NEUROIMAGE **5,** 83–96 (1997)
ARTICLE NO. NI960248

Combining Spatial Extent and Peak Intensity to Test for Activations in Functional Imaging

J-B. Poline,* K. J. Worsley,† A. C. Evans,‡ and K. J. Friston*

**Wellcome Department of Cognitive Neurology, Institute of Neurology, London, United Kingdom; †Department of Mathematics and Statistics, McGill University, Montreal, Canada; and ‡Montreal Neurological Institute, McGill University, Montreal, Canada*

Received April 22, 1996

Within the framework of statistical mapping, there are up to now only two tests used to assess the regional significance in functional images. One is based on the magnitude of the foci and tends to detect high intensity signals, while the second is based on the spatial extent of regions defined by a simple thresholding of the statistical map, a test that is more sensitive to extended signals. The aim of this paper is to combine the two tests into a single test that is more sensitive to a wider range of signals. This combined test is based on an analytical approximation of the distribution of these two parameters (size and height) and is applied in the context of statistical maps. The risk of error in noise-only 2D or 3D volumes is assessed under a wide range of experimental conditions obtained by varying both the resolution of the map and the threshold at which clusters are defined. In addition, we have investigated this new test on simulated signals, and applied it to an experimental PET dataset. The experimental risk of error is close to the predicted one, and the overall sensitivity increases when analyzing a volume containing different types of signals. © 1997 Academic Press

1. INTRODUCTION

Statistical analysis of functional images often employs pixel-based analyses. Statistical parametric maps (SPMs) are constructed to reflect the probability of change at each voxel. The significance of regional activation in PET studies and more recently fMRI studies is usually assessed using two kinds of tests. The first uses the magnitude of the SPM value, in other words, the peak height of a cluster in SPM (Friston *et al.*, 1991; Worsley *et al.*, 1992). Because increases in activity do not necessarily conform to a sharp peak but can appear as more broad spatially distributed increases, some tests based on the spatial extent of suprathreshold regions were introduced, at first using empirical distributions (Poline and Mazoyer, 1992, 1993; Roland *et al.*, 1993) and then using results from the Gaussian random field theory (Friston *et al.*, 1994). These tests are generally more sensitive at the expense of dropping the risk of error from the pixel level to the "cluster" level. Unfortunately, spatial-extent-based analyses do not necessarily increase the sensitivity in all cases. Clearly, sharp localized activations might occur (for instance in small structures) and would be missed by the use of spatial extent tests. Regions showing increased activity are likely to have various shapes and extents in real data.

To address this issue, some new strategies have been considered. First, peak height detection could be enhanced if the data were subject to a filter that matches the signal to be detected. A multifiltering strategy was therefore proposed and has been shown to be generally more sensitive (Poline and Mazoyer, 1994a; Worsley *et al.*, 1996). While efficient and robust, this strategy requires larger computer resources and is limited by the use of large filters that degrade the spatial resolution.

A second strategy consists of testing for both the spatial extent and the peak height and was investigated using a test based on Monte Carlo simulations (Poline and Mazoyer, 1994b). Monte Carlo simulations are costly in terms of computer resources and have to be repeated for different analysis parameters in order to properly assess the distribution under the null hypothesis, especially the thresholds defining the clusters.

While these two tests (peak intensity and spatial extent) are currently available, it is not valid to use them both at the same time without correcting for the implicit multiple testing.

We propose in this paper to combine peak height and spatial extent tests into a single test, using a theoretical model for the joint bivariate distribution of these two variables under the null hypothesis. We have investigated specificity using some simulated noise volumes and sensitivity using (i) some known simulated signals (in two and three dimensions) and (ii) an actual PET language activation study.

1053-8119/97 $25.00

Copyright © 1997 by Academic Press
All rights of reproduction in any form reserved.

2. MATERIALS AND METHODS

2.1. Combined Test: Optimizing the Intensity Threshold

The two tests that we propose to combine appear to be quite different; the peak height test sets a high-intensity threshold, such that any voxel above it can be considered significant; the spatial extent test requires two thresholds: a low-intensity threshold and a size threshold, such that any cluster with a spatial extent greater than, or equal to, the size threshold is declared significant. In fact, the first test can be regarded as a special case of the second with a size threshold of 1 (i.e., a peak is above the intensity threshold if, and only if, its spatial extent is greater than zero). Therefore, we are really combining two variants of the same test.

This leads to the more general question of how to choose the intensity threshold to maximize the sensitivity of the spatial extent test. This question has been partly answered by Friston *et al.* (1994), who considered a simple model of white noise plus Gaussian signal, both smoothed by a Gaussian point spread function (PSF). With this model, they show that for signals wider than the full width at half-maximum (FWHM) of the PSF, the intensity threshold should be low; for signals sharper than the FWHM of the PSF, the image threshold should be high.

Figure 1 illustrates why this is so. It turns out that the critical cluster size can be obtained approximately as follows. Construct a signal identical to the PSF, smooth it with the PSF, to give a Gaussian function with $\sqrt{2}$ times the FWHM of the PSF. Set the height to the intensity threshold for a 5% test of peak height, found using the formula in Adler (1981), reported in Worsley *et al.* (1992). Call this the *critical function.* Now threshold the critical function at an arbitrary lower level (t) and measure its spatial extent; this is then the approximate size threshold we seek for a 5% spatial extent test at an intensity threshold of t (Fig. 1a).

Now, if the FWHM of a Gaussian signal is less than that of the PSF, it will be sharper than the critical function; it therefore follows that the image threshold must be set high in order for the spatial extent of the signal to exceed the spatial extent of the critical function (Fig. 1b). Conversely, if the FWHM of a Gaussian signal is larger than that of the PSF, then the smoothed signal is wider than the critical function and so the intensity threshold should be set low in order to be best detected (Fig. 1c).

The above discussion shows that sensitivity to Gaussian signals depends on the choice of intensity thresholds: low thresholds are best for wide signals, and high thresholds are best for sharp signals. This suggests that sensitivity can be maximized by searching over a range of intensity thresholds. This is analogous to scale-space searches (i.e., multifiltering), originally proposed by Poline and Mazoyer (1994a) and then developed in a theoretical framework by Siegmund and

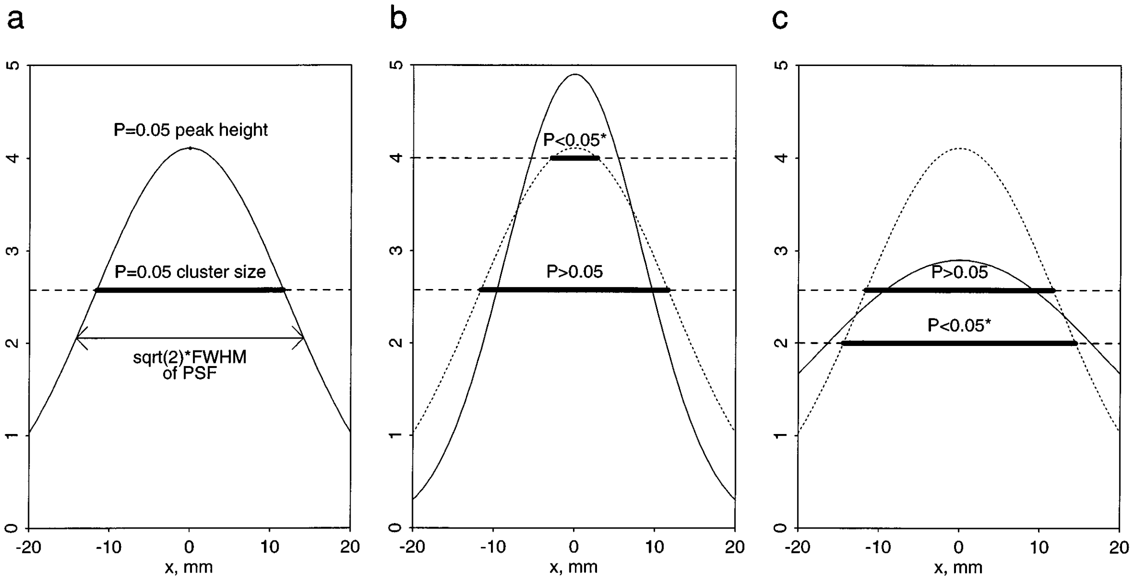


FIG. 1. (a) The critical function (solid curve), constructed by smoothing a signal identical to the point spread function (PSF), with the PSF, to give a Gaussian function with FWHM equal to $\sqrt{2}$ times the FWHM of the PSF. The height equals the intensity threshold for a 5% test based on peak height. The spatial extent of the critical function at any lower level is the approximate size threshold for a 5% spatial extent test. (b) If the FWHM of the Gaussian signal (dotted curve) is less than that of the PSF, its image will be sharper than the critical function; it follows that the image threshold must be set high (*) in order for the spatial extent of the signal to exceed the spatial extent of the critical function. (c) Conversely, if the FWHM of the Gaussian signal (dotted curve) is larger than that of the PSF, then the smoothed signal is wider than the critical function and the intensity threshold should be set low (*) for greatest sensitivity.

