THE ROYAL SOCIETY

Statistical limitations in functional neuroimaging II. Signal detection and statistical inference

Karl Magnus Petersson[1*], **Thomas E. Nichols**[2,3], **Jean-Baptiste Poline**[4,5] **and Andrew P. Holmes**[5,6]

[1]*Cognitive Neurophysiology R2-01, Department of Clinical Neuroscience, Karolinska Institute, Karolinska Hospital, S-171 76 Stockholm, Sweden (karlmp@neuro.ks.se)*
[2]*Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA (nicholst@stat.cmu.edu)*
[3]*Center for the Neural Basis of Cognition, Carnegie Mellon University and University of Pittsburgh, 4400 Fifth Avenue, Pittsburgh, PA 15213, USA*
[4]*Commissariat l'Energie Atomique, Direction de la Recherche Médicale, Service Hospitalier Frederic Joliot, 4 Pl General Leclerc, 91406 Orsay, France (poline@shfj.cea.fr)*
[5]*Wellcome Department of Cognitive Neurology, Functional Imaging Laboratory, 12 Queen Square, London WC1N 3BG, UK (andrew@fil.ion.ucl.ac.uk)*
[6]*Robertson Centre for Biostatistics, Department of Statistics, University of Glasgow, Boyd Orr Building, University Avenue, Glasgow G12 8QQ, UK (andrew@stats.gla.ac.uk)*

CONTENTS PAGE

The field of functional neuroimaging (FNI) methodology has developed into a mature but evolving area of knowledge and its applications have been extensive. A general problem in the analysis of FNI data is finding a signal embedded in noise. This is sometimes called signal detection. Signal detection theory focuses in general on issues relating to the optimization of conditions for separating the signal from noise. When methods from probability theory and mathematical statistics are directly applied in this procedure it is also called statistical inference. In this paper we briefly discuss some aspects of signal detection theory relevant to FNI and, in addition, some common approaches to statistical inference used in FNI.

* Author for correspondence.

Low-pass filtering in relation to functional–anatomical variability and some effects of filtering on signal detection of interest to FNI are discussed. Also, some general aspects of hypothesis testing and statistical inference are discussed. This includes the need for characterizing the signal in data when the null hypothesis is rejected, the problem of multiple comparisons that is central to FNI data analysis, omnibus tests and some issues related to statistical power in the context of FNI. In turn, random field, scale space, non-parametric and Monte Carlo approaches are reviewed, representing the most common approaches to statistical inference used in FNI. Complementary to these issues an overview and discussion of non-inferential descriptive methods, common statistical models and the problem of model selection is given in a companion paper. In general, model selection is an important prelude to subsequent statistical inference. The emphasis in both papers is on the assumptions and inherent limitations of the methods presented. Most of the methods described here generally serve their purposes well when the inherent assumptions and limitations are taken into account. Significant differences in results between different methods are most apparent in extreme parameter ranges, for example at low effective degrees of freedom or at small spatial autocorrelation. In such situations or in situations when assumptions and approximations are seriously violated it is of central importance to choose the most suitable method in order to obtain valid results.

Keywords: functional neuroimaging methods; PET; fMRI; signal detection; image filtering; statistical inference

1. INTRODUCTION

Some aspects of signal detection theory relevant to functional neuroimaging (FNI) and common approaches to statistical inference used in FNI are discussed in this review. In a companion paper (Petersson *et al.*, preceding paper), non-inferential descriptive methods and statistical models for FNI data, as well as some issues related to the problem of model selections relevant to FNI, are reviewed. The field of FNI methodology has developed into a mature but evolving area of knowledge, illustrating the need for validated and effective descriptive and inferential methods. Non-inferential descriptive methods are used to characterize signals present in the data, while inferential methods are commonly used to test hypotheses and determine confidence intervals. Because of the danger of making incorrect assertions in an emerging area of neuroscience, the emphasis in the analysis of FNI data has so far been on statistical methods protecting against false-positive results.

In this and the companion paper (Petersson *et al.*, preceding paper) the focus is on the assumptions, approximations and limitations of the methods reviewed. This delineates the limits of applicability and is essential in order to use the available methods optimally and to interpret FNI results appropriately. The methods described have been selected as representative and because of their widespread application; in general they perform well when the assumptions and limitations are taken into account. For technical details and a more complete picture of these methods, their benefits and examples of their applicability we refer to the original literature that describes these aspects well. The inferential methods used in FNI differ in the assumptions made about the properties of data and in the approximations used in the statistical theory. Approximations and assumptions about data need to be critically examined. However, it is not the assumptions or approximations themselves that are of critical importance. Rather, the central issues are how well these are fulfilled by empirical data, how robust the methods are to departures from the assumptions and the effect of the approximations made. This highlights the importance of empirical validation and the explicit characterization of the inherent limitations of a given method, themes which are developed in this review. We will concentrate mainly on the most common methods, giving an updated and concise conceptual overview.

Most FNI methods are based on voxel data, an approach pioneered by Fox *et al.* (1988; Fox & Mintun 1989) and Friston *et al.* (1990, 1991). Fox *et al.* (1988; Fox & Mintun 1989) based their approach on intersubject averaging and change-distribution analysis of subtracted positron emission tomography (PET) images, representing an omnibus approach (cf. § 3(c)). Friston *et al.* (1990, 1991) introduced the concept of statistical parametric mapping underscoring the need for strong control of the familywise error rate in hypothesis testing (cf. § 3(b)). Alternatively, FNI approaches can also be based on regions of interest (ROIs). Defining ROIs can be problematic and is, to a certain extent, arbitrary, in particular when defined directly on FNI data unless specific regional hypotheses are given prior to inspection of the data. When defined on co-registered anatomical images this approach has some attractive features, in particular avoidance of the severe multiple comparisons problem associated with a global search of statistic images.

The primary FNI data are commonly pre-processed (e.g. realigned, anatomically normalized and low-pass filtered), a statistical model and a test statistic are chosen and model parameters are estimated for statistical inference taking into account multiple non-independent comparisons and possible temporal autocorrelation. Anatomical normalization is performed to adjust for gross anatomical differences when data are averaged across subjects, but has limits at different anatomical levels. At a basic level, it is unlikely that a unique point-to-point transformation can be defined from one brain to the other in a meaningful way. For example, some anatomical feature may exist in one brain but not in another. This makes assessing the adequacy of different transformations difficult since no gold standard is given (Grachev *et al.* 1999). In certain situations and in particular for functional magnetic resonance imaging (fMRI) data, an ROI approach may offer a way around some of the limitations

with anatomical normalization (Mazoyer *et al*. 1993; Crivello *et al*. 1995). Anatomically defined ROIs can offer a good solution when the anatomical regions are well defined, as, for example, in the case of subcortical structures or when there are natural and well-defined, prior, regionally specific hypotheses. In combination with several single-subject fMRI studies, a meta-analytic approach may be used on ROI data without the need to normalize the primary data. There would also be a limited need for image smoothing or spatial filtering (cf. § 2). The definition of ROIs is driven by prior anatomical information, while the arbitrariness of anatomical normalization procedures is often less guided. Automatic segmentation procedures for the sulci, gyri, grey matter and white matter have become important tools in defining ROIs. Increased and detailed knowledge of the actual anatomy of the human brain is required if the aim of relating any region of a normal brain to its counterpart in another normal brain is to be achieved (Mangin *et al*. 1995; Regis *et al*. 1995).

The use of sufficiently well-fitting statistical models and making sure that the assumptions made are fulfilled is generally necessary for the validity of the subsequent statistical inference. If the assumptions are violated or ill-fitting models are used, then this may make the ensuing inference (statistically) invalid. This indicates the importance of proper model selection and the verification of assumptions (cf. Petersson *et al*., preceding paper). In this paper, therefore, we will assume that appropriate model selection has been performed and that the chosen model fits sufficiently well.

This review is broadly organized as follows. First, some aspects of signal detection, filtering of FNI data and bias of the detection sensitivity are discussed. These issues are related to functional–anatomical variability and across-subject averaging of data. Next follows a brief overview of some general issues with implications for statistical inference. Of particular interest in the context of FNI is the need to characterize the signal present in data when the null hypothesis is rejected, the multiple comparisons problem and statistical power. Finally, three methods of statistical inference in common use in FNI, parametric random field (RF), non-parametric and Monte Carlo approaches, are reviewed.

2. SPATIAL FILTERING

In this section some implications of functional–anatomical variability on across-subject averaging, strategies to handle this variability and the biasing of detection sensitivity by spatial filtering are discussed.

(a) *Functional–anatomical variability*

In cognitive activation experiments, the changes in neuronal activation due to experimental manipulation are often moderate compared to the background noise. However, modern imaging systems have improved the signal-to-noise ratio significantly. To improve signal to noise further the experimental manipulation can be repeated several times on the same or different subjects and the acquired data averaged over repetitions (Fox *et al*. 1988). Since experimentally induced responses can be quite variable there are statistical motivations for acquiring multiple scans per subject and also for including multiple subjects in a study. Repeated measurements allow generalization to average effects and to properly estimated different within- and between-subjects sources of variability.

In the case of intersubject averaging, it is first necessary to realign the data from each individual and to standardize (i.e. spatially normalize) and transform the data anatomically into a standard stereotactic space. There are limits on both the anatomical normalization procedures and the precision with which an anatomic correspondence between different brains can be defined. It is an open question whether functional–anatomical variability can always be effectively normalized or if this variability sometimes is too large (Hunton *et al*. 1996). Several current FNI methods assume that brain function can be mapped to structure uniquely at the resolution of the imaging system and the spatial pre-processing (i.e. realignment, spatial normalization and optional low-pass filtering). For PET data, this assumption seems to work sufficiently well. However, for fMRI data, the spatial resolution makes it possible to detect appreciable residual anatomical and structure–function variability between subjects. This opens up for three possibilities: first, to study population generalizations of functional anatomy down to a given resolution, second, to study the variability in localization of a given function in relation to a well-defined anatomical landmark and, third, we may also be interested in studying a given function and its actual location in an individual subject. These aims require different information and different types of spatial pre-processing.

There have been some attempts to assess residual functional–anatomical variability, mostly in realigned, spatially normalized and more or less low-pass filtered PET data. These attempts have often used the variability in location of the local maximum statistic (peak location). However, the variability in peak location may not always be an ideal measure of functional–anatomical variability. For example, the local maximum of the *t*-statistic image is known to have appreciable intrinsic variability in location, in particular when the degrees of freedom are low relating to noise in the estimated variance image (Taylor *et al*. 1993; Worsley *et al*. 1993; Holmes 1994; Holmes *et al*. 1996). Thus, there may be an inherent variability in peak location that adds to the true residual functional–anatomical variability. The intersubject variability estimated in this way may thus be overestimated. Several studies have estimated intersubject standard deviations of the peak coordinates to be in the order of 5–10 mm (Fox & Pardo 1991; Hunton *et al*. 1996; Ramsey *et al*. 1996; Hasnain *et al*. 1998). When PET data from different laboratories are compared this variability increases (Poline *et al*. 1996; Senda *et al*. 1998) and there are some indications that activation foci that are less than 10 mm apart may not always be reliably distinguished with PET (Grabowski *et al*. 1996). The intra-individual variability may also be significant, even for robust primary motor activations (Hunton *et al*. 1996). The residual intra-individual variability is likely to reflect in part inherent variability in peak location.

The interindividual residual variability generally exhibits spatial structure and is dependent on the algorithm used for anatomic normalization. Simulation studies have

indicated that a reduction of misregistration error and minimizing the residual anatomic variability can significantly improve the signal detection sensitivity (Worsley *et al*. 1996*b*). It should be noted that the effect of intersubject averaging, when there is a residual functional–anatomical variability, amounts to a spatial filtering effect. In general, using a voxel-based analysis, it is important to reduce the impact of misregistration and interindividual residual functional–anatomical variability. A common strategy is to low-pass filter the data either at reconstruction or with a suitably chosen convolution kernel (e.g. an isotropic Gaussian kernel). Alternative approaches have been suggested (Coulon *et al*. 1997) which may be more stable in terms of interindividual functional–anatomical variability. This is based on the idea that it is difficult to define appropriate point-to-point transformation between brains but that three-dimensional (3D) landmarks and anatomical structures may be more easy to relate (Mangin *et al*. 1995; Regis *et al*. 1995). Whether it is possible to identify 3D landmarks that can be reliably matched across different brains and to what extent such landmarks can constrain spatial normalization and correspond to functional units is an open question. To the extent such procedures can be validated, knowledge about stable landmarks may be used to match the signal across subjects in a structural sense.

Spatial filtering, which in effect is a local weighted averaging procedure, may also increase the equivalence of voxel data across measurements and, thus, the validity of voxel-based statistical models. On the other hand, too much filtering may average functionally different signals yielding a functional–anatomic blurring effect. Note that there is a fundamental distinction between fMRI and PET in the sense that it is generally possible to obtain reliable within-subject signals with fMRI, while radioactive dose limitation often precludes this in PET. This opens up the possibility of using different pre-processing strategies for the two imaging modalities.

