NEUROIMAGE **40,** 223–235 (1996)
ARTICLE NO. 0074

Detecting Activations in PET and fMRI: Levels of Inference and Power

K. J. FRISTON, A. HOLMES, J-B. POLINE, C. J. PRICE, AND C. D. FRITH

The Wellcome Department of Cognitive Neurology, Institute of Neurology, Queen Square, London WC1N 3BG United Kingdom

Received May 8, 1996

This paper is about detecting activations in statistical parametric maps and considers the relative sensitivity of a nested hierarchy of tests that we have framed in terms of the level of inference (voxel level, cluster level, and set level). These tests are based on the probability of obtaining *c*, or more, clusters with *k*, or more, voxels, above a threshold *u*. This probability has a reasonably simple form and is derived using distributional approximations from the theory of Gaussian fields. The most important contribution of this work is the notion of *set-level inference*. Set-level inference refers to the statistical inference that the number of clusters comprising an observed activation profile is highly unlikely to have occurred by chance. This inference pertains to the set of activations reaching criteria and represents a new way of assigning *P* values to distributed effects. Cluster-level inferences are a special case of set-level inferences, which obtain when the number of clusters $c = 1$. Similarly voxel-level inferences are special cases of cluster-level inferences that result when the cluster can be very small (i.e., $k = 0$). Using a theoretical power analysis of distributed activations, we observed that set-level inferences are generally more powerful than cluster-level inferences and that cluster-level inferences are generally more powerful than voxel-level inferences. The price paid for this increased sensitivity is reduced localizing power: Voxel-level tests permit individual voxels to be identified as significant, whereas cluster- and set-level inferences only allow clusters or sets of clusters to be so identified. For all levels of inference the spatial size of the underlying signal *f* (relative to resolution) determines the most powerful thresholds to adopt. For set-level inferences if *f* is large (e.g., fMRI) then the optimum extent threshold should be greater than the expected number of voxels for each cluster. If *f* is small (e.g., PET) the extent threshold should be small. We envisage that set-level inferences will find a role in making statistical inferences about distributed activations, particularly in fMRI.

INTRODUCTION

This paper concerns statistical inference about activation profiles in functional neuroimaging, particularly functional magnetic resonance imaging (fMRI). We present a taxonomy of tests that pertain to different levels of inference for an activation profile, namely, a voxel, a cluster of voxels, and a set of clusters. We then consider the relative sensitivity of the ensuing tests in terms of power and how that power varies as a function of resolution and the nature of the underlying signal. This paper is concerned primarily with distributed signals that have no *a priori* anatomical specification.

Activations in position emission tomography (PET) and fMRI are almost universally detected using some form of statistical mapping. The statistical processes that ensue (i.e., statistical parametric maps or SPMs) are usually characterized in terms of regional excursions above some threshold and a *P* value is assigned to these excursions. In this paper we reexamine the nature of this statistical inference and ask "what is the most powerful way to proceed?" with special reference to high-resolution data, such as that obtained with fMRI. The problem addressed, by the more advanced approaches, is that of the multiple *dependent* comparisons embodied in the SPM. This dependency (of one voxel's value on its neighbors) is a result of smoothness or autocorrelations in the data (or more strictly the error terms). Put simply, the statistic at any voxel can be used as the basis of statistical inference if, and only if, that particular region was predicted in advance. However, if the alternate hypothesis was anatomically open (i.e., activations have occurred somewhere), then we need to assess the statistics at all voxels individually and simultaneously, while ensuring that the probability of a false positive is less than the specified test level α. In statistical terminology this is a multiple comparisons problem. At each voxel we have a null hypothesis of no activation. Considering all voxels together we have a family of hypotheses and we wish to assess the omnibus hypothesis that all the voxel (null) hypotheses are true, while controlling the probability of

1053-8119/96 $18.00

a false positive or a type 1 error. There two forms of control over family-wise type 1 error (FWE), weak and strong, which determine the level at which departures from the omnibus hypothesis can be reported (see Holmes, 1994, for a fuller discussion).

A test procedure controls FWE in the weak sense if the probability of false rejection of the omnibus hypothesis is less than α. A procedure with only weak control has no "localizing power." If the null hypothesis is rejected then all that can be said is that there is some departure from the null hypothesis at some voxel(s). Here the "level of inference" is the whole volume analyzed. A procedure controls FWE in the strong sense (at the voxel level) if the probability of a false positive over any set of voxels, for which the null hypothesis is true, is less than α, regardless of the truth of the null hypothesis elsewhere. This more stringent criterion gives localizing power: Voxels identified by such a procedure may be declared individually significant and the inference is at the voxel level.

The simplest multiple comparisons procedure which maintains strong control over FWE is based on the Bonferroni inequality: Here the voxel-level *P* values are corrected for the number of voxels. However, for even mild dependencies between the voxels, this method is excessively conservative. This is important because, even in fMRI data, physiological autocorrelations render nearby voxels correlated and a simple Bonferroni correction becomes inappropriate. The most successful solutions to the problem of statistical inference in smooth spatially extended statistical processes are predicated on the theory of Gaussian fields. Early work was based on the theory of level crossings (Friston *et al.*, 1991) and differential topology (Worsley *et al.*, 1992). These approaches control FWE strongly, allowing for inference at the voxel level: A corrected *P* value is assigned to a voxel using the probability that the observed voxel value, or a higher one, could have occurred by chance in the volume analyzed. There have been a number of interesting elaborations at this level of inference [e.g., searching scale-space and other high-dimensional SPMs (e.g., Siegmund and Worsley 1994)] and results for many statistics now exist (e.g., Worsley, 1994). The next development, using the theory of Gaussian fields, was to use the spatial extent of a cluster of voxels defined by a height threshold (Friston *et al.*, 1994; see also Poline and Mazoyer, 1993, and Roland *et al.*, 1993). These procedures control FWE strongly at the cluster level, permitting statistical inference about each cluster, and are based on the probability of getting a cluster of the size observed (defined by a height threshold), or a larger one, in the volume analyzed. In this paper we introduce a third level, namely, the set level, which is based on the probability of getting the observed number of clusters (defined by a height threshold and an extent threshold), or more, in the volume analyzed. This inference is about the set of clusters (contiguous regions above some height and size thresholds) or more simply about the excursion *set*. Because there is only one "set" there is no multiple comparison problem at this level of inference. The objective of this work was to compare the relative power of these different levels of inference under different conditions. In the sense that all these inferences are based on corrected *P* values we consider only the case where no *a priori* knowledge about the anatomy of the activations is available. We reiterate that if the activated region is predicted in advance the use of the above "corrected" *P* values is unnecessary and inappropriately conservative. In this instance we would recommend a simple Bonferroni correction for the number of regions predicted.