Worsley (1995) and Worsley and co-workers (1996), in which sensitivity is maximized by searching over a range of smoothing kernel FWHMs. In other words, intensity thresholds play an analogous role to FWHM; both affect the sensitivity for different size signals.

However, there is one major difference between scale-space searches and intensity threshold searches, which is apparent from Fig. 1. If the signal is wider than the PSF, then the intensity threshold should be set as low as possible; if the signal is narrower than the PSF, then the threshold should be set as high as possible, with nothing in between. This is not the case for scale-space searches; there the optimal FWHM matches the FWHM of the signal (by the Matched Filter Theorem), which opens up the possibility of estimating the FWHM of the signal by the FWHM of the optimal filter.

No such estimator appears to be possible via intensity threshold and spatial extent tests. Instead, the optimal image threshold takes one of two values: either as low as possible for wide signals or as high as possible for narrow signals. The generally more powerful test would then simply take the best of the two image thresholds, that is, the minimum of the two P values. Now the highest possible threshold, as the above discussion shows, corresponds simply to the peak height test. Thus the best test should take the minimum of the P value for the low intensity threshold spatial extent test and the peak height test. This is the very test that we propose in this paper.

To summarize:

- The sensitivity of the spatial extent test for different signal sizes depends on the intensity threshold;
- This suggests searching over values of the intensity threshold (the image threshold defining the clusters), by analogy to scale-space searches;
- However, the optimal intensity threshold for Gaussian shaped signals flips between very low and very high, depending on whether the signal width is larger or smaller then the PSF width (unlike scale-space searches, where optimal FWHM matches that of the signal);
- The highest possible image threshold corresponds to the peak height test;
- Thus the combination of the spatial extent test (for a low image threshold) and the peak height test, via the minimum P value of the two tests, should provide a better detection for signals of all sizes.

2.2. Combined Test: Theory

In this section we specify a combined test based on two parameters (peak height and spatial extent).

First, an approximation for the probability of a given cluster having a spatial extent S greater than s_0 and maximum intensity or peak height H greater than h_0 is derived in the Appendix using results from Gaussian random field theory (see Eq. (14)). This distribution is a function of the image threshold and spatial resolution and is derived under the hypothesis of a Gaussian autocovariance function (see Section 4). Second, a way of combining the spatial extent and the maximum intensity is chosen in order to select events (an occurrence of a cluster) that will be rejected at a given risk of error under the null hypothesis of pure noise. There are an infinite number of possibilities for this step. In Poline and Mazoyer (1994b), two clusters (s_0, h_0) and (s_1, h_1) would be rejected with the same risk of error if

$$P(S \geq s_0, H \geq h_0) = P(S \geq s_1, H \geq h_1).$$

In other words, the rejection area is defined by the isocumulative curve

$$P(S \geq s_0, H \geq h_0) = \text{constant}$$

(see Fig. 2). For our proposed combined test, the risk of error is simply defined as the minimum of the risk for spatial extent and the risk for maximum peak height. This gives a rejection area defined by

$$\min\{P(S \geq s_0), P(H \geq h_0)\} = \text{constant}$$

(see Fig. 2 and Subsection 2.3).

The test of Poline and Mazoyer (1994b) has a smooth boundary; ours has a rectangular boundary. The main difference is as follows. If one statistic, say spatial extent, is more significant than the other (peak height), then our test will give the same significance irrespective of the peak height; the test of Poline and Mazoyer (1994b) will be more conservative if the evidence against the null hypothesis from peak height is weak. On the other hand, if both spatial extent and peak height are equally significant, our test will be more conservative and that of Poline and Mazoyer (1994b) will give more significant results. In practice the difference between the two tests is likely to be small; ours has the distinct advantage that it is based on the simple concept of the minimum of the P values of spatial extent and peak height.

Third, the conditional probability above is used to compute the unconditional probability. This is simply done since, for a high threshold t, the clusters are independent and the number of clusters C above t follows approximately a Poisson law with mean m (Adler, 1981) given by

$$m(t) = V|\Lambda|^{1/2}(2\pi)^{-(D+1)/2} t^{D-1} e^{-t^2/2}, \qquad (1)$$

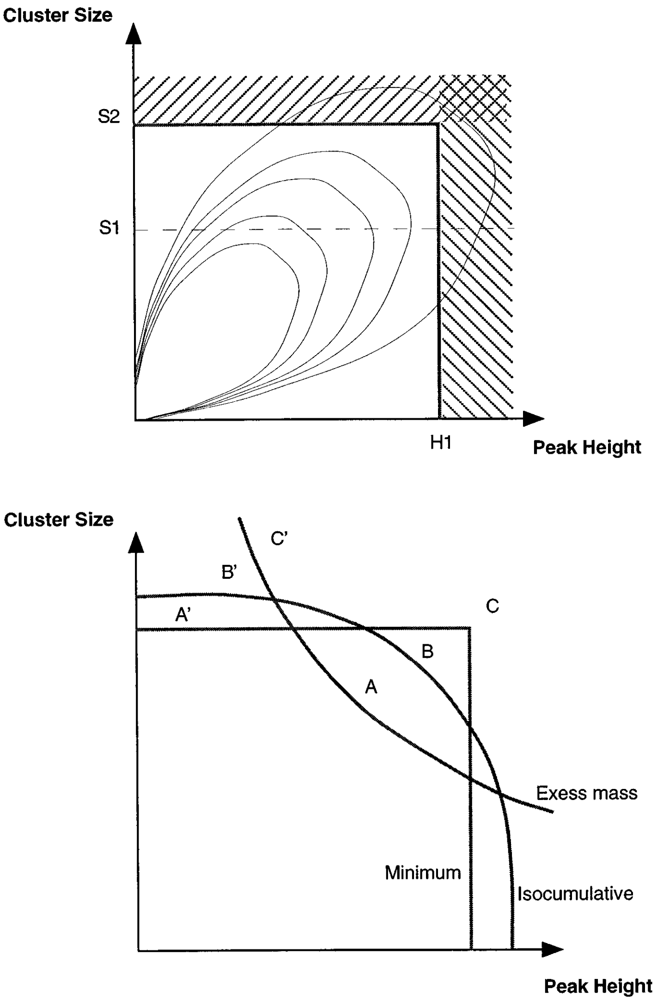


FIG. 2. (Top) The rejection area for a cluster with height H_1 and size S_1 based on the minimum P value of the two established tests. In this example the minimum probability is found using the peak height H_1. S_2 corresponds to the cluster with a similar P value for either the cluster size test or the peak height test. This defines our 2D rejection boundary (thick continuous line). All the clusters falling on this line have the same probability of being rejected, i.e., the integral over the striped area. Hairlines represent isointensity curves of the 2D probability density function. Also shown are the rejection regions for the spatial extent and peak height tests alone (stripes going up and down, respectively). (Bottom) Three different shapes for the rejection boundary: iso-cumulative contours (Poline and Mazoyer, 1994b), minimum P value (our proposed combined test), and volume of the cluster between the intensity threshold and the image itself (excess mass). For example, cluster (C) would be detected by the three tests, cluster (B) by the "iso-cumulative" test and the excess mass test only, and cluster (A) by the excess mass test only. On the other hand, cluster (A′) is detected by the "minimum" test and not by the two others, (B′) is detected by the "minimum" and isocumulative test, and (A′) by all three tests.

where V is the volume of the search region and Λ is the variance matrix of the derivatives of the image in each dimension. For an image generated by white noise smoothed by a Gaussian PSF,

$$V|\Lambda|^{1/2} = \text{RESELS}(4 \log_e 2)^{D/2},$$

where RESELS is the volume of the search region divided by the product of the FWHMs of the PSF in each dimension (see Worsley *et al.*, 1992).

