pp. 3–19, 1993.

[7] S. D. Forman, J. D. Cohen, M. Fitzgerald, W. F. Eddy, M. A. Mintun, and D. C. Noll, "Improved assessment of significant activation in functional magnetic resonance imaging fMRI: Use of a cluster size threshold," *Magnetic Resonance Med.*, vol. 33, pp. 636–647, 1995.

[8] R. S. J. Frackowiak, S. Zeki, J. B. Poline, and K. J. Friston, "A critique of a new analysis proposed for functional neuroimaging," *Eur. J. Neuroscience*, vol. 8, pp. 2229–2231, 1996.

[9] J. B. Poline, K. J. Worsley, A. C. Evans, and K. J. Friston, "Combining spatial extent and peak intensity to test for activations in functional imaging," *NeuroImage*, vol. 5, pp. 83–96, 1997.

[10] P. E. Roland and B. Gulyas, "Assumptions and validations of statistical tests for functional neuroimaging," *Eur. J. Neuroscience*, vol. 8, pp. 2232–2235, 1996.

[11] S. Rabe-Hesketh, E. T. Bullmore, and M. J. Brammer, "Analysis of functional magnetic resonance images," *Statistical Methods Med. Research*, vol. 6, pp. 215–237, 1997.

[12] R. A. Fisher, "The coefficient of racial likeness and the future of craniometry," *J. R. Anthropological Soc.*, vol. 66, pp. 57–63, 1936.

[13] E. J. G. Pitman, "Significance tests which may be applied to samples from any populations," *J. Royal Statistical Soc.*, vol. 4, pp. 119–130, 1937.

[14] E. S. Edgington, *Randomization Tests*. New York: Dekker, 1980.

[15] B. J. F. Manly, *Randomization and Monte Carlo Methods in Biology*. London, U.K.: Chapman and Hall, 1991.

[16] P. J. Good, *Permutation Tests*. New York: Springer-Verlag, 1994.

[17] R. C. Blair and W. Karniski, "Distribution-free statistical analyzes of surface and volumetric maps," in *Functional Neuroimaging: Technical Foundations*. New York: Academic, 1994, pp. 19–28.

[18] S. Arndt, T. Cizadlo, N. C. Andreasen, D. Heckel, S. Gold, and D. S. O. Leary, "Tests for comparing images based on randomization and permutation methods," *J. Cerebral Blood Flow Metabolism*, vol. 16, pp. 1271–1279, 1996.

[19] E. T. Bullmore, M. J. Brammer, S. C. R. Williams, S. Rabe-Hesketh, N. Janot, A. S. David, J. D. C. Mellers, R. Howard, and P. Sham, "Statistical methods of estimation and inference for functional MR image analysis," *Magnetic Resonance Med.*, vol. 35, pp. 261–277, 1996.

[20] A. P. Holmes, R. C. Blair, J. D. G. Watson, and I. Ford, "Nonparametric analysis of statistic images from functional mapping experiments," *J. Cerebral Blood Flow Metabolism*, vol. 16, pp. 7–22, 1996.

[21] M. J. Brammer, E. T. Bullmore, A. Simmons, S. C. R. Williams, P. M. Grasby, R. J. Howard, P. W. R. Woodruff, and S. Rabe-Hesketh, "Generic brain activation mapping in fMRI: A nonparametric approach," *Magnetic Resonance Imaging*, vol. 15, pp. 763–770, 1997.

[22] "The ICD-10 classification of mental and behavioral disorders. Diagnostic criteria for research", World Health Org., Geneva, Switzerland, 1993.

[23] J. M. Swanson, J. A. Sergeant, E. Taylor, E. J. S. Sonuga-Barke, P. S. Jensen, and D. P. Cantwell, "Attention deficit hyperactivity disorder and hyperkinetic disorder," *Lancet*, vol. 351, pp. 429–433, 1998.

[24] J. Suckling, M. J. Brammer, A. Lingford-Hughes, and E. T. Bullmore, "Removal of extracerebral tissues in dual-echo magnetic resonance images via linear scale space features," *Magnetic Resonance Imaging*, vol. 17, pp. 247–256, 1999.

[25] E. T. Bullmore, M. J. Brammer, G. Rouleau, B. S. Everitt, A. Simmons, T. Sharma, S. Frangou, R. M. Murray, and G. Dunn, "Computerized brain tissue classification of magnetic resonance images: A new approach to the problem of partial volume artefact," *NeuroImage*, vol. 2, pp. 133–147, 1995.

[26] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley, 1992.

[27] J. Talairach and P. Tournoux, *A Coplanar Stereotactic Atlas of the Human Brain*. Stuttgart, Germany: Thieme Verlag, 1988.

[28] R. W. Cox, "Analysis and visualization of 3D fMRI data," in *Proc. 3rd Scientific Meeting Soc. Magnetic Resonance*, 1995, vol. 2, p. 834.

[29] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge: Cambridge Univ. Press, 1992.

[30] A. J. Dobson, *Introduction to Statistical Modeling*. London, U.K.: Chapman & Hall, 1983.

[31] R. G. Miller, *Beyond ANOVA. Basics of Applied Statistics*. London, U.K.: Chapman & Hall, 1997.