NeuroImage **9,** 311–329 (1999)
Article ID nimg.1999.0402, available online at http://www.idealibrary.com on IDEAL®

ROC Analysis of Statistical Methods Used in Functional MRI: Individual Subjects

Pawel Skudlarski, R. Todd Constable, and John C. Gore

Department of Diagnostic Radiology, Yale University School of Medicine, New Haven, Connecticut 06520-8042

Received December 24, 1997

The complicated structure of fMRI signals and associated noise sources make it difficult to assess the validity of various steps involved in the statistical analysis of brain activation. Most methods used for fMRI analysis assume that observations are independent and that the noise can be treated as white gaussian noise. These assumptions are usually not true but it is difficult to assess how severely these assumptions are violated and what are their practical consequences. In this study a direct comparison is made between the power of various analytical methods used to detect activations, without reference to estimates of statistical significance. The statistics used in fMRI are treated as metrics designed to detect activations and are not interpreted probabilistically. The receiver operator characteristic (ROC) method is used to compare the efficacy of various steps in calculating an activation map in the study of a single subject based on optimizing the ratio of the number of detected activations to the number of false-positive findings. The main findings are as follows: *Preprocessing.* The removal of intensity drifts and high-pass filtering applied on the voxel time-course level is beneficial to the efficacy of analysis. Temporal normalization of the global image intensity, smoothing in the temporal domain, and low-pass filtering do not improve power of analysis. *Choices of statistics.* the cross-correlation coefficient and *t*-statistic, as well as nonparametric Mann–Whitney statistics, prove to be the most effective and are similar in performance, by our criterion. *Task design.* the proper design of task protocols is shown to be crucial. In an alternating block design the optimal block length is be approximately 18 s. *Spatial clustering.* an initial spatial smoothing of images is more efficient than cluster filtering of the statistical parametric activation maps.

1. INTRODUCTION

Functional neuroimaging is usually based on the premise that the differences between images of the brain obtained in different mental or functional states can reveal the differential involvement of various brain structures in particular activities. One leading technique in neuroimaging is functional magnetic resonance imaging (fMRI) (Ogawa *et al.*, 1993), which has become very popular due to the wide availability of suitable instrumentation, superior performance over previous techniques, and its relative ease of use. However, although in concept the implementation of fMRI is straightforward, there remain several important issues regarding the analysis of fMRI data that remain unresolved. For example there is little consensus on the proper methods of statistical analysis that should be used, and this makes it difficult to compare and evaluate results between the growing number of sites working with fMRI. The weakness of fMRI signals recorded in studies of complex cognitive functions, and the arbitrariness of the choice of data analytic strategies, raises concerns that published results may become significantly skewed to fit the expectation of the neuroscience community—that is, the results of statistical analyses which conform to expectations are more likely to be believed, accepted for publication, and quoted. Attempts to establish the best available statistical procedures are important not only to increase the power of the technique but also to limit the experimental freedom in the choice of processing strategies and thereby eliminate bias in selecting those that detect the "right activations."

The statistical problems faced in fMRI may seem at first sight to be relatively simple, so that it should be possible to derive optimal processing techniques on a theoretical basis. However, there is little agreement between statisticians working in this field upon the choice of the best strategy. The main reason for this is that fMRI signals are in reality quite complex in their structures. For example, they include various nonuniform sources of noise and artifact that cannot be easily described and accounted for in general statistical models. The observations are not independent, either in time or in space. Attempts at statistical analyses of fMRI data from first principles usually rely on several simplifying assumptions that are difficult to establish

1053-8119/99 $30.00

and are usually not satisfied. The importance of deviations from such theoretical models is poorly understood. Simple analysis of the rate of false-positive activations found in practice shows that such models do not correctly predict the significance of observations. For this reason, even if a satisfactory statistical theory can be constructed, it will have to pass a test of experimental confirmation before being widely applied.

We propose an alternative practical approach to the evaluation of fMRI processing methods by making comparisons between different techniques of analysis using the receiver operator characteristic (ROC) method (Skudlarski *et al.*, 1997). Using data obtained in a real fMRI study, we create data sets in which activation foci are artificially added so that their intensity and spatial extent are known. We then apply various methods of data analysis to this set of images and measure how accurately each method can recognize the presence and locations of activations. This enables us to compare the accuracy of outcome of each analysis with the known distribution of artificially added activations. We have previously used (Constable *et al.*, 1995) this approach to compare different implementations of *t*-statistical tests.

In this paper we look at various other statistical measures used in fMRI analysis but consider here only their performance for detecting activations, without attempting to assign any probabilistic interpretations to the significance of the results. We compare them using a single criterion—the ability to detect most of the real activations while minimizing the detection of false activations.

ROC analysis was adopted for this purpose (Constable *et al.*, 1995) and has been used by others (Forman and Cohen, 1995; Sorenson, 1995; Xiong *et al.*, 1996) for similar purposes but mostly to validate particular approaches used in fMRI, and usually using computer simulated data sets with noise of a specific stochastic nature. However, we believe this approach can be misleading. The general validity of a particular method based on a theory that assumes noise of certain characteristics cannot be established by applying it to a data set with noise with precisely those properties. In our approach to simulation we use actual data from real fMRI experiments, which should therefore more closely match the noise encountered in practice. We then add artificial activations that realistically simulate typical fMRI activations.

The goal of this paper is to provide an objective way of choosing optimal methods and parameters in fMRI analysis to increase its power and reduce subjective elements that otherwise influence the results obtained. We consider all the steps that are typically involved in the analysis of fMRI data of a single subject.

We begin with a description of our implementation of the ROC technique and define criteria for assessing the efficacy of processing steps. This technique is then applied to compare different methods that can be chosen in subsequent steps of the fMRI analysis. The efficacy of several preprocessing steps such as temporal normalization, drift subtraction, and frequency filtering are analyzed. Next the efficacy of different statistical methods that can be used to create statistical parametric maps (SPM) from data obtained during single imaging runs are compared. The design of the study, the size, the number, and the patterns of blocks of activation and control tasks are found to be crucial and are further analyzed. Different methods of using the spatial correlation of expected activations, such as cluster filtering of statistical maps, smoothing of the raw images, and smoothing of final maps, are compared. Finally, methods of creating one composite result SPM for a study of a single subject consisting of multiple separate imaging runs are considered. Two of the above simulations (comparison of statistics used for creating single SPM and the effects of the temporal normalization) were performed additionally on different data sets using two methods of motion correction, yielding results very similar to those obtained without motion correction. Appendices describe in detail the methods used for linear drift removal (Appendix A) and spatial multifiltering (Appendix B). Finally, Appendix C presents some statistical measures obtained from the data sets we investigated. This may be useful to compare our results with results obtained on other data.

2. METHODS

2.1. ROC Method

The application of ROC methods to the analysis of fMRI processing techniques was introduced by Constable *et al.* (1995). It has been used extensively as a tool for objective comparisons of various strategies (Skudlarski *et al.*, 1995; Friston *et al.*, 1996; Xiong *et al.*, 1996). The basic premise of this method relies on adding artificial activations to a set of raw images and applying each method being studied to this altered data set. For the proper application of ROC methods the MRI images should contain noise and artifacts representative of fMRI data obtained in practice (a feature that is sometimes neglected (Sorenson and Wang, 1996; Xiong *et al.*, 1996)). The locations and intensities of detected activations can then be compared to the known pattern of the added activations to thereby measure the accuracy of detection. The relationship between the true-positive ratio (proportion of correctly detected activations to all added activations) and the false-positive ratio (proportion of pixels that were incorrectly recognized as active in all pixels without added

activations) describes the power of the technique. If the detection procedure has a parameter that controls its sensitivity then by adjusting that criterion (usually the threshold level) the resultant curve shows the relationship between the proportion of true-positive and the proportion of false-positive activations. The precise shape of the ROC curve depends on the characteristics of the probability distributions of the signals and noise and the degree to which they overlap, but does not make assumptions about these distributions.

2.1.1. Obtaining single "value of merit" from the ROC curve. In a situation when many different aspects of each detection algorithm are compared, it is desirable to produce a single quantitative figure of merit for each ROC curve. Several methods have been proposed (Metz, 1978; Swets, 1988), including the integral of the full ROC curve and its best operating point (point furthest from the diagonal). In this paper the mean of the ROC curve over the limited range of false-positive ratio between 0 and 0.1 is used as such a measure. This somehow arbitrary value of 0.1 was chosen as an upper limit for the false-positive ratio that is used in the fMRI studies. By limiting this integral to low (but realistic) false-positive rates (high thresholds only) we limit the scope of our analysis to the cases that are of primary interest in fMRI, when the ratio of false activations is much smaller than the ratio of real activations. The importance of this limit is obvious when analyzing the efficiency of some cluster filtering techniques as represented on the left panel of Fig. 1. In one case a "neighborhood filter" pixel is considered active only when a certain number of its closest neighbors are active as well. In this approach pixels on the border of activated regions will not be treated as activated even for an extremely low threshold level. In such circumstances, the true-positive ratio will be always significantly smaller than one. This may offset the high efficiency of such a filter in the more interesting regime of a more realistic and higher threshold. As the threshold is changed the ROC curves of different techniques