We denote by P_{rej} the probability that the spatial extent and peak height probability of a single cluster falls in the rejection area. Then if k clusters occur in the volume V, the probability that at least one cluster will be rejected is simply

$$P(\text{rejection}|C = k) = (1 - (1 - P_{\text{rej}})^k).$$

Summing over k, weighted by the probability of $C = k$, we get

$$\begin{aligned} P(\text{rejection}) &= \sum_{k=0}^{\infty} (1 - (1 - P_{\text{rej}})^k) m^k e^{-m}/k! \\ &= 1 - e^{-m(t)P_{\text{rej}}}. \end{aligned} \quad (2)$$

2.3. Combined Test: Summary

The procedure for finding the P value of a cluster of size S_0 and a peak of height H_0 above a threshold t (so that the total peak height is $t + H_0$) is as follows.

1. Find the marginal probability for spatial extent for a given cluster:

$$P(S \geq S_0) = e^{-(|\Lambda|^{1/2} t^D S_0/ac)^{2/D}}, \quad (3)$$

where a is given by Eq. (6) in the Appendix and c is given by Eq. (12) in the Appendix (see also Friston *et al.*, 1994).

2. Find the marginal probability for peak height above threshold for a given cluster by dividing $m(t + H_0)$ by $m(t)$ (see Eq. (1)) to give

$$P(H \geq H_0) = (1 + H_0/t)^{D-1} e^{-tH_0 - H_0^2/2}, \quad (4)$$

which can be well approximated by e^{-tH_0} for large t and small H_0 (Friston *et al.*, 1994).

3. Find the minimum probability

$$\mu = \min\{P(S \geq S_0), P(H \geq H_0)\}.$$

In the next two steps, we find values s_0 and h_0 such that the cluster level marginal probabilities are both equal to μ:

$$\mu = P(S \geq s_0) = P(H \geq h_0).$$

4. If $\mu < P(S \geq S_0)$, then put $h_0 = H_0$ and equate Eq. (3) to μ and solve for s_0 to give

$$s_0 = ac|\Lambda|^{-1/2} t^{-D}(-\log_e \mu)^{D/2}.$$

5. If $\mu < P(H \geq h_0)$, then put $s_0 = S_0$ and equate Eq. (4) to μ and solve for h_0; there is no exact expression, but for large t and small h_0, $h_0 \approx (-\log_e \mu)/t$.

6. Use Eq. (14) in the Appendix to find the combined probability of rejection for the spatial extent and peak height at the cluster level:

$$\begin{aligned} P_{\mathrm{rej}} &= P(S \geq s_0) + P(H \geq h_0) - P(S \geq s_0, H \geq h_0) \\ &= 2\mu - P(S \geq s_0, H \geq h_0). \end{aligned}$$

7. Use Eq. (2) to correct this for searching over clusters, to get the final P value of the combined test:

$$P(\text{rejection}) = 1 - e^{-mP_{\mathrm{rej}}}.$$

2.4. Validation Using Monte Carlo Simulations

2.4.1. Noise Simulations

We performed straightforward simulations under the null hypothesis to validate the risk of error as computed using Eq. 2. Random Gaussian fields were simulated using white Gaussian noise convolved with a Gaussian kernel, using two-dimensional (128 × 128 pixels) or three-dimensional (64 × 64 × 32 voxels) processes. The agreement between the experimental false positive rate and the predicted (theoretical) one was assessed over various thresholds (t ranging from 2 to 3.5) and various smoothness parameters (Gaussian kernels with FWHM ranging from 10 to 20 mm in the xy plane and from 8 to 16 mm in the z direction, with the convention that a voxel is 2 × 2 × 2 mm^3). The number of simulations was, respectively, 10^4 for the 2D case and 5×10^3 for the 3D case. This ensures a good precision on the risk of error found on the simulated noise fields (for example, for a 5% risk of error, we have (for 5×10^3 simulations) $\sigma = \sqrt{(.05 * .95)/5 \times 10^3} = 0.3\%$). In 3D, the cluster detection was performed in a periodic manner to avoid edge effects, that is, clusters cut by edges. (Obviously, in actual brain volumes, the data are not periodic. Because of the edges of the brain, clusters found in noise only situation will tend to be smaller on average, and therefore the comparisons performed here might be too stringent).

2.4.2. Signal Simulations

The simulations presented here are not intended to provide a full-power analysis of the new test, simply because we do not have any good model for true signals. We therefore limited our analyses to some specific examples. These examples, however, should demonstrate the ability of the new test to overcome the limitations of previous strategies. We simulated three kinds of signal: (1) a sharp peak with high intensity (which should be detected by the peak height test), (2) a large signal with low intensity (which should be detected by the spatial extent test), and (3) a signal that would have approximately the same probability of being detected with either of the two established tests.

TABLE 1

Spatial Extent and Height of Simulated Signals in Two and Three Dimensions

Signal	1	2	3
2D			
Spatial extent S, mm^2	5.25	116.00	11.25
Peak height H	3.76	1.91	3.71
3D			
Spatial extent S, mm^2	4.87	224.25	15.12
Peak height H	4.46	2.35	3.81

All three signals were either 2D squares or 3D cuboids placed in Gaussian white noise and convolved with Gaussian kernels (FWHM = 14.10 mm or 7.05 pixels in the x and y direction and 11.77 mm in the z direction). The size and the maximum peak value of these three signals (in 2D and in 3D) are summarized in Table 1.

A signal was considered to be detected when at least one local maximum was found within an area corresponding to the voxels above 75% of the maximum of the original filtered signal.

2.5. Application to a Real (PET) Dataset

We present an application of the combined test on an experimental dataset acquired from six subjects who took part in a verbal fluency activation study. Subjects were all right-handed males aged between 20 and 60. During the stimulation task, subjects were asked to silently generate verbs related to nouns at the rate of one noun every 6 s. The control condition was a silent rest condition. Six scans were acquired per subject (three controls and three stimulations) on an Ecat 951 CTI PET scanner. Scans were stereotactically normalized to the Talairach space (Friston *et al.*, 1995) and analyzed using a complete randomized block design. A T statistic map was constructed for the contrast condition effect (verb generation minus rest), correcting for the global activity in the brain using proportional scaling leaving 25 degrees of freedom. The T map was then transformed to a Z map (see http://www.fil.ion.ucl.ac.uk, SPM short course for more information on the T to Z transformation), which was tested for significant regions by peak height, suprathreshold cluster size ($t = 2.81$), and by the combined test. We smoothed our data with a narrow kernel (6 mm FWHM in the three directions) to balance significance levels for the spatial extent and the peak height tests (more smoothed data would show much higher significance for the peak height test). The resulting Z map had a resolution of

9.8, 11.2, and 12.5 in the x, y, and z directions, respectively. Since we wanted to compare two valid methodologies, we corrected the P value obtained with the two original tests as if they were independent. Although this is clearly not true and leads to conservative results, it is the one way to correct for this "two tests" procedure.

3. RESULTS

3.1. Validation Using Monte Carlo Simulations

3.1.1. Shape of the Bivariate Distribution

We compare the bivariate distribution of the spatial extent and peak height of clusters found above the threshold $t = 3$ with the theoretical approximation obtained in Eq. (14). Figure 3 shows the results for the 3D simulations (results are, in fact, better in the 2D case, as seen in Section 3.1.2) with, from top to bottom, the theoretical (top) and experimental (middle) distributions and the difference (bottom), all on a logarithmic scale. We observe that although the agreement is good, the match is not perfect (maximum absolute difference is 0.034, found for objects of 1 pixel and intensity $H = .06$). However, we note that the 2D test involves summations over this distribution and therefore the actual results of the test are more robust than one might expect from Fig. 3. This point will be confirmed by the simulations below (Section 3.1.2).

