NEUROIMAGE **8,** 108–112 (1998)
ARTICLE NO. NI980375

COMMENTS AND CONTROVERSIES

Comments on a Monte Carlo Approach to the Analysis of Functional Neuroimaging Data

Karl Magnus Petersson

Cognitive Neurophysiology R2-01, Department of Clinical Neuroscience, Karolinska Institute, Karolinska Hospital, S-171 76 Stockholm, Sweden,
E-mail: karlmp@neuro.ks.se

INTRODUCTION

Functional neuroimaging is probably always going to be methodologically pluralistic. There are several reasons for this. For example, brain functions or processes can be characterized at different levels and scales, and it may be the case that there is no fundamental processing level but that different phenomena are optimally described at different scales or levels. Methods developed and validated at specific spatiotemporal scales and for certain parameter ranges (e.g., degrees of freedom, amount of filtering) may not be applicable at other spatiotemporal scales or parameter ranges. Furthermore, the rapid development in fMRI and MEG/EEG illustrates the need for descriptive, exploratory, and inferential methods. Descriptive and exploratory methods are useful in characterizing the nature of the signal that is present in the data, while inferential methods are used to test hypotheses and to determine confidence intervals.

Basically, there are three inferential approaches to the analysis of functional imaging data: theoretical parametric (e.g., Friston *et al.*, 1995; Worsley *et al.*, 1992, 1996) approaches, nonparametric approaches (e.g., Holmes *et al.*, 1996), and Monte Carlo or simulation approaches (e.g., Forman *et al.*, 1995; Poline and Mazoyer, 1993). These approaches differ in the assumptions made about the properties of data and the approximations used in their statistical analyses. What is of importance is not the number of assumptions or the characteristics of the approximations made, but how well these assumptions and approximations are fulfilled by empirical data and the robustness of the method if the assumptions are not fully met. This notion emphasizes the importance of empirical validation and explicit characterization of the inherent limitations of a given method.

Progress in and the credibility of a scientific field are critically dependent on the long-term consistency and convergence of empirical results. Discussion and critical evaluation of the methods used in a given scientific field are of vital importance in this process. An example of such a critical evaluation and discussion is summarized below.

Recently, functional neuroimaging studies have been published in *Nature* (Geyer *et al.*, 1996), *Science* (Kinomura *et al.*, 1996), and the *Proceedings of the National Academy of Sciences of the USA* (Roland *et al.*, 1998) using a cluster analysis method described by Roland *et al.* (1993). This method has been criticized by Frackowiak *et al.* (1996) and subsequently defended by Roland and Gulyas (1996). In this issue of *NeuroImage*, Roland and colleagues (Ledberg *et al.*, 1998) return to some of the issues previously raised. Ledberg *et al.* (1998) describe a revised version of the Roland *et al.* (1993) method acknowledging the critique of Frackowiak *et al.* (1996). This illustrates the importance of proper empirical validation of any proposed method before it is accepted and applied to experimental data. The constructive result of this critical evaluation is the significant improvement of the method (Ledberg *et al.*, 1998).

The reason for the closer examination of the Roland *et al.* (1993) method and the consequent discussion in the *European Journal of Neuroscience* was diverging results and interpretations of data relating to the functional neuroanatomy of vision, in particular color perception (Frackowiak *et al.*, 1996). Two general topics are at issue. The first relates to the functional anatomy of vision, which is not discussed in Ledberg *et al.* (1998) and will not be discussed here. The second issue relates to methodology and is independent of the first, whereas conclusions about the functional anatomy of vision are most certainly dependent on the method used.

A Monte Carlo approach to the analysis of PET data using cluster size as the test statistic was proposed by Poline and Mazoyer (1993). A similar approach has been applied to the analysis of fMRI data (Forman *et al.*, 1995). In general, Monte Carlo approaches are critically dependent on adequately characterizing the image noise, using sufficient numbers of simulated realizations (since the tails of the observed probability

1053-8119/98 $25.00

distributions must be determined with high precision), and formulating statistically relevant events or null hypotheses to determine the critical level for the chosen test statistic.