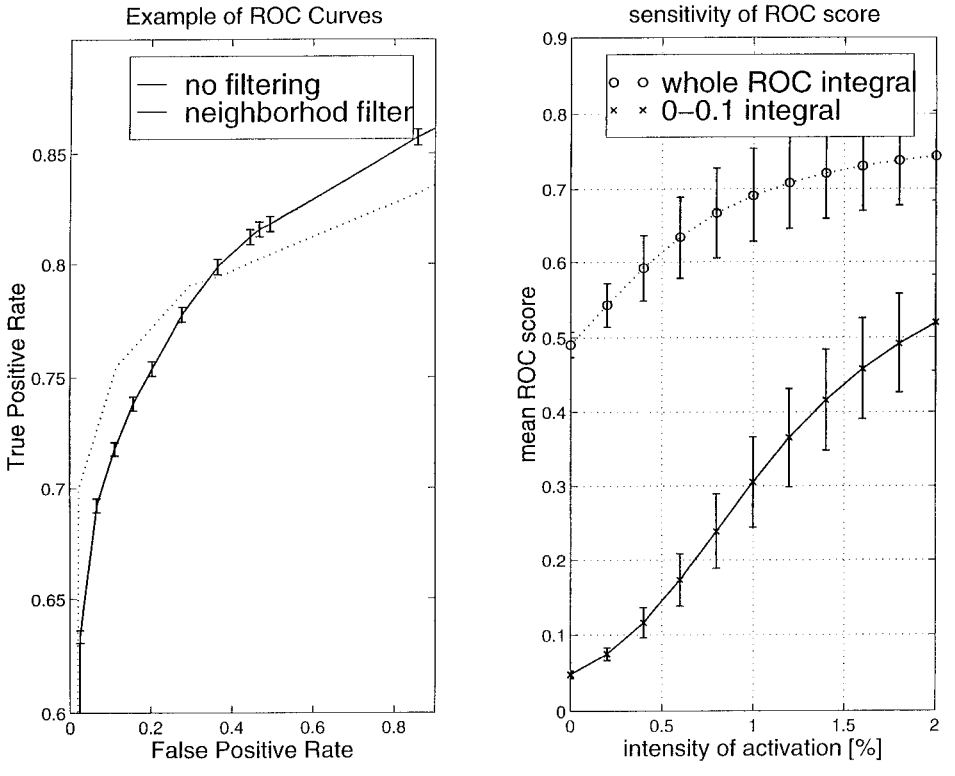

FIG. 1. The left panel presents two representative ROC curves. The solid line was obtained using *t*-statistics without spatial processing while the dotted line was obtained using the neighborhood filter. Activations were added in large (25 voxel) foci. The curves cross. The curve obtained with the neighborhood filter, although obviously better in the low false-positive regime (the working regime of fMRI), would be considered inferior if the integral of the whole curve is used as a measure of accuracy. The right panel presents the mean measure of the ROC curve as a function of intensity of activation. The dotted curve was obtained from the whole ROC curve while the solid curve was obtained using only 0–0.1 region as used in this paper. This relation helps to interpret the differences in ROC power of various techniques by translating them to an equivalent change in the contrast to noise ratio (CNR). In the range of intensities used in this paper, an increase of ROC measure by 0.01 is roughly equivalent to a 4% increase in CNR. The limited ROC integral proves to be more sensitive to the intensity of added activation. A statistically significant difference in the ROC power, as defined here, can be observed for 30% change of strength of activation, while a 100% change is necessary to gain significant change in the ROC power using the full ROC integral. In this simulation we used only one imaging run in a study so that the estimated error is larger than in the later results.

may cross and it is important to choose the one that is higher in the regime that is most relevant for practical fMRI applications: that of low false-positive ratios. In addition we find that our criterion for judging ROC curves performs better than using the entire ROC curve integral as can be seen in Fig. 1 (right panel). The mean ROC score is plotted there as a function of the intensity of the added activation. Using the whole ROC integral we need to change the intensity of the added activation by more than a factor of 2 to find a significant (larger than estimated error) change in the ROC score. Our limited integral of 0–0.1 regime is sensitive to changes in the added activation intensity of 30%.

Because the ROC curves contribute to our analysis only through their integral our results are essentially equivalent to a conventional power analysis that is averaged for a range of Type-I error levels (alpha between 0 and 0.1). This averaging is performed for two reasons. This makes our results less dependent on a particular alpha level, which can be chosen differently for various studies. Depending on the study size (in terms of the number of subjects and imaging series and the required final significance) the required alpha level for the analysis of individual series may vary widely. The other more pragmatic reason is that this averaging stabilizes the results of random error in our simulations and thus produces a more precise estimate of the relative power of different methods of analysis.

In the examples that follow, all of the results for the power of different techniques will be given in terms of the parameter *P*, the mean value of the ROC curve over the region in which the false-positive ratio lies between 0 and 0.1. This *P* value is always between 0 and 1. The latter would reflect a perfect technique that recognized all true activations without returning any false-positive findings. For a completely random method (guessing as a way of detecting activations) $P = 0.05$. The right panel of Fig. 1 presents the values of *P* as a function of the intensity of added activations. This curve (calculated using *t*-value as the activation detection tool) can be used later to interpret the significance of gains in the statistical power of different statistical methods. An increase of 0.01 in the ROC score is seen to be roughly equivalent to an increase in the signal contrast by 4%, or the same reduction of the noise. Taking into account the fact that the error in the ROC score (see below) varies between 0.005 and 0.1 we can state that we are able to detect gains in the power to detect activations equivalent to an increase in the ratio between noise- and stimulus-dependent change of signal intensity of 2–4%.

2.1.2. Simulation activations. The activations we added were defined as sinusoidally varying in time at various frequencies. We believe that this procedure provides data sets with known artificial activations that are similar to true fMRI data sets. The onset and decrease of fMRI signal with activation produce signals that may resemble sine waves. The amplitude of added activations was varied between 0.3 and 3% of the image intensity, equivalent to a range of 0.1 to 1 of the standard deviation of the signal intensity for individual pixels. This is a range common for fMRI activations in cognitive tasks, which are generally weaker than sensory-motor activations that are more robust and easier to detect. The results were stable with respect to the intensity of activations in this range and the results presented later were obtained for a single intensity of 1.5% (amplitude of signal difference between peak On and Off conditions equal to 0.5 of the average noise standard deviation).

Except for the simulation for temporal normalization, the synthesized activation signal was added to 10% of pixels. Activated pixels were grouped in randomly spaced clusters of about 10 pixels (the size of clusters was varied in the analysis of spatial smoothing/clustering algorithms). The high number of activated pixels was chosen to increase the power of our analysis. However, in the analysis of the effects of time normalization using the whole image intensity, when a large number of activations may distort the results, a smaller number of activated pixels (1, 2, and 4%) was used.

One problem with our procedure is that activations are added in the same positions in the image, rather that in the same position in the brain. This means that the effect of motion in masking activations is omitted. This may lead to underestimation of the statistical power of techniques that are significantly better in treating motion. Our approach is thus not applicable to analyzing the performance of motion correction algorithms. Since the motion correction may affect the relative merits of various techniques, we also present some results calculated with and without motion correction. These results cannot be used to directly compare the power of analysis with and without motion correction but they show that motion correction does not significantly change the results of our analysis.

2.1.3. Analysis of accuracy of detection rather than estimate of significance. It must be emphasized that in this paper we do not address the issue of calculating the statistical significance of activations that are detected. We concentrate on finding the most powerful strategy and do not attempt to estimate the absolute significance of its results. Such calculations are to this date highly problematic in fMRI due to the complicated and variable nature of the noise. Our approach allows us to compare the losses and gains in the statistical power produced by isolated steps of the data analysis. This can be done with far less stringent (and thus more realistic) assumptions about the characteristics of the signal and noise distributions than those necessary to estimate the statistical significance of findings. We believe that currently the best method for estimating

the statistical significance of fMRI findings in order to estimate *P* values of activations is obtained by careful randomization of actual MRI images used in the same study, the so called "bootstrap technique" (Arndt *et al.*, 1996; Bullmore *et al.*, 1996; Skudlarski and Gore, 1996).

It should be noted that due to several violations of common statistical assumptions (mainly the assumptions of independence in both space and time domains) the values of *t*-statistics that we calculate should not be directly interpreted to have their typical statistical meaning. We use this and other "statistics" merely as measures that reflect in some way the intensity of activations. Our study is devoted only to finding which measure is the best at detecting true activations in the presence of the noise.

2.2. Imaging

We used images taken from 8 subjects from an fMRI study of attention (Peterson *et al.*, 1997) in which runs of 128 images/slice were taken while subjects performed the Stroop task. In this study, the four periods of active condition were interleaved with four periods of rest in each imaging series. We have chosen slices from the superior regions of the brain that did not produce significant reproducible activations in those tasks. The pattern of artificially contrived activations was always different from the pattern of real activity so that any images containing real activations were assigned to both the "active" and the "control" group. Before adding activations, the sets of images assigned to be "activated" were not statistically different from those assigned as "control." We have chosen to use these data instead of a series of blank images taken with the subject resting in the magnet because our experience (Skudlarski *et al.*, 1995) shows that data sets taken during the performance of a real fMRI study differ significantly in the amount of variance from data sets taken while subjects are resting during the entire imaging series. Most probably this difference can be attributed to differences in the amount of microscopic motion—that is motion smaller than 0.3 mm (less than one tenth of the pixel size).

Each study was performed on a GE 1.5 T Signa MR unit equipped with echo planar imaging (EPI) (Advanced NMR, Wilmington, MA). The imaging parameters were as follows: $\alpha = 60°$; echo time, TE = 45 ms; repetition time, TR = 1500 ms; field of view, FOV = 40 ∗ 20 cm; slice thickness, 8 mm; matrix size, 128 ∗ 64; and Nex = 1. Analyzed studies were screened for motion by analysis of the center of mass: no gross (larger than 0.5 pixel) movements were observed, and no motion correction was performed, as every motion correction procedure available changes significantly the structure of signal and its spectral distribution. In the last section we present results recalculated for a study analyzed with two methods of motion correction.

2.3. Model Scheme of Data Analysis

In a typical study such as the one considered here the subject is imaged in several identical imaging runs (here four runs). During each imaging series two tasks (A, B) are interleaved at various frequencies between 1 (AB), 1.5 (ABA), and up to 10 on/off cycles per imaging series. The data analysis is performed as follows (the details of each of these steps will be discussed below):

- Data is preprocessed using spatial smoothing, drift elimination, temporal normalization or temporal filtering (with a high-pass filter).
- For each imaging series one statistical parametric map (SPM) is created.
- The SPMs from identical series from the same subjects are combined into one SPM representative for this subject.
- Activation maps are thresholded and cluster filtered.
- *t*-Value was used as activation measure unless it was specifically noted.

3. PREPROCESSING IN THE TIME DOMAIN

3.1. Time Normalization

Since the intensity of the MRI images may change during an imaging run it is common to employ time normalization to eliminate variance due to changes in the global intensity. In this procedure the overall intensity of every image is multiplied by a factor that estimates the scanner instability. Such a normalization is justified if the variation of global intensity is due to some global mechanism and is uniform across the whole image, but can produce deleterious effects if the apparent change in the mean intensity comes from localized variations. The potential advantage of this procedure is to eliminate one source of possible artifacts. There are two possible disadvantages: real activations may affect the intensity of the whole image (and thus may be decreased through normalization), or the estimate of noise (such as one used in calculating *t*-statistics) will be distorted by the selective removal of part of the variance.