(b) *Image smoothing and signal detection*

Filtering data may or may not increase the signal-to-noise ratio, depending on the relationship between the size and shape of the signal and the convolution kernel. Some of the effects of image smoothing may be understood in the light of the matched filter theorem (Rosenfeld & Kak 1982). This theorem states that a signal in a background of white noise is detected with optimal sensitivity if a convolution kernel, which matches the size and shape of the signal, is used. It should be noted that the situation is slightly more complicated when the noise component is coloured, that is autocorrelated, which is most often the case with FNI data. If the spatial extent of the signal is large compared to the extent of the spatial autocorrelation, then the result of the matched filter theorem may serve as a good approximation. However, if this is not the case, the choice of an optimal filter is more complicated than just choosing a matched filter. In this case the autocorrelation has to be taken into account. This can be illustrated in the stationary case; assuming that the data can be whitened with a filter W, then the matched filter theorem can be applied to the whitened data generating a matched filter S. This is equivalent to applying the convolution of W and S, $W*S$, as a filter directly. Since the whitening filter W is related to the autocorrelation, this makes the dependence of the optimal filter on the autocorrelation explicit. However, we will reason informally here, describing the qualitative effects of filtering, assuming that the matched filter theorem is a good approximation.

Empirical and simulation studies indicate that the choice of filter kernel affects detection sensitivity and the results of subsequent statistical analysis (Poline & Mazoyer 1991, 1994*c*; Andreasen *et al*. 1995; Worsley *et al*. 1996*b*). Image smoothing amounts to applying a spatial low-pass filter that affects the data or signal itself by attenuating high-frequency components and possibly (slightly) displacing the location. Note that PET and fMRI are measuring haemodynamic indicators of neuronal activity, that is in effect imaging a smooth fluid process, and FNI data are essentially always filtered. In PET, image reconstruction (e.g. filtered back projection) applies a filter to reconstruct image data from projection data, which corresponds to a ramp filter with a chosen window (e.g. Hanning or Butterworth). In fMRI, the frequency spectrum in k-space is approximated with a function of compact support. This is equivalent to the application of a band-pass filter, that is, each voxel is a filtered version of the true T_2^* signal. This implies that the choice not to filter the data is equivalent to accepting the implicit filter of the imaging system. There are some indications that accepting the detection bias inherent in the imaging system (i.e. no image smoothing) may sometimes decrease the signal-to-noise ratio relative to when smoothing is used (Andreasen *et al*. 1995). In addition, unless the variability in functional–anatomical localization and misregistration error can be reduced to zero, averaging measurements within and across subjects corresponds to a low-pass filter.

The matched filter theorem indicates that the detection sensitivity is biased towards signals similar to the smoothing kernel. An alternative to the single-filter approach takes the mirror consequence of this theorem as its starting point. Specifically, given that a matching filter optimizes sensitivity, this also implies that the choice of a particular filter biases the detection sensitivity towards signals of a certain shape and size (or scale). For sufficiently different signals the filter choice will be suboptimal with reduced detection sensitivity as a consequence. A solution to the problem of detection bias is to use several different filters and search for signals at different scales (Poline & Mazoyer 1994*a*,*c*). The multi-filtering approach has recently been integrated with RF theory (RFT), the so-called scale space approach (Worsley *et al*. 1996*b*, 1997*a*; Worsley 1999). The scale space approach is discussed below in § 5.

A second alternative is based on the concept of signal and image restoration (e.g. Andrews & Hunt 1977; Geman & Geman 1984). Ideally, the restoration process smoothes the image noise while the signal is preserved. Markov RF (MRF) theory (Besag 1974; Chellapa & Jain 1993) can be used to incorporate prior information via prior distributions in a Bayesian approach (Holmes & Ford 1993). There is a close relationship between MRF and Gibbs RFs (via the Hammersley–Clifford theorem) (Besag 1974; Geman & Geman 1984; Chellapa & Jain 1993). The Bayesian approach has been successfully used

to tackle problems of image registration, reconstruction, restoration and segmentation (Geman & Geman 1984; Chellapa & Jain 1993; Gee *et al*. 1995; Aschburner *et al*. 1997). However, signal restoration and smart filtering techniques have so far not been extensively applied because the impact on the subsequent statistical inference is often unclear. Although spatial smoothing biases the results of an analysis, the effect is better understood than the way a strong prior might bias the results, for example the effect of edge-respecting filters.

In this context, the Bayesian approach may be viewed as stochastic regularization closely related to standard regularization theory (Tikhonov & Arsenin 1977; for interesting connections to statistical learning theory see Vapnik (1998) and Wahba (1995)). Recently an application of the MRF–Bayesian regularization approach to fMRI signal restoration was described (Descombes *et al*. 1998), using spatio-temporal MRFs in combination with simulated annealing optimization (Kirkpatrick *et al*. 1983). The prime difficulty with this approach is in specifying adequate prior information in terms of a Bayesian prior. The use of a specific restoration prior may introduce image artefacts and, if the wrong order of the MRF is chosen, high bias may result (Rangarajan & Chellappa 1995; Descombes *et al*. 1998). Extensive simulations and validation may shed light on these issues. There are some indications that the restoration approach using tuned priors may suppress spatio-temporal noise without spoiling the signal, resulting in better spatio-temporal delineation of the fMRI signal (Descombes *et al*. 1998). However, the implications on the subsequent statistical inference are at present unknown.

In closing this section, which has mainly focused on signal detection in relation to image smoothing, it should be noted that the arguments are general and also apply to signals in the temporal domain. In particular, this relates to whether it is beneficial to smooth or filter fMRI time-series in the temporal dimension. If there is prior knowledge of the characteristics of the temporal signal, this could be used to construct a matching filter, thus optimizing signal detection. For example, it has been suggested that fMRI data should be temporally filtered with the haemodynamic response function (Friston *et al*. 1994*a*; Frackowiak *et al*. 1997). In the case of limited prior knowledge of the temporal signal, a multifiltering approach may bias the detection sensitivity less than a single-filter approach at the cost of a more extensive search (cf. § 5). Finally, it should also be pointed out that brain activations can bias automatic realignment and anatomic normalization procedures. Even if this bias can be small, since it is highly correlated with the experimental paradigm, it may introduce false-positive activations. In addition, stimulus-correlated motion may introduce systematic intensity changes in fMRI data (Hajnal *et al*. 1994). Note also that global normalization may be biased by activation introducing paradigm correlated errors (this is further discussed in Petersson *et al*. (preceding paper)).

3. STATISTICAL INFERENCE IN FUNCTIONAL NEUROIMAGING

In the following sections, some general aspects of statistical inference, hypothesis testing and the determination of confidence intervals are discussed. We also describe the multiple comparisons problem and weak and strong control of the familywise error rate, which is the relevant measure of the false-positive rate in a multiple comparisons situation. In addition, we describe some omnibus tests that have been used in FNI, with weak control over the familywise error rate. We also review some issues relating to statistical power and discuss the three main approaches to statistical inference used in FNI, that is RF, non-parametric and Monte Carlo approaches, in some detail.

(a) *Hypothesis testing and determination of confidence intervals*

In § 3 of the companion paper, we described the ways in which parameters of interest and nuisance parameters are related to the data and means of estimating those parameters (Petersson *et al*., preceding paper; see also Frackowiak *et al*. 1993). Ostensibly, we are interested in the parameters themselves, such as the magnitude of an activation. However, the overriding concern in the field has been the avoidance of making false claims. Hence, the parameters are always assessed relative to their uncertainty in a statistical hypothesis-testing framework. Informally, we wish to know whether the magnitude of the parameter (or contrast of parameters) is substantial with respect to its uncertainty. In this section, hypothesis testing and its use in FNI are reviewed.

Hypothesis testing proceeds as such. The null hypothesis is assessed with a test statistic, a function of the data which is sensitive to departures from the null hypothesis and reflects the effects of interest; the observed statistic is compared to its distribution under the null hypothesis, producing a p-value. A small p-value is interpreted as indicating that there is little support for the null hypothesis, though its interpretation is more subtle. The p-value is the probability of observing a statistical value as large or larger, under an identical replication of the experiment and under the assumption that the null hypothesis is true. Hence, the p-value is a statement about the data under the null hypothesis, not the null hypothesis itself. In the decision theoretic framework of hypothesis testing, a pre-specified level of significance is used to accept or reject the null hypothesis. Alternatively, the smallness of the p-value may be viewed as a measure of the strength of the empirical evidence against the null hypothesis (Edgington 1995), representing a smooth transition between empirical evidence interpretable as indicating the alternative hypothesis and empirical evidence in favour of the null hypothesis.

If one rejects the veracity of the null hypothesis whenever the p-value is below a critical value α, then a valid test will control the false-positive rate at α. The false-negative rate β is closely related to the statistical power $1-\beta$. The statistical power is the probability of rejecting the null hypothesis when it is false. While it would seem natural to focus attention on the power of the test, the power is a function of the unknown alternative and the best that can be done is to use test statistics that maximize power over all alternatives (relative to all other tests of the same class). Most standard test statistics satisfy this requirement. In general, the power of tests increases with the sample size (strictly, the degrees of freedom).

The regression approach in FNI fits univariate models at every voxel and effects of interest are tested in each individual model by generating and assessing a statistic image. Usually an image regression approach is used, which implies that the same univariate model is fitted at each voxel. The choice of test statistic is central and should represent a valid measure of the phenomenon of interest. While the details of some common test statistics are considered below, the common test procedures in FNI conform to the standard structure of hypothesis testing. If a particular, pre-specified voxel or ROI is of interest, then standard univariate theory can be applied. Otherwise the statistic image is searched for, for example voxels of significant magnitude using the local maximum statistic or, given an intensity threshold, significant clusters using the suprathreshold cluster size statistic. However, the notion of characterizing the signal when the null hypothesis is rejected and issues similar to it regarding the status of the alternative hypothesis have often been neglected under the standard hypothesis-testing framework.

(i) *Some general inference issues*

Statistical inference in FNI has focused on hypothesis testing, which is primarily concerned with the null hypothesis. This is convenient, since the distribution of the test statistic is only needed under the null hypothesis. However, once the null hypothesis is rejected it is useful to try to characterize the signal present in the data. In traditional univariate statistics this amounts to using confidence intervals instead of p-values. In FNI, the dimensionality of the statistical image precludes straightforward use of confidence intervals. However, there is a clear need to characterize the alternative hypothesis. While there are multivariate approaches for characterizing the alternative hypothesis (cf. Petersson *et al.*, preceding paper), there is still a need to characterize the signal, that is a departure from the null hypothesis statistically at a regional or voxel level. For example, if two clusters are declared significant, a natural question is whether one cluster is significantly larger than the other. As another example, consider two statistical images each with significant activation foci in approximately the same region, with a few significant overlapping voxels but separated maxima. An interesting question is whether these two foci represent the same anatomical location given the variability of the intersubject registration and other sources of variability. Currently, there are no methods that address such questions and in general there are few methods to characterize the signal present in the data.

A further problem is that a significant effect may not be a relevant one. Since statistical power increases with independent measurements, with sufficient observations a hypothesis test has the power to detect minuscule changes. The limiting case is colloquially known as the hypothesis-testing fallacy: since the null hypothesis of exactly no change is essentially never true for any continuous system, as the number of observations tends to infinity a hypothesis test will always reject the null hypothesis given enough observations. Arguably, any consistent change may be of interest, but such minuscule effects may simply reflect unmodelled idiosyncrasies of the particular experiment or effects not specifically related to the task. This is particularly a problem for single-subject (single-session) fMRI experiments, where effects are commonly assessed relative to the scan-to-scan error variance (i.e. using a fixed effects model) for which a large number of observations can be obtained.

The implicit interpretation of significant regions as important regions implicated in processing the particular mental task of interest may become increasingly problematic as the sensitivity of the test increases with sample size. In essence, this exposes the mismatch between the statistical hypothesis of no difference and that of the experimental hypothesis that the region is not involved in processing the given task. Given that no part of the brain remains exactly the same and that some FNI systems are beginning to have the power to measure even very small changes reliably, the question becomes one of spatio-temporal modelling and interpretation, rather than the simple 'is there a difference' hypothesis-testing approach. This also points towards the need for quantitative spatio-temporal models linking neuronal dynamics to the haemodynamic effects observable with FNI techniques.

Finally, an additional concern is the opposite of the above. Without knowledge of the power of our testing procedures, we have no idea of the imagewise false-negative error rate, the sensitivity, save vague notions from past experience. This problem is compounded by the erroneous tendency to interpret lack of significance as evidence of no change, when in fact the lack of significance could be due to low sensitivity due to low subject numbers. Recent clinical applications of FNI, such as presurgical planning, highlight a related issue. In these situations it is commonly type II errors (false negatives) that are most important. For example, consider presurgical planning for a resection on an epileptic subject, where functional experiments are conducted to see whether a specific brain region supports important functions. The tragic consequence of a false negative, if acted on, may be that an important brain region is damaged.

(b) *The multiple comparisons problem*

Statistical analysis of FNI data often implies that many hypotheses are tested on the same data set. Central to the multiple (e.g. voxel-by-voxel) hypothesis-testing approach is an adequate handling of the multiple comparisons problem, that is, it is necessary to control the false-positive rate appropriately. The overall false-positive rate increases with multiple testing unless care is taken. For example, if each voxel is assessed with a univariate test at level $\alpha = 0.01$, corresponding to thresholding a Gaussian statistic image at $Z = 2.33$, then one would expect *ca.* 1% of the voxels to appear above the threshold by chance even if there is no activation anywhere. It should also be noted that these false-positive voxels may be clustered due to spatial autocorrelation. Ideally, the statistical inference procedure should handle the multiple comparisons problem effectively, avoiding any unnecessary loss of sensitivity and statistical power.

Given the null hypothesis H_0 and a test statistic $\mathcal{T}(X)$ of the data X, the test is said to be liberal, conservative or exact if, for any given level α and rejection region $R(\alpha)$, the probability that $\mathcal{T}(X)$ belongs to the rejection

region $R(\alpha)$, $P[\mathcal{T}(X) \in R(\alpha))|H_0]$, is greater than, less than or equal to α, respectively. Appropriate control of the false-positive rate requires an exact or conservative test. Usually the more conservative the test the less the sensitivity. (However, this does not always have to be the case since sensitivity is a function of the actual alternative.)