The Theory of Gaussian Fields

The results mentioned above, and described more fully below, derive from the theory of Gaussian fields. The assumptions implicit in this approach are (i) that the SPMs are reasonable lattice representations of underlying continuous Gaussian fields, (ii) that the components of the fields have a multivariate Gaussian distribution and are wide sense stationary, and (iii) that the height thresholds employed are high. Wide sense stationary simply means that the multivariate probability distributions of nearby points do not change with position in the field. This implies that spatial autocorrelations are invariant across the field. These are reasonable assumptions in neuroimaging as long as the voxel size is small relative to the smoothness. There has been some interest in revising spatial extent approaches in the context of fMRI (where the voxel sizes are larger in relation to resolution) using Monte Carlo simulations and adjustments to the smoothness estimators (e.g., Forman *et al.*, 1995). Usual estimates of smoothness (e.g., Friston *et al.*, 1991; Worsley *et al.*, 1992) fail when the reasonable lattice assumption is violated. In our work we sidestep this issue by simply interpolating the data to reduce voxel size or smoothing the data to increase smoothness. It is generally accepted that the voxel size should be about half the full-width half maximum (FWHM) of the smoothness, or less, for the reasonable lattice assumption to hold.

In statistical parametric mapping there is a caveat in reference to the multivariate Gaussian assumptions. The expressions presented below pertain to Gaussian fields (as opposed to fields composed of statistics other than *Z*). In many instances, the alternative hypothesis is tested with a parametric map of the *t* statistic (i.e., SPM$\{t\}$). The SPM$\{t\}$ is usually transformed to a Gaussian field (i.e., SPM$\{Z\}$) using some suitable transformation. This "Gaussianized" SPM$\{t\}$ only approximates a true SPM$\{Z\}$ if the (effective) degrees freedom of the underlying *t* statistic are reasonably high. Fortuitously,

this is usually assured in single-subject fMRI and multisubject PET studies. The good lattice and Gaussian assumptions can be further ensured by slight spatial smoothing of the data, which, in addition, usually increases the sensitivity of the ensuing analysis.

The high threshold requirement stems from the fact that many of the distributional approximations used are asymptotic and are only exact as thresholds tend to infinity. In practice this means that the use of low thresholds (e.g., $u < 1.64$) should be avoided (or at least validated using simulations).

Strong vs Weak Control, Levels of Inference, and Regional Specificity

There is a fundamental difference between rejecting the null hypothesis of no activation at a particular voxel and rejecting the null hypothesis over the entire volume analyzed. As noted above the former requires the strongest control over FWE and the latter the weakest. One way of thinking about this difference is to note that if an activation is confirmed at a particular point in the brain then, implicitly, the hypothesis of an activation somewhere is also confirmed (but the converse is not true). The distinction between weak and strong control, in the context of statistical parametric mapping, relates to the level at which the inference is made. The stronger the control, the more regional specificity it confers. For example a voxel-level inference is stronger than a cluster-level inference because the latter disallows inferences about any component voxel. In other words cluster-level inferences maintain strong control at the cluster level but only weak control at the voxel level. Similarly set-level inferences are weaker than cluster-level inferences because they refer to the set of regional activations but not any individual region or cluster in that set. Procedures with the weakest control over FWE assess only the omnibus hypothesis and have been referred to as "omnibus" tests (e.g., the γ_2 test, Fox *et al.*, 1989) and frame the alternative hypothesis in terms of voxel-level effects anywhere in the brain. These hypothesis are usually tested using all the voxels above some threshold (e.g., exceedence proportion tests, Friston *et al.*, 1991) or use all the voxel values (e.g., quadratic tests, Worsley *et al.*, 1995). A weaker control over FWE, or high-level inference, has less regional specificity but remains a valid way of establishing the significance of an activation profile. Intuitively one might guess that the weaker procedures provide more powerful tests because there is a trade-off between sensitivity and regional specificity. This is what we found using a power analysis (see below). In this paper we focus on the weaker hypotheses and consider voxel-level and cluster-level inferences subordinate to set-level inferences. This allows us to ask which is the most powerful approach for detecting brain activations.

This paper is divided into three sections. The first section describes the distributional approximations used to make statistical inferences about a SPM and frames the results to show that all levels of inference can be regarded as nested, special cases of a single general probability (namely, the probability of getting c, or more, clusters with k, or more, voxels above height u, in a volume S of smoothness W). The second section describes the details and rationale behind the power analysis employed in the subsequent section. The final section deals with the relative power of voxel-level, cluster-level, and set-level inferences and its dependency on signal characteristics, namely, the spatial extent of the underlying hemodynamics and the signal to noise ratio.

THEORY AND DISTRIBUTIONAL APPROXIMATIONS

In this section we introduce the basic results from the theory of Gaussian fields that are used to provide a general expression for the probability of getting any set of clustered voxels. We then show that voxel-level, cluster-level, and set-level inferences are all special cases of this general formation and introduce some special cases that have not been considered before.

The General Expression

In what follows we assume that a D-dimensional SPM conforms to a reasonable lattice representation of a Gaussian field of volume S voxels and smoothness W. For a Gaussian random field W is related to the FWHM of the resolution [$W = \mathrm{FWHM}(4 \ln 2)^{-1/2}$]. Equivalently $W = |\Lambda|^{-1/(2D)}$, where Λ is the covariance matrix of the field's first partial derivatives. An excursion set is defined as the set of voxels that exceeds some threshold u. This excursion set comprises m clusters each with n voxels. At high thresholds m approximates the number of maxima and has been shown to have a Poisson distribution (Adler, 1981, Theorem 6.9.3, page 161).

$$P(m = c) \approx \lambda(c, E\{m\}) = 1/c!E\{m\}^c e^{-E\{m\}}, \qquad (1)$$

where $E(m)$ is the expected number of maxima (i.e., clusters), where (Hasofer 1978):

$$E\{m\} \approx S(2\pi)^{-(D+1)/2} W^{-D} u^{D-1} \exp(-u^2/2). \qquad (2)$$

The number of voxels n comprising a cluster is distributed according to:

$$P(n \geq k) \approx \exp(-\beta k^{2/D}),$$

where

$$\beta = [\Gamma(D/2 + 1).E\{m\}/(S \cdot \Phi\{-u\}))]^{2/D} \qquad (3)$$

(Friston *et al.*, 1994). In this formulation n is a continuous variable measuring the volume of a cluster in voxels, $\Phi\{-u\}$ is the cumulative density function for the unit Gaussian distribution, and Γ denotes the gamma function. With these results it is possible to construct an expression for the probability of getting c, or more, clusters of size k, or more, above a threshold u:

$$P_W(u, k, c) \approx 1 - \sum_{i=0}^{c-1} \sum_{j=i}^{\infty} P(m = j). \binom{j}{i} .P(n \geq k)^i .P(n < k)^{j-i}$$

$$\approx 1 - \sum_{i=0}^{c-1} \lambda(i, E\{m\}.P(n \geq k)). \quad (4)$$

This expression assumes that the m and n are independent (i.e., the number of clusters in a volume and their size are independent), which in turn depends on the stationariness assumption. A full derivation of this equation can be found in the appendix. Eq. (4) can be interpreted, in an intuitive sense, in the following way: Consider clusters as "rare events" that occur in a volume according to the Poisson distribution with expectation $E\{m\}$. Now the proportion of these rare events that meet the spatial extent criterion will be $P(n \geq k)$. These criterion events will themselves occur according to a Poisson distribution with expectation $E\{m\} \cdot P(n \geq k)$. The probability that the number of events will be c or more is simply 1 minus the probability that the number of events lies between 0 and $c - 1$ [i.e., the sum in Eq. (4)].