3.1.1. Methods for time normalization compared. We have compared three methods of performing time normalization using three different normalizing coefficients. We consider the mean intensity of the whole image and the mean intensity within the brain (this will exclude possible artifacts outside of the brain from affecting the normalization parameters). In the third approach we calculate the histogram of intensity within

an image, fit its central part with a gaussian, and then find the position of its peak. For each of these methods of determining the normalizing parameters each image intensity is divided by this parameter so that the appropriate estimate of mean intensity is constant after normalization. This normalization is performed separately for each slice, so that changes of intensity in one slice do not affect the intensity of the others. One possible sensitive issue is the extent of activation: if a strongly activated area is large (such as in the case of visual stimulation) it can affect the mean intensity of the whole image. In such a case the normalizing procedure will reduce the real effect. To safeguard against this possibility in our simulation we considered 1, 2, 4, or 8% of pixels to be activated. This fraction is important because the time normalization procedure is based on the assumption that activation changes the global image intensity only slightly. The larger the activated area, the less likely time normalization will be appropriate, because the effect of stimulus is likely to be present in the measure used for the temporal normalization. If no global sources of noise are present the activation covering *x*% of the image will decrease its intensity by *x*% in result of temporal normalization.

3.1.2. Time normalization does not improve accuracy of detection. The results for the realistic yet conservative 4% case are presented in Table 1. One can see that normalization based on the mean intensity of the whole image or mean intensity of the brain significantly decreases the power of the analysis, whereas the more sophisticated histogram fitting technique gives results that seem to be slightly worse but are not significantly different than with no time normalization. Only for the case when 2% of the pixels are activated, in which the area activated is smaller the time normalization improved efficiency, but this difference was not statistically significant. For 1% of added activations the estimated error of analysis was even higher and the (not statistically significant) advantage of time normalization minuscule. Thus we can establish an upper limit for the possible gain due to the temporal normalization. Even if temporal normalization is helpful for detecting localized activations (covering less that 4% of analyzed brain tissue), it may increase the power to detect activations by the equivalent of no more than a 2% increase of signal to noise ratio. We therefore conclude that temporal normalization of the global image intensity should not be performed. The time normalization was also not helpful when combined with the removal of the low frequency drift discussed in the following section. The results of our measurements of efficacy for various stimulation frequencies, presented in Fig. 2, suggest that the fMRI studies used in our work do exhibit linear drifts. The fact that time normalization was not helpful can be most likely explained by postulating that the drift was not uniform across a whole image. If the drift was created by variations in the B_0 field uniformity this would be visible as minute linear motions in the imaging plane that in turn cause nonuniform intensity drifts that depend on the local gradient in the image intensity.

TABLE 1

	Freq = 1.5 (ABA)	Freq = 3.5 (ABABABA)	Freq = 7.5 (ABABABAB-ABABABA)
t-stat	0.29 ± 0.012	0.42 ± 0.012	0.52 ± 0.012
Paired *t*-stat	0.16 ± 0.009	0.31 ± 0.009	0.53 ± 0.009
Skewed *t*-stat	0.28 ± 0.012	0.42 ± 0.012	0.54 ± 0.012
Boxcar correlation	0.29 ± 0.011	0.42 ± 0.012	0.52 ± 0.012
Exact correlation	0.33 ± 0.013	0.45 ± 0.014	0.54 ± 0.014
Percentage difference	0.17 ± 0.014	0.27 ± 0.015	0.44 ± 0.014
Skewed percentage difference	0.19 ± 0.014	0.28 ± 0.013	0.39 ± 0.015
Fourier	0.15 ± 0.015	0.26 ± 0.016	0.30 ± 0.014
Mann–Whitney	0.29 ± 0.013	0.42 ± 0.012	0.52 ± 0.014
Split-2 *t*-stat	0.2 ± 0.03	0.39 ± 0.01	0.505 ± 0.01
Split-3 *t*-stat	0.13 ± 0.03	0.37 ± 0.01	0.49 ± 0.01
Split-4 *t*-stat	0.12 ± 0.02	0.29 ± 0.01	0.41 ± 0.01

Note. The ROC measured power of various statistics applied for three different study designs, with various lengths of individual stimulus presentations. The *t*-statistics, Mann–Whitney, and correlation methods prove to be best and very similar in their effectiveness. The effect of frequency of task switching is clearly visible—faster task switching helps us eliminate strong low frequency noise.

Our experience with large numbers of subjects analyzed in semiautomatic fashion shows that some studies show some unusually strong shifts in the global image intensity. In such cases use of time normalization may save otherwise unusable data. However, our previous results show that it would be unwise to apply time-normalization as a general procedure—the average time course of overall image intensity should be monitored (this can be easily done together with motion and ghost artifact analysis) to alert for unusual drift artifacts and to allow for the time normalization to be applied selectively.

3.2. Removal of Low Frequency Drift/High-Pass Filtering

An alternative technique of preprocessing in the temporal domain to eliminate artifactual low frequency drifts is treat each voxel time course individually (without assuming, as in the previous case of time normalization, that the low frequency noise of every pixel has an identical global time course) (Biswal *et al.*, 1996). The rationale for this technique is based on the observation that the power of the fMRI noise (in the time domain) is concentrated at the low frequency end of the spectrum, and as long as the expected response has a high temporal frequency, the low frequency components of the spectrum can be filtered out. Of course this method can be used only for experimental

designs that include several switches between task/control conditions within each imaging run. The advantage of this procedure is the possibility of selective elimination of some noise without removing genuine fMRI signal. A disadvantage is that this procedure may skew the distribution of data points and this may lead to an underestimation of the data variance and in the possibility of removing part of the true signal. Similar results can be obtained using a high pass filter that attenuates the low frequency part of the spectrum. The filter parameters should be chosen so that they remove as much of the unwanted low frequency changes as possible without affecting signal variation at the frequency of the stimulus.

In this simulation we assumed activation occurred at a frequency of 3.5 cycles for the 128 image long series (ABABABA design). We compared a high pass filter with methods of drift subtraction in which the linear, quadratic, or cubic component of each voxel time-course was fit and subtracted. For high-pass filtering we applied Butterworth filters with cutoff frequencies between 0.7 and 0.2 of the stimulus frequency. Table 1 presents samples of our results. The linear drift removal increases efficiency, the removal of the quadratic polynomial makes a further improvement, but the cubic polynomial does not help any more. The best overall result is obtained by use of high pass Butterworth filtering at the frequency of 0.35 of the stimulus frequency. Combining the technique of the time normalization with the quadratic (or high-pass filter) drift removal does not further increase the ROC power.

3.3. Smoothing in the Temporal Domain

Some groups (Frackowiak *et al.,* 1997) recommend smoothing in the temporal domain with a filter whose width is defined by the width of the hemodynamical response curve. We have compared the efficiency of three different statistics (*t*-statistics, cross-correlation with the activation time-course and Mann–Whitney statistic) for three different frequencies of added signal using temporal smoothing of different widths.

3.3.1. Smoothing in temporal domain decreases efficiency of analysis. The results for the cross-correlation analysis are presented in Table 2. We can see that temporal smoothing drastically decreases the power to detect activations. The greater the smoothing the larger the loss in ROC power. The nonparametric Mann–Whitney statistic is least affected by temporal smoothing but it is still degraded by the use of temporal smoothing. The numerical values of the statistical measures used (*t*-values, correlation parameters or Mann–Whitney parameter), even if calculated with correction for the decreased effective number of degrees of freedom, were largely increased by the temporal smoothing because the variance decreased. However, they were increased for both active and, even more, for inactive pixels and thus the power to detect real activations dropped. This was probably due to the fact that the estimate of noise used by each of those statistics was degraded by the use of smoothing. We conclude that temporal smoothing is not only not beneficial in detecting activations, but it may also lead to gross overestimation of the significance of fMRI findings.

TABLE 2

	Freq = 3.5 (ABABABA) no motion corr.	Freq = 7.5 (ABABABABABABABA) no motion correction
No temp. smooth.	0.42 ± 0.01	0.58 ± 0.01
FWHM = 1.5 image	0.38 ± 0.01	0.54 ± 0.01
FWHM = 3 images	0.30 ± 0.01	0.44 ± 0.01
FWHM = 6 images	0.23 ± 0.01	0.35 ± 0.01

Note. The power of cross-correlation with the exact time course of activations, and nonparametric methods are compared with temporal smoothing of three different widths: 1.5, 3, and 6 images and with no temporal smoothing. Simulation was performed with three different frequencies of activation patters, 3.5 and 7.5 task pairs in the imaging run. In all cases the temporal smoothing significantly decreases the power to detect activations measured by the integral of the ROC curve. The same simulation for *t*-statistics and Mann–Whitney statistics shows similar behavior, but the Mann–Whitney statistics is less sensitive to temporal smoothing.

3.3.2. Temporal correlation in the time domain is responsible for the failure of temporal smoothing. The results of simulations showing that temporal smoothing decreases the power of analysis comes as a surprise since it contradicts previous views and thus requires a more complete analysis. Temporal smoothing can be beneficial for detecting signal buried in white noise that has been convoluted with a known response function (as justified for example by the "matched filter theorem" (Frackowiak *et al.,* 1997)), but it clearly diminishes our ability to distinguish between real and false-positive activations in the fMRI study. This further enforces our view that the direct analysis of fMRI data (e.g., based on the randomization, bootstrap technique) cannot be substituted by theoretical predictions based on simplifying assumptions about noise structure. To find out precisely which violation of statistical assumptions is responsible for this effect, we performed a similar analysis with scanner noise replaced by white, uncorrelated gaussian noise. In such a model temporal smoothing proved to be beneficial. However, if this noise was altered to introduce a slight temporal autocorrelation (it was smoothed in the time domain before adding activations) the benefit of temporal smoothing disappeared and the analysis without any smoothing proved to be superior. Thus we conclude that it is the correlation present between consecutive MRI images that is responsible for this effect. To ensure that this correlation was not introduced by movement effects we

performed the same analysis with motion correction. The details of this analysis are presented in section 8. Results for the analysis of temporal smoothing are presented in Table 6. The motion correction performed with and without movement decorrelation does not change the relative efficiency of analysis with and without temporal smoothing.

For completeness we included a low pass filter (at frequencies of 2 and 5 times the stimulus frequency) into our analysis but the results in these cases, as seen in Table 1, still show no benefit for temporal smoothing.