(i) *Weak and strong control over the familywise error rate*

In order to handle the multiple comparisons problem (Hochberg & Tamhane 1987) appropriately, the rejection criteria have to be chosen so that the probability of rejecting one or more of the null hypotheses when the rejected null hypotheses are actually true is sufficiently small. Let the search volume $\Omega = \{v_1, \ldots, v_K\}$ consist of K voxels $v_1, \ldots, v_K$ and let $H_1, \ldots, H_K$ be the null hypotheses for each voxel. The omnibus null hypothesis H_Ω is the (logical) conjunction of $H_1, \ldots, H_K$, that is, $H_\Omega = H_1 \cap \ldots \cap H_K$. To test $H_1, \ldots, H_K$ we use a family of tests, $\mathcal{T}_1, \ldots, \mathcal{T}_K$. For all $j \in \{1 \ldots K\}$ let E_j be the event that the test $\mathcal{T}_j$ incorrectly rejects H_j, that is, $E_j = [\mathcal{T}_j \in R(\alpha_j)]$, where $R(\alpha_j)$ is the corresponding rejection region at the level α_j. Suppose the test is exact or possibly conservative, i.e. $P[E_j|H_\Omega] \leqslant \alpha_j$.

In the context of the family $\mathcal{T}_1, \ldots, \mathcal{T}_K$ of tests, the familywise error rate (FWE) is defined as the probability of falsely rejecting any of the null hypotheses $H_1, \ldots, H_K$. Let E_Ω be the event that the omnibus hypothesis is rejected, that is, $E_\Omega = E_1 \cup \ldots \cup E_K$. Weak control of the FWE requires that the probability of falsely rejecting the omnibus null hypothesis H_Ω is, at most, the test level α, that is, $P[E_\Omega|H_\Omega] \leqslant \alpha$. Evidence against the omnibus hypothesis H_Ω indicates the presence of some activation somewhere. This implies that the test has no localizing power, meaning that the false-positive rate for individual voxels is not controlled. Tests that have only weak control over the FWE are called omnibus tests and are useful in detecting whether there is any experimentally induced effect at all, regardless of location. If, on the other hand, there is interest in not only detecting an experimentally induced signal but also reliably locating the effect, a test procedure with strong control over the FWE is required.

Strong control over the FWE requires that the FWE be controlled, not just under H_Ω, but also under any subset of hypotheses. Specifically, for any subset of voxels $\omega \subseteq \Omega$ and corresponding omnibus hypothesis H_w, $P[E_w|H_w] \leqslant \alpha$. That is, all possible subsets of hypotheses are tested with weak control over the FWE. This ensures that the test is valid at every voxel and that the validity of the test in any given region is not affected by the truth of the null hypothesis elsewhere. Thus, a test procedure with strong control over the FWE has localizing power.

(ii) *Spatial autocorrelation and multiple non-independent comparisons*

One way to achieve strong FWE control, to control the otherwise increasing level of false-positive results with an increasing number of comparisons, is to adjust the level of significance at which the different hypotheses $H_1, \ldots, H_K$ are tested. The single-step Bonferroni correction is an illustrative example of such a strategy. Suppose that $H_1, \ldots, H_k$ are tested at an equal level, say b, i.e. $P[E_1|H_\Omega] = b, \ldots, P[E_K|H_\Omega] = b$. If all voxels have the same marginal distribution, then testing them at an equal level amounts to thresholding the statistical image, giving a single threshold test. In general, $P[E_\Omega|H_\Omega] = P[E_1 \cup \ldots \cup E_K|H_\Omega] \leqslant P[E_1|H_\Omega] + \ldots + P[E_K|H_\Omega] = K \times b$. If b is chosen so that $K \times b = \alpha$, i.e. $b = \alpha/K$, it follows that $P[E_\Omega|H_\Omega] \leqslant \alpha$. This so-called Bonferroni correction will be conservative when the tests are correlated, since then $P[E_\Omega|H_\Omega]$ will be smaller than $P[E_1|H_\Omega] + \ldots + P[E_K|H_\Omega]$. For a large number of correlated tests, the Bonferroni correction results in a conservative overall procedure and an unnecessary loss of statistical power. There are other and more refined multistep procedures for handling the general multiple comparisons problem (Hochberg & Tamhane 1987).

FNI data are often characterized by spatial autocorrelation, that is closely spaced voxels are correlated. The spatial autocorrelation is partly due to the point spread function of the imaging system but physiological factors are also important. The spatial autocorrelation is commonly less extensive in fMRI than in PET data. It should also be noted that image smoothing introduces spatial autocorrelation. Given non-trivial spatial autocorrelation in the statistic image this implies multiple non-independent comparisons and a simple Bonferroni correction would be conservative. Instead, an effective solution of the multiple non-independent comparisons problem is central to the voxel approach. The different approaches to statistical inference described below attempt to solve this problem in different ways. Broadly speaking, these divide into parametric, non-parametric and Monte Carlo simulation approaches (cf. below). In general, parametric approaches are exact when no approximations are made and the assumptions made are fulfilled. The parametric approaches used in FNI are usually based on some type of RFT (e.g. Adler 1981, 1998; Worsley *et al.* 1992, 1996*a*; Worsley 1994; Friston *et al.* 1995) generating distributional approximations. The results of the Monte Carlo approaches are always approximations and whether the Monte Carlo results are good approximations depends critically on whether all significant sources of variability under null hypothesis conditions are sufficiently well modelled. The non-parametric approaches solve the multiple comparisons problem exactly.

(c) *Omnibus tests*

There are some similarities between multivariate approaches developed for assessing non-focal distributed change in activation pattern (for a review, see Worsley *et al.* 1995) and several statistical tests with weak FWE control. Experimentally induced distributed changes consist of changes in a subset of voxels in the search volume. In principle, these changes may be detected by univariate voxel statistics if sufficient statistical power is available. Since tests with strong FWE control aim at localization of experimental effects these are (usually) less powerful than tests with weak control for testing the existence of any experimental effect. At present, the statistical power of FNI studies is most often unknown and there are indications that some FNI studies are lacking in power. In the context of low statistical power, omnibus tests with weak FWE control are interesting since they can (often) detect whether there is any experimental

effect at all with a greater sensitivity than tests with strong control (Friston *et al*. 1994*b*, 1996).

One of the earliest proposed omnibus tests was the γ^2-statistic (Fox *et al*. 1988; Fox & Mintun 1989) based on the idea that activations may increase the number of local maxima outliers, implying increased kurtosis of the distribution of local maxima. The γ^2-statistic is therefore relatively more sensitive to focal changes than other proposed omnibus tests, making it potentially less useful in detecting distributed changes. No theoretical basis for the specificity of the test has so far been derived (Worsley *et al*. 1995), but the specificity of the γ^2-statistic has been investigated empirically. Another early proposal was the suprathreshold exceedence proportion test, that is, the proportion of voxels of the search volume that passed a given but arbitrary intensity threshold (Friston *et al*. 1990). Originally, it was assumed that the voxels in the statistic image were independent, leading to an underestimation of the variance of the statistic, increasing the false-positive rate compared to the nominal level. This has recently been corrected and the correct limiting distributions derived and described (Worsley & Vandal 1994; Worsley *et al*. 1995). As an alternative, Worsley *et al*. (1995) proposed the S-statistic for comparing distributed differences between two states. The S-statistic is equal to the squared Z-score averaged over the search volume representing a mean sum of squares test. If the search volume is large enough then the distribution of the S-statistic is well approximated by a χ^2 distribution with estimable effective degrees of freedom. This approximate result is based on the theory for smooth stationary Gaussian RFs (cf. below; Worsley & Vandal 1994). The detection sensitivity of the S-test is optimized when the smoothing kernel matches the underlying activation (Worsley & Vandal 1994) representing a special case of the matched filter principle. The sensitivity of the S-statistic is also biased towards distributed signals and there is a corresponding lack of power in detecting focal changes (Worsley *et al*. 1995). Since most effects of interest in FNI are focal in nature, a suitable univariate approach may be preferred when sufficient statistical power is available. However, if there is prior information indicating that the signal is likely to be subtle and distributed in nature, the S-statistic may yield a gain in detection sensitivity. However, a univariate test with strong control can detect distributed changes and omnibus tests with weak control can pick up focal changes. Which is the most suitable test in a given situation is a question of what kind of experimentally induced changes are expected, the statistical power available and what types of changes a certain test is most sensitive to, that is, the detection bias of the test statistic. It is worth remembering that the null hypothesis is the same whether the testing procedure has strong or weak FWE control, only the localizing power differs. Lastly, it should be noted that the S-statistic has recently been generalized and used in a multivariate linear modelling framework (Worsley *et al*. 1997*b*). This is further discussed in the multivariate section of the companion paper (Petersson *et al*., preceding paper). In addition, the concept of set-level inference was recently introduced (Friston *et al*. 1996). Set-level inference represents an omnibus test based upon the number of activated clusters. Here the number of activated clusters is defined by two thresholds, an intensity or height threshold for the test statistic, defining the voxels contained in a suprathreshold cluster and a threshold for the clusters size. A distributional approximation for this test, based on stationary smooth Gaussian RFT and some additional assumptions (high thresholds and a distributional approximation for the number of voxels in a suprathreshold cluster), has been derived (Friston *et al*. 1996).

(d) *Statistical power in functional neuroimaging*

Several authors have pointed to the problem of statistical power in FNI (Andreasen *et al*. 1995; Grabowski *et al*. 1996; Vitouch & Glück 1997; Van Horn *et al*. 1998). Most studies are interpreted with an informal notion of power based on what has been detected in previous studies and these considerations have been used to determine the sample sizes. The need to specify the spatio-temporal characteristics of the signal completely (i.e. the alternative hypothesis) has precluded general power analyses. Lack of statistical power may add a complication to interpretation of the results. Lack of power can be reflected in limited reproducibility of results across similar experimental studies. For example, an activation that does not reproduce across studies may be the result of a false-positive detection, an effect specific to the particular subjects studied or a false negative due to lack of statistical power. There are some indications that the number of subjects commonly used in FNI studies may need to be increased to obtain reproducibility even at nominally high significance levels (Strother *et al*. 1997). A natural strategy for increasing statistical power is to increase the sample size (Vitouch & Glück 1997; Van Horn *et al*. 1998). In the case of fixed effect studies and fMRI this is less of a problem. However, in random effects analyses (cf. Petersson *et al*., preceding paper), the number of subjects is a major determinant of statistical power. In particular, when the interindividual variability is large relative to the intra-individual variability, it is still the number of subjects included in the study that is most important when seeking to generalize the results to the population sampled (even though many scans may be available per subject).

4. RANDOM FIELD (RF) APPROACHES

RFT has been extensively developed and used in assessing the significance of signals present in FNI data. The RFT approach provides a way of handling the problem of multiple non-independent comparison in large FNI data sets. Essentially, the RFT approach allows for spatial correlation between voxels in the statistical image when correcting for multiple comparisons, thereby improving on the Bonferroni correction. With a single threshold test, where all suprathreshold voxels are declared activated, the omnibus hypothesis is rejected as soon as a voxel with maximum value exceeds the threshold. Thus, when a single threshold test is used, the distribution of interest is that of the maximal value (global maximum) in the random field. RFT is used to deduce an approximate distribution characterizing the distribution of the global maximum, using results on the expectation of topological characteristics of excursion sets, the set of points with values above a given threshold.

Examples of topological characteristics that have been very useful in smooth RFT are the Euler characteristic and the Hadwiger characteristic. At high thresholds u, the Euler characteristic counts the number of connected components minus the number of holes plus the number of hollows (Worsley 1996*b*). For higher thresholds, the holes and hollows occur with low probability, that is, they tend to disappear and the Euler characteristic will approximately count the number of local maxima. For even higher thresholds near the global maximum M, the Euler characteristic counts 1 if $M \geqslant u$ and 0 otherwise, such that the expected Euler characteristic approximates the p-value of M. This p-value approximation is used for significance evaluation of the local maximum statistic. The fundamental results for Gaussian RFs (GRFs) can be found in the work of Nosko (1969), Hasofer (1976) and Adler (1981).

Applications of RFT to FNI data were first described in the seminal papers of Friston *et al.* (1991) and Worsley *et al.* (1992). They presented methods of testing for the intensity of the signal using the maximum test statistic in statistical images in two and three dimensions, respectively, using smooth GRFs. GRF theory was later used to derive approximate distributions for the size of the largest suprathreshold cluster, where clusters are defined by a primary threshold at predetermined height (Friston *et al.* 1994*b*). One reason for using the size of clusters above a given threshold is that this may be a more sensitive test for spatially extended activations in FNI data, since localizing power at the voxel level is not demanded. The local maximum statistic is not always the most sensitive way of characterizing the signal in FNI data. Further work provided the means of computing the probability of getting N clusters of size greater than K in a given search volume (Friston *et al.* 1994*b*). More recently, with some additional assumptions, an approximate distribution for the conjoint distribution of the size K and the maximum value in a cluster were also derived (Poline *et al.* 1997). The search for local maxima over different scales has also been made possible in the GRF context (Siegmund & Worsley 1995; Worsley *et al.* 1996*b*).