SUMMARY AND COMMENTS ON THE DISCUSSION IN THE *EUROPEAN JOURNAL OF NEUROSCIENCE*

The method described by Roland *et al.* (1993) is a Monte Carlo method based on cluster size as the test statistic. This method was critically examined by Frackowiak *et al.* (1996). In brief, they point out that the probabilistic statement used by Roland *et al.* (1993) to determine critical levels for cluster size is incorrect (incidentally, the reformulation in Roland and Gulyas (1996) is also incorrect), that the number of simulated realizations are too few, and that Roland *et al.* applied the simulated distributions given in the tables of their (1993) paper as if the results generalized irrespective of study design or population (see also Ledberg *et al.*, 1998). Given the large standard deviations for tail events in the tables of Roland *et al.* (1993), it is clear that the simulations were too small and Roland and Gulyas (1996) concede this point. However, Roland and Gulyas (1996) maintained that extended simulations performed subsequently indicated that the results generalized irrespective of study design. It is clear that *t*-statistic fields, including Gaussianized versions, and the spatial autocorrelation of *t* fields are dependent on the degrees of freedom (see further Worsley *et al.*, 1992, in particular the Appendix). In addition, the spatial autocorrelation structure of image data may change from one data set to another. Therefore, simulations for each new population or new study design must be performed. The method proposed by Ledberg *et al.* (1998) accepts this point. Frackowiak *et al.* (1996) also criticize the use of high α or significance levels (see, for example, Gulyas and Roland, 1994, 1995, 1996, in which significance levels of $P \leq 0.5$–0.6 are used). In response to this criticism, Roland and Gulyas (1996) maintain that the high α levels used might be a reasonable balance between false positives and false negatives. Frackowiak *et al.* (1996) also point out that it is not possible to ascertain from the original methods paper (Roland *et al.*, 1993) whether physiological autocorrelations are disregarded when determining a critical cluster size. In this context, it is unclear why Roland *et al.* (1993, pp. 10 and 19), when creating noise images to determine the spatial autocorrelation function, twice subtract low-pass-filtered (10-mm FWHM) versions of the noise from the original noise images (Roland *et al.*, 1993, p. 19). This procedure retains only the high-frequency components of the noise and disregards those at low frequency, which would tend to make the spatial extent of any significant autocorrelations and the critical cluster size artifactually small.

In addition to the above remarks, the comments of Roland and Gulyas (1996) contain a critical examination of the assumptions underlying the SPM method developed by Friston and colleagues (e.g., Friston *et al.*, 1994, 1995) and a preliminary report of reproducibility of their previous color vision data (Gulyas *et al.*, 1994). While some of the critical remarks (Roland and Gulyas, 1996) relating to SPM appear to be based on misunderstandings of SPM or the recent developments of SPM (a detailed peer-reviewed commentary is available upon request, they implicitly point out some of the inherent limits and limitations of the Gaussian random field ("standard SPM") approach.

Finally, all methods for analyzing functional neuroimaging data must be empirically validated since they depend on a number of approximations or approximately valid assumptions. Validation of a method ideally uses some type of cross-validation, i.e., using an independent methodology. In the Gulyas and Roland (1994) original study, 29 activations related to color are reported as pure (sub)modality specific (p. 1816 and Table 4). These activations were located in the occipital, parietal, temporal, precentral, and prefrontal cortices as well as the cerebellum. In contrast, the preliminary reproducibility report in Roland and Gulyas (1996) is restricted to the occipital and occipitotemporal regions. Since activations in these regions could be predicted based on previous data (see, e.g., Corbetta *et al.*, 1991; Lueck *et al.*, 1989; Zeki *et al.*, 1991), the preliminary report in Roland and Gulyas (1996) has the character of a sensitivity analysis. In this context, a specificity analysis on noise images or a formal power analysis may have provided additional information.

In a recent (preliminary) reevaluation of previously presented and new data, Gulyas and Roland and colleagues (Gulyas *et al.*, 1997) indicate that only occipital (V1, V2, and lateral/inferior occipital gyri) and fusiform gyri are involved in color-related perceptual operations consistent with other color perception data (Corbetta *et al.*, 1991; Lueck *et al.*, 1989; McKeefry and Zeki, 1997a; Zeki *et al.*, 1991). A study of color perception using fMRI was recently reported (McKeefry and Zeki, 1997b) which replicates and extends previous results (Lueck *et al.*, 1989; Zeki *et al.*, 1991).