We now consider various ways in which Eq. (4) can be used to make inferences about brain activations. In brief, if the number of clusters $c = 1$ then the probability reduces to that of getting one, or more, clusters with k, or more, voxels. This probability can be used to estimate the P value for a single cluster of volume k. This corresponds to a cluster-level inference. Similarly if $c = 1$ and the number of suprathreshold voxels $k = 0$, the resulting cluster-level probability (i.e., the probability of getting one or more excursions of any volume above u) can be applied at the voxel level. In other words the existing tests are special (limiting) cases of set-level inferences. Note that k is a continuous measure of the volume; although k is expressed in units of voxels, it pertains to a continuous Gaussian field.

Voxel-Level Inferences

Consider the situation in which the threshold u is the statistic upon which we base our inference. In this instance, the size k and the number of clusters c can only take the values 0 and 1 (i.e., there is at least one point at or above threshold with an unspecified size). The corresponding probability $P_W(u, 0, 1)$ is the corrected P value for the voxel in question where:

$$P_W(u, 0, 1) \approx 1 - \exp(-E\{m\}) \quad (5)$$

by Eq. (1) and because $P(n > 0) = 1$. This, of course, is simply the corrected probability based on the expected number of maxima as employed in the early days of statistical mapping in PET (e.g., Friston *et al.*, 1991) and elaborated using the expected Euler characteristic as an alternative to $E\{m\}$ (Worsley *et al.*, 1992).

Cluster-Level Inferences

Consider now the case in which we choose to base our inference on spatial extent k. k is defined only by specifying a height threshold U. In this instance c can only take the value 1 (i.e., there is at least one cluster of size k or more). The corresponding probability $P_W(U, k, 1)$ is the corrected P value based on spatial extent (Friston *et al.*, 1994):

$$P_W(U, k, 1) \approx 1 - \exp(-E\{m\}.P(n \geq k)) \quad (6)$$

and has proved to be more powerful than voxel-based inference when applied to high-resolution data (see below).

Set-Level Inferences

Now consider the instance where inference is based on cluster number c. In this case both height U and extent K threshold need to be specified before the statistic c is defined. The corresponding probability $P_W(U, K, c)$ is given by Eq. (4) and is the corrected P value for the set of activation foci reaching these joint criteria. This level of inference has not been routinely used in neuroimaging and one aim of this paper is to evaluate its potential power. Current tests are based on cluster-level inferences (i.e., $c = 1$). Therefore, we want to know whether the power of the analysis can be enhanced by allowing c to be greater than 1 and, if so, what are the best thresholds to adopt. This is the key question addressed by the power analyses below.

A Special Case of Set-Level Inferences—An Omnibus Test

There is a conceptual relationship between set-level inferences and nonlocalizing tests based on the exceedence proportion (i.e., the total number of voxels above a threshold U). Exceedence proportion tests (e.g., Friston *et al.*, 1991) and thresholdless quadratic tests (Worsley *et al.*, 1995) have been proposed that test for activation effects over the volume analyzed in a spatially omnibus sense. These tests have not been widely used because they have no localizing power and do not pertain to a set of well-defined activations. In this sense

these tests differ from set-level tests because the latter do refer to a well-defined set of activation foci. However, in the limiting case of a small spatial extent threshold k the set-level inference approaches an omnibus test:

$$P_W(U, 0, c) \approx 1 - \sum_{i=0}^{c-1} \lambda(i, E[m]). \quad (7)$$

This test simply compares the expected and observed number of maxima (or clusters at high thresholds) in an SPM using the known Poisson distribution under the null hypothesis. This is a remarkably simple test that could be considered alongside existing omnibus tests (subject to the usual constraint of a relatively high height threshold).

A POWER ANALYSIS

In this section we describe the model upon which the power analysis was based and how we derived expressions for power or sensitivity. The specificity of a test is defined as the probability of correctly rejecting the null hypothesis (i.e., the probability of a true negative $= 1 - \alpha$, where α corresponds to the probability of a false positive). The sensitivity of a test is, conversely, the probability of correctly accepting the alternative hypothesis for a given specificity. A plot of sensitivity against α corresponds to a receiver operator characteristic (ROC) curve. Examples of these curves will be provided below. In order to determine the power of different tests analytically it is necessary to define the exact nature of the signal implied by the alternative hypothesis. In this paper we consider a simple model (first presented in Friston *et al.*, 1994) which assumes that the activations are spatially distributed with no predilection for one anatomical area or another. Although this model is used for mathematical convenience it is not physiologically unreasonable and embodies an ignorance of where the activations will be found. More specifically, it models activations that are distributed throughout the volume and the power analysis below applies to this, and only this, model. Different models (i.e., a single activation focus) would yield different results. In this paper we focus on a "distributed" model, where we expect set-level inferences to be more sensitive.

Suppose the "signal" comprises Gaussian foci, of random height, distributed continuously throughout the volume. The shape of the signal is characterized by the width (f) of these foci expressed as a proportion of W. This signal can be modeled by a continuous ensemble of kernels with randomly distributed heights or equivalently by convolving an uncorrelated Gaussian random process with a kernel of the same height. Let the signal (following convolution with the point spread function) have a standard deviation σ, where σ corresponds to the amplitude of the measured signal. Following Friston *et al.* (1994) the threshold and smoothness for the process under the alternative hypothesis are

$$u^* = u.(1 + \sigma^2)^{-1/2} \quad (8)$$
$$W^* = W.[(1 + \sigma^2)/(1 + \sigma^2/(1 + f^2))]^{1/2}.$$

The specificity of any test based on u, k, and c is simply 1 minus the probability α of rejecting the null hypothesis when the null hypothesis is correct, where

$$\alpha = P_W(u, k, c). \quad (9)$$

The sensitivity (or power) $\gamma(\alpha)$, for a given specificity $(1 - \alpha)$, is the probability of accepting the alternative hypothesis when it is correct. Under the alternative hypothesis

$$\gamma(\alpha) = P_{W^*}(u^*, k, c). \quad (10)$$

By varying u, k, or c we obtain sensitivity as a function of α. This is the ROC curve. The sensitivity at a particular specificity [e.g., $\gamma(0.05)$], or the area under this curve, indexes the relative power of the test under consideration. In what follows we will compare the power of voxel-, cluster-, and set-level inferences for signals of different sizes to identify the most powerful sorts of inference.

Voxel-Level Inferences

In this instance the only parameter that can be varied is the threshold u (i.e., $k = 0$ and $c = 1$): An example of an ROC curve for voxel-level inferences is seen in Fig. 1 (top). Unless otherwise stated, in this and subsequent examples; $\sigma = 0.3$, $f = 2$, $S = 64^D$, $D = 3$, and W corresponded to an isotropic smoothness of 3 voxels at FWHM. The influence of signal parameters on power is shown in the lower panel by plotting $\gamma(0.05)$ as a function of amplitude σ and size f. It can be seen that power is a strong function of signal amplitude σ for all sizes of signal. It is also evident that higher thresholds are slightly more powerful, when the signals are smaller than the resolution ($f < 1$). High-resolution fMRI and optical imaging data suggest that hemodynamic changes are typically expressed on a spatial scale of 5–8 mm. This is around or below the resolution of PET (especially when the data are smoothed); however, it is greater than the resolution of fMRI data, which, before any interpolation or smoothing, can be equated with voxel size (e.g., 3 mm). In other words in PET $f < 1$ and in fMRI $f > 1$. This distinction is important and has profound implications for the sensitivity of tests considered below.