RFT has been further developed to accommodate other statistic fields, such as t-, χ^2- and F-fields (Worsley 1994; Cao 1999). These non-GRFs are constructed from GRFs. The χ^2-fields are constructed as the sum of squared GRFs, F-fields are constructed as the quotient between two χ^2-fields and t-fields are constructed as the quotient between a Gaussian and the square root of a χ^2-field. In general, there are fewer results for non-GRFs. However, these fields are fundamental when the local variance cannot be considered constant across the volume (t-fields) or when testing for a number of effects at each location (χ^2- or F-fields). Recently, RFT has been extended to include theory for different covariance fields, so-called autocorrelation, cross-correlation and homologous correlation fields (Cao & Worsley 1999; cf. § 3(e) in Petersson *et al.*, preceding paper).

In the original work of Worsley *et al.* (1992) it was assumed that the excursion sets did not touch the boundary of the search volume, limiting the results to infinite search volumes. The results are reasonable approximations for finite search volumes provided the search volume is large relative to the surface area and the smoothness of the field. These constraints have recently been relaxed and a unified approach described; the RF is transformed to an isotropic RF and then the volume, surface area and diameter are estimated in the 'resel space' (resel, resolution element; Worsley *et al.* 1996*a*). In this section, we review the assumptions on which the use of smooth RFT is based and discuss some aspects of the robustness of these methods. The more general assumptions are presented first. Tests that require more constraining assumptions are presented later.

RFs are often introduced as underlying smooth RFs, which are assumed to be well approximated by the lattice representation, the discrete (voxellated) statistic image process. The RFs in question are the component fields, that is, the normalized error fields. Since the variance field is unknown the residual images cannot be normalized to obtain realizations of the component field process (unless the variance can be assumed constant across the volume and considered known by the high degrees of freedom implicit in global pooling). Instead, the residual images are viewed as approximate realizations of the error field process and the estimated variance is used to standardize the residual images to obtain estimates of the component fields. Alternatively, one may take the opposite perspective that the discrete statistic image is approximated with a smooth RF (of the same size and smoothness). Note that, for GRFs, the component field is the GRF itself, but for t-, F- and χ^2-fields the component fields are the GRFs that are used in generating the t-, F- and χ^2-fields, respectively. Since the smoothness is defined on the component fields, the smoothness is not simply related to the autocovariance function (ACF) of the field itself but to the ACF of the component fields.

(a) *Stationarity assumptions*

Stationarity (also called translational invariance) requires that the covariance structure of the RF does not depend on its location in the field. In other words, it is assumed that the ACFs of the unobservable component fields approximated by normalized error fields, from which the t-, F- or χ^2-fields are constructed, are not dependent on their location in the field. For fMRI data with many scans acquired per subject, it may be possible to detect departures from stationarity (given adequate statistical power) since it is possible to estimate a map of local smoothness. Investigating the robustness of the RFT approximations and results with respect to departures from the stationarity assumption in simulated fields is of some importance. However, some results for RFs with local non-stationarities are under development (Worsley *et al.* 1999).

It should be noted that voxel variance estimation is a characteristic of statistical parametric mapping (Friston *et al.* 1991, 1994*b*, 1995). To leverage results for GRFs, the resulting t-statistic image can be transformed pointwise into a Z-statistic image via a probability integral transform. The resulting Z-image is known as a Gaussianized t-statistic image. After a local correction and with a sufficient number of degrees of freedom, the Z-image can be considered well approximated as a stationary GRF (Worsley *et al.* 1992). Sufficient image smoothing should ensure that the local covariance structure is

approximately stationary. However, there are signal detection issues and other problems associated with too much filtering (cf. § 2). In general, RFT models the stationary covariance component, which in practice is often local. In addition, there may be a non-local, non-stationary covariance structure (e.g. distant voxels may be correlated). Some empirical results indicate that the p-values for local maxima may not be greatly affected by local non-stationarity. In contrast, the p-values for cluster sizes were more sensitive to this non-stationarity (Worsley *et al*. 1999).

(b) *Regularity conditions, lattice representation, sampling issues and smoothness estimation*

Another general assumption in the application of smooth RFT to discrete statistical images is that the statistical image can be considered as a well-sampled version of the smooth RF or, conversely, that the smooth RF is a good approximation of the statistic image. All the tests described that are applied to the discrete statistic images require that this assumption is a reasonable approximation. In theory, the tests are applicable to smooth RFs requiring that some regularity conditions be fulfilled. In particular, the smooth RF should be differentiable (i.e. L^2 differentiable; cf. Adler 1981; Worsley 1994, 1995). In practice, these assumptions are reasonable for a sufficiently large number of degrees of freedom and a sufficient amount of image smoothing during pre-processing (Worsley *et al*. 1992, 1996*a*). In general, the frequency spectrum of the stochastic process is not bounded, but in experimental data the observable frequencies are limited (i.e. only the frequencies below half the frequency of the sampling process are observable by the Shannon–Nyqvist sampling theorem). The sampling issue becomes particularly important in the context of 'smoothness' estimation.

Smoothness estimation amounts to the estimation of a parameter related to the spatial ACF. Note that no assumption is made on the shape of the ACF, the only requirement being that it is twice differentiable at the origin. It can be shown that this second derivative is always negative and equals minus the variance of the derivative of the process (Adler 1981). It should also be noted that the smoothness estimation in RFT relates to the spatial autocorrelation of the statistic image, which is described by the smoothness parameter and this is different from image smoothing or filtering applied to the data during pre-processing. Specifically, the ACF of interest is that of the statistic image and this is in principle different from the ACF of a smoothing filter applied to the data during pre-processing. The smoothness estimate is based on the determinant of the variance–covariance matrix of the partial derivatives of the component RF (Worsley 1996*a*). If it is assumed that the principle axes of the covariance function are aligned with the axes of the statistical image then all off-diagonal elements of this matrix are zero and only the diagonal elements need to be estimated. However, this assumption is not required (Worsley *et al*. 1992). The approach proposed by Worsley *et al*. (1992) was limited to cases where the excursion set did not touch the boundary of the search region and in Worsley *et al*. (1996*a*) this assumption is relaxed.

The sampling or voxel size issue is an important constraint in relation to the smoothness estimation in the statistic image and this issue has become increasingly important with fMRI data, which in general have a better spatial resolution than PET data. Smoothness estimation has been the subject of several investigations, particularly for images with limited spatial auto-correlation. The smoothness of the statistic image is a central parameter in the assessment of the probability of occurrence of a maximum above a given value and for the approximate distribution of the size of regions above some threshold. In case the estimated smoothness, measured in FWHM (full width at half maximum), is less than three times the voxel size it cannot be expected that the RF is well sampled. In other words the statistic image is not well approximated by a smooth RF. It has been observed that the smoothness is overestimated at low FWHM (Forman *et al*. 1995; Xiong *et al*. 1995; Ledberg *et al*. 1998). This leads to a conservative test and loss of power for the suprathreshold cluster size statistic. One may think that this also would lead to an increased false-positive rate of the local maximum statistic, but at small smoothness values this effect is counteracted by the fact that the smooth RF has features at the subvoxel level which are not evident in the statistic image. Specifically, the subvoxel resolution structure becomes increasingly likely as the smoothness of the RF decreases (which is just another way of saying that as smoothness decreases the RF is less well sampled). This effect will (at some point) more than compensate for the effect of overestimated smoothness. It has been observed (Worsley 1997) that the RFT correction of the local maximum statistic is conservative and may in certain cases be more stringent that the Bonferroni correction (at low smoothness). Instead, Worsley (1997) proposed a continuity correction in the special case of a GRF with a Gaussian ACF to bridge the gap between the case of independent voxels (for which the Bonferroni correction is accurate) and the case of large smoothness : voxel size ratio for which the RFT is accurate. There are several other suggestions to correct the inflated measure of smoothness at low FWHM, particulary in the context of suprathreshold cluster size tests (Forman *et al*. 1995; Xiong *et al*. 1995). Note that, unless the estimated smoothness in FWHM is greater than three times the voxel size, the good lattice representation condition required by the RF approach cannot be considered fulfilled. Instead, the smooth RF approach becomes increasingly inadequate at lower smoothness and other approaches to statistical inference have to be applied.

Xiong *et al*. (1995) derived an estimate of the smoothness parameter based on heuristic arguments. At low smoothness, this estimate shows less discrepancy with the expected value than the estimate used for well-sampled RFs. Xiong *et al*. (1995) also proposed an empirically estimated correction of the approximate distribution of suprathreshold cluster sizes. Both the estimated smoothness and the estimated distributional correction introduce variability in the estimated p-values (Poline *et al*. 1995). Xiong *et al*. (1995) used these estimates in combination with a GRF approach. This lacks theoretical foundation since the theory for smooth RF requires that the condition of good lattice representation be reasonably fulfilled. However, the approach was validated on phantom data

and on simulated two-dimensional (2D) Gaussian white noise processes convolved with a Gaussian kernel giving rise to a Gaussian ACF (which seems to be a good approximation for appropriately filtered fMRI data). There was fair agreement between the estimated and observed results (Xiong *et al.* 1995). Related work has been carried out by Forman *et al.* (1995). They derived a modified estimate of the smoothness parameter that was used in combination with a Monte Carlo approach (cf. § 7(c)). The modified estimate shows less discrepancy to the expected value than the estimate used for well-sampled RFs. However, as already indicated, smoothness is a concept from the theory of smooth RFs and the use of this concept in situations when this theory becomes increasingly inadequate requires careful validation.

Recent developments in RFT are related to the issue of continuity corrections. Results have been derived for GRFs while the results for *t*- and *F*-fields are still unknown (K. Worsley, personal communication). In the case of GRFs, the results are not much different from the minimum of the Bonferroni and the Adler threshold (K. Worsley, personal communication). The results from RFT are applicable under the assumption of good lattice representation. To achieve this, one possibility is to filter the data and another is to use a supersampling interpolation approach (Friston *et al.* 1996). If the underlying biological signal can be considered continuous, the sampled volume can be interpolated to ensure that the FWHM is three times larger than the voxel size, a value for which the original smoothness estimation is accurate. The interpolation kernel used should reflect the *a priori* hypotheses about the smoothness of the underlying process, although it is unlikely that the choice of the interpolation kernel will make very much difference in actual analyses. In the case of fMRI data, the best way is sinc interpolation, but this does not quite maintain stationarity. Instead, the highest possible frequency should first be filtered out and then sinc interpolated so that exact stationarity is preserved (K. Worsley, personal communication).

(i) *Robustness of the smoothness estimate*

The estimation of the smoothness parameter should be independent of experimentally induced effects. Smoothness estimation should generally be made on the residual images. Worsley *et al.* (1992) estimated the smoothness on the difference images minus the average difference image divided by the global variance estimate (i.e. the residuals). Under the assumption that the global variance estimate can be regarded as the true global variance (because of the large number of degrees of freedom implicated), these difference images then have the same variance structure as the underlying component fields. In a general linear model context, the residual image corresponds to the original image data minus all model effects and after standardization to unit variance at each voxel. Since the estimated voxel variance is used to standardize the residual fields resulting in additional noise in the estimated component fields, a correction for the degrees of freedom is necessary (Worsley *et al.* 1992; Worsley 1996*a*).

It is important to note that the smoothness estimate itself is the realization of a random variable (Poline *et al.* 1995). Poline *et al.* (1995) gave an approximate expression for the variance of this estimator. When estimated on a single image, the variability of the resulting corrected *p*-value is found to be moderate (i.e. s.d.$(p)/E[p]$ is of the order of 20%). Averaging the estimation over several residual images can reduce this variability. In fMRI where the number of volumes per subject can be very large, it is necessary for practical reasons to limit the number of residual images on which this estimation is performed. The estimate is expected to be very stable when performed on more than 60 residual images. However, the appropriateness of the estimate depends on the adequate fit of the statistical model at a sufficient number of voxels.

(c) *The Gaussian assumption and Gaussianized t-fields*

The Gaussian assumption is fundamental to the results derived by Adler (1981) and Worsley *et al.* (1996*a*,*b*). The RFs are assumed to be smooth stationary standard GRFs. That is, the marginal distributions are assumed to be multivariate Gaussian with zero mean, unit variance and with the same covariance structure for any pattern of locations regardless of position within the field. Even if the expected value for the Hadwiger characteristic is given for any RF (Worsley 1995), the actual computation of this requires a known parametric form of the marginal distributions of the RF. The corrections proposed in the unified approach (Worsley *et al.* 1996*a*) assume either Gaussian, χ^2-, *F*- or *t*-fields (see also Worsley 1994). In some sense, all these are based on the multivariate normal assumption since χ^2-, *F*- and *t*-fields are constructed from GRFs.

The multivariate Gaussian assumption is difficult to verify for FNI data. However, sufficient image smoothing, a sufficient number of effective degrees of freedom and the multivariate central limit theorem (Billingsley 1995) lend support to the approximation. On the other hand, it is fairly simple to test for normality in each voxel of the residuals. These tests have so far not indicated significant departure from the Gaussian assumption (Poline & Mazoyer 1993; Holmes 1994; Aguirre *et al.* 1998;), but the issue is to test for the multivariate Gaussian behaviour of any set of voxels. As suggested by the central limit theorem (Billingsley 1995) and the fact that the PET reconstruction process (filtered back projection) implies summations of a large number of Poisson distributed count data, the regional activity in reconstructed PET images is expected to be approximately Gaussian distributed. Further, it has recently been shown that, as the projection counts approach infinity, the reconstructed images will become multivariate and normally distributed, given that PET projection data are Poisson distributed (Maitra 1997). However, in the case of low-count PET data there may be departures from normality. An attractive alternative is to cross-validate the smooth RF approach through comparison with non-parametric procedures (Nichols & Holmes 1999).