COMMENTS ON THE MONTE CARLO APPROACH AND THE METHOD OF LEDBERG AND COLLEAGUES

The method of Ledberg *et al.* (1998) is also a Monte Carlo approach using cluster size as the test statistic. The method described represents a revised version of the Roland *et al.* (1993) approach. In brief, the method of Ledberg *et al.* (1998) uses the general linear model on which SPM is predicated (called the mixed or covariance model in Ledberg *et al.* (1998)) to model data in each voxel, thus generating signal images. Balanced

noise images (i.e., scans from the same condition were subtracted from one another and then summed with other subtracted images; cf. Ledberg *et al.*, 1998) are generated from PET data and transformed into pseudo-normal images (i.e., t images transformed into Z images; the finite dimensional marginal distributions are not necessarily multivariate normal; cf. Ledberg *et al.*, 1998) from which the spatial autocorrelation function (ACF) is estimated. Simulated normal white-noise images are convolved with a convolution kernel K that is closely related to the estimated ACF (i.e., $K = \mathrm{IFT}(\mathrm{sqroot}|\mathrm{FT}(\mathrm{ACF})|)$, where FT and IFT are the Fourier transform and the inverse Fourier transform, respectively), and simulated distributions for the cluster size statistic are generated. The method of Ledberg *et al.* (1998) differs in several respects from the method of Roland *et al.* (1993). More specifically, the number of simulated realizations was substantially increased, the low-frequency component of the noise images was not subtracted, the spatial ACF was estimated directly on pseudo-normal transformed balanced noise images, and the distribution of cluster sizes was estimated for each new data set.

Monte Carlo approaches are critically dependent on adequately characterizing and modeling noise. The way in which noise images are generated is of critical importance for the adequacy of the approach. A possible alternative method to the way Ledberg *et al.* (1998) generate balanced noise images would be to first fit a general linear model, and given that this is a good model for the signal in the data, the residual image may be used as a noise image. A crucial point, particularly when using cluster size as the test statistic, is the estimation of the ACF in the statistical image process. Underestimating the spatial extent of significant autocorrelation or the variability of the ACF would tend to make the critical levels artificially low or unreliable. One way to investigate this potential problem would be to characterize how sensitive the estimated critical levels are to changes in the spatial extent of significant autocorrelation in conjunction with a characterization of the variability of the ACF in real functional imaging data, i.e., a robustness analysis. There are some indications of a sensitive dependence of the critical levels on the ACF (Roland *et al.*, 1993; Roland and Gulyas, 1996). This issue needs further investigation.

The estimator of spatial ACF chosen by Ledberg *et al.* (1998) is asymptotically unbiased and it is known to underestimate the spatial extent of significant spatial autocorrelations (Ledberg *et al.*, 1998; Yaglom, 1986). This will tend to make the critical thresholds too small. Ledberg *et al.* (1998) suggest that this problem may be handled by inflating the convolution kernel. In addition Ledberg *et al.* (1998) truncated the estimated ACF at large lags. These manipulations of the convolution kernel and the estimated ACF are not well characterized and need further investigation (Ledberg *et al.*, 1998). However, the estimation of the probability as a function of cluster size on simulated data indicates that this may be less of a problem. These estimations also indicate that the greatest gains in sensitivity are achieved at high degrees of freedom. A tentative alternative strategy for estimating the ACF would be to try to approximate the ACF directly with an anisotropic Gaussian kernel or indirectly by estimating smoothness in the noise images (for an example see Worsley, 1996).

Ledberg *et al.* (1998) characterize the variability in the ACF estimator by pointwise estimates of the ACF variance on simulated data. This may be a valid estimation of the ACF variability in PET data, but the approach implicitly assumes that there are no other variability/variance sources of the ACF in PET data than those well modeled as an interaction between the convolution kernel and white-noise images. Variability sources that may not conform sufficiently well to such a model are structured noise introduced by filtering back-projected low-count, Poisson distributed data and different neurophysiological variance sources.