3.4. Summary

Temporal normalization on the whole image level does not increase the power of fMRI analysis, suggesting that image intensity variations have a nonuniform spatial structure and cannot be removed or even decreased by application of a global correction. On the other hand temporal detrending of individual voxel time-courses is highly beneficial—the best method appears to be high-pass filtering with a cut off frequency of about 0.35 of the stimulus switching frequency, but it is only marginally better than removal of quadratic drift components from the signal. Temporal smoothing and low pass filtering decrease the ability to detect real activations. These results suggest that better metrics (statistics) should be developed that incorporate the temporal correlation present in the fMRI signal in error estimations. Clearly the functions that incorporate some sort of error estimation (such as *t*-statistics and Mann–Whitney statistics) are superior to those that rely on the signal change only (e.g., percentage of difference in the mean signal intensity), but functions whose variance estimate would remain unbiased by the temporal correlation's present in data should perform even better.

4. STATISTICS: CALCULATION OF SPMS

4.1. Description Here we compare the power of different analytic techniques to create SPMs from a single imaging series. We compare the following:

- *t*-statistics
- paired *t*-statistics
- skewed *t*-statistics (with correction for linear drift)
- time-course correlation (using the block task/control function as a correlate)
- time-course correlation (using the actual sinusoidal activation curve as a correlate)
- Fourier spectral analysis
- simple subtraction
- subtraction with correction for the linear drift
- nonparametric Mann–Whitney tests.

The main advantage of the time-correlation method is that it allows the possibility of incorporating our knowledge of the precise pattern of the signal change into the data analysis. We may then search not only for the signal increase or decrease, but also for changes that can be expected from our knowledge of the hemodynamical and neuronal response. To quantify this advantage the time-correlation approach was analyzed in two ways; via correlation with a boxcar function and by correlation to the actual sinusoidal variation of the added activations. The first method is very similar to using a *t*-statistic (since in the *t*-test we assume constant level of activations), while the second is used to find the upper bound for the gain obtained by incorporating the proper time-course of activation. Here we correlate the fMRI signal time course to the exact shape of the activation added—which is more than can be hoped for in any real experiment since our knowledge of the hemodynamic response in various parts of the brain is only approximate. Our implementation of the Fourier method measures the power of the signal component at the frequency of the stimulus divided by the mean signal power in the wide frequency band surrounding this frequency as a measure of activation. We have also compared the split/2, split/3, and split/4 *t*-statistics (Constable *et al.*, 1995). These statistics divide the data set into 2, 3, or 4 equal blocks, calculate *t*-statistics for each block separately, and return the smallest value. In effect they require that for a given threshold a specified number of blocks reaches this threshold. This can be seen as a form of internal replication approach. All of these techniques were applied with three different frequencies of the stimulus (1.5, 3.5, 7.5 task/control periods within an imaging run of constant length and imaging time).

In Appendix A we present the details of the "skew corrected" version of the *t*-statistic. In this approach the test is applied to compare the mean intensities of images in Task and Control states but is not affected by the effect of any uniform drift present in the data.

4.2. Results

Table 3 presents the mean ROC power (P) for each of the tests. For all statistics, the higher the frequency of the task switching, the greater the power of the method. This is due to the fact that most of the noise in fMRI data lies in a low frequency range. In section 5 below the implications of this for optimal task design will be further discussed.

The Fourier method and simple subtraction (with or without drift correction) performed significantly worse than the other techniques. The paired *t*-statistic performed poorly when the task and control presentation was low frequency (image pairs are far apart and thus

TABLE 3

No preprocessing	0.424 ± 0.005
Time norm. (whole image mean)	0.403 ± 0.006
Time norm. (brain mean)	0.402 ± 0.006
Time norm. (histogram peak)	0.416 ± 0.006
Linear term subtracted	0.425 ± 0.005
Quad. term subtracted	0.434 ± 0.005
Cubic term subtracted	0.434 ± 0.005
Linear term subt. (with time norm.)	0.402 ± 0.006
Quad. term subt. (with time norm.)	0.413 ± 0.005
Cubic term subt. (with time norm.)	0.412 ± 0.005
High-pass filter best cutoff frequency (frequency of 0.35 of stimulus frequency)	0.435 ± 0.005
High-pass filter (frequency of 0.25 of stimulus frequency)	0.422 ± 0.006
High-pass filter (frequency of 0.7 of stimulus frequency)	0.412 ± .006
Low-pass frequency filter (frequency of 2 of stimulus frequency)	0.403 ± 0.007
Low-pass frequency filter (frequency of 5 of stimulus frequency)	0.414 ± 0.008

Note. Comparisons of various methods of preprocessing the time course of voxel intensity. In this example activation was added to 4% of brain pixels. Time normalization based on normalization of the whole image intensity proves *not* to be efficient. High pass filtering significantly improves the analysis but caution has to be applied with the choice of the cutoff frequency. The removal of the quadratic estimate of image drift has similar effect in enhancing the power of analysis.

not really correlated) but is one of the best methods for fast switching paradigms. Split *t*-statistics perform worse than regular *t*-statistics—the more splits the worse the performance. This may seem to be contradictory to the results of our previous work (Constable *et al.*, 1995) that found the split-4 method to be one of the best techniques, but the explanation is simple. Splitting data into subgroups and performing *t*-statistics separately is justified only if there is a significant additional variance between those groups. This happens when data are collected in separate individual series and then combined together as was the case with Constable *et al.* (1995). In this situation, variations between imaging runs are significantly larger than within runs and split statistics are still useful, as will be shown in Section 7. In the simulation presented in Table 3 we consider data taken during one imaging run, and there is no justification for splitting them into subseries.

The skew correction does not improve the power of *t*-statistics. This is the case because our study design AB . . . A (frequency of 1.5, 3.5, and 7.5 per imaging series) is not susceptible to the effect of linear drift. Only if this design cannot be used (e.g., pharmacological studies when the effects of the drug wash away too slowly) should the skew correction be applied, as shown in Section 3 and on Fig. 2.

The cross-correlation, *t*-statistic, and Mann–Whitney do not differ in performance significantly. Using the exact time course of the added activation does increase the power of the correlation method but only by a small margin.

4.3. Summary

The cross-correlation method is very similar in power to the *t*-statistic and Mann–Whitney method and these all are significantly better than using percentage difference measures and the Fourier method.

Cross-correlation to a boxcar function is as good as simple *t*-statistics, while cross correlation to the exact response function gives a slight but statistically significant increase in the statistical power.

The use of skew corrected *t*-statistics is not helpful for this study design, which begins and ends with the same task (AB . . . A). We will see later (Section 5.3 and Fig. 2) that it is justified in the even (AB . . . AB) designs.

Paired *t*-statistics perform poorly when the stimulus period is long (in this case pairs of images are distant in time), while it approaches the power of regular *t*-statistics at high task/control switching frequency (in this case the paired images are close in time and paired *t*-statistics gains because it disregards the long scale signal variations).

Split statistics should not be used for data collected during one imaging run. Methods for combining the data collected in separate imaging series will be discussed in Section 7.

The most striking effect noticeable in Table 3 is that the effect of task design and the frequency of stimulus alternation are much more important than the effect of the choice of statistics used. This trend occurs with all statistical methods. We analyze this effect more precisely in the following section.

5. TASK DESIGN—FREQUENCY AND PHASE UNCERTAINTY

5.1. Description

Comparisons of the statistical power obtained at different stimulus frequencies confirm the well known (Weiskoff *et al.*, 1993) fact that in fMRI it is the low frequency signal variations that generate most of the false activations, so it is useful to switch task/control stimulus conditions frequently. This assumes however that the fMRI response is in phase with the stimulus. If the phase delay of the measured BOLD response is known it can be accounted for using correlation (by convoluting the target wave form with the hemodynamical response function) and *t*-statistic (by reassigning some images between tasks and/or dropping some images at the transition between tasks) analysis. How-

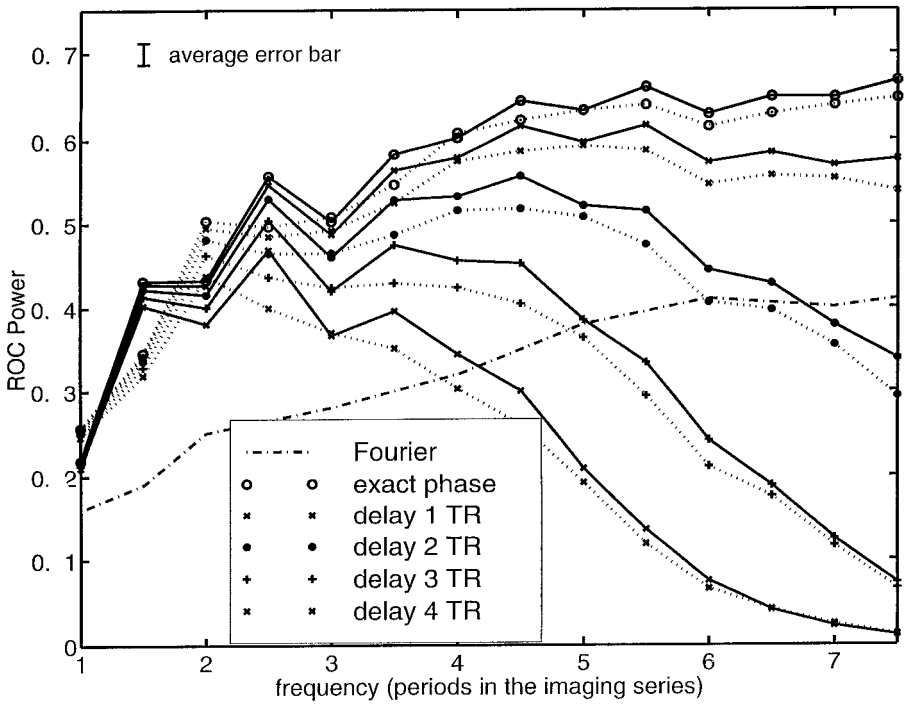

FIG. 2. The ROC power of analysis is presented here as a function of the task/control stimulus switching frequency for various shifts (phase mismatches) between the introduced signal and observed (expected) fMRI response. The dotted line presents the skew-corrected *t*-statistics, the solid line presents the cross correlation with the sinusoidal reference function. The dashed dotted line presents the power of the Fourier power spectrum method that (in our implementation) is independent of the response phase. For a response uncertainty larger than 1 TR (1500 ms in our study) the efficiency peaks at 3–4 periods per 128 image long imaging series. The response for imaging runs of various length is analyzed on Fig. 3. The solid curve presenting results obtained using correlation method is exhibiting spikes for every noninteger value of frequency. This represents the uneven (AB . . . A) study design that is much less sensitive to the presence of the linear drift. Both correlation analysis and normal *t*-statistics (not shown here) prove to be much better for this balanced study design. The skew-corrected *t*-statistics (described in the Appendix A) is not sensitive to linear drift and thus performs much better for the unbalanced (AB . . . AB) study design, this advantage disappears in the balanced study design makes it less sensitive to the presence of linear drift. This proves that the studies used in our simulation were exhibiting significant intensity drift. While skewed *t*-statistics perform generally worse they are significantly better for data from studies with integer frequency of task switching. Those points represent the uneven (AB . . . AB) task design that is more susceptible for the temporal drift artifacts.

ever, problems may arise when this delay is nonuniform within the brain or when it is unknown (for example, due to any additional significant delay in the performance of more complex cognitive tasks) (Alperin *et al.*, 1996). In cases when that phase is unknown the higher the frequency of the task/control changes, the bigger the error due to the phase mismatch between stimulus and response. Our simulation was designed to understand the consequences of phase delays in tasks of different frequencies and to find an optimal study design by striking a balance between these two opposing factors.