From the point of view of voxel-level inferences, power increases as f decreases. Therefore, power can be

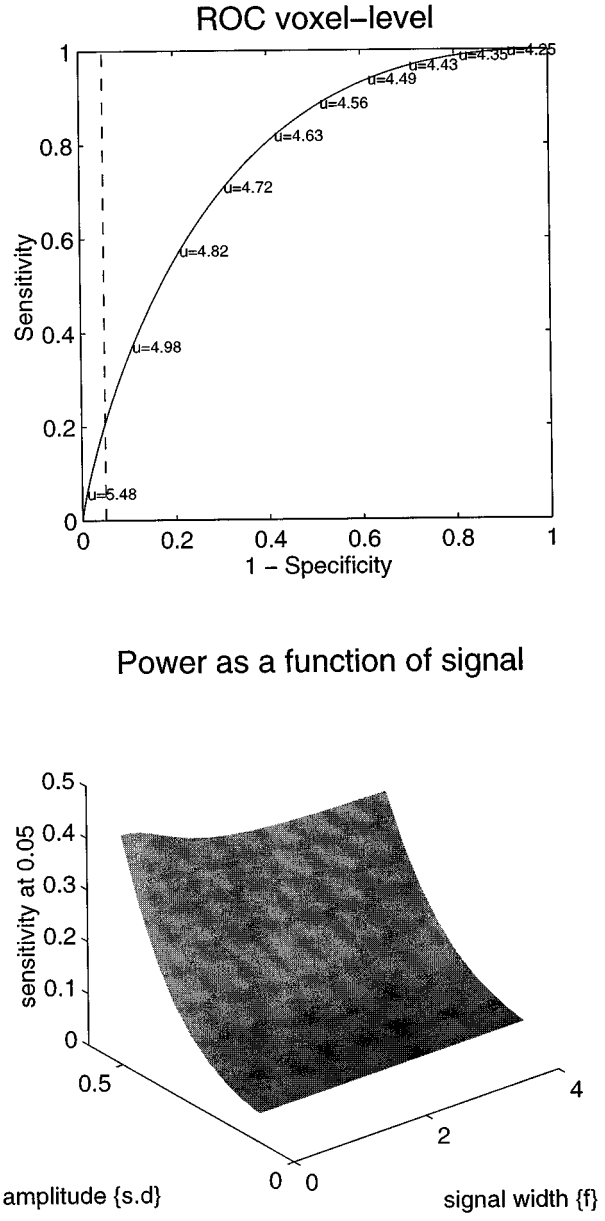


FIG. 1. (Top) ROC curve for voxel-level inference $P_W(u, 0, 1)$, where W corresponds to a FWHM of 3 voxels and the volume $S = 64^3$. Signal amplitude $\sigma = 0.3$ and width $f = 2$. The dotted line corresponds to $\alpha = 0.05$. In this, and subsequent, ROC curves, values of the threshold used to vary specificity are provided on the curves. (Bottom) Three-dimensional plot of power ($\alpha = 0.05$) as a function of signal amplitude σ and width f for the same smoothness and volume.

increased by smoothing the data (i.e., increasing W or, equivalently, decreasing f). This is a fundamental observation and it is worth noting that high degrees of smoothing have been employed in PET for many years. *It is important to note that this also holds for fMRI* when (and only when) using voxel-level inferences. This argument discounts partial volume effects in the sense that the model adopted here, for the alternate hypothesis, does not include the impact on signal amplitude σ of smoothing to decrease f. This impact can be marked if the signals are small in relation to the smoothing kernel employed.

Cluster-Level Inferences

In this section we reiterate a previous observation (made in Friston *et al.*, 1994) that cluster-level inferences are generally more powerful than voxel-level inferences, with the exception that voxel-level inferences are more powerful when $f < 1$ (as is often the case for PET). It is useful to remember that voxel-level inferences are a special case of cluster-level inferences that obtain when $k = 0$. Figure 2 (top) shows the ROC curves for a cluster-level inferences at threshold $U = 2.8$ (solid line). This curve was calculated by varying the cluster threshold k in Eqs. 9 and 10 with $u = U$. The equivalent ROC curve from the previous analysis is also shown (broken line). The lower panel of Fig. 2 demonstrates the effect of different signal sizes (for a fixed amplitude of $\sigma = 0.3$). This represents a plot of $\gamma(\alpha)$ as a function of u and f. It is immediately obvious that for small signals (i.e., low resolution) the most powerful tests obtain when the threshold is high and k tends to 0. A vanishingly small k corresponds to $P_W(u, 0, 1)$, a voxel-level inference. Conversely, when $f > 1$ the more powerful tests are associated with a low height threshold u and implicitly a high extent threshold k. In practical terms these result suggest that voxel-level inferences are best in PET and cluster-level inferences should supervene in fMRI. This conclusion appears to be consistent with the anecdotal experience of those developing fMRI analysis strategies.

Set-Level Inferences

In this section we observe that set-level inferences are generally more powerful than cluster-level inferences and that this holds irrespective of relative signal size f. We then proceed to ask what is the optimum extent threshold for set-level inferences. In brief, we observe that for $f < 1$, small values of k are most powerful and conversely when $f > 1$, larger values are appropriate.

Figure 3 shows the ROC curve that obtains by varying the number of clusters c for a fixed threshold $U = 2.8$ and an extent threshold $K = 16$ (the remaining parameters are the same as those in the preceding sections). It can be seen that the set-level inference (solid line) is much more powerful than either the cluster-level (dashed line) or voxel-level (broken line) inferences. To determine whether there are any special cases ($c = 1$) of the set-level test (i.e., cluster or voxel level) that are more powerful than the general case ($c > 1$) we computed $\gamma(0.05)$ by allowing k to vary for different values of u and c. The lower panels of Fig. 3 shows the result of this analysis and demonstrate that