3.1.2. Noise Simulations

Varying the threshold t. The risk of error observed in 2D and 3D is compared to the expected risk of error (type I error) for a series of thresholds $t = 2$, 2.5, 3, and 3.5 at a constant resolution (FWHM = 14.1 mm in x and y and 11.8 mm in z). On the same graphs we plotted the risk of error found with the previously validated tests (spatial extent, dashed line; peak height, dotted line). Recall that the threshold t here is the base threshold that defines the clusters, while H is the peak height of these clusters.

Figures 4 and 5 show the results for two and three dimensions, respectively. We present the results for the two low thresholds, where the theory is least accurate, and for the highest threshold ($t = 4$). Our results indicate the good agreement between expected and predicted risks of error with high thresholds, with the peak height test being slightly too conservative in 2D and 3D for all simulations. At low thresholds ($t = 2$), the combined test is too conservative for small risks of error in 2D, but not conservative enough for risks of error above 20% in 3D. The cluster size test also failed to control properly the type I error in 3D for $t = 2$. This effect should be strongly attenuated when using actual datasets that have edges. Overall, the combined test is

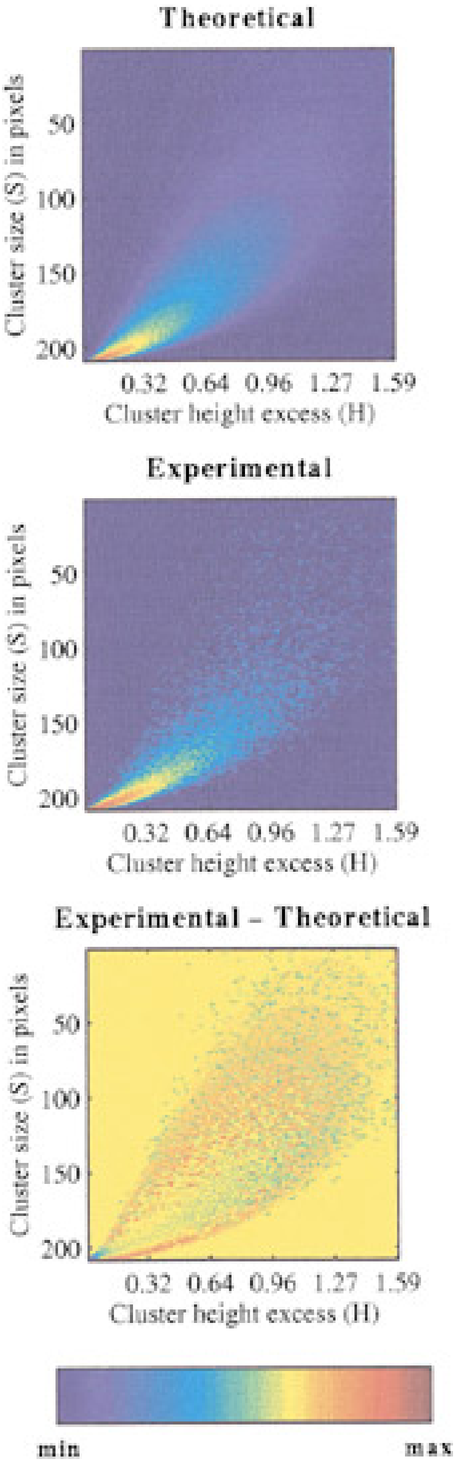
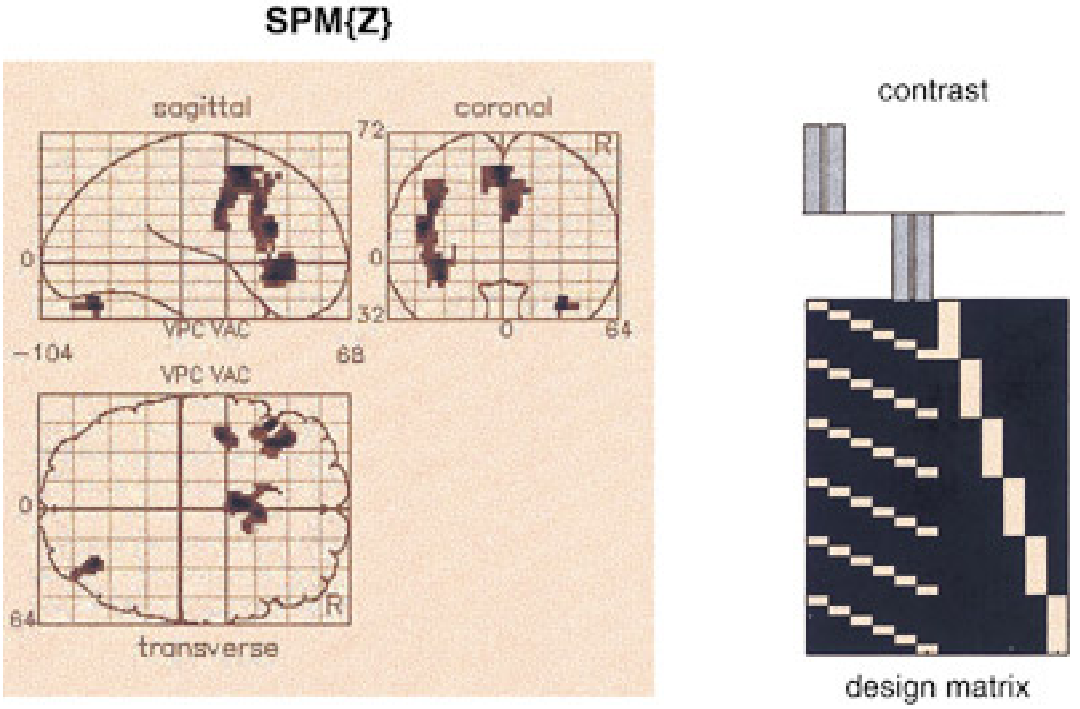


FIG. 3. (Top) Theoretical (predicted) bivariate distribution of spatial extent and peak height for regions occurring above an image threshold of $t = 3$ in a $64 \times 64 \times 32$ volume ($128 \times 128 \times 64$ mm^3) with resolution 17.5 mm in x and y and 12.5 mm in z. Data intensity is presented in a log scale to increase the visibility of the tail of the distribution. (Middle) Observed bivariate distribution of spatial extent peak height under the same conditions as those described above. (Bottom) Difference between the two. On the two top graphs, blue is a zero intensity (maximum intensity is 1) while on the bottom graph, null values are in yellow.



P values & statistics:

height & extent {n,Z}	extent {n}	peak height {Z}	location {mm}
0.044 (79, 4.78)	**0.275 (79)**	**0.072 (4.78)**	**−46 24 20**
0.007 (238, 4.68)	**0.003 (238)**	**0.106 (4.68)**	**−2 8 48**
		0.316 (4.20)	4 16 32
		0.981 (3.50)	12 10 40
0.040 (143, 4.61)	**0.040 (143)**	**0.144 (4.61)**	**−36 24 −8**
		0.284 (4.24)	−36 32 0
		1.000 (3.03)	−28 24 8
0.144 (45, 4.33)	**0.704 (45)**	**0.386 (4.33)**	**32 −74 −24**
		0.999 (3.31)	38 −84 −24
0.113 (99, 4.23)	**0.148 (99)**	**0.534 (4.23)**	**−40 −4 32**
		0.699 (3.88)	−42 −6 24

Height threshold {u} = 2.80, p = 0.003
Extent threshold {k} = 40 voxels
Expected voxels per cluster, E{n} = 11.9
Expected number of clusters, E{m} = 0.8

Volume S = 53132 voxels or 625.0 Resels
Degrees of freedom due to error = 25
Smoothness = 9.8 11.2 12.5 mm {FWHM}
= 4.1 4.8 5.3 {voxels}

FIG. 8. Experimental dataset analyzed with the three different tests. The first panel (top left) shows the maximum intensity projection of the Z map in the saggital coronal and axial orientations. The volume has been previously normalized to correspond to the Talairach atlas. In the right panel is the design matrix of the experiment for a two-way analysis of covariance. (see Friston *et al.* (1995) for a full description of the elements of the design matrix). The table presents for each region (clusters defined with an image threshold of $t = 2.8$ and the extent threshold of 40 voxels) the P value given by the combined test (in parenthesis: number of pixels (size) and peak height (Z)), the P value given by the spatial extent test, the P value given by the peak height test, and the location of the peak. The P values for the peak height and the spatial extent tests were corrected for a two-test procedure (see Section 2.5).