In our simulation we compared the statistical power obtained for frequencies of the task stimulus presentation between 1.5 and 10, and for various lengths of the imaging series (between 32 and 128 images) with an unaccounted phase shift between 0 and 5 images (0–7.5 s). Our goal was to find the optimal frequency for different degrees of the phase uncertainty. The simulation was performed using two methods: the skewed *t*-statistic and by correlation to the actual sinusoidal stimulus time-course. These were also compared to the power of the Fourier method (which performs weakly when there is no phase error, but is insensitive to such a phase shift). Of course, phase can be taken into account in the Fourier analysis but in this case we are interested in a method that is not affected by erroneous assumptions about phase.

5.2. Results

Figure 2 presents the power of the ROC analysis for various task frequencies in which the added signal response is delayed in relation to the "expected" signal response.

The dotted line presents results obtained using the skew-corrected *t*-statistics, while the solid line was obtained by correlation to the actual sinusoidal activation curve. The dashed-dotted line shows the results of the phase insensitive Fourier method. The results

presented in this plot were obtained using an imaging run of 128 images. Without the phase error the power curve saturates at about 5 cycles (18 s and 12 images for task). With the added phase error the power curves still increase at low frequency but then peak and decline as the task periods become so short that the phase error starts to interfere with accuracy.

Even if the phase uncertainty is as large as 4.5 s (3 images), the best frequency lies between 3 and 4 cycles (30–22 s per task presentation); while 4–6 cycles (15–20 s for each task presentation) seems to be generally appropriate for studies with phase uncertainty bigger than 2 s. The Fourier method (whose main advantage is its insensitivity to phase) still performs significantly worse (unless the response error is larger than 5 s with task blocks shorter than 10 s).

On the two panels of Fig. 3 we present the results of additional simulations performed for various lengths of imaging series of 32, 64, 96, or 128 images. The ROC power is plotted as a function of frequency (left panel) or as a function of the length of each single task condition (right panel). Both panels present curves obtained with a response delay of two images (3 s) (solid line) and no response delay (dotted line). For various lengths of imaging series the maximum power of the analysis can be found at different frequencies, but always at the same task length of about 12 images (18 s). This seems to be a desirable length for each task presentation for cognitive studies. This optimal length may, however, be much shorter for sensory-motor studies in which virtually all the response delay is due to the hemodynamical response and this can be estimated with accuracy better than 3 s (Frahm *et al.*, 1992).

5.3. Summary

The optimal length of one task block is independent of the length of the imaging run and is about 18 s (12 images at TR = 1500 ms). This value was assuming that the response uncertainty is relatively large (3 s). In simpler sensory motor tasks this delay may be estimated better and the optimal block lengths will then be shorter. To eliminate linear drift artifacts it is important to begin and end each imaging run with the same task (AB . . . A) design. For imaging runs with more than five task blocks it is even beneficial to drop the last task block to obtain this uneven design.

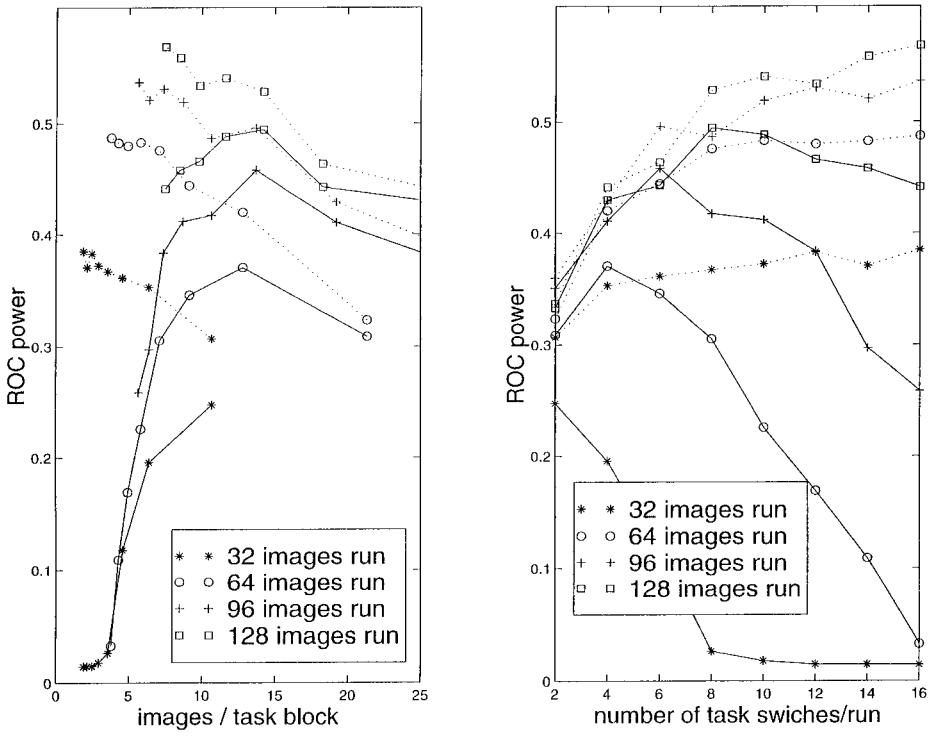

FIG. 3. The ROC power for various task switching paradigms and various lengths of imaging series is presented. Solid lines were obtained with 2 image (3 s) delays between the expected and the added signal response. The dotted curve assume exact knowledge of hemodynamic and neuronal responses. The left panel presents the data as a function of the length of each task presentation period, while the right panel shows the results as a function of task switching frequency. For various lengths of imaging series, the peak efficiency proves to be a function of the task presentation length rather than the task switching frequency. For a phase uncertainty of 3 s the optimal task length is about 12 images (18 s). This is a good assumption for complex cognitive studies. The optimal task length will be shorter if the response can be estimated with better accuracy—possible in sensory-motor tasks when the response delay is only of hemodynamical origin and thus can me modeled with an accuracy better than 1 s.

6. SPATIAL DOMAIN CORRELATION

6.1. Description

It is generally believed that for a wide range of cognitive studies activation observed using fMRI occurs over relatively large cortical volumes and is not confined to individual voxels. This assumption leads to the use of a processing strategy that will detect preferentially such large volumes. The analytical method that eliminates the possibility of observing small activation foci enables us to lower the threshold of activation detection without allowing too many false activations to be present. This reasoning is based on the assumption that white noise produced by MRI devices has no spatial correlation. The use of an implied large spatial correlation in the "real" activations can be made using cluster filtering of activation maps or by spatially smoothing individual images or final SPMs.

The advantages of the use of spatial correlation to increase the power of detection of large activation foci has been discussed extensively (Worsley *et al.*, 1992; Friston *et al.*, 1994; Forman and Cohen, 1995; Poline *et al.*, 1995; Skudlarski *et al.*, 1995; Xiong and Jia-Hong Gao, 1995). It is widely believed that both cluster techniques and smoothing of the images are beneficial for detecting sizable activations and should be applied to fMRI data. In this work for the first time we directly compare the efficiency of both techniques, applying them to real fMRI data.

6.1.1. Methods compared. We analyze the efficiency of four methods of using the spatial correlations between activations: (1) spatial smoothing applied to the raw data before the creation of activation maps; (2) spatial smoothing applied to the statistical map; (3) cluster filters applied to the thresholded activation map. In a cluster filter of size N only activation foci larger than the assigned cluster size are left in the thresholded SPM; all active pixels that do not belong to a contiguous cluster of N pixels are dropped out; (4) neighborhood filters applied to the thresholded activation maps, leaving only voxels that have a sufficient number of active neighbors. For each active pixel, its active neighbors are counted: counting 2 for each wall neighbor and 1 for each corner, and only if this score is larger or equal than a chosen filter parameter N, is the pixel treated as active.

The simulation was performed with different distributions of artificially added activations. We consider activation foci of various sizes (radius ranging between 1 and 4 pixels) and spatial smoothing with gaussian filters of width FWHM (0.6 . . . 3) pixel. Smoothing with a median filter gave results analogous to the gaussian filter of width 1.5 and is not included in the results presented.

We additionally consider a combination of the above described procedures. The multifiltering analysis based on the proposal of Poline and Mazoyer (1994) was applied to the raw data. In our implementation (described in Appendix B) we calculate statistical maps from both the raw data set and the initially smoothed data set. Those statistical maps are later averaged to create the final map. Since a spatial filter is applied to the data and averaging is performed on the level of SPMs this procedure differs significantly from using a spatial filter of different widths. SPMs calculated from the smoothed data are able to pick up large areas of relatively weak activations while the unsmoothed SPM is sensitive to isolated strong foci; averaging those maps lets us observe both kinds of activations with significant power.

6.2. Results

Figures 4 and 5 present the results of applying four different analytical methods, each with varying smoothing/filtering parameters. Figure 4 presents methods based on gaussian smoothing of images (left panel) or SPMs (right panel). Figure 5 presents two kinds of filters applied to the thresholded SPM: cluster filter (left panel), and use of the neighborhood filter (in the right panel). In Fig. 4 the multifiltering approach is also presented by open circles.

Both gaussian smoothing techniques proved to be significantly better than cluster/neighborhood filtering.

The optimal width of the filter used for smoothing was dependent on the size of the activation foci. This dependence was much less profound with the use of the multifiltering approach.

The best SPMs (highest power measured by the ROC curve) were obtained using the multifiltering technique: by adding two SPMs: one created from original data and the other from smoothed data. This method is most robust, providing a gain in power for a wide range of activation sizes. Since the statistics are performed separately on filtered and unfiltered data, this method is significantly different from smoothing data with any individual filters and preserves the advantages of "both worlds," so that both significant isolated activations and large slightly activated regions are detected.