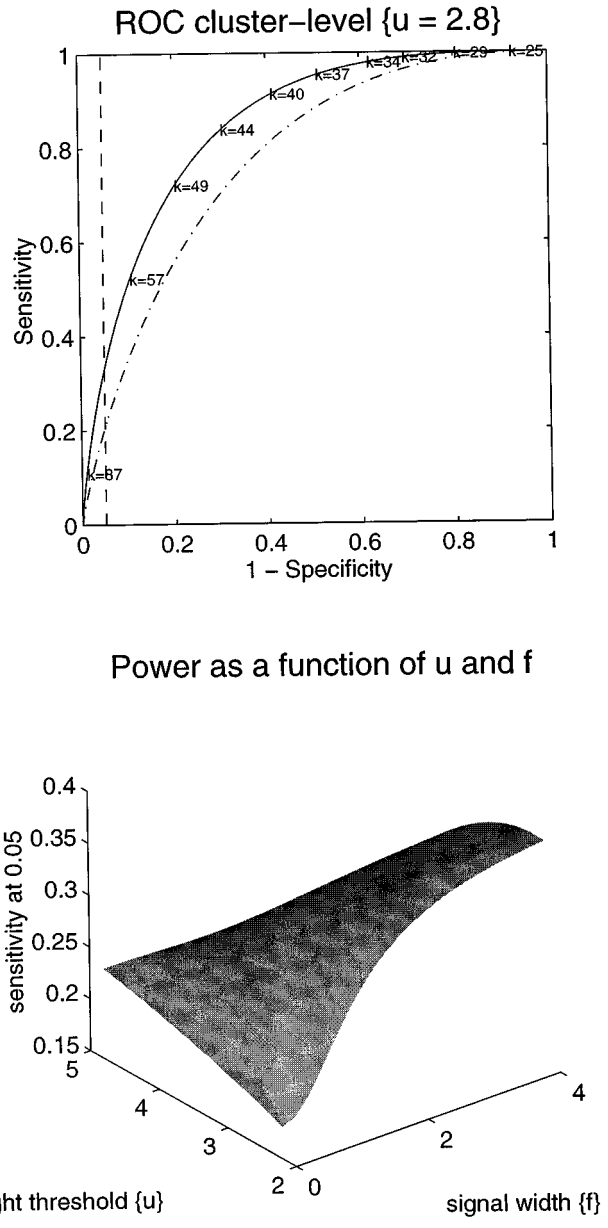


FIG. 2. (Top) ROC curve for cluster-level inference $P_W(2.8, k, 1)$, where W corresponds to a FWHM of 3 voxels and the volume $S = 64^3$. Signal amplitude $\sigma = 0.3$ and width $f = 2$. The broken line corresponds to the equivalent voxel-level ROC curve of the previous figure. (Bottom) Three-dimensional plot of power ($\alpha = 0.05$) as a function of signal width f and threshold u, for the same smoothness, volume, and σ.

the most powerful tests result when $c > 1$ (i.e., set level) over the range of thresholds u (and implicitly k) employed. This is the case for both low- and high-resolution data (left and right lower panels, respectively). It can be seen that when $c = 1$ the best threshold is a high one when $f < 1$ and a low one when $f > 1$. This is the observation made in the previous section. The main conclusion here however is that set-level inferences ($c > 1$) are generally the most powerful for the model of signal adopted in this analysis.

It now remains to identify the thresholds that give the most powerful set-level inference for distributed signals. For this analysis we have used A (the area under the ROC curve) for a range of parameters (see the top panel of Fig. 4). Intuitively, for a fixed height threshold and $f > 1$, one might conjecture that the extent threshold should be designed to filter out clusters that are due to noise, but should be low enough to catch all the activations. If, on the other hand, the signal size is less than the resolution ($f < 1$) it is not possible to discriminate between foci due to signal and noise and the most powerful approach is that with the weakest control (i.e., an omnibus test based on the total number of maxima when $k = 0$). Roughly speaking, this is what we found. The lower panel in Fig. 4 shows A as a function of f and k (for a threshold $u = 2.8$). The solid vertical line (in white) is the expected number of voxels per cluster $E\{n\}$ under the null hypothesis [$E\{n\} = \Gamma(D/2 + 1) \cdot \beta^{-D/2}$, see Friston *et al.* (1994)]. It can be seen that when the signal is greater than the resolution ($f > 1$) the most powerful set-level inferences obtain with extent thresholds that are about the same size as, or larger than, the expected cluster size due to noise ($E\{n\}$). Conversely, when the signal is smaller than the resolution ($f < 1$) the most powerful extent threshold is small. More simply, the power is maximized when the extent threshold matches, roughly, the size of the activation. These results suggest that for PET ($f < 1$) a small extent threshold is most appropriate, whereas for fMRI an extent threshold of at least $E\{n\}$ should be used. This conclusion was verified by repeating this analysis for a range of signal amplitudes and height threshold (results not shown). We do not consider a more exhaustive characterization of the optimum threshold in this paper because (i) in real life signals may vary greatly in their spatial extent (i.e., there is no single optimum) and (ii) regional specificity should not always be subordinated to sensitivity. For example, it may be more biologically meaningful to make a less powerful inference about a small set of large activation foci that have some clear relationship to anatomical or functional brain systems. This consideration relates to the interpretation of the set of clusters.

The Interpretability of Set-Level Inferences

Are set-level inferences really useful? Clearly if one obtained 6 clusters with a P value of less than 0.05, and 4 clusters would have been expected by chance (i.e., $E\{m\} = 4$), then it would be inappropriate to infer that the clusters observed represented a distributed effect. In this instance the set-level inference can only be used in an omnibus sense (i.e., the stimulation was sufficient to activate the brain in some way, but we cannot impute

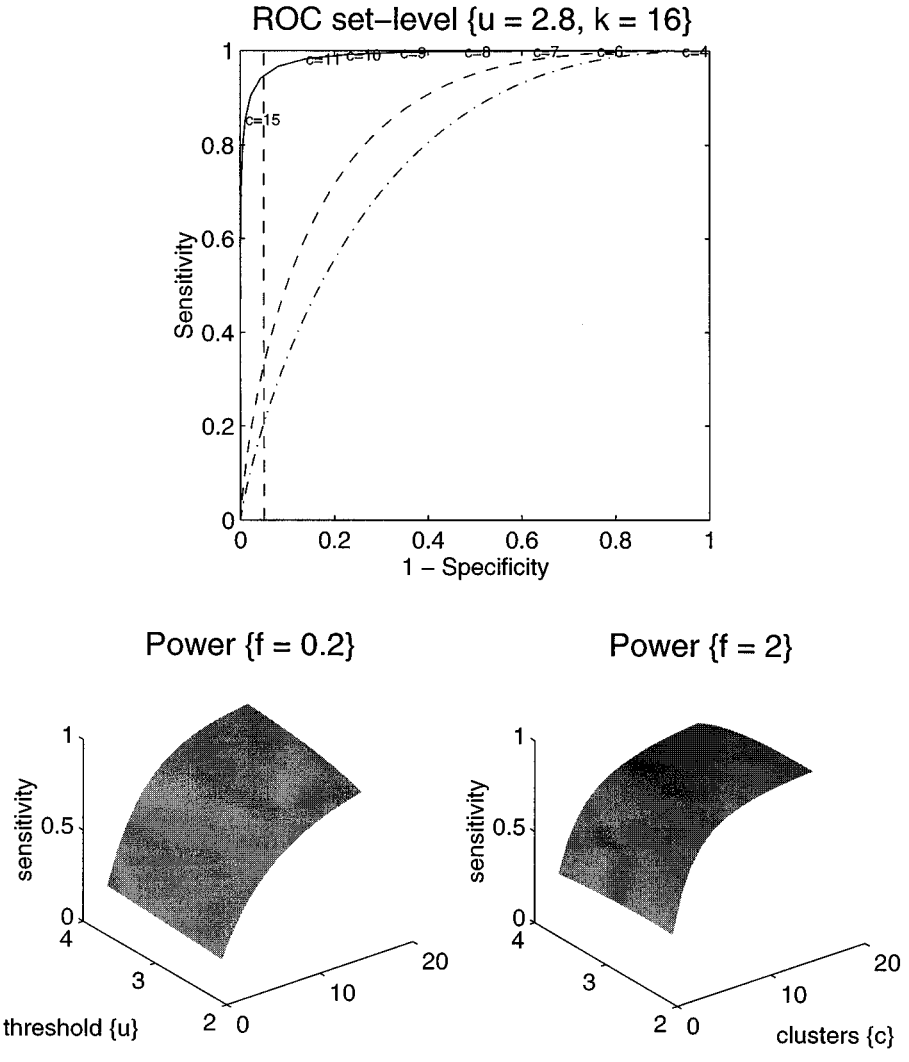


FIG. 3. (Top) ROC curve for set-level inference $P_W(2.8, 16, c)$, where W corresponds to a FWHM of 3 voxels and the volume $S = 64^3$. Signal amplitude $\sigma = 0.3$ and width $f = 2$. The dashed and broken lines corresponds to the equivalent cluster- and voxel-level ROC curves of the previous figures, respectively. (Bottom) Three-dimensional plot of power ($\alpha = 0.05$) as a function of cluster number c and threshold u, for the same smoothness, volume, and σ. Left $f = 0.2$ and right $f = 2$.