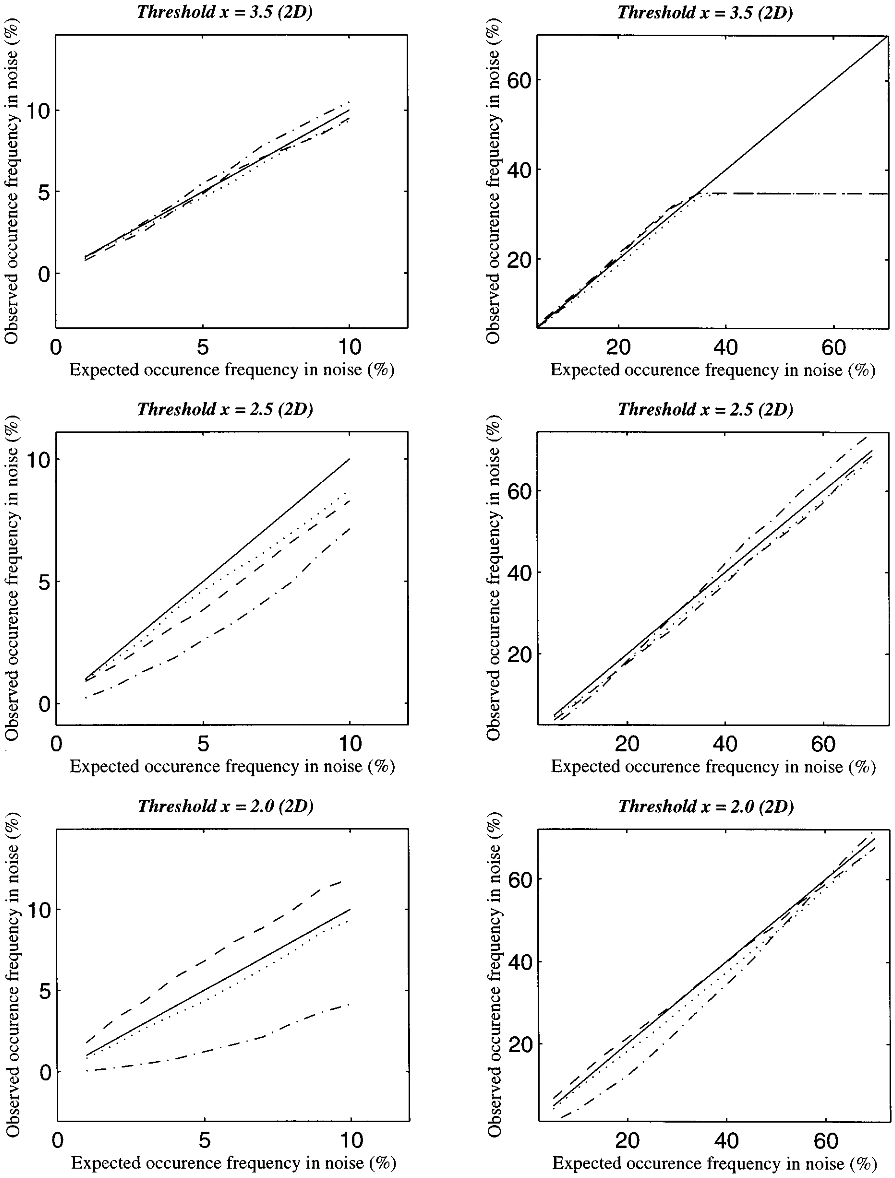


FIG. 4. Expected vs observed risk of error with various thresholds in 2D volumes (128 × 128 pixels or 256 × 256 mm^2) at a fixed resolution (FWHM$_{xy}$ = 14.1 mm). (Left) Risk of error between 1 and 10%. (Right) Risk of error varying between 10 and 70%. (Top) High intensity threshold (t = 3.5). (Middle) Low intensity threshold (t = 2.5). (Bottom) Very low threshold (t = 2.0). The dashed line shows the results from the spatial extent test, the dotted line from the peak height test, and the dot and dashed line from the combined test. The solid line corresponds to the $y = x$ line. Results were assessed using 7 × 10^3 simulations.

conservative except for low thresholds and high risks of error (greater than 20%), but the vast majority of applications should avoid these domains.

Varying the width of the convolution kernel (resolution). Varying the kernel resolution (constant threshold t = 3) did not have any major effects on the recorded risk of error: in all situations (2D or 3D, low resolution or high resolution) the observed risk was very close to the predicted one for the combined test (see Fig. 6; results in 2D are not shown as they are very similar to the 3D results). The peak height test again showed a tendency to be overly conservative over the range of resolutions employed.

3.1.3. Simulated Signals

Figure 7 presents the percentage of detection of the three signals defined previously at four different risks of error for the 3D signals. Similar results were found in 2D. For the first two "extreme" signals (sharp peak and spatially extended signals) the combined test results were in all cases close to the best results obtained

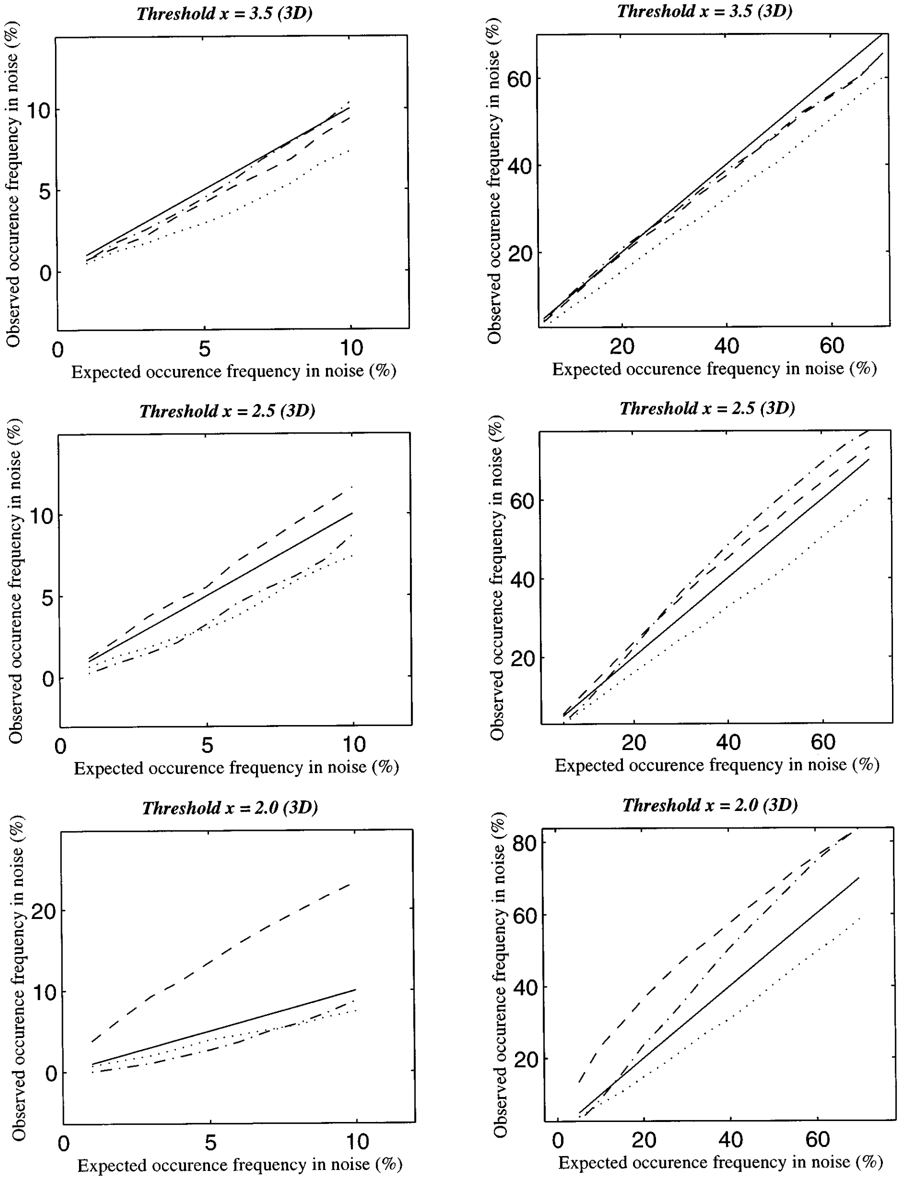


FIG. 5. Expected vs observed risk of error with various thresholds in 3D volumes (64 × 64 × 32 voxels or 128 × 128 × 64 mm^3) at a fixed resolution (FWHM$_{xy}$ = 14.1 mm and FWHM$_z$ = 11.8 mm). (Left) Risk of error between 1 and 10%. (Right) Risk of error varying between 10 and 70%. (Top) High intensity threshold (t = 3.5). (Middle) Low threshold (t = 2.5). (Bottom) Very low threshold (t = 2.0). The dashed line shows the results from the spatial extent test, the dotted line from the peak height test, and the dot and dashed line from the combined test. The solid line corresponds to the $y = x$ line. Results were assessed using 3 × 10^3 simulations.