We find that while cluster filtering is beneficial for large activation regions, the multifiltering approach similar to one proposed by Poline and Mazoyer (1994) outperforms the other techniques.

The initial smoothing of the data or our multifiltering technique can be combined with clustering or smoothing of the final t-maps. We have found that the combination of filtering of data and SPMs does not improve the results compared to either technique used individually. The multifiltering technique seems to perform best and

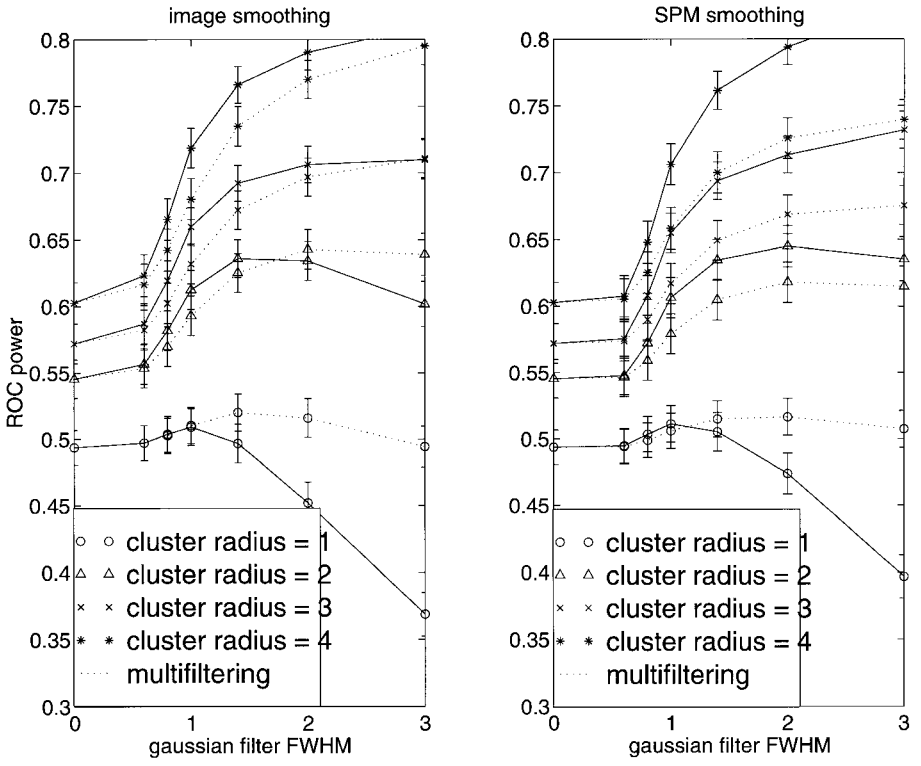

FIG. 4. Two methods of eliminating isolated activation foci (and thus increasing our power to detect spatially extended activations) are presented. The left panel presents ROC power obtained while the raw MRI images are smoothed prior to statistical processing. The right panel was calculated when the SPMs were smoothed. The four curves present results obtained with various sizes of activation foci added. The activation size parameters describe (in units of pixel size) the radius of the added activation clusters. Dotted lines were obtained using multifiltering (SPMs calculated with and without smoothing were added together). Both methods have comparable power. Multifiltering techniques are slightly worse when applied to large activation foci, but they are definitely superior if activation foci of various sizes maybe present.

is not further improved by adding cluster filtering on top of it and is least sensitive to filtering of the SPMs.

The estimate of detectability here was based only on the integral of the ROC curve and thus considers only the number of true and false positive findings, not their distributions. The SPM maps obtained with cluster filtering are "smoother," with less isolated activations, which may make them "more believable," but this subjective factor should not be taken into account.

6.3. Summary

The smoothing of raw MRI images with a gaussian filter of FWHM between 1 and 2 pixels (3–6 mm) proves to be better than any version of cluster filtering of the final SPM. Adding SPMs obtained from smoothed data and SPM obtained from unsmoothed data is the best approach, especially if activation foci of various sizes are present. Smoothing of the final SPMs with the same gaussian filter is nearly as good as smoothing of the images, but the benefits of multifiltering are more significant if the individual images are smoothed.

7. COMBINATIONS OF DATA FROM INDIVIDUAL RUNS

7.1. Description

In a typical fMRI study the same task paradigm is repeated over several imaging series. As we have shown earlier (Constable *et al.*, 1995) instead of performing statistics on the whole data set it is often beneficial to divide the imaging set into multiple subsets, perform statistics on each subset separately, and to later combine the results from each subset. This procedure is effective because the intraseries variance is lower than the interseries variance. Thus, it is justified to perform statistics on each imaging run separately so that the time-course analysis performed with *t*-statistics, correlation analysis, or other means is not impaired by global changes of intensity between separate imaging runs. Here we wish to find the best way of combining data obtained from subsets. In each study we have four imaging series, each of which gives an individual SPM. We create the combined SPM by taking the largest, second, third largest, or the smallest of the *t*-values in

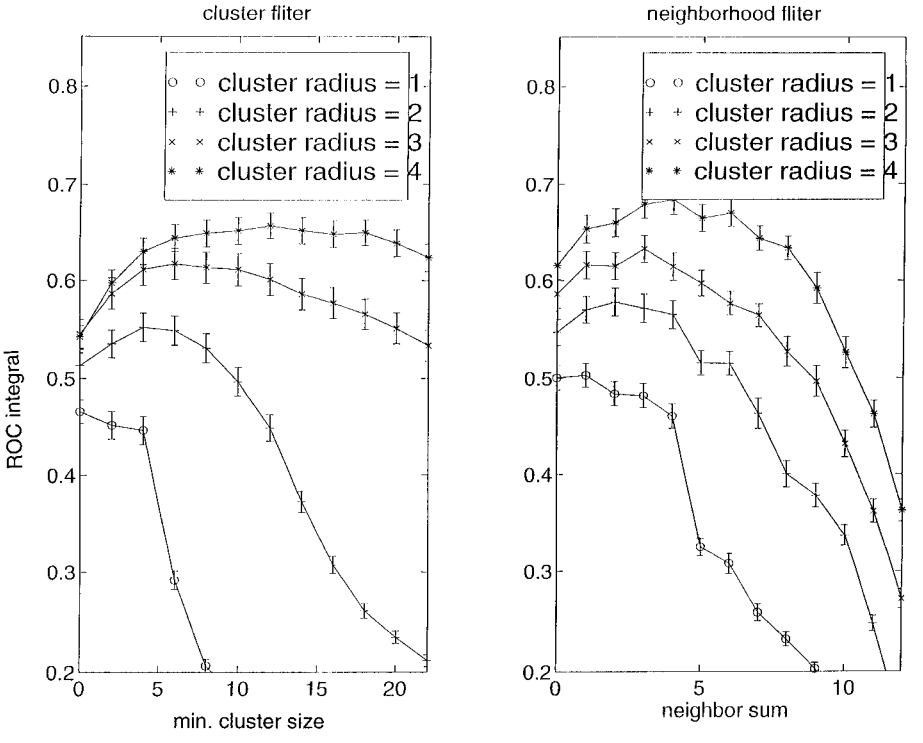

FIG. 5. Two methods of eliminating isolated activation foci (and thus increasing power to detect spatially extended activations by filtering of thresholded SPMs) are presented. Simulations were performed with various sizes of activation clusters added. The activation size parameters describe (in units of pixel size) the radius of added activation clusters. The cluster size (left panel) and neighbor sum (right) panel used on the horizontal axis describe the filter parameter used—the higher number the more restrictive spatial filter was used. The ROC powers obtained with the use of cluster filters (only active clusters composed than more than *N* voxels survive) are presented on the left panel; the right panel presents results for the neighborhood filter (only active voxels with enough active neighbors—counting 2—for each wall neighbor and 1 for corner—survive). The neighborhood filter proves to be better than the cluster filter for all sizes of activations, but both methods are significantly worse than methods based on gaussian smoothing of images or SPMs presented on Fig. 6.

the individual maps (this is what was called 1/4, 2/4, 3/4, 4/4 in Constable, *et al.* (1995)). These maps are compared with the mean, and median of those maps and with a map calculated by running *t*-statistics on the whole data set (not divided between series) and with the average ROC power obtained from using only one imaging series. If the presentation of the paradigm was exactly identical in all series one can combine series by averaging the raw data before calculating statistics and calculated statistics for one averaged series only. This data reduction step makes sense because the larger variance between data series should not influence our estimate of noise in measuring the difference between On and Off images taken within the same run. The effects of different splittings of the whole data set (split/2 or split/3) are also presented.

7.2. Results

Table 4 presents our results. The map calculated as an SPM calculated from the averaged imaging series proves to be best. This method has an disadvantage that it can be applied only if imaging series are

TABLE 4

	Freq = 1.5 (ABA)	Freq = 3.5 (ABABABA)	Freq = 7.5 (ABABABAB-ABABABA)
Max	0.354 ± 0.007	0.475 ± 0.007	0.535 ± 0.007
2nd	0.454 ± 0.007	0.544 ± 0.007	0.576 ± 0.007
3rd	0.443 ± 0.007	0.540 ± 0.007	0.574 ± 0.007
Min	0.347 ± 0.007	0.507 ± 0.007	0.566 ± 0.007
Mean	0.500 ± 0.007	0.570 ± 0.007	0.587 ± 0.007
Median	0.483 ± 0.007	0.560 ± 0.007	0.582 ± 0.007
Averaged series	0.514 ± 0.01	0.583 ± 0.01	0.596 ± 0.01
All combined	0.319 ± 0.007	0.475 ± 0.007	0.553 ± 0.007
Individual series	0.293 ± 0.007	0.420 ± 0.007	0.500 ± 0.007
Split-2	0.314 ± 0.007	0.456 ± 0.007	0.558 ± 0.007
Split-3	0.293 ± 0.007	0.426 ± 0.007	0.549 ± 0.007

Note. Comparison between the different methods of combining the data from several identical imaging series. The SPM obtained with averaging data series proved to be best. If the imaging series are not identical and so they cannot be averaged, the mean of SPMs calculated from individual series proved to be the second most powerful method. It is significantly better than the conservative approach of using the lowest value, while both methods are better than combining the data from different imaging series and calculating a single SPM.

identical (the presentation of stimuli cannot be alternated) and none of the involved series can have images missing due to scanning artifacts or motion. The next best method is the calculating the mean of SPMs calculated separately for each of the imaging series. This method has an advantage because the statistical interpretation of this mean *t*-statistic is independent of the number of series averaged, which is important if in longer experiments some imaging series have to be discarded due to motion or artifacts.