When applicable, validated nonparametric approaches may in certain respects be viewed as benchmark methods. Ledberg *et al.* (1998) compare their cluster simulation method with SPM (SPM96, Friston *et al.*, 1995) and a nonparametric method described by Holmes *et al.* (1996) adapted for the cluster-size statistic on a PET data set at different Z-score thresholds. PET images are usually preprocessed in several steps, one step being low-pass filtering. In Ledberg *et al.* (1998), the PET images are Gaussian (isotropic) filtered at 6 mm (=3 voxels) FWHM and the data are modeled using the general linear model giving 40 residual degrees of freedom (df). Assuming the method of Holmes *et al.* (1996) as a benchmark, Ledberg *et al.* (1998) conclude under Results that their method gives reliable estimations of the probability distributions for Z thresholds above 2.58. In contrast, SPM is judged to be too conservative. There are basically three reasons for this result. First, the number of residual df is rather low and the SPM approximation of the Gaussianized discrete field with a smooth Gaussian field becomes increasingly better at higher df (cf. the recommendation $df = 120$ of Worsley *et al.*, 1996). Second, the amount of filtering used by Ledberg *et al.* (1998) is rather low. When using SPM, it is more common to filter in the range 12–16 mm for several reasons. The voxel distributions are better approximated by normal distributions, the discrete field is better approximated with a smooth Gaussian random field (i.e., sufficiently good lattice representation), smoothness in the statistic image increases, and importantly, residual anatomical variability after spatial normalization is compensated for. Furthermore, when low levels of filtering are used, i.e.,

below or of the same order as the FWHM of the PET scanner, the spatially varying point-spread function of the PET scanner may have to be taken into account. Finally, Ledberg *et al.* (1998, see Appendix and Fig. 8) indicate that the SPM result overestimates smoothness at low spatial extents of the ACF. More specifically, if the ACF is Gaussian, the smoothness is significantly overestimated at 3 voxels (corresponding to 6 mm) but not at 6 voxels (12 mm) FWHM.

In general, statistical power studies are complicated by the fact that the nature of the signal implied by the alternative hypothesis must be known. A simulation study assuming a spatially distributed signal with no predilection for one anatomical area or another indicates that it is more powerful to use the local maximum-based statistic than the cluster size statistic with SPM analysis of PET data (Friston *et al.*, 1996). However, if the nature of the signal is sufficiently different from this assumption, other test statistics may be more powerful.

CONCLUSION

Critical evaluation of methods used in a scientific field are of vital importance for progress. The constructive result of the methodological discussion between Frackowiak *et al.* (1996) and Roland and Gulyas (1996) has been a significant improvement of Roland's (1993) method (Ledberg *et al.*, 1998). Results obtained with the earlier method are difficult to interpret and may benefit from reevaluation with a more completely validated technique. Ledberg *et al.* (1998) indicate that their method is more sensitive than the SPM cluster size statistic at low amounts of filtering and low/intermediate residual *df*. The method could also be compared with SPM in parameter ranges that are more commonly used (i.e., high *df* and appropriate filtering). Ledberg *et al.* (1998) have made a promising advance, but the effects of manipulation of the convolution kernel and the ACF, and the robustness of the method, need further investigation.

Nonparametric approaches hold great potential, making minimal assumptions about functional imaging data. Additionally, Monte Carlo methods have changed the field of statistics, taking problems that were intractable and providing straightforward solutions. However, Monte Carlo approaches are critically dependent on adequate characterization and modeling of image noise. Both approaches offer flexibility in the choice of parameter ranges and test statistics. Statistical power studies will indicate how much more sensitivity can be achieved for a given level of specificity in the analysis of functional images. Recording of large fMRI time series may mean that statistical power becomes less of an issue in these circumstances.

An important theoretical point is that discrete Gaussianized *t* fields are fundamentally different from discrete Gaussian fields. In principle, this implies that Gaussianized *t* fields cannot be accurately simulated by Gaussian fields. In practice, there are two solutions to this problem, either to simulate *t* fields themselves or to approximate the Gaussianized *t* field with a Gaussian field. This approximation improves progressively at high *df* (Worsley *et al.*, 1996), but may give conservative results when spatial extent is used as the test statistic (Cao, J., The size of the connected components of excursion sets of c2, *t* and *F* fields. *Advances in Applied Probability*, accepted for publication). Alternatively, direct application of the theoretical results for *t* fields is possible and straightforward (Cao, op. cit.; Worsley *et al.*, 1996).