either by the spatial extent test or by the peak height test, demonstrating the ability of the test to detect both kinds of signal with very little loss in sensitivity. Indeed, if both these signals were present in the same volume, the overall gain in sensitivity would be of the order of 20 to 30%.

The third signal (detected with a similar probability by either the spatial extent or peak height test) was also easily detected both in 2D and in 3D. Note that in any case the significance value given by the 2D test should always be less than the minimum of the P value given by the two one parameter tests, since this is how we constructed the rejection area. Occasionally, because we are working with approximate distributions, the significance of the combined test might be greater (lower P values) when compared to the two other tests. However, this disparity should not be great (see, for instance, the simulation results on the "balanced" signal in 3D for risks of error greater than 10%).

3.2. 2D Test Results on the Verbal Fluency Dataset

On the experimental dataset the combined test performed as predicted (see Fig. 8). We deliberately chose a

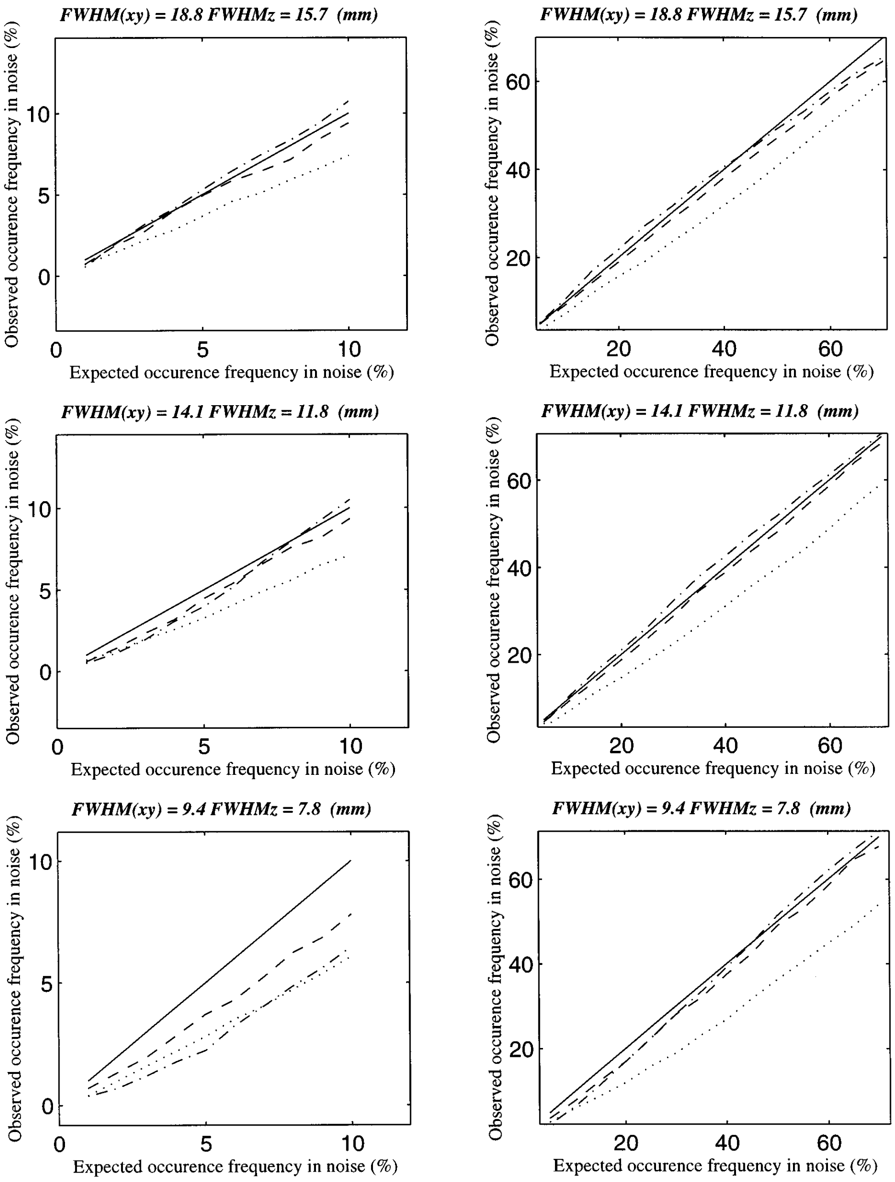


FIG. 6. Expected vs observed risk of error with various resolution in 3D volumes (64 × 64 × 32 voxels or 128 × 128 × 64 mm^3) at a fixed threshold ($t = 3$). (Left) Risk of error between 1 and 10%. (Right) Risk of error varying between 10 and 70%. (Top) Low resolution (FWHM$_{xy}$ = 18.8 mm and FWHM$_z$ = 15.7 mm). (Middle) Medium resolution (FWHM$_{xy}$ = 14.1 mm and FWHM$_z$ = 11.8 mm). (Bottom) Very high resolution (FWHM$_{xy}$ = 9.4 mm and FWHM$_z$ = 7.8 mm). The dashed line shows the results from the spatial extent test, the dotted line from the peak height test, and the dot and dashed line from the combined test. The solid line corresponds to the $y = x$ line. Results were assessed using 3 × 10^3 simulations.

contrast showing relatively weak results in terms of significance (strong results would be significant with any kind of test). The results demonstrate the versatility of the combined test: while some areas were detected because of the peak height (e.g., first cluster at the 0.072 risk of error), others were significant because of their large spatial extent (second and third cluster, $P = 0.003$ and $P = 0.040$ respectively), while all these clusters were found significant using the combined test. Overall, the combined test showed a greater sensitivity, although this benefit is emphasized by the correction for the two-tests procedure which is too stringent.

4. DISCUSSION

We have proposed a test based on both the spatial extent and the peak height of clusters in statistical images which under the null hypothesis are well approximated by a smooth Gaussian random field. The test uses an analytic approximation of the bivariate

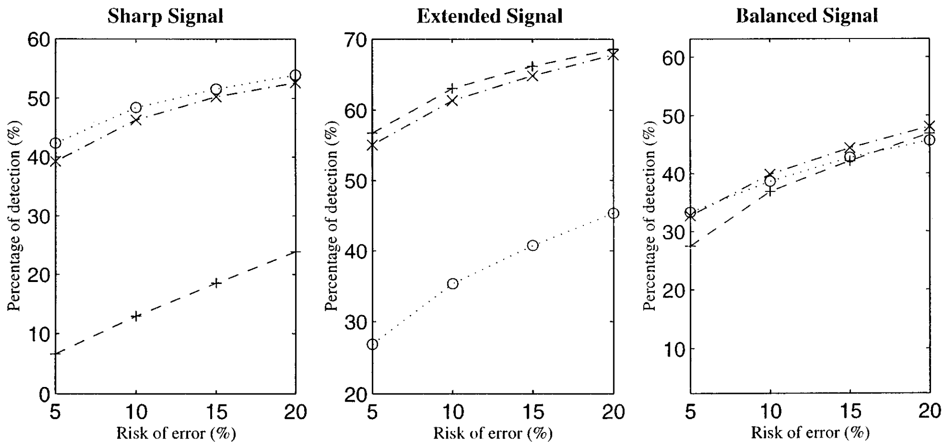


FIG. 7. Percentage of detected signal versus risk of error for three types of signal in 3 × 10^3 3D volumes (64 × 64 × 32 voxels or 128 × 128 × 64 mm^3). (Left) Sharp peak. (Middle) Extended signal. (Right) "Balanced" signal. The dashed line shows the results from the spatial extent test, the dotted line from the peak height test, and the dot and dashed line from the combined test. Volume resolution was $\text{FWHM}_{xy} = 14.1$ mm and threshold $t = 3$. See Table 1 for the characteristics of the signals.

probability distribution function of these two parameters. The test was designed to be as general as possible, being able to detect both high, sharp signals and broad, extensive signals. The test proposed in this work should provide greater sensitivity for a wider range of signals, with a controlled risk of error over the entire search region.