The *t*-statistic calculated for the whole (undivided) data set behaves worse than for any of the combination methods and only slightly better than the statistical power from a single imaging series. If one of the individual SPMs is used as the final result, the *t*-value that is 2nd and 3rd value is significantly better than either the smallest or the largest. Different splittings of the whole data set (split/2 or split/3 statistics) applied to the whole data set are not efficient. This result confirms that data should be split into subsets only in agreement with the natural division of the experiment (to eliminate the additional variance due to change of imaging run).

7.3. Summary

The data in the fMRI set should be analyzed in blocks containing images taken during one imaging run only. If imaging series are identical then the best method of combining images is to average the images from all the imaging series and then to calculate the SPM from those averaged images. If imaging series do differ in the length or order of task presentation, then individual maps should be obtained from separate runs and averaged to create the final SPM. The more conservative approach of taking the smallest value (requiring the pixel to be active in every imaging run) is less powerful in detecting activations. Another important point is that only mean maps calculated from studies with different numbers of imaging series are comparable.

8. MOTION CORRECTION

8.1. Description

The full analysis of the various effects of motion correction methods is complicated enough to justify a separate study. However, since motion correction is now routinely applied in many fMRI studies we felt compelled to check if our results remain valid in a study analyzed with motion correction. We do not attempt to compare the efficiency of analysis with and without motion correction, but we want to evaluate the power of some processing steps analyzed above that may be most sensitive to motion correction.

8.2. Methods

We used two versions of motion correction algorithm from the SPM package: one including motion correction only and the other with additional decorrelation that removes the component of the signal correlated with motion estimates. Activations were added after motion correction. This makes it difficult to compare directly analyses performed with and without motion correction, but nevertheless it should be helpful in comparing the efficiency of different processing strategies used with the same motion correction approach. We applied this technique to two of the simulations presented above: the efficiency of temporal normalization (presented in Section 3.3) and the comparison between various statistics (Section 4). Data analyzed here were obtained in a different study of olfactory processing (Fulbright *et al.*, 1998). We used imaging series consisting of 80 images, with activation added with a frequency of 2.5 cycles per series (ABABA design). We used data from three subjects with six imaging series obtained from each subject.

8.3. Results

Actual values of the ROC power obtained in this simulation differ from those obtained in the main study but these differences can be attributed to changes in the study design. Tables 5 and 6 summarizes our results.

Both simulations show that both techniques of motion correction do not change the relative efficiency of the steps in the data analysis that we compared.

The relative powers of various statistics do not change significantly due to motion correction. The slight decrease in the efficiency of skew corrected techniques suggests that the drift that is removed by this technique is mostly caused by real or apparent movement of the subject head in the imaging plane. The advantage of cross-correlation and Mann–Whitney techniques seem to be enhanced by the motion correction with decorrelation.

9. CONCLUSIONS

This study shows that ROC based techniques can be used as an efficient method for estimating the relative effectiveness of various individual steps in fMRI data analysis. Based on the simulations reported, our specific recommendation for the fMRI processing strategies can be summarized as follows:

- Time normalization does not in general increase overall efficiency but it may possibly be useful in individual cases to rescue certain flawed studies distorted by significant intensity drift.
- Subtraction of the linear and quadratic components from the signal improves the effectiveness of data

analysis, and removal of higher order components is not more beneficial. High pass filtering with a cutoff frequency of 0.35 of the stimulus frequency is the most efficient preprocessing filter in the time domain.

- Temporal smoothing does not improve our ability to detect activations. The gain in the perceived significance of activations (true positives) detected is overtaken by an even larger increase in the analogous statistical measures for non activated pixels (false positives), and thus temporal smoothing not only does not improve the fMRI analysis but, if not corrected for, may lead to overestimation of the significance of fMRI findings.
- Cross-correlation, *t*-statistics, Mann–Whitney test are all excellent statistics and yield comparable results. Skew correction is helpful only for even (highly susceptible to drift) AB . . . AB study designs.
- Results improve as the frequency of task/control switching between stimulus condition increases. Realistically, taking into account the uncertainty of the exact timing of the fMRI response, task switching with about 15–20 s per condition is advised.
- Gaussian smoothing of the raw fMRI images is better than cluster or neighborhood filtering of thresholded statistical maps. Multifiltering (achieved by adding maps obtained from filtered and unfiltered data) can increase efficiency even more—especially if the activation foci are of variable or unknown size.
- The data from identical but separate imaging runs should be analyzed by averaging analogous images from individual series. If series are not identical (some images are missing or task order has been changed) they should be analyzed separately and later combined using the mean of the individual SPMs. This is the most powerful and the most convenient (especially if the number of usable runs varies between subjects) way of combining data obtained in a series of consecutive imaging runs from a single subject.

9.1. Limits of validity of this study

All the simulations performed in this study were performed on the data obtained using the same 1.5 T scanner. Since the main purpose of this study is to assess methods of analysis using actual data with real signal and noise, our results may in principle be specific to this scanner only. However, while some of the data presented here may not be typical, the approach to assessing techniques using the ROC method and simulated activations added to real data, is widely applicable. Appendix C provides several statistical characteristics of data sets used in this study which may be useful for others using different MRI systems to see if our results can be applied to the data from their studies.

The fact that motion was not realistically included into our simulations may result in underestimating the advantages of some techniques that are especially good in treating motion. The simulations performed using two methods of motion correction yielded results that are very close to those obtained without motion correction, which builds our confidence that the recommendations given in this study should be valid if motion correction is being used. In an ideal simulation the activations would be added not in a fixed locations in an image but in fixed locations in the brain and thus would move in the image space with movement of the subjects.

10. APPENDIX A: CORRECTION FOR THE LINEAR DRIFT IN THE DATA

Quite often fMRI data contain uniform linear drifts with the intensities of certain voxels slowly rising or falling during the whole imaging series. We believe that one source of drift is an instability in the B_0 field. This produce an apparent linear motion, which manifests itself as an intensity drift proportional to the spatial gradient of the signal. Our observations on a GE system suggest that this drift is actually not linear but sinusoidal with a period of several minutes.

Several groups (Bandettini *et al.*, 1993) have added a step removing this drift into their analysis. In the presence of activations, the drift has to be removed separately from the data obtained in each condition and this makes this process quite difficult and prone to create artifacts.

Our approach is to take the drift into account during calculation of the SPM, replacing the *t*-value by a skew-corrected *t*-value.

The calculation of *t*-value can be seen as fitting the time-course data by a step function:

$$f(t) = a\theta_{ON}(t) + b\theta_{OFF}(t),$$

where θ_{ON}, θ_{OFF} are characteristic functions of the ON and OFF conditions. *t*-Values are calculated as $(a–b)$ normalized by the deviation of the real time-course from this fitting function.

We replace this fit by a function that can take into account the linear slope of the time course:

$$f_{SKEW}(t) = a\theta_{ON}(t) + b\theta_{OFF}(t) + kt.$$

Our corrected value will be given by the difference $(a–b)$ normalized by the deviation from this fit. Parameters a, b, k of the fit are calculated by the method of least squares.

The same procedure can be applied to calculating the skew corrected version of different SPMs such as percent (or absolute) difference of signal intensity. Figure 6 presents sample time-course and fits that are used while performing comparisons using *t*-statistics and or skew corrected *t*-statistics. In this case the

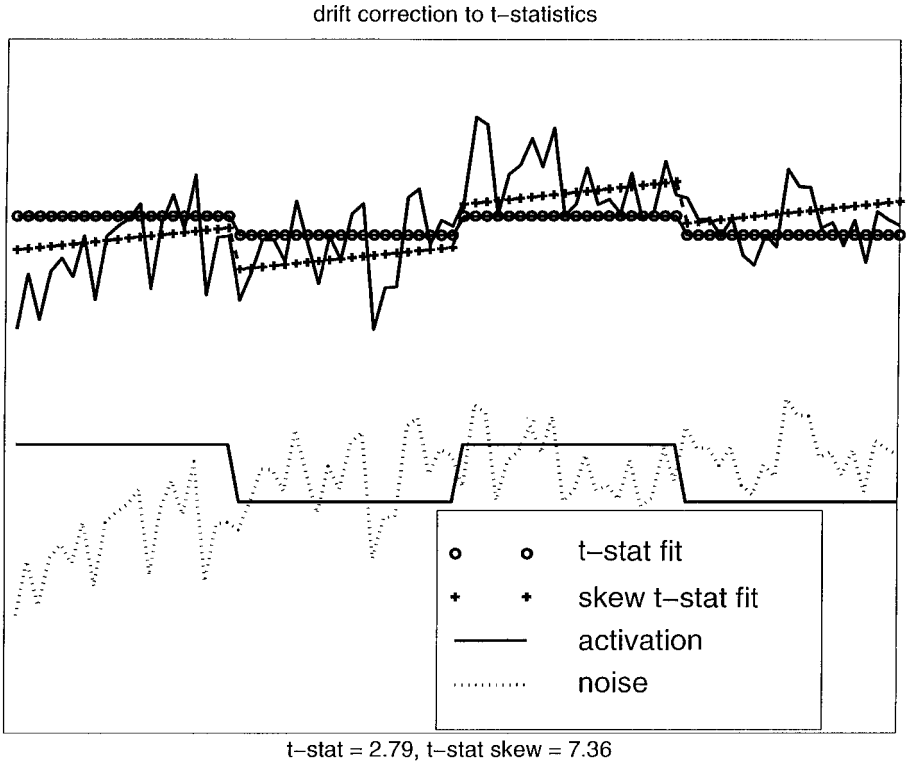

FIG. 6. The mechanism of skew correction for the linear drift is presented. The sample time course is presented as a combination of random noise, linear drift and response to the On–Off task. The data fits used with *t*-statistics and with the skew *t*-statistics are overlaid, showing how the linear drift can diminish the estimated response if *t*-statistics are used. The difference would be virtually eliminated if the study started and end with the same task condition, so that linear drift would affect each state in the same way. In this example, the use of skew corrected *t*-statistic increases the calculated *t*-value from 2.79 to 7.36. Of course other scenarios in which the skew correction would actually hurt the data analysis by obscuring real activations or detecting false activations can be easily designed.