Empirical validation of the RFT results for *t*-, *F*- and χ^2-fields with large degrees of freedom is computationally demanding. Many GRFs have to be simulated and then combined, this process being replicated a great number of times to obtain a reasonable estimate of the number of level crossings for high thresholds, so that the tails of the

probability distributions are well estimated. In this context, it should be noted that Gaussianized *t*-fields are fundamentally different from GRFs (Worsley *et al*. 1992, 1996*a*; Worsley 1994; Worsley & Vandal 1994; Cao 1999), since in general the marginal distributions are not multivariate Gaussian. In principle, this implies that Gaussianized *t*-fields cannot be directly simulated by GRFs. Approximating Gaussianized *t*-fields with GRFs, without appropriate corrections will lead to an increased false-positive rate of the local maximum statistic for small to moderate degrees of freedom (i.e. 10–40 d.f.; cf. Worsley 1994, 1997; Worsley *et al*. 1996*a*), while the opposite is the case for the suprathreshold cluster size statistic (Cao 1999). For example, at 40 d.f. and a search volume of 1000 cc, a nominal 5% false-positive rate of the local maximum statistic corresponds to an actual test size of 6.9% (i.e. the true false-positive rate). This implies that at low degrees of freedom it is preferable to use the results on *t*-fields directly rather than attempt to Gaussianize the *t*-field.

(d) *Global variance pooling*

The assumption of unequal voxel variance across the brain volume has not always been rejected in PET data (Worsley *et al*. 1992, 1996*a*). However, there is evidence that this assumption is not generally tenable (Holmes *et al*. 1996; Worsley *et al*. 1996*a*), particularly for fMRI data (Worsley *et al*. 1997*a*). This would imply the use of *t*-field results but, more often than not, these will be catered for by GRF theory after Gaussianization when the number of effective degrees of freedom is sufficiently large. As a curious observation, the global pooling of variance may have adaptive signal detection properties, that is, possible underestimation of variance in activated areas in combination with overestimation in non-activated areas, enhancing the contrast between activated and non-activated regions. There are some indications that the stationary variance assumption may be less problematic in relation to the false-positive rate for PET data (Grabowski *et al*. 1996; Worsley *et al*. 1996*a*). In an empirical study, no false-positive activations were observed and the replicability (as measured in this study) was no lower when compared to other methods investigated (Grabowski *et al*. 1996). The pooled variance approach seems to tolerate variations in the voxel variance of *ca*. 8%, while the local approach seems to tolerate variations in the variance between experimental conditions of *ca*. 6% reasonably well (Worsley *et al*. 1996*a*). In the case of low degrees of freedom it would be attractive to be able to use the pooled variance estimate, thereby obtaining a more reliable variance estimate. There are some indications that using a pooled variance estimate may give results that are more reproducible compared to using voxel variance estimates (Hunton *et al*. 1996; Strother *et al*. 1997). Since a statistic image is constructed by dividing the estimated signal image by the estimated standard deviation, any noise in the variance image is propagated to the statistic image. In particular, this is a problem for *t*-statistic images with low degrees of freedom, even though the signal image may be smooth. Since the variance image is noisy at low degrees of freedom, this implies instability in the location of the local maximum statistic (Taylor *et al*. 1993; Grabowski *et al*. 1996; Holmes *et al*. 1996; Hunton *et al*. 1996). The properties of such noisy statistic images are not well approximated by those of a smooth RF with the same smoothness, since the smooth RF may have subvoxel resolution features. The net result is that the RF approach is conservative for voxel level inference at smaller degrees of freedom. A preferable strategy in this case would be to pool the variance estimates locally, effectively smoothing the variance image. However, this requires a non-parametric approach to statistical inference (Holmes *et al*. 1996).

(e) *The high threshold assumption*

As previously noted, the expected Euler characteristic approximates the probability of detecting a local maximum above a given threshold, when the given threshold is high (Worsley 1996*b*). Thus, it is necessary to use high thresholds in order for the *p*-value used for significance evaluation to be a good approximation. The approximation appears to be accurate for high thresholds such that the *p*-value is less than 0.2 (Worsley *et al*. 1996*a*). At lower thresholds the expected Euler characteristic approximates the expected number of local maxima. In fact, the approximation of the expected number of local maxima by the expected Euler characteristic turns out to be a very good approximation (Adler 1998).

(f) *Cluster size tests*

One may argue that suprathreshold cluster size tests are useful when areas of neural activity will give rise to signal changes in contiguous voxels. In this situation, the suprathreshold cluster size tests may be more sensitive than the local maximum statistic, since cluster size tests do not have localizing power at the voxel level. To derive the approximated distribution for the cluster size above a given threshold, that is the probability of getting c or more clusters of size k or larger, additional assumptions have to be made (Friston *et al*. 1994*b*). First, to derive the approximate distribution of the suprathreshold cluster size under the null hypothesis, the occurrence of clusters is approximated by a Poisson process, with the expected number of clusters given as the expectation of the Euler characteristic (Adler 1981, theorem 6.9.3). An approximation of the distribution of cluster size for high thresholds was given by Nosko (1969), parameterized to match the known expected region size. Combining these two results yields the distribution of the maximal suprathreshold cluster size. Simulation studies indicate that the empirical results are well approximated by the theoretical results (Friston *et al*. 1994*b*). Simulations have only been carried out at one threshold defining the clusters and for relatively large smoothness values. Since boundary corrections have so far not been developed in this context, it is important that the search volume is large compared to the smoothness of the statistical image. However, the results are likely to be conservative, since if the volume is small then the boundaries will tend to reduce the size of clusters. The assumptions for the set level of inference are essentially the same (Friston *et al*. 1996).

(i) *The bivariate suprathreshold conjoint test*

The bivariate conjoint test was designed to detect signals that are either focal or have a large spatial

extent. It can be viewed as a way to correct for a two-tests procedure (maximal intensity and spatial extent). The derivation of an approximate bivariate distribution for the maximum value and the suprathreshold spatial extent of a cluster defined by thresholding require the additional hypothesis of a Gaussian ACF for the GRF (Poline *et al.* 1997). It uses the form of the distribution of the peak height above a threshold and an approximation of the shape of the RF around local maxima. Simulation studies indicate that the derived approximate distribution works reasonably well (Poline *et al.* 1997). The Gaussian ACF approximation is expected to perform well in the vicinity of local maxima; however, its robustness to variability of the assumed kernel shape remains to be demonstrated.

(g) *RFT: a short summary*

To summarize, the RFT has proved versatile in testing a number of statistics, such as the local intensity (maximum), the size of regions above a given threshold or the number of regions larger than a given size. The approach makes no or little assumption about the shape of the covariance function but assumes a stationary multivariate Gaussian distribution for the component fields, generally assessed on standardized residual images. Automatic procedures to verify such assumptions may take the form of multivariate analysis and tests on the standardized residuals (Poline *et al.* 1998). The smooth RFT approach has been extensively validated on simulated data, which fulfil the assumptions of the approach and, hence, assess the theoretical approximations made. Validation using simulated data fulfilling the assumptions does not indicate the robustness of the method and it is seldom investigated how well a given FNI data set fulfils the assumptions of the RFT approach, the principle problem being the generation of enough representative real data characterizing a given null hypothesis. However, empirical studies using real null data have been reported indicating that the RFT approach gives reasonably accurate results (Aguirre *et al.* 1997; Zarahn *et al.* 1997). In addition, investigations of the robustness and characterization of inherent limitations of the RFT approach, with respect to the various assumptions and parameters, have been carried out, for example degrees of freedom (Worsley *et al.* 1992), smoothness estimation (Poline *et al.* 1995) and variance heterogeneity (Worsley *et al.* 1996*a*). Alternatively, non-parametric methods may be used as benchmarks for cross-validation of the RFT approach (Nichols & Holmes 1999).

5. SCALE SPACE APPROACHES

The matched filter theorem indicates that a matching filter optimizes detection sensitivity (cf. § 2(b)). This also implies that a particular choice of filter biases the detection sensitivity towards signals of a certain shape and size. For sufficiently different signals the filter choice will be suboptimal and reduces detection sensitivity. Note also that the detection sensitivity of the local maximum statistic may in general be more sensitive to an optimal filter choice than, for example, the suprathreshold cluster size statistic. Empirical studies have indicated that filtering or image smoothing may modulate both the observed effect size and statistical power (Van Horn *et al.* 1998), while theoretical work indicates that an inadequate filter choice can drastically reduce the detection sensitivity (Poline & Mazoyer 1994*b*; Worsley *et al.* 1996*b*). In addition, experimentally induced signals at different locations of the brain often differ both in size and shape, precluding the existence of a single optimal filter. A multifiltering approach has been suggested (Poline & Mazoyer 1994*c*) as a solution to these problems. The multifiltering approach commonly uses a scale or parameter family of filters (e.g. isotropic Gaussian kernels parameterized by FWHM or filter width). This creates a scale space, that is, a scale dimension is added to the common Euclidean 3D space and the resulting four-dimensional scale space is then searched for signals. The multifiltering approach is less biased with regard to detection sensitivity than a single-filter approach; it may increase the overall detection power (unless prior information on the signal size or shape is available) and provide information on the actual signal size. Multifiltering implies an extended search compared to the single-filter approach; therefore, the critical thresholds need to be adjusted to control the false-positive rate appropriately, accounting for the additional non-independent multiple comparisons over the scale dimension.

The multifiltering approach of Poline & Mazoyer (1994*c*) depends on the prior selection of a finite number of filter widths. Poline & Mazoyer (1994*a*,*b*,*c*) also suggested combining the multifiltering with a hierarchical decomposition technique and a Monte Carlo approach to statistical inference (cf. § 7). An alternative scale space approach for PET data has been described by Worsley *et al.* (1996*b*). This approach solves the inference problem for a smooth stationary GRF, assuming approximately homogenous voxel variance and using the pooled variance estimate. In addition, the spatial ACF is initially approximated as a Gaussian estimate with subsequent non-Gaussian corrections. Extensions of this approach to *t*-fields appear theoretically difficult (Worsley *et al.* 1997*a*; Worsley 1999). Worsley *et al.* (1996*b*) suggested that a work around may be to create a *t*-image at the highest resolution, Gaussianize by pointwise transformation and then smooth this Z-image to the various resolutions. The notion is that the smoothed Gaussianized *t*-field may be adequately approximated by smooth stationary GRFs. Since the images are supposed to be sampled versions of a continuous process, an interpolating supersampling approach may improve the approximation. This latter approach may be applicable to fMRI data. Since the extent of the inherent spatial ACF of fMRI data is commonly limited compared to PET data (Forman *et al.* 1995; Frackowiak *et al.* 1997), the ACF of the smooth statistical image should be well approximated by a Gaussian ACF. However, recent developments in RFT (Worsley 1999) extend the scale space approach to χ^2-fields. This approach has applications to fMRI data with spatially varying haemodynamics (Worsley *et al.* 1997*a*) and offers a parametric alternative to the approach suggested by Bullmore *et al.* (1996).

In summary, there may be situations in which the loss of overall detection sensitivity with the single-filter approach due to variable signal size may be comparable

to the cost of testing a range of filters. However, to take advantage of this requires prior information. When there is limited prior information on the shape or spatial extent of the activation, a scale space approach allows a less biased detection sensitivity and the overall detection sensitivity can increase (Poline & Mazoyer 1994*b*,*c*; Worsley *et al.* 1996*b*). The scale space approaches suggested so far have used 3D Gaussian convolution kernels that may not be optimal for all regions of the brain. Future developments may include nonlinear scale space approaches, allowing for adaptive or anisotropic filtering optimized for the underlying anatomical structure. Different measures in linear scale space have been suggested, accounting for different anatomies (Coulon *et al.* 1997). It should be noted that filtering increases the detectability at the expense of resolvability, that is, filtering may slightly dislocate activations or fuse adjacent activations. However, fused activations may be detected as bifurcations in scale space (Worsley *et al.* 1996*b*). Currently, no non-parametric scale space approaches have been proposed, but in principle could offer the relaxation of specific assumptions, including Gaussian ACF and equal voxel variance.

6. NON-PARAMETRIC APPROACHES

As previously described, statistical hypothesis testing requires knowing the distribution of the test statistic under the null hypothesis. There are situations where the assumptions that justify a specific form or parametric family of null distributions are untenable or difficult to verify. In those cases non-parameteric tests are an important alternative. Instead of assuming a particular parametric form, non-parametric tests derive the null distribution empirically. Fisher (1935), arguably the progenitor of modern statistics, introduced the randomization test early this century as a fundamental tool of statistical inference. However, the computational demands of the randomization test were then too great for all but the simplest problems and, hence, they enjoyed little use until recent advances in computing made them more accessible. In this section we will review non-parametric tests, the permutation and randomization test in particular and indicate how the non-parametric approach has been applied in FNI.

The hypothesis-testing framework is the same as described above. A statistical model is defined, the parameters estimated, a null hypothesis defined and a test statistic calculated. The distinction between parametric and non-parametric tests lies in calculation of the p-value. For the parametric test the assumptions provide a parametric or analytical null distribution, while for the non-parametric test the null distribution is determined from the data. The fundamental idea is that of exchangeability under the null hypothesis. Data are exchangeable if permutation of data or its labels does not change the distribution characterizing the experiment. Specifically, a set of random variables is exchangeable if (and only if) their joint distribution is invariant with respect to permutations of the random variables. Thus, if the data are exchangeable under the null hypothesis, each test statistic value computed from each permutation of the data is equally likely. A distribution of equally likely statistical values is then used to assess the extremity of the observed statistic value: the p-value is the proportion of the permutation distribution greater than or equal to the observed statistic. An important characteristic of randomization or permutation tests is that they are exact if the assumption of exchangeability is fulfilled under the null hypothesis.