How do we now report the results of this test? Because it incorporates a spatial extent test, the risk of error associated with the combined test is at the cluster level rather than at the voxel level. This is the case for all spatial extent methods. We suggest that if the combined test detects a cluster via the spatial extent component, then the cluster as a whole is declared "significant." In other words, this test provides for a cluster-level inference.

Obviously, the higher the intensity threshold at which the clusters are defined, the greater the regional specificity of the inference, since the clusters will be smaller. We have seen that the test should be applied with thresholds greater than 2.5. However, Section 2.1 shows that the intensity threshold should be as low as possible for a Gaussian shaped signal (unless one wants to have a high regional specificity). We suggest the use of threshold around $t = 3$. The resolution of the map seems to have little influence on the risk of error (see Fig. 6). The only noticeable departure of the observed risk, compared to the predicted risk, was an underestimation for the peak intensity test and for all the three tests at low resolution and low risk. The restriction that the volume is well sampled compared to the resolution still holds, and this might be a limitation when analyzing raw (noninterpolated, nonfiltered) fMRI data.

We note that a high resolution is, in theory, not an issue. It is always possible to interpolate the data (with sinc or spline interpolation) to emulate a good sampling compared to the resolution. Obviously, this implies increasing the amount of data to process.

Unlike search over scale or Monte Carlo simulations, the computational cost of the combined test is very small. Using matlab (MathWorks Inc.), actual computation for a cluster is almost instantaneous on a Sun Spark 20 (Sun Inc.) workstation.

Sensitivity and the Shape of the Rejection Area: Isocumulative Curves versus Minimum Significance

Because we defined our rejection area in a bivariate space, as stated in the methods section, there is no unique definition for this area (there is no order relation in the bivariate space of size and height). In theory, one could tailor the test when a priori knowledge is available on the expected signal types. For instance, in fMRI, the spatial resolution is better and therefore the test could be adapted to detect spatially restricted signals or indeed to reject artifacts (e.g., isolated voxels with very high intensity values). The definition of the rejection area in this work should lead to a test slightly more sensitive to clusters with extreme values (very high intensity but small spatial extent or large spatial extent with relatively small heights) and less sensitive to regions with more "balanced" parameters) when compared to the definition used in previous work (see Fig. 2).

In summary the assumptions made when applying the combined test are (1) a good lattice approximation to a Gaussian field, (2) a large volume relative to the resolution, (3) a Gaussian autocorrelation function, and (4) a sufficiently high image threshold ($t \geq 25$). Although all these assumptions are made in deriving the test, in practice the combined test should be robust to violations of some of these assumptions. In particular because of the way in which the bivariate distribution was constructed (see Appendix), we expect some robustness with respect to the shape of the point spread function (PSF), even if in the case of PET data, the original scanner PSF and the Gaussian filter usually applied both tend to produce a PSF well approximated by a Gaussian function. Moreover, the theory only depends on the curvature of the autocorrelation function at zero lag, not on the whole function (cf. Eq. (9)). Note also that there is a variability introduced in the statistical results when the smoothness of the map is assessed (which is generally the case when analyzing statistical maps, (Poline *et al.*, 1995)).

The application of nonparametric statistics as developed by Holmes *et al.* (1996) is an attractive alternative when any or some of the above assumptions are violated. However the principle of the combined test can be retained. This would combine the freedom given by nonparametric approaches with the sensitivity of the combined test.

An alternative to the combined test would be to use a function of the spatial extent and the peak height, such as the "volume" (or excess mass) above a given threshold. This is equivalent to a rejection region with a hyperbolic shape (see Fig. 2), since, assuming a paraboloid model, volume is proportional to the product of spatial extent and height. This rejection area is similar to, but more extreme than, the iso-cumulative contours test of Poline and Mazoyer (1994b). This test might miss a cluster with large spatial extent if its peak height was low, while our combined test should detect such a cluster (see Fig. 2).

In a sense, the search over scale space has aims and features similar to those of the combined test. The common aim is to detect in a single procedure various kinds of signal (different sizes). Note that this implies occasional loss of sensitivity when the foci are better detected by one or the other test. The multifiltering approach assumes a search with isotropic kernels independent of the underlying anatomy. In other words, the combined test should be more sensitive to signals of large size with an irregular shape. These kind of signals may be expected given the complexity of the underlying anatomy.

5. CONCLUSION

Because it is essential to have both accurate control over the risk of error and good sensitivity over a variety of responses, we have proposed a test that addresses these two aspects. This test can be seen as an extension of the test based on spatial extent that is sensitive to sharp localized signals as well, and so it should have applications to statistical maps produced with functional MRI.

6. APPENDIX

As a first approximation, we model the shape of the image $Z(\mathbf{x})$ near a local maximum at $\mathbf{x}_0$ as an inverted paraboloid using a simple Taylor series expansion (Adler, 1981, Chap. 6):

$$Z(\mathbf{x}) \approx Z + (\mathbf{x} - \mathbf{x}_0)'\ddot{\mathbf{Z}}(\mathbf{x} - \mathbf{x}_0)/2,$$

where Z and $\ddot{\mathbf{Z}}$ are the values of $Z(\mathbf{x})$ and its second derivative evaluated at the local maximum. If $H = Z - t$ is the peak height above a threshold t, then the extent S above t, found using this model, is approximately

$$S \approx aH^{D/2}/|-\ddot{\mathbf{Z}}|^{1/2}, \tag{5}$$

where

$$a = \frac{(2\pi)^{D/2}}{\Gamma\left(\frac{D}{2} + 1\right)} \tag{6}$$

is the volume of a unit D-dimensional sphere, multiplied by $2^{D/2}$. It can be shown that H has approximately an exponential distribution with mean $1/t$ (see Adler, 1981, Chap. 6); it is necessary to find the distribution of $\ddot{\mathbf{Z}}$ conditional on H, or equivalently, conditional on Z. Let $\rho(\mathbf{x})$ be the correlation function of $Z(\mathbf{x})$ and let

$$\mathbf{\Lambda} = \mathrm{Var}\ (\dot{\mathbf{Z}}) = -\ddot{\rho}(\mathbf{0}).$$

Then it can be shown that Z and $\ddot{\mathbf{Z}}$ have a multivariate normal distribution with zero expectation and second moments

$$\mathrm{Var}\ (Z) = 1,$$

$$\mathrm{Cov}\ (Z, \ddot{\mathbf{Z}}) = -\mathbf{\Lambda},$$

$$\mathrm{Var}\ (\ddot{\mathbf{Z}}) = \rho^{(4)}(\mathbf{0}).$$

The distribution of $\ddot{\mathbf{Z}}$ conditional on $\mathbf{Z}$ is then multivariate Gaussian with moments

$$\mathrm{E}(\ddot{\mathbf{Z}}|Z) = -\mathbf{\Lambda}Z, \qquad \mathrm{Var}\ (\ddot{\mathbf{Z}}|Z) = \rho^{(4)}(\mathbf{0}) - \mathbf{\Lambda} \otimes \mathbf{\Lambda}.$$

We can therefore write, conditional on Z,

$$-\ddot{\mathbf{Z}} = \mathbf{\Lambda}Z + \Delta, \tag{7}$$

where Δ is multivariate normal with mean zero and variance (7). Since Δ is small relative to $\mathbf{\Lambda}Z$ then we can use the standard approximation for the determinant based on the first term in a Taylor series expansion:

$$\log|-\ddot{\mathbf{Z}}| \approx \log|\mathbf{\Lambda}| + D\log Z + \epsilon/(2Z). \qquad (8)$$