In summary, though the field of functional neuroimaging is still in rapid development, there is a body of well-described theories and validated methods that provides a framework for collecting reliable neuroscientific data and making biologically plausible inferences. This provides a background for the development of new analytical theories and experimental methods that can be cross-validated against the existing knowledge in probability theory, statistics, and neuroscience.

ACKNOWLEDGMENTS

This work was supported by grants from the Swedish Medical Research Council (8276) and the Karolinska Institute.

REFERENCES

Corbetta, M., Miezen, F. M., Dobmeyer, S., Shulman, G. L., and Petersen, S. E. 1991. Selective and divided attention during visual discrimination of shape, color and speed: Functional anatomy by positron emission tomography. *J. Neurosci.* **11:**2383–2402.

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

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

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

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

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

Geyer, S., Ledberg, A., Schleicher, A., Kinomura, S., Schormann, T., Burgel, U., Klingberg, T., Larsson, J., Zilles, K., and Roland, P. E. 1996. Two different areas within the primary motor cortex of man. *Nature* **382:**805–807.

Gulyas, B., Larsson, J., Amunts, K., Zilles, K., and Roland, P. E. 1997. Cortical regions in the human brain systematically participating in the processing and analysis of color. *NeuroImage* **5:**S2.

Gulyas, B., and Roland, P. E. 1994. Processing and analysis of form, color and binocular disparity in the human brain—Functional anatomy by positron emission tomography. *Eur. J. Neurosci.* **6:**1811–1824.

Gulyas, B., and Roland, P. E. 1995. Visual cortical fields participating in spatial frequency and orientation discrimination: Functional anatomy by positron emission tomography. *Hum. Brain Mapping* **3:**133–152.

Gulyas, B., and Roland, P. E. 1996. Erratum: Gulyas, B., and Roland, P. E. Visual cortical fields participating in spatial frequency and orientation discrimination: Functional anatomy by positron emission tomography. *Hum. Brain Mapping* **3:**133–152. *Hum. Brain Mapping* **4:**91.

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

Kinomura, S., Larsson, J., Gulyas, B., and Roland, P. E. 1996. Activation by attention of the human reticular formation and thalamic intralaminar nuclei. *Science* **271:**512–515.

Ledberg, A., Åkerman, S., and Roland, P. R. 1998. Estimation of the probability of 3D clusters in functional brain images. *NeuroImage,* **8:**111–126.

Lueck, C. J., Zeki, S., Friston, K. J., Deiber, M.-P., Cope, P., Cunningham, V. J., Lammertsma, A. A., Kennard, C., and Frackowiak, R. S. J. 1989. The colour centre in the cerebral cortex in man. *Nature* **340:**386–389.

McKeefry, D. J., and Zeki, S. 1997a. Mapping and topographic organization of the visual field in human area V4 as revealed by fMRI. *NeuroImage* **5:**S1.

McKeefry, D. J., and Zeki, S. 1997b. The position and topography of human colour centre as revealed by functional magnetic resonance imaging. *Brain* **120:**2229–2242.

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

Roland, P. E., and 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., and Åkerman, S. 1993. Three-dimensional analysis of clustered voxels in 15-*O*-butanol brain activation images. *Hum. Brain Mapping* **1:**3–19.

Roland, P. E., O'Sullivan, B., and Kawashima, R. 1998. Shape and roughness activate different somatosensory areas in the human brain. *Proc. Natl. Acad. Sci. USA* **95:**3295–3300.

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

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

Worsley, K. J. 1996. *An Unbiased Estimator for the Roughness of a Multivariate Gaussian Random Field.* Technical Report, Department of Mathematics and Statistics, McGill University, Montreal, Canada.

Yaglom, A. M. 1986. *Correlation Theory of Stationary and Related Random Functions, Vol. 1, Basic Results,* Springer-Verlag, New York.

Zeki, S., Watson, J. D., Lueck, C. J., Friston, K. J., Kennard, C., and Frackowiak, R. S. J. 1991. A direct demonstration of functional specialization in human visual cortex. *J. Neurosci.* **11:**641–649.