There are two possible justifications of exchangeability, either from randomization in the experimental design or by post hoc assumption. Obviously the former is preferred, but it is not always possible to achieve. When there is no randomization the test procedure is called a permutation test. (We will use the term permutation test from here on unless a distinction is needed.) In some cases there may be implicit sources of non-exchangeability even under the null hypothesis. For example, if there is a temporal trend in the data, the data would not be exchangeable, even if no experimental effect or activation is present, since subtracting early data from late data will create a larger expected activation than a more temporally balanced subtraction. Another important example of non-exchangeability under the null hypothesis is the temporal autocorrelation of fMRI data (Zarahn *et al.* 1997). It should be noted that, in a design with few replications or a multisubject analysis with few subjects, there are few ways to permute the data and, hence, the permutation distribution will be discrete or coarse. For example, if there are just 20 possible permutations, then the smallest possible p-value is $1/20 = 0.05$. A practical limitation of the permutation test for PET data is the need for numerous permutations.

Another key strength of non-parametric methods is that they allow the use of non-standard test statistics with unknown parametric forms. Holmes *et al.* (1996) demonstrated this with their permutation test for PET data. In particular, by using the distribution of maximal statistics their test gave strong control of the image-wide false-positive rate without any RF assumptions (cf. § 3(b) for a description of strong control). They also used the smoothed variance t-statistic, called the pseudo t-statistic. Since PET data often have rather limited degrees of freedom, this will result in poor variance estimation, resulting in a noisy variance image. The noisy variance image translates into a noisy t-image, even if the mean difference image is smooth. The pseudo t-statistic smoothes the variance image locally, decreasing the uncertainty of the variance estimates (possibly) at the cost of increased bias. Whilst there are RFT results for Gaussian, t-, F- and χ^2-images, there are none available for the pseudo t-image. However, a non-parametric approach makes it possible to use the pseudo t-statistic and assess the significance of local maxima in pseudo t-images. Recently, this work has been expanded to encompass more experimental designs, as well as the supra-threshold cluster size statistic (Nichols & Holmes 1999).

Application of the permutation test to fMRI is hampered by temporal autocorrelation. A possible approach to temporally correlated data is to model the temporal autocorrelation appropriately. For example, Bullmore *et al.* (1996) used a parametric first-order autoregressive model to account for the autocorrelation and then decorrelated or whitened the data. Since null data were found to fit known parametric null distributions poorly, a non-parametric method was used. Since

whitened data are approximately exchangeable, permutation distributions were created by permuting the whitened data. (The whitened data are not perfectly exchangeable because the temporal autocorrelation parameter used to whiten the data is just an estimate, instead of true, unknown autocorrelation.) In addition, the assumption that the null distribution is the same for all voxels was introduced, the results pooled over the whole search volume and, hence, only ten permutations were calculated at each voxel. This assumption is questionable, particularly in the light of more recent work (cited below). Furthermore, the central problem of multiple comparisons was not addressed, except to report the image-wide false-positive rate in pixels. Finally, Bullmore *et al*. (1996) presented a means of assessing the pixelwise power of the method, which is a rare finding in the FNI literature.

Locascio *et al*. (1997) suggested time-series analysis in the time domain in combination with resampling methods, following an approach similar to Bullmore *et al*. (1996). In brief, temporal autocorrelation was estimated using a parametric model and, based on this estimate, the data were whitened; the whitened data were then submitted to a permutation test for statistical inference. Specifically, the parametric model used was a combination of the general linear model and an autoregressive model of arbitrary order; through model selection techniques they found that different autoregressive orders were necessary in different voxels. In order to handle the multiple comparisons problem and gain strong control of the imagewise false-positive rate, they used the maximal *t*-statistic and a permutation test.

Finally, it has been noted that exchangeability of the experimental labels is sufficient for a valid permutation test. For example, Liu *et al*. (1998) analysed fMRI data from an 'oddball' paradigm study, where stimuli are presented rapidly and the target stimuli are presented infrequently. Since the pattern of target stimuli was random, exchangeability of the experimental labels of 'target' or 'non-target' were guaranteed regardless of the temporal autocorrelation structure. This approach also applies to block-design fMRI studies when the blocks are randomized.

In general, non-parametric approaches make minimal assumptions on data and offer great flexibility in the choice of test statistics. In fact, any test statistic may be used; Bullmore *et al*. (1996) used the experimental frequency power and a non-parametric approach because there was no satisfactory parametric form available. Likewise, Holmes *et al*. (1996) used the pseudo *t*-statistic and a non-parametric approach because there was no parametric form at all. The cost of non-parametric tests is a computational one, though for PET data this is not excessive (Nichols & Holmes 1999). In addition, since there may be implicit sources of non-exchangeability even under the null hypothesis, care has to be taken to ensure that data are at least approximately exchangeable.

7. MONTE CARLO APPROACHES

In this section the Monte Carlo approach to statistical inference is introduced and some necessary conditions for this approach to be valid are discussed. Three different Monte Carlo methods that have been applied to FNI data are described. In general, Monte Carlo methods have changed the field of statistics, taking problems that were intractable and providing straightforward approximate and pragmatic solutions. The Monte Carlo or empirical simulation approach to statistical inference is conceptually simple. A test statistic that characterizes the phenomenon of interest is chosen and the noise distribution, that is, the distribution under the null hypothesis, is approximated on the basis of simulated data. This presupposes that all relevant aspects of noise or null conditions are captured in the simulations. The simulated null distribution is then used to determine (estimated) critical thresholds for hypothesis testing and significance assessment. In FNI applications of this approach, discrete statistic images are simulated which are assumed to approximate the real statistic fields under the null hypothesis. These simulated realizations are then used to characterize the (simulated) distribution of the test statistic by plotting frequency data. Like non-parametric approaches, this approach offers great flexibility in the choice of parameter ranges (e.g. degrees of freedom and smoothness) and test statistics.

In general, Monte Carlo approaches are critically dependent on accurate characterization and modelling of null hypothesis conditions. Since the extreme tails of the simulated probability distributions have to be estimated with high precision, a large number of simulated realizations of the null hypothesis conditions are needed. High precision in the tails is necessary in order to determine the critical level for the chosen test statistic with sufficient accuracy. The validity of the Monte Carlo approach to statistical inference depends crucially on how well the simulated distributions approximate the real null distributions. In order to yield valid approximations, all the important sources of variability represented in the imaging process under null conditions have to be characterized and modelled appropriately. In the case of FNI this includes physical, physiological and cognitive sources of variability. The principal problem is thus to characterize the null conditions adequately. The difficulties lie in determining what are the relevant null conditions and how to characterize these and measure the relevant aspects of the null conditions in terms of variability sources.

The applications of the Monte Carlo approach to FNI data described have attempted to simulate stationary discrete statistic images, of which the marginal distributions and the spatial ACF are assumed to match image noise or null conditions. In effect, the first two moments characterizing the random image are matched and the form of the marginal distributions and the spatial ACF have either been assumed or estimated from noise images. In the case of PET, there is often not enough null condition data to estimate the relevant properties of null data for a given experiment with a particular set of subjects reliably. Instead, simplifying assumptions or approximations become necessary. However, in the case of fMRI, it may be possible to generate enough null data under suitable conditions. Furthermore, since data from different studies are likely to show different characteristics, new simulations are required for each new experiment or study population, with appropriately matched

parameters estimated on null data relevant to the new experiment or study population. In addition, robustness issues relevant to the Monte Carlo approaches have so far not been investigated in any detail.

(a) *Suprathreshold cluster statistics, hierarchical decomposition and multifiltering*

One of the first Monte Carlo approaches applied to PET data used the suprathreshold cluster size as the test statistic (Poline & Mazoyer 1993, 1994*a*,*b*,*c*). Poline & Mazoyer (1993) obtained the (approximate) distribution of the maximal suprathreshold cluster size by simulations using a Poisson approximation for the occurrence of suprathreshold clusters and assuming a Poisson distribution for cluster sizes at high thresholds. Other simulation approaches using the suprathreshold cluster size statistic have also been described for PET (Roland *et al.* 1993; Ledberg *et al.* 1998) and fMRI data (Forman *et al.* 1995). It is of course possible to use other test statistics, for example any interesting characteristic of the suprathreshold cluster could be used for a suprathreshold cluster test. An example is the excess mass statistic, that is the sum of the voxel values of the suprathreshold cluster (Holmes 1994). In general, when a cluster characteristic C has been chosen, based on what is deemed a relevant characteristic of the signal, the straightforward approach is to study the maximum of C, $\max[C]$. This would automatically handle the multiple comparisons problem and also reduces the computational load. None of the proposed Monte Carlo approaches have used the $\max[C]$-statistic but have instead simulated the distributions related to C.

One of the drawbacks with the suprathreshold cluster size statistic is that the magnitude of the signal is not considered (Poline & Mazoyer 1994*a*,*b*). With the cluster size statistic, clusters of the same size will be considered equally likely independent of the magnitude of the voxel values making up the clusters. This implies that a low-intensity cluster of the same size as a high-intensity cluster will be judged as occurring equally often by chance (i.e. under the null hypothesis) which in general is not the case. Instead, high-intensity clusters are generally less likely to occur by chance than low-intensity clusters. This problem is addressed, for example, by the excess mass statistic and was the rationale for introducing the bivariate suprathreshold cluster statistic (Poline & Mazoyer 1994*a*). The rejection region for the bivariate cluster statistics in the 2D parameter space turns out to be difficult to estimate. In order to solve this problem the Poisson approximation is invoked again. This approximation can only be expected to work well at 'high' iso-cumulative curves (corresponding to high thresholds in the univariate case), since the Poisson approximation depends on the assumption of rare events. A simpler solution, making the Poisson assumption unnecessary, has been suggested (Holmes 1994, Appendix H, p. 226).

A general problem with any suprathreshold cluster statistic is the fact that an arbitrary choice of threshold determines what type of activations are possible or most sensitively detected. For example, in the case of the cluster size statistic, at low thresholds the critical cluster size will tend to be large and focal activations are not detected; likewise at high thresholds low-level activations will be missed. The maximum excess mass and bivariate suprathreshold cluster statistics partly solve this problem by remaining sensitive to focal activations at low thresholds. However, part of the problem remains, as no activation can be detected below the chosen threshold. In response to the fact that the arbitrary choice of threshold determines which type of activation is most sensitively detected, Poline & Mazoyer (1994*a*) suggested the combined use of an image segmentation method called hierarchical decomposition and multifiltering (Koenderink 1984; Lifschitz & Pizer 1990; Ter Haar Romeny *et al.* 1991). The approach is complicated by the fact that the characteristics of the objects are recursively defined by the decomposition procedure and are thus not independent (Holmes 1994) which is also the rational for choosing a Monte Carlo approach to statistical inference.

Another characteristic of suprathreshold cluster tests is that they only have strong control over false positives at the cluster level but not at the voxel level. This implies that suprathreshold cluster statistics have reduced localizing power: when a cluster is significantly activated this does not give information on which voxels of the cluster are activated. This is particularly problematic when low thresholds are used, since the clusters will be larger, that is, localizing power generally decreases with lower thresholds, but there may be a gain in statistical power (cf. § 3(c)) under certain circumstances. Simulations have indicated that this can be the case with fMRI data (Friston *et al.* 1996). However, the same study also indicated that this might not always be the case with PET data. It is important to note that, in general, the results of power studies are dependent on the nature of the signal present in the data and the method used to detect it. The study of Friston *et al.* (1996) assumed a spatially distributed signal with no predilection for any particular region of the search volume.

(b) *The cluster simulation method and spatial autocorrelation estimation*

In addition to the general limitations of the Monte Carlo approach to statistical inference, the approach of Roland *et al.* (1993) has several additional shortcomings that have been thoroughly reviewed (Frackowiak *et al.* 1996; Petersson 1998). There seems to be a consensus that this approach is inadequate and it has recently been revised (Ledberg *et al.* 1998). Briefly, in the revised approach (called the cluster simulation method) the primary data is smoothed with a 3D isotropic Gaussian filter and balanced noise images (cf. Ledberg *et al.* 1998) are generated from PET data. The one-dimensional marginal (voxel) distributions are assumed to be approximately Gaussian and a t-image is created from the balanced noise images. The t-image is then transformed to a Gaussianized t-image (which is called a pseudo-normal Z-image; cf. Ledberg *et al.* 1998). This Gaussianized t-image is assumed to be stationary and the spatial ACF is estimated directly from this statistical image under the assumption of approximate stationarity. The estimated ACF is used to derive a suitable convolution kernel K. Simulated normal white-noise Z-images are then convolved with the kernel K and simulated distributions for the suprathreshold cluster size statistic are generated. In effect, stationary Gaussian

Z-images are simulated and assumed to approximate the relevant characteristics of the Gaussianized t-image.