Where $\epsilon = \mathrm{tr}\,(\mathbf{\Lambda}^{-1}\Delta)$. The distribution of ϵ is multivariate Gaussian with expectation zero and variance (from Eq. (7)):

$$\sigma^2 = \mathrm{tr}\,((\mathbf{\Lambda}^{-1} \otimes \mathbf{\Lambda}^{-1})\rho^{(4)}(\mathbf{0})) - D^2. \qquad (9)$$

The simplest way of evaluating σ^2 is to transform the voxel coordinates $\mathbf{x}$ to $\mathbf{x}^* = \mathbf{\Lambda}^{-1/2}\mathbf{x}$. Defining the transformed correlation function $\rho^*(\mathbf{x}^*) = \rho(\mathbf{x})$, we get from Eq. (9)

$$\sigma^2 = \sum_{i=1}^{D}\sum_{j=1}^{D} \left.\frac{\partial^4\rho^*(\mathbf{x}^*)}{\partial x_i^{*2}\partial x_j^{*2}}\right|_{\mathbf{x}^*=\mathbf{0}} - D^2.$$

For the case of a Gaussian shaped correlation function

$$\rho(\mathbf{x}) = \exp\,(-\mathbf{x}'\mathbf{\Lambda}\mathbf{x}/2),$$

$$\rho^*(\mathbf{x}^*) = \exp\,(-\mathbf{x}^{*\prime}\mathbf{x}^*/2),$$

$$\left.\frac{\partial^4\rho^*(\mathbf{x}^*)}{\partial x_i^{*2}\partial x_j^{*2}}\right|_{\mathbf{x}^*=\mathbf{0}} = \begin{cases} 3 & i = j \\ 1 & i \neq j \end{cases},$$

$$\sigma^2 = 2D.$$

Putting Eqs. (5), (6), and (8) together, $\log(S)$ conditional on H can be modeled by

$$\log a - (1/2)\log|\mathbf{\Lambda}| - (D/2)\log Z + (D/2)\log H - \log(1 + \epsilon/2Z)$$

with ϵ normal. Since H is small relative to t then we can replace Z in the above mean and standard deviation by t.

Comparing the theoretical distribution and the one assessed using our simulations, we found that instead of approximating ϵ by a Gaussian distribution, it was better to approximate $1 + \epsilon/2Z$ by a multiple of a χ^2 distribution with degrees of freedom ν chosen so that $\mathrm{var}\,(1 + \epsilon/2Z) = \mathrm{var}\,(\chi^2/\nu) = 2/\nu$. This gives $\nu = 8t^2/\sigma^2 = 4t^2/D$ for the Gaussian correlation function, and the approximation

$$\nu a|\mathbf{\Lambda}|^{-1/2}t^{-D/2}H^{D/2}S^{-1} \sim \chi^2_\nu. \qquad (10)$$

Finally, we can adjust the constant a to ensure that the expected total region size above t agrees with the sum of the individual components S, a technique employed in Friston *et al.* (1994). The expected cluster size is the expected total region size divided by the expected number of clusters m from Eq. (1):

$$\begin{aligned} \mathrm{E}(S) &= V\Phi(t)/m \\ &= |\mathbf{\Lambda}|^{-1/2}t^{-(D-1)}(2\pi)^{(D+1)/2}e^{t^2/2}\Phi(t), \end{aligned} \qquad (11)$$

where V is the search volume and $\Phi(t)$ is the P value of a voxel above t:

$$\Phi(t) = \int_t^\infty e^{-z^2/2}\Big/\sqrt{2\pi}\,dz.$$

From our direct derivation (10) and the fact that the marginal distribution of H is exponential with mean $1/t$, we get

$$\mathrm{E}(S) = a|\mathbf{\Lambda}|^{-1/2}t^{-D/2}\mathrm{E}(H^{D/2}) = |\mathbf{\Lambda}|^{-1/2}t^{-D}(2\pi)^{D/2}.$$

Therefore we can correct our theoretical approximation such that the expected region size matches Eq. (11). The correction factor is

$$c = \sqrt{2\pi}\,te^{t^2/2}\Phi(t), \qquad (12)$$

which approaches one as t tends to infinity. This suggests the more accurate approximation

$$\nu ac|\mathbf{\Lambda}|^{-1/2}t^{-D/2}H^{D/2}S^{-1} \sim \chi^2_\nu, \qquad (13)$$

in which a is multiplied by c. Integrating this over the marginal distribution of H, we get:

$$\begin{aligned} &\mathrm{P}(S \geq s_0, H \geq h_0) \\ &\quad \approx \int_{h=h_0}^\infty \Psi_\nu\{\nu ac|\mathbf{\Lambda}|^{-1/2}t^{-D/2}h^{D/2}/s_0\}te^{-th}dh, \end{aligned} \qquad (14)$$

where Ψ_ν is one minus the χ^2 distribution function with degrees of freedom $\nu = 4t^2/D$, given by

$$\Psi_\nu(x) = \int_x^\infty \frac{u^{\nu/2-1}e^{-u/2}}{2^{\nu/2}\Gamma(\nu/2)}\,du.$$

ACKNOWLEDGMENT

J.B.P. was funded by an European Union grant, Human Capital and Mobility, No. ERB4001GT932036. K.J.W. was supported by the Natural Sciences and Engineering Research Council of Canada and the Fonds pour la Formation des Chercheurs et l'Aide à la Recherche de Québec. K.J.F. was funded by the Wellcome Trust. We are very grateful to Rik Vandenberghe and his colleagues for providing the PET data.

REFERENCES

Adler, R. J. 1981. *The Geometry of Random Fields,* Wiley, New York.

Friston, K. J., Frith, C. D., Liddle, P. F., and Frackowiak, R. S. J.

1991. Comparing functional (PET) images: The assessment of significant change. *J. Cereb. Blood Flow Metab.* **11:**690–699.

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

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. B., Frith, C. D., and Frackowiak, R. S. J. 1995. Statistical parametric maps in functional imaging: A general approach. *Hum. Brain Map.* **2:**189–210.

Friston, K. J., Ashburner, J., Frith, C. D., Poline, J.-B., Heather, J. D., and Frackowiak, R. S. J. 1995. Spatial registration and normalization of images. *Hum. Brain Map.* **2:**165–189.

Holmes, A. P., Blair, R. C., Watson, J. D. G., and Ford, I. 1996. Non-parametric analysis of statistic images from functional mapping experiments. *J. Cereb. Blood Flow Metab.* (in press).

Poline, J.-B., and Mazoyer, B. M. 1992. Cluster analysis of individual PET activation maps. *Proc. IEEE Med. Imag. Conf. Orlando.*

Poline, J.-B., and Mazoyer, B. M. 1993. Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise ratio pixel clusters. *J. Cereb. Blood Flow Metab.* **13:**425–437.

Poline, J.-B., and Mazoyer, B. M. 1994a. Enhanced detection in brain activation maps using a multifiltering approach. *J. Cereb. Blood Flow Metab.* **14:**639–641.

Poline, J.-B., and Mazoyer, B. M. 1994b. Analysis of individual brain activation maps using hierarchical description and multi-scale detection. *IEEE Tran. Med. Imag.* **13:**702–710.

Poline, J.-B., Worsley, K. J., Holmes, A. P., Frackowiak, R. S. J., and Friston, K. J. 1995. Estimating smoothness in statistical parametric maps: Variability of P values. *J. Comp. Assist. Tomogr.* **19**(5):788–796.

Roland, P. E., Levin, B., Kawashima, R., and Åkerman, S. 1993. Three-dimensional analysis of clustered voxels in ^{15}O-butanol brain activation images. *Hum. Brain Map.* **1:**3–19.

Siegmund, D. O., and Worsley, K. J. 1995. Testing for a signal with unknown location and scale in a stationary gaussian random field. *Ann. Stat.* **23:**608–639.

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

Worsley, K. J., Marrett, S., Neelin, P., and Evans, A. C. 1996. Searching scale space for activation in PET images. *Hum. Brain Map.* (in press).