any regionally defined effects). Conversely, consider the case where 10 clusters were seen (at $P < 0.05$) and only 2 clusters were expected under the null hypothesis. In this case it would be reasonable to equate the observed set of clusters with a distributed system activated by the task in question. This situation distinguishes set-level inferences from omnibus tests and highlights their potential usefulness. The key issue here is the critical number of clusters relative to the number expected by chance (i.e., c_α relative to $E\{m\}$, where $P_W(U, K, c_\alpha) = \alpha$). This ratio is a strong function of the thresholds defining the clusters (i.e., U and K). Generally higher height thresholds give higher ratios: For example, Fig. 5 shows the values of c_α, $E\{m\}$, and their ratio for $S = 32^3$ voxels, $D = 3$, and W equivalent to a FWHM of 3 voxels. It can be seen that for $U > 3.2$ the ratio is sufficiently high to support an interpretation of the clusters obtained in terms of a distributed activation profile. The stepped nature of the curves in Fig. 5 reflects the fact that the critical values of c are discrete numbers. The main point here is that set-level inferences should be qualified by noting the expected number of clusters that would have occurred by chance. To characterize a set as representing a "distributed system" requires the observed number of clusters to greatly exceed the expected number. This will only be the case when higher thresholds are employed; otherwise, the set-level inference confers no regional specificity. This argument is about how one interprets a significant set of clusters (with a small P value). Some sets may be significant, but the number of clusters may only marginally exceed that expected by chance; other sets may be equally significant but comprise many more clusters than expected under the null hypothesis. Only in the

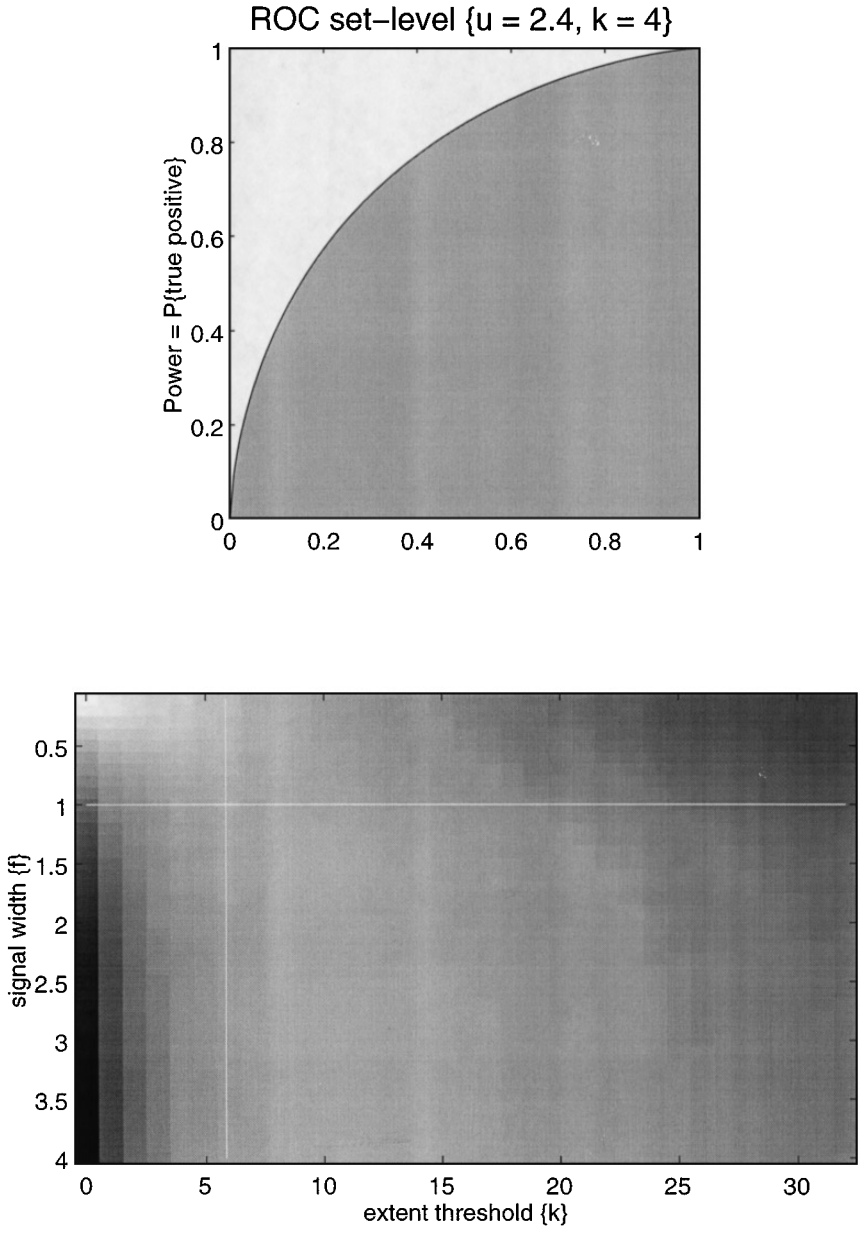


FIG. 4. (Top) ROC curve for set-level inference $P_W(2.4, 4, c)$, where W corresponds to a FWHM of 3 voxels and the volume $S = 64^3$. Signal amplitude $\sigma = 0.2$ and width $f = 2$. The shaded area corresponds to A and is a reflection of power over all specificities. (Bottom) Image format representation of A as a function of signal width f and extent threshold k, for the same smoothness, volume, and σ. The vertical line is the expected number of voxels per cluster under the null hypothesis ($E[n]$). The horizontal line is $f = 1$. The minimum of this gray scale is 0.59 and the maximum is 0.84.

later case can the observed set of clusters be interpreted, in any meaningful way, as a set of true activations.