The estimated kernel K and, hence, the simulated distributions are affected by bias and variance of the ACF estimate. The estimator used to estimate the ACF is asymptotically unbiased and underestimates the spatial extent of the true ACF on finite samples (Yaglom 1986; Ledberg *et al*. 1998). The estimated ACF is truncated at large lags where the ACF is judged not to be significantly different from zero (Ledberg *et al*. 1998). It should be noted that underestimating the spatial extent of the ACF when the cluster size statistic is used will tend to underestimate the critical levels. If, in addition, the variability of the ACF is appreciable or if the ACF is not reliably estimated, then the critical levels will be unreliable. There are some indications that the critical levels depend sensitively on the ACF (Roland *et al*. 1993; Roland & Gulyas 1996) and how reliable the ACF estimate is in PET data is an open question. It was suggested that this problem could be handled by inflating the convolution kernel (Ledberg *et al*. 1998). No procedural criteria for these modifications of the convolution kernel are given and the effects of these manipulations have not been characterized. The variability in the estimation of the ACF is characterized by point estimates of the ACF variance on simulated data. This approach implicitly assumes that there are no other variability sources of the ACF than those that can be well modelled as an interaction between the convolution kernel and white-noise images. Variability sources that possibly do not conform sufficiently well to such a model are structured noise introduced by filtering back-projected, low-count, Poisson-distributed data and different physiological variability sources. This implies that the variability of the ACF estimate on simulated data may differ from the variability of the ACF estimate on real null PET data. However, when the cluster simulation method is compared with simulated reference data, the estimated probability as a function of cluster size appears conservative. There may be several reasons for this. First, the estimated convolution kernel was inflated by adding a constant at small lags. If this manipulation results in an overestimated spatial extent of the convolution kernel, then conservative results are to be expected. Second, Gaussianized t-fields are fundamentally different from GRFs (Worsley *et al*. 1992, 1996*a*; Worsley 1994; Cao 1999). In principle, this implies that Gaussianized t-fields cannot be directly simulated by GRFs. Note that the cluster simulation method simulates discrete GRFs, while the reference distributions were determined on simulated discrete Gaussianized t-fields. Approximating Gaussianized t-fields with GRFs without appropriate corrections will lead to an overestimated false-positive rate of the suprathreshold cluster size statistic and, therefore, lead to a conservative test (Cao 1999), in particular for small to moderate degrees of freedom.

The cluster simulation method was also validated on real PET data. When applicable, non-parametric approaches (making minimal assumptions) may be viewed as benchmark methods for cross-validation of other methods (Good 1994; Edgington 1995). The cluster simulation method was compared with the non-parametric method described by Holmes *et al*. (1996) (adapted for the cluster size statistic) on a real PET data set. This comparison indicated that the cluster simulation method tended to underestimate the critical cluster size at lower thresholds, while it tended to overestimate the critical cluster size at higher thresholds (Ledberg *et al*. 1998). Since new simulations have to be performed for each new data set and these simulations are at least as computationally intensive as the non-parametric method of Holmes *et al*. (1996) and since the non-parametric approach was used for validation, it seems that the non-parametric method should be preferred when applicable. Finally, in order to extend the cluster simulation method to fMRI data, the method for generating noise images has to be adapted to the fact that fMRI time-series commonly show temporal autocorrelation (Weisskoff *et al*. 1993; Friston *et al*. 1994*a*; Worsley & Friston 1995; Zarahn *et al*. 1997; Purdon & Weisskoff 1998). If appropriate, null conditions can be defined, an alternative possibility may be to generate real fMRI null data (Aguirre *et al*. 1997; Zarahn *et al*. 1997).

(c) *A Monte Carlo approach to fMRI data*

Forman *et al*. (1995) described a Monte Carlo method for analysing fMRI data, both for spatially autocorrelated and uncorrelated data. Here we comment on the autocorrelated case, since fMRI data often show some spatial autocorrelation. Briefly, the spatial autocorrelation is approximated by a 2D isotropic Gaussian ACF and 2D isotropic Gaussian Z-images are simulated by convolving white-noise Z-images with a suitable isotropic Gaussian convolution kernel. In order to connect to real fMRI data, the filter width of the convolution kernel is estimated from noise t-images, which are generated through random pairings within experimental conditions. Finally, the simulated distribution of the suprathreshold cluster size statistic under the null hypothesis is generated based on the simulated data. Signal t-images are generated through random pairings across experimental conditions and the observed cluster sizes are compared with the simulated distribution.

Since the smoothness estimate of Friston *et al*. (1991) tends to overestimate the kernel width at low spatial autocorrelation of the statistic image, Forman *et al*. (1995) presented a modified formula to improve the estimation of filter widths close to or below the pixel dimensions (cf. § 4(b)). It appears that the filter width is estimated directly from the noise t-image under the assumption that the t-image is sufficiently well approximated as stationary and isotropic (i.e. the spatial ACF is translational as well as rotationally invariant). It is unclear whether it is assumed that the number of effective degrees of freedom is large enough to approximate the t-image with an isotropic Gaussian Z-image.

Forman *et al*. (1995) validated their approach against real fMRI data by comparison of the cumulative cluster size distribution from experimental (16 noise t-images in two subjects) and simulated data. Although there seems to be a fair agreement between the experimental and the simulated cumulative distributions, there are apparent systematic deviations (Forman *et al*. 1995). Finally, Forman *et al*. (1995) emphasized that the approach (as described) does not distinguish true signal from systematic artefactual sources of signal. Such sources (e.g. biorhythms and artefacts related to motion) have to

be removed or accounted for by some other means, for example pre-processing or modelling.

In summary, Monte Carlo approaches to statistical inference are critically dependent on characterizing and modelling null data adequately. In order to yield valid approximations, all significantly contributing variability sources that are represented in the image process under null hypothesis conditions must be sufficiently well modelled (including physical or instrumental, physiological and cognitive sources). This may put limits on the generalizability of the results from Monte Carlo approaches, since particular characteristics of the null data related to the sample will be simulated. Since the extreme tails of the simulated probability distributions have to be estimated with high precision, a large number of simulated realizations of the null hypothesis conditions are needed. The applications of the Monte Carlo approach to FNI data described have attempted to simulate stationary discrete statistic images, the marginal distributions and spatial ACF of which are assumed to match image noise or null conditions. In effect, the first two moments characterizing the random image are matched and the form of the marginal distributions and the spatial ACF have either been assumed or estimated from noise images. The spatial ACF has either been estimated in its entirety (Ledberg *et al.* 1998) or under the assumption that it belongs to some suitable predefined class of functions (Poline & Mazoyer 1993; Forman *et al.* 1995). In particular, when the suprathreshold cluster size statistic is used, underestimating the spatial extent of significant autocorrelation or the variability of the ACF will tend to make the critical levels for the cluster size statistic artefactually small or unreliable. The robustness of Monte Carlo approaches in relation to different assumptions and the characteristics of real data are at present unknown. In this context, the way null data are characterized and noise images are generated is of central importance.

8. CONCLUSION

Functional neuroimaging methods provide experimental access to the living human brain and a framework of well-described theories and empirically validated methods is available. The field of FNI methodology has developed into a mature but evolving area of knowledge and applications have been extensive. In the companion paper (Petersson *et al.*, preceding paper) we discuss some aspects of the complex problem of model selection. In general, model selection is an important central prelude to subsequent statistical inference which depends on sufficiently well-fitting models. Assessing model fit and verification of assumptions are challenging tasks. The scientific process has many features in common with the classic data analytic strategy; there is an intense and fruitful interaction between exploration, model selection and critical inference. Progress in a scientific field is dependent on formulating and describing relevant problems as well as long-term consistency and convergence of empirical results. In this process, discussion and evaluation of the methods used in a scientific field are of central importance. In this paper, we have reviewed and discussed some aspects of signal detection theory and statistical inference relevant to the analysis of FNI data. As in the companion paper (Petersson *et al.*, preceding paper), the emphasis has been on assumptions and inherent limitations. Most of the methods described here generally serve their purposes well when the inherent assumptions and limitations are taken into account. It should also be noted that many of the methods presented yield similar, although not identical, results when applied to the same data set pre-processed in the same way. Significant differences in results may be most apparent in extreme parameter ranges, for example at low effective degrees of freedom or at small spatial autocorrelation. In such situations or in situations when assumptions and approximations are seriously violated, it is of central importance to choose the most suitable method in order to obtain valid results. The inferential methods used in FNI differ in the assumptions and approximations made. The central issues are how well these are fulfilled by the data being analysed and how robust the methods are if assumptions or approximations are not fully met. So far the emphasis in the analysis of FNI data has been on statistical methods protecting against false-positive results and there is a need to develop further effective methods for characterizing the signal present in data when the null hypothesis has been rejected.

Future progress in functional neuroimaging is partly dependent on the further development of the FNI methods used, at all stages of data processing, that is pre-processing (e.g. realignment, anatomic normalization, image segmentation and spatio-temporal filtering) and model building, as well as descriptive exploratory tools and methods for statistical inference. For example, if the cortical surface (and subcortical structures) can be extracted from anatomical images (Dale *et al.* 1999; Fischl *et al.* 1999) and well registered to functional images, then it is natural to process the data and detect activation signals in relation or restricted to these surfaces instead of in three dimensions. Spatial filtering of images directly on the cortical sheet (2D) would not mix data from regions that are close in the 3D space but far away when considering the geodesic distance on the cortical surface. Important for such an endeavour are recent developments making it possible to analyse locally non-stationary RFs. If applied to fMRI data, allowing for statistically valid single-subject results, the distortions between the anatomical $\mathcal{T}_1$- and $\mathcal{T}_2^*$-weighted functional images have to be accounted for. Another example is to try to achieve better characterization of individual subjects and then aggregate the results with non-image-based methods, including meta-analytic approaches. More comprehensive models of the fMRI signal should produce greater specificity of the BOLD signal, giving higher quality within-subject images. Models parameterizing the activation foci (e.g. by size and location) may be introduced and a hierarchical model could combine these foci across subjects, incorporating uncertainty in the intersubject registration.

K.M.P. was supported by grants from the Swedish Medical Research Council (8276) and the Karolinska Institute and A.P.H. was supported by the Wellcome Trust for part of this work. The authors are grateful for many fruitful discussions concerning methodological issues with Jesper L. R. Andersson,

Karl J. Friston, Anders Ledberg, Anthony R. McIntosh, John Ollinger and Keith Worsley.

REFERENCES

Adler, R. J. 1981 *The geometry of random fields*. New York: Wiley.

Adler, R. J. 1998 *On excursion sets, tube formulae, and maxima of random fields*. (Submitted.) See http://iew3.technion.ac.il:8080/Home/Users/FIRST.phtml?show+Robert+Adler.

Aguirre, G. K., Zarahn, E. & D'Esposito, M. 1997 Empirical analyses of BOLD fMRI statistics. II. Spatially smoothed data collected under null-hypothesis and experimental conditions. *NeuroImage* **5**, 199–212.

Aguirre, G. K., Zarahn, E. & D'Esposito, M. 1998 A critique of the use of the Kolmogorov–Smirnov (KS) statistic for the analysis of BOLD fMRI data. *Magn. Reson. Med.* **39**, 500–505.

Andreasen, N. C., Arndt, S. A., Cizadlo, T., O'Leary, D. S., Watkins, G. L., Boles Ponto, L. L. & Hichawa, R. D. 1995 Sample size and statistical power in [O15] $H_2{}^{15}O$ studies of human cognitiion. *J. Cerebr. Blood-Flow Metab.* **16**, 804–816.

Andrews, H. & Hunt, B. 1977 *Digital image restoration*. Englewood Cliffs, NJ: Prentice Hall.

Aschburner, J., Neelin, P., Collins, D. L., Evans, A. & Friston, K. 1997 Incorporating prior knowledge into image registration. *NeuroImage* **6**, 344–352.

Besag, J. 1974 Spatial interactions and statistical analysis of lattice systems. *J. R. Statist. Soc.* B **36**, 192–236.

Billingsley, P. 1995 *Probability and measure*, 3rd edn. New York: Wiley.

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

Cao, J. 1999 The size of the connected components of excursion sets of χ^2, t and F fields. *Adv. Appl. Prob.* (In the press.)

Cao, J. & Worsley, K. J. 1999 The geometry of correlation fields with an application to functional connectivity of the brain. *Ann. Appl. Prob.* (In the press.)

Chellapa, R. & Jain, A. 1993 *Markov random fields: theory and practice*. Boston, MA: Academic Press.

Coulon, O., Bloch, I., Frouin, V. & Mangin, J.-F. 1997 Multiscale measures in linear scale-space for characterizing cerebral functional activations in 3D PET difference images. *Lecture Notes Comput. Sci.* **1252**, 188–199.

Crivello, F., Tzourio, N., Poline, J.-B., Woods, R. P., Maziotta, J. C. & Mazoyer, B. M. 1995 Intersubject variability in functional neuroanatomy of silent verb generation: assessment by a new activation detection algorithm based on amplitude and size information. *NeuroImage* **2**, 253–263.

Dale, A. M., Fischl, B. & Sereno, M. I. 1999 Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage* **9**, 179–194.

Descombes, X., Kruggel, F. & Von Cramon, D. Y. 1998 fMRI signal restoration using a spatio-temporal Markov random field preserving transitions. *NeuroImage* **8**, 340–349.

Edgington, E. S. 1995 *Randomization tests*, 3rd edn (revised). New York: Marcel Dekker.

Fischl, B., Sereno, M. I. & Dale, A. M. 1999 Cortical surface-based analysis. II. Inflation, flattening, and a surface-based coordinate system. *NeuroImage* **9**, 195–207.

Fisher, R. A. 1935 *The design of experiments*. Edinburgh: Oliver Boyd.

Forman, S. D., Cohen, J. D., Fitzgerald, J. D., Eddy, W. F., Mintun, M. A. & 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. & Mintun, M. A. 1989 Non-invasive functional brain mapping by change distribution analysis of averaged PET images of $H_2{}^{15}O$ tissue activity. *J. Nucl. Med.* **30**, 141–149.

Fox, P. T. & Pardo, J. V. 1991 Does intersubject variability in cortical functional organization increase with neural 'distance' from the periphery? *Ciba Found. Symp.* **163**, 125–144.

Fox, P. T., Mintun, M. A., Reiman, E. M. & Raichle, M. E. 1988 Enhanced detection of focal brain responses using intersubject averaging and change-distribution analysis of subtracted PET images. *J. Cerebr. Blood-Flow Metab.* **8**, 642–653.

Frackowiak, R. S. J., Zeki, S., Poline, J.-B. & Friston, K. J. 1996 A critique of a new analysis proposed for functional neuroimaging. *Eur. J. Neurosci.* **8**, 2229–2231.

Frackowiak, R. S. J., Friston, K. J., Frith, C. D., Dolan, R. J. & Mazziotta, J. C. 1997 *Human brain function*. San Diego: Academic Press.

Friston, K. J., Frith, C. D., Liddle, P. F., Dolan, R. J., Lammertsma, A. A. & Frackowiak, R. S. 1990 The relationship between global and local changes in PET scans. *J. Cerebr. Blood-Flow Metab.* **10**, 458–466.

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

Friston, K. J., Jezzard, P. & Turner, R. 1994*a* Analysis of functional MRI time-series. *Hum. Brain Mapp.* **1**, 210–220.

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

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P. & Frackowiak, R. S. J. 1995 Stastistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* **2**, 189–210.

Friston, K. J., Holmes, A., Poline, J.-B., Price, C. J. & Frith, C. D. 1996 Detecting activations in PET and fMRI: levels of inference and power. *NeuroImage* **4**, 223–235.

Gee, J. C., Le Briquer, L. & Barillot, C. 1995 Probabilistic matching of brain images. In *Information processing in medical imaging* (ed. Y. Bizais), pp. 113–125. Dordrecht: Kluwer Academic Publishers.

Geman, S. & Geman, D. 1984 Stochastic relaxation, Gibbs distribution, and Bayesian restoration of images. *IEEE Trans. Patt. Anal. Mach. Intell.* **6**, 721–741.

Good, P. 1994 *Permutation tests*. New York: Springer.

Grabowski, T. J., Frank, R. J., Brown, C. K., Damasio, H., Boles Ponto, L. L., Watkins, G. L. & Hichwa, R. D. 1996 Reliability of PET activation across statistical methods, subject groups and sample sizes. *Hum. Brain Mapp.* **4**, 23–46.

Grachev, I. D., Berdichevsky, D., Rauch, S. L., Heckers, S., Kennedy, D. N., Caviness, V. S. & Alpert, N. M. 1999 A method for assessing the accuracy of intersubject registration of the human brain using anatomic landmarks. *NeuroImage* **9**, 250–268.

Hajnal, J., Myers, R., Oatridge, A., Schwieso, J. E., Young, I. R. & Bydder, G. M. 1994 Artifacts due to stimulus correlated motion in functional imaging of the brain. *Magn. Reson. Med.* **31**, 263–291.

Hasnain, M. K., Fox, P. T. & Woldorff, M. G. 1998 Intersubject variability of functional areas in the human visual cortex. *Hum. Brain Mapp.* **6**, 301–315.

Hasofer, A. M. 1976 The mean number of maxima above high levels in Gaussian random fields. *J. Appl. Prob.* **13**, 377–379.

Hochberg, Y. & Tamhane, A. C. 1987 *Multiple comparisons procedures*. New York: Wiley.

Holmes, A. P. 1994 Statistical issues in functional brain mapping. PhD thesis, University of Glasgow.

Holmes, A. P. & Ford, I. 1993 A Bayesian approach to significance testing for statistic images from PET. In

Quantification of brain function: tracer kinetics and image analysis in brain PET (ed. K. Uemura, N. Lassen, T. Jones & I. Kanno), pp. 521–531. Amsterdam: Excerpta Medica.

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

Hunton, D. L., Miezin, F. M., Buckner, R. L., Van Mier, H. I., Raichle, M. E. & Petersen, S. E. 1996 An assessment of functional-anatomical variability in neuroimaging studies. *Hum. Brain Mapp.* **4**, 122–139.

Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. 1983 Optimization by simulated annealing. *Science* **220**, 671–680.

Koenderink, J. J. 1984 The structure of images. *Biol. Cybernet.* **50**, 363–370.

Ledberg, A., Åkerman, S. & Roland, P. R. 1998 Estimation of the probability of 3D clusters in functional brain images. *NeuroImage* **8**, 113–128.

Lifschitz, L. M. & Pizer, S. M. 1990 A multiresolution hierarchical approach to image segmentation based on intensity extrema. *IEEE Trans. Patt. Anal. Mach. Intell.* **12**, 529–540.

Liu, C., Raz, J. & Turetsky, B. 1998 An estimator and permutation test for single-trial fMRI data. In *ENAR Meeting of the International Biometric Society.* Pittsburgh, March 1998.

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

Maitra, R. 1997 Estimating precision in functional images. *J. Comput. Graph. Stat.* **6**, 132–142.

Mangin, J.-F., Règis, J., Bloch, I., Frouin, V., Samson, Y. & Lûpez-Krahe, J. 1995 A Markovian random field based random graph modelling the human cortical topography. In *First International Conference on Computer Vision, Virtual Reality and Robotics in Medicine* (Lecture Notes in Computer Science), pp. 177–183. New York: Springer.

Mazoyer, B. M., Tzourio, N., Poline, J.-B., Levrier, O., Petit, L., Raynaud, L. & Joliot, M. 1993 Anatomical regions of interest versus stereotactic space: a comparaison of two approaches for brain activation maps analysis. *Ann. Nucl. Med. PET Brain* **93**.

Nichols, T. E. & Holmes, A. P. 1999 Nonparametric permutation tests for functional neuroimaging experiments: a primer with examples. (In preparation.)

Nosko, V. P. 1969 Local structure of Gaussian random fields in the vicinity of high level shines. *Soviet Math. Dokl.* **10**, 1481–1484.

Petersson, K. M. 1998 Comments on a Monte Carlo approach to the analysis of functional neuroimaging data. *NeuroImage* **8**, 108–112.

Poline, J.-B. & Mazoyer, B. J. 1991 Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise-ratio pixel clusters. *J. Cerebr. Blood-Flow Metab.* **11** (Suppl. 2), S564.

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

Poline, J.-B. & Mazoyer, B. J. 1994*a* Analysis of individual positron emission tomography activation maps using hierarchical description and multi-scale detection. *IEEE Trans. Med. Imaging* **13**, 702–710.

Poline, J.-B. & Mazoyer, B. J. 1994*b* Cluster analysis of individual functional brain images: some new techniques to enhance the sensitivity of activation detection methods. *Hum. Brain Mapp.* **2**, 103–111.

Poline, J.-B. & Mazoyer, B. J. 1994*c* Enhanced detection in brain activation maps using a multifiltering approach. *J. Cerebr. Blood-Flow Metab.* **14**, 639–642.

Poline, J.-B., Worsley, K. J., Holmes, A. P., Frackowiak, R. S. & Friston, K. J. 1995 Estimating smoothness in statistical parametric maps: variability of *p* values. *J. Comput. Assist. Tomogr.* **19**, 788–796.

Poline, J.-B., Vandenberghe, R., Holmes, A. P., Friston, K. J. & Frackowiak, R. S. J. 1996 Reproducibility of PET activation studies: lessons from a multi-center European experiment. *NeuroImage* **4**, 34–54.

Poline, J.-B., Worsley, K. J., Evans, A. C. & Friston, K. J. 1997 Combining spatial extent and peak intensity to test for activations in functional imaging. *NeuroImage* **5**, 83–96.

Poline, J.-B., Van de Moortele, P.-F., Paradis, A.-L., Frouin, V. & LeBihan, D. 1998 Analyses of model responses and residuals in an event related functional MRI experiment using multivariate techniques. *NeuroImage* **7**, S762.

Purdon, P. L. & Weisskoff, R. M. 1998 Effect of temporal autocorrelation due to physiological noise stimulus paradigm on voxel-level false-positive rates in fMRI. *Hum. Brain Mapp.* **6**, 239–249.

Ramsey, N. F., Kirby, B. S., Gelderen, P. V., Berman, K. F., Duyn, J. H., Frank, J. A., Mattay, V. S., Van Horn, J. D., Esposito, E., Moonen, C. T. W. & Weinberger, D. R. 1996 Functional mapping of human sensorimotor cortex with BOLD fMRI correlates highly with $H_2{}^{15}O$ PET rCBF. *J. Cerebr. Blood-Flow Metab.* **16**, 755–764.

Rangarajan, A. & Chellappa, R. 1995 Markov random field models in image processing. In *The handbook of brain theory and neural networks* (ed. M. A. Arbib), pp. 564–567. Cambridge, MA: MIT Press.

Regis, J., Mangin, J. F., Frouin, V., Sastre, F., Peragut, J. C. & Samson, Y. 1995 Generic model for the localisation of the cerebral cortex and preoperative multimodal integration in epilepsy surgery. *Stereotact. Funct. Neurosurg.* **65**, 72–80.

Roland, P. E. & Gulyas, B. 1996 Assumptions and validations of statistical tests for functional neuroimaging. *Eur. J. Neurosci.* **8**, 2232–2235.

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

Rosenfeld, A. & Kak, A. C. 1982 *Digital picture processing.* Orlando, FL: Academic Press.

Senda, M., Ishii, K., Oda, K., Sadato, N., Kawashima, R., Sugiura, M., Kanno, I., Ardekani, B., Minoshima, S. & Tatsumi, I. 1998 Influence of ANOVA design and anatomical standardization on statistical mapping for PET activation. *NeuroImage* **8**, 283–301.

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

Strother, S. C., Lange, N., Anderson, J. R., Schaper, K. A., Rehm, K., Hansen, L. K. & Rottenberg, D. A. 1997 Activation pattern reproducibility: measuring the effect of group size and data analysis models. *Hum. Brain Mapp.* **5**, 312–316.

Taylor, S. F., Minoshima, S. & Koeppe, R. A. 1993 Letter to the editor: instability of localization of cerebral blood flow activation foci with parametric maps. *J. Cerebr. Blood-Flow Metab.* **13**, 1040–1041.

Ter Haar Romeny, B. M., Florack, L. M. J., Koenderink, J. J. & Viergever, M. A. 1991 Scale space: its natural operators and differential invariants. In *Information processing and medical imaging* (ed. A. C. F. Colchester & D. J. Hawkes), pp. 238–255. Berlin: Springer.

Tikhonov, A. N. & Arsenin, V. Y. 1977 *Solutions to ill-posed problems.* Washington, DC: Winston & Sons.

Van Horn, J. D., Ellmore, T. M., Esposito, G. & Berman, K. F. 1998 Mapping voxel-based statistical power on parametric images. *NeuroImage* **7**, 97–107.

Vapnik, V. N. 1998 *Statistical learning theory.* New York: Wiley.

Vitouch, O. & Glück, J. 1997 'Small group PETting': sample sizes in brain mapping research. *Hum. Brain Mapp.* **5**, 74–77.

Wahba, G. 1995 Generalization and regularization in nonlinear learning systems. In *The handbook of brain theory and neural networks* (ed. M. A. Arbib), pp. 426–430. Cambridge, MA: MIT Press.

Weisskoff, R. M., Baker, J., Belliveau, J., Davis, T. L., Kwong, K. K., Cohen, M. S. & Rosen, B. R. 1993 Power spectrum analysis of functionally weighted MR data: what's in the noise? *Proc. Soc. Magn. Reson. Med.* **1**, 7.

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. 1995 Estimating the number of peaks in a random field using the Hadwiger characteristic of excursion sets, with applications to medical images. *Ann. Stat.* **23**, 640–669.

Worsley, K. J. 1996*a* *An unbiased estimator for the roughness of a multivariate Gaussian random field.* Montreal: Department of Mathematics and Statistics, University of McGill.

Worsley, K. J. 1996*b* The geometry of random images. *CHANCE* **9**, 27–40.

Worsley, K. J. 1997 An overview and some new developments in the statistical analysis of PET and fMRI data. *Hum. Brain Mapp.* **5**, 254–258.

Worsley, K. J. 1999 Testing for signals with unknown location and scale in a χ^2 random field, with an application to fMRI. *Adv. Appl. Prob.* (Submitted.)

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

Worsley, K. J. & Vandal, A. C. 1994 *Quadratic tests for local changes in random fields with applications in medical images.* Montreal: Department of Mathematics and Statistics, University of McGill (Technical report).

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

Worsley, K. J., Evans, A. C., Marrett, S. & Neelin, P. 1993 Letter to the editor: authors reply. *J. Cerebr. Blood-Flow Metab.* **13**, 1041–1042.

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

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

Worsley, K. J., Marrett, S., Neelin, P. & Evans, A. C. 1996*b* Searching scale space for activations in PET images. *Hum. Brain Mapp.* **4**, 74–90.

Worsley, K. J., Wolforth, M. & Evans, A. C. 1997*a* Scale space searches for a periodic signal in fMRI data with spatially varying hemodynamic response. (Submitted.)

Worsley, K. J., Poline, J.-B., Friston, K. J. & Evans, A. C. 1997*b* Characterizing the response of PET and fMRI data using multivariate linear models. *NeuroImage* **6**, 305–319.

Worsley, K. J., Andermann, M., Koulis, T., MacDonald, D. & Evans, A. C. 1999 Detecting changes in non-stationary images via statistical flattening. *NeuroImage.* (Submitted.)

Xiong, J., Gao, J.-H., Lancaster, J. L. & Fox, P. T. 1995 Clustered pixels analysis for functional MRI activation studies of the human brain. *Hum. Brain Mapp.* **3**, 287–301.

Yaglom, A. M. 1986 *Correlation theory of stationary and related random functions.* I. *Basic results.* New York: Springer.

Zarahn, E., Aguirre, G. K. & D'Esposito, M. 1997 Empirical analyses of BOLD fMRI statistics. I. Spatially unsmoothed data collected under null-hypothesis conditions. *NeuroImage* **5**, 179–197.