DISCUSSION

We have addressed the sensitivity of various tests for activation foci in the brain by considering a hierarchy of test, framed in terms of the level of inference (voxel level, cluster level, and set level). All these nested levels of inference are based on a single probability of obtaining c, or more, clusters with k, or more, voxels above a threshold $u[P_W(u, k, c)]$. The higher the level of the inference the weaker the hypothesis tested in terms of regional specificity. The weakest case of set-level inferences are based on the total number of maxima above a threshold (i.e., $k = 0$) and correspond to the old omnibus tests used in PET. Cluster-level inferences are a special case of set-level inferences that obtain when the number of clusters $c = 1$. Similarly, voxel-level infer-

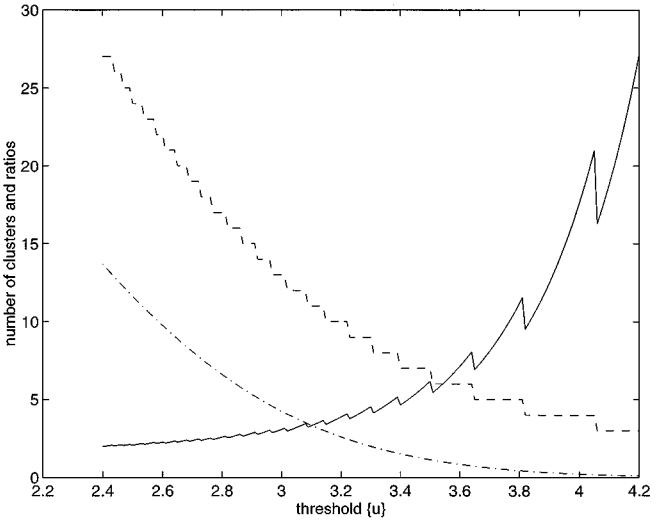


FIG. 5. The critical ($P < 0.001 = \alpha$) cluster number c_α, the expected number $E\{m\}$, and their ratio as a function of threshold U for a process of $S = 32^3$, $D = 3$, and W equivalent to a FWHM of 3 voxels.

ences are special cases of cluster-level inferences that result when the cluster has an unspecified volume (i.e., $k = 0$).

Levels of Inference and Power

On the basis of an analytical power analysis we concluded that set-level inferences are generally more powerful than cluster-level inferences and cluster-level inferences are generally more powerful than voxel-level inferences for distributed signals. For all levels of inference the size of the underlying signal f (relative to resolution W) determines the most powerful threshold to adopt. For set level inferences, if $f > 1$ (e.g., fMRI) then the optimum extent threshold should be greater than the expected number of voxels for each cluster under the null hypothesis. If $f < 1$ (e.g., PET) the extent threshold should be smaller if the most powerful test is desired. For cluster-level inferences of $f > 1$ (e.g., fMRI) then the height threshold u should be as low as possible (without violating the assumptions of high thresholds); we would recommend a value of around 2.4. If on the other hand, $f < 1$ (e.g., PET) then a high threshold is more powerful (e.g., 3.6). For voxel-level inferences the most powerful tests obtain when $f < 1$.

We envisage that the main advantage of set-level inferences will be found in fMRI. We have repeatedly observed results that comprise a set of activation foci that are very sensible in neurobiological terms; however, no single cluster is large enough, and no voxel is high enough, to survive a correction for the large volumes of high-resolution data analyzed. It is hoped that set-level inferences will provide more latitude in making inferences about these sets of activation foci.

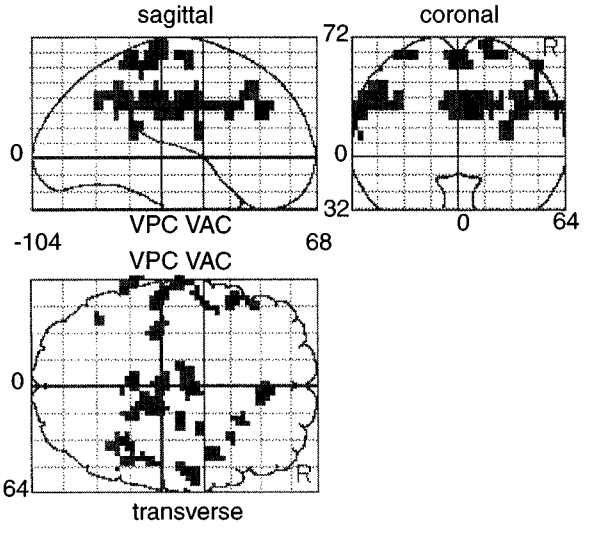
The power analyses presented below examine sensitivity as a function of the various thresholds employed to define the excursion set. This is conceptually distinct from maximizing sensitivity by smoothing the data in space (e.g., Worsley *et al.*, 1994) or time (e.g., Friston *et al.*, 1995) by appeal to the matched filter theorem. The results presented here pertain to inferences about excursion sets that could be applied to four-dimensional processes that include a dimension of "smoothing" or indeed a dimension of "thresholds" (i.e., a search of smoothing or scale-space or an explicit search through threshold-spaces).

The Nature of the "Signal"

The analyses in this paper are predicated on a particular model of "signal." One important aspect of this model is that the underlying hemodynamic responses are distributed throughout the volume analyzed. We appeal here to neurobiology to establish the validity of this model, in the sense that most brain processes are implemented by the integration of many functionally segregated areas, which are anatomically distributed. This balance between topographic segregation and functional integration is probably one of the key characteristics of complex biological systems like the brain. It should be noted, however, that many experiments discount functional integration and attempt to elicit activity in one area that is functionally specialized for a single sensorimotor or cognitive process. In this instance the signal may well comprise one (or a small number of) foci and set-level inferences may not be appropriate. In defence of set-level inferences most functional anatomy studies of cognitive function show that many separable cognitive components are instantiated in distributed neuronal systems and therefore the set of activation foci that ensue are probably more comprehensive descriptors of the evoked responses. The second point that can be made here is that set-level inferences do not preclude lower-level inferences. We develop this point below.

Which Level of Inference

When confronted with the task of characterizing an unknown and probably distributed activation profile, set-level inferences should clearly be considered, provided the implicit loss of regional specificity is appropriate. However voxel-, cluster-, and set-level inferences can be made concurrently. For example, using thresholds of $u = 2.4$ and $k = 6$ allows for a set-level inference in terms of the clusters reaching criteria. At the same time each cluster in that set has associated with it a corrected P value based on its size and the cluster-level inference. Similarly, each voxel in that cluster can be identified with a corrected P value based on the voxel-



set–level {c}	cluster–level {k,Z}	voxel–level {Z}	uncorrected	location {mm}
0.000 (20)	**0.041 (18, 5.56)**	**0.000 (5.56)**	**0.000**	**–3 –45 60**
		0.802 (3.53)	0.000	6 –51 54
	0.016 (24, 4.96)	**0.007 (4.96)**	**0.000**	**6 33 24**
	0.005 (32, 4.83)	**0.013 (4.83)**	**0.000**	**3 –9 30**
		0.018 (4.76)	0.000	–9 –15 30
	0.116 (12, 4.74)	**0.019 (4.74)**	**0.000**	**15 –30 66**
	0.048 (17, 4.70)	**0.024 (4.70)**	**0.000**	**54 –12 30**
	0.001 (47, 4.65)	**0.029 (4.65)**	**0.000**	**–54 –27 36**
		0.130 (4.26)	0.000	–63 –18 24
		0.436 (3.86)	0.000	–51 –3 30
	0.168 (10, 4.65)	**0.029 (4.65)**	**0.000**	**27 –6 18**
	0.009 (28, 4.56)	**0.042 (4.56)**	**0.000**	**36 –48 30**
		0.480 (3.82)	0.000	48 –48 36
		0.645 (3.68)	0.000	36 –57 36
	0.246 (8, 4.38)	**0.083 (4.38)**	**0.000**	**24 –15 60**
	0.168 (10, 4.35)	**0.095 (4.35)**	**0.000**	**45 –39 54**
		0.222 (4.10)	0.000	48 –30 54
	0.363 (6, 4.35)	**0.095 (4.35)**	**0.000**	**–57 –48 36**
	0.440 (5, 4.28)	**0.121 (4.28)**	**0.000**	**–63 –42 18**
	0.000 (55, 4.26)	**0.129 (4.26)**	**0.000**	**12 –36 36**
		0.327 (3.97)	0.000	0 –27 30
		0.614 (3.71)	0.000	18 –48 30

Height threshold {u} = 3.20, p = 0.001
Extent threshold {k} = 5 voxels
Expected voxels per cluster, E{n} = 2.4
Expected number of clusters, E{m} = 0.6
Volume {S} = 14476 voxels or 569.2 Resels
Degrees of freedom due to error = 44
Smoothness = 10.8 10.9 11.7 mm {FWHM}
= 4.6 4.6 5.0 {voxels}

FIG. 6. (Top) (SPM[*Z*]) This is a maximum intensity projection of the SPM[*Z*]. The display format is standard and provides three views of the brain from the front, below, and the lefthand side. Data are presented only for clusters and regions that survive the height and extent threshold detailed in the figure's footnotes. The gray scale is arbitrary and the space conforms to that described in the atlas of Talairach and Tournoux (1988). (Bottom) Tabular data characterizing the activation profile in terms of a set-level inference, or *P* value, based on the number of clusters *c*, for each cluster a *P* value reflecting the cluster-level inference based on the number of voxels comprising the cluster *k*, and *P* values corresponding to voxel-level inferences based on the *Z* score of selected maxima within each cluster. The uncorrected *P* value and location (*x*, *y*, *z* mm) of these voxels are also provided. The footnotes specify the thresholds used and parameters relating to this particular analysis.

level inference (i.e., its *Z* value). The nested taxonomy presented here allows for all levels to be reported, each higher level providing protection for the lower level. As long as the level of inference is clearly specified we can see no reason why different levels cannot be employed in characterizing the significance of the results obtained. If the inferences are made in a step-down fashion there should be no increase in FWE. Put simply, if the number of clusters is significant, then those clusters that are significantly large can also be identi-

fied. Inferences about voxels within significantly large clusters can then be made with impunity (this, of course, does not preclude the use of any one level on its own). An example of such a characterization is seen in Fig. 6, which represents a standard (SPM96) analysis of fMRI data. The experimental design and construction of the SPM$\{Z\}$ is not important here; suffice it to say that the SPM reflects hemodynamic activations due to intrinsically generated movements (relative to cued movements). The key thing to focus on is the tabular characterization of significant activations. u and k are, here, the user-defined height and extent thresholds, c is the number of observed clusters, k is the number of voxels in each cluster, and Z are maxima in these clusters. On the left the set-level inference suggests that the 12 clusters that comprise the activation profile are conjointly significant ($P < 0.001$). Within this set some regions can be considered significant at the cluster level (e.g., the cluster with a maximum at -24, 45, 12 mm, $P < 0.034$), whereas others are not (e.g., the last cluster with 24 voxels). The last cluster is only significant at a set level, i.e., when considered in the context of the remaining 11 clusters. In a similar vein, the voxel at -24, 45, 12 can be considered significant in its own right because the voxel-level P value is $P < 0.001$. Other voxels in this cluster, although part of a significant cluster, are not significant at the voxel level (e.g., the voxel at -25, 32, 12 mm, $P = 0.13$). This example highlights the potential benefit of set-level inferences; in that the entire activation profile can be described anatomically and characterized as significant, therein providing a complete and comprehensive picture of the activations, without having to omit activations that fail to reach cluster- or voxel-level significance.

Conclusion

We have reviewed current methods of inference using the theory of Gaussian fields in statistical parametric mapping and introduced a third level of inference that may be useful in characterizing distributed activation with functional neuroimaging.

APPENDIX

This appendix presents the derivation of Eq. (4). Let $q = E\{m\}$ and $p = P(n > k)$, then

$$\sum_{j=1}^{\infty} P(m = j).\binom{j}{i}.P(n \geq k)^i.P(n < k)^{j-i}$$

$$= \sum_{v=0}^{\infty} P(m = v + i).\frac{(v + i)!}{v!.i!}\,p^i(p - 1)^v$$

where $v = j - i$. Substituting the Poisson form for $P(m = v + I)$,

$$= \sum_{v=0}^{\infty} \frac{q^{v+i}e^{-q}}{v!.i!}\,p^i(p - 1)^v = \frac{1}{i!}.(p.q)^i e^{-q} \sum_{v=0}^{\infty} (q(1 - p))^v/v!,$$

noting that the sum on the right is a Taylor expansion of $e^{q(1-p)}$ we have,

$$= \frac{1}{i!}(p.q)^i e^{-pq} = \lambda(i, pq)$$

and finally,

$$P_W(u, k, c) \approx 1 - \sum_{i=0}^{c-1} \lambda(i, E\{m\}.P(n \geq k)).$$

ACKNOWLEDGMENTS

The authors were funded by the Wellcome trust. We thank all our colleagues for invaluable discussion, in particular Keith Worsley. J.B.P. was supported by EU: Human Capital and Mobility Grant No. ERB4001GT932036. *Note.* The theory described in this paper has been implemented in SPM96. This software (which runs in MATLAB, Sherborn, MA, under UNIX) is freely available from the authors.

REFERENCES

Adler, R. J., and Hasofer, A. M. 1981. *The Geometry of Random Fields.* Wiley, New York.

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M., and Noll, D. C. 1995. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. *Magn. Reson. Med.* **33:**636–647.

Fox, P. T., and Mintun, M. A. 1989. Noninvasive functional brain mapping by change distribution analysis of averaged PET images of $H^{15}O_2$ tissue activity. *J. Nucl. Med.* **30:**141–149.

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 Mapping* **1:**214–220.

Friston, K. J., Holmes, A. P., Poline, J.-B., Grasby, P. J., Williams, S. C. R., Frackowiak, R. S. J., and Turner, R. 1995. Analysis of fMRI time-series revisited. *NeuroImage* **2:**45–53.

Hasofer, A. M. 1978. Upcrossings of random fields. *Suppl. Adv. Appl. Prob.* **10:**14–21.

Holmes, A. P. 1994. Ph.D. thesis.

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.

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

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

Worsley, K. J. 1994. Local Maxima and the expected Euler characteristic of excursion sets of χ^2, F and t fields. *Adv. Appl. Prob.* **26:**13–42.

Worsley, K. J., Poline, J.-B., Vandal, A. C., and Friston, K. J. 1995. Tests for distributed, nonfocal brain activations. *NeuroImage* **2:**183–194.

Siegmund, D. O., and Worsley, K. J. 1994. Testing for a signal with unknown location and scale in a stationary Gaussian random field. *Ann. Stat.* **23:**608–639.

Talairach, J., and Tournoux, P. 1988. *A Co-planar Stereotaxis Atlas of a Human Brain*. Thieme, Stuttgart.