ABAB activation was partially obscured by the linear drift and so regular *t*-statistics return a *t*-value of 2.79, while the skew corrected version gives 7.36. The distribution of *t*-values obtained without activations with or without use of skew correction does not differ, so that their meanings are comparable.

Data presented in the Table 1 show that for this balanced (AB . . . A) study design when each imaging series begins and ends with the same task the skew correction does not improve efficiency. However, it is very useful in an (AB . . . AB) design that is more prone to artifacts due to the linear drift.

Figure 2 presents the comparison of the statistical power of analysis using skewed *t*-statistics with the correlation method, which due to the use of the exact shape of the activation response curve is better than for a regular *t*-statistic. We notice that at each integer frequency (AB . . . AB design) the correlation method has a significant drop in efficiency in comparison to the half integer frequency (AB . . . A design); for those conditions the skew corrected *t*-statistic is significantly better than other methods. These results prove that the

TABLE 5

	No motion correction	Motion corrected (no decorrelation)	Motion corrected (with decorrelation)
t-stat	0.362 ± 0.01	0.361 ± 0.01	0.365 ± 0.01
Skewed *t*-stat	0.361 ± 0.01	0.357 ± 0.01	0.360 ± 0.01
Boxcar correlation	0.365 ± 0.01	0.361 ± 0.01	0.365 ± 0.01
Exact correlation	0.387 ± 0.01	0.389 ± 0.01	0.395 ± 0.01
Percentage difference	0.173 ± 0.01	0.167 ± 0.01	0.163 ± 0.01
Skewed percentage difference	0.208 ± 0.01	0.217 ± 0.01	0.207 ± 0.01
Mann–Whitney	0.368 ± 0.01	0.374 ± 0.01	0.381 ± 0.01

Note. Comparison between different statistics (as in Table 1) performed for data analyzed without motion correction, with motion correction, and with motion correction and decorrelation (using the SPM package for motion correction). Columns with ROC values should not be compared directly (see Section 8), but the relative power of statistics varies only slightly depending on the use of motion correction. The slight decrease in the power of skew corrected techniques suggests that an important component of the intensity drift is created by (real or apparent) motion. The advantages of cross-correlation and Mann–Whitney techniques seem to be enhanced by the motion correction with decorrelation.

TABLE 6

	No motion correction	Motion corrected (no decorrelation)	Motion corrected (with decorrelation)
No temp. smooth.	0.372 ± 0.01	0.369 ± 0.01	0.369 ± 0.01
FWHM = 1.5 image	0.283 ± 0.01	0.279 ± 0.01	0.283 ± 0.01
FWHM = 3 images	0.212 ± 0.01	0.206 ± 0.01	0.211 ± 0.01
FWHM = 6 images	0.153 ± 0.01	0.144 ± 0.01	0.152 ± 0.01

Note. Temporal smoothing of the fMRI data significantly decreases our power to detect activations, both with and without motion correction. This suggest that the temporal correlation responsible for this effect (see Section 3.3) is not created by real or apparent motion.

skew correction is a necessary step only if for some reason we cannot balance our task design, but it is still helpful for balanced study.

11. APPENDIX B: MULTIFILTERING TECHNIQUE OF SPATIAL SMOOTHING

In our study we used a simple version of the multifiltering approach (Poline and Mazoyer, 1994). During the data analysis we produce two versions of the data set—raw images and images that were smoothed with a gaussian filter of an appropriate FWHM. As in normal gaussian filtering the width of this filter depends on the size of the activation foci that we want to enhance. Unlike the standard approach the analysis still sustain some sensitivity to strong focal activations. This smoothed data set is reduced in spatial resolution by a factor of 2 for memory efficiency.

Each SPM is calculated on both data sets and the resulting maps are later averaged (added) to create the final SPM, that can be thresholded and cluster filtered to obtain an activation map at the desired significance/sensitivity level. This technique differs significantly from regular spatial filtering with any form of filter. The smoothed data set has significantly reduced variance so that even small changes in the intensity that extend far enough to survive filtering produce large *t*-values. Small localized activations that are smoothed in the spatial filtering produce large *t*-values in the nonsmoothed data set so that both types of activations can be seen in the final map.

12. APPENDIX C: SOME STATISTICAL CHARACTERISTICS OF fMRI DATA SET USED

In this Appendix and Table 7, we present several basic statistical characteristics of the fMRI data set used in this study. We calculated the mean intensity, standard deviation, skewness, and kurtosis and the absolute value of the estimated linear drift (slope *k* parameter used in skew statistics in Appendix A). We consider small blocks of pixels located in four distinct areas located in the gray matter (medial Superior Frontal Gyrus), white matter (corona radiata), Cerebrospinal Fluid (lateral ventricle), and outside the head. To estimate better the effect of possible signal drift we calculate those variables for a whole imaging series of 128 images and for subseries of the first 64 images. If the noise can be characterized as white noise the length of the series should not change those characteristics; if there is more power in the low frequency part of the spectrum (as in the case of the drift) the standard deviation will increase with the length of the series.

Those numbers are provided mainly to allow comparisons between different scanners.

One can notice that the standard deviation always increases for longer data series but this increase is most dramatic in the white matter and nearly nonexistent in the outside air.

This, together with the highest value of the estimated linear drift, suggest that the drift is more apparent in the white matter than in gray matter. The differences are smaller than the estimate of error but

TABLE 7

	Series length	Gray matter	White matter	CSF	Air
Mean	64	1517 ± 70	1310 ± 70	2067 ± 130	34.4 ± 3
	128	1522 ± 70	1288 ± 70	2052 ± 130	34.5 ± 3
Standard deviation	64	30.0 ± 3	32.5 ± 6	36.2 ± 3	17.9 ± 1
	128	31.0 ± 3	37.9 ± 6	39.1 ± 3	18.8 ± 1
Skewness	64	0.004 ± 0.004	0.021 ± 0.01	0.026 ± 0.01	0.48 ± 0.07
	128	−0.004 ± 0.004	−0.023 ± 0.01	0.035 ± 0.01	0.51 ± 0.07
Kurtosis	64	2.95 ± 0.06	2.83 ± 0.1	2.83 ± 0.1	2.86 ± 0.2
	128	2.99 ± 0.06	2.85 ± 0.1	2.91 ± 0.1	2.99 ± 0.2
	64	0.11 ± 0.05	0.30 ± 0.1	0.23 ± 0.1	0.01 ± 0.01
Absolute value of estimated linear drift	128	0.06 ± 0.05	0.25 ± 0.1	0.21 ± 0.1	0.01 ± 0.01

this error reflects mainly the variance between different studies and imaging runs and not the differences between series of different length. More work is necessary to understand this phenomenon precisely.

The estimated linear drift is larger for short data series, which suggests that it has a significant random component, which is averaged out for longer series.

The skewness is very small for all regions except for the outside air where the distribution of intensities is Raleigh rather than Gaussian, the intensity is always positive and thus it is skewed.

REFERENCES

Alperin, N., Vikingstad, E. M., *et al.* 1996. Hemodynamically independent analysis of cerebrospinal fluid and brain motion observed with dynamic phase contrast MRI. *Magn. Reson. Med.* **35**(5):741–755.

Arndt, S., Cizadlo, T., *et al.* 1996. Tests for comparing images based on randomization and permutation methods. *J. Cereb. Blood Flow Metab.* **16:**1271–1279.

Bandettini, P. A., Jesmanowicz, A., *et al.* 1993. Processing strategies for time-course data sets in functional MRI of the human brain. *Magn. Reson. Med.* **10:**161–173.

Biswal, B., DeYoe, E. A., *et al.* 1996. Reduction of physiological fluctuations in fMRI using digital filters. *Magn. Reson. Med.* **35**(1):107–114.

Bullmore, E., Brammer, M., *et al.* 1996. Statistical methods of estimation and inference for functional MR image analysis. *Magn. Reson. Med.* **35:**261–277.

Constable, R. T., Skudlarski, P., *et al.* 1995. An ROC approach for evaluating functional brain MR imaging and postprocessing protocols. *Magn. Reson. Med.* **34**(1):57–64.

Forman, S. D., Cohen, M. F. J. D., Eddy, W. E., Mintun, M. A., Noll, D. C. 1995. Improved assesment 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., Firston, K. J., *et al.* 1997. *Human Brain Function*. Academic Press, San Diego.

Frahm, J., Bruhn, H., *et al.* 1992. Dynamic MRI of human brain oxygenation during rest and photic stimulation. *J. Magn. Reson.* **2:**501–505.

Friston, K. J., Holmes, A., *et al.* 1996. Detecting activations in PET and fMRI: Levels of inference and power. *Neuroimage* **4:**223–235.

Friston, K. J., Worsley, K. J., *et al.* 1994. Assessing the significance of focal activations using their spatial extent. *Human Brain Mapp.* **1:**210–220.

Fulbright, R. K., Skudlarski, P., *et al.* 1998. Functional MR imaging of regional brain responses to pleasant and unpleasant odors. *Am. J. Neuroradiol.* **19**.

Metz, C. E. 1978. Basic principle of ROC analysis. *Semin. Nuclear Med.* **VIII**(4):283–298.

Ogawa, S., Menon, R. S., *et al.* 1993. Functional brain mapping by blood oxygenation label-dependent contrast magnetic resonance imaging. *Biophys. J.* **64:**803–812.

Peterson, B., Skudlarski, P., *et al.* 1997. A functional magnetic resonance imaging study of tic surpression in Tourette's syndrom. *Arch. Gen. Psychiatry,* **in press.**

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

Poline, J.-B., Mazoyer, B. M. 1994. Enhanced detection in brain activation maps using a multifiltering approach. *J. Cereb. Blood Flow Metab.* **14**(4):639–642.

Poline, J. B., Worsley, K. J., *et al.* 1995. Estimating smoothness in statistical parametric maps: Variability of p values. *J. Comput. Ass. Tomogr.* **19**(5):788–796.

Skudlarski, P., Constable, R. T., *et al.* 1995. *ROC Based Analysis of Cluster Filtering in the Functional MRI of the Human Brain.* SMRM Meeting, Nice.

Skudlarski, P., Constable, R. T., *et al.* 1995. *Spatial Correlation in The Functional MRI of the Human Brain.* SMRM Meeting, Nice.

Skudlarski, P., Constable, R. T., *et al.* 1997. *ROC Analysis of Statistical Methods in the fMRI.* SMRM Meeting, Vancouver.

Skudlarski, P., Gore, J. C. 1996. *Assessment of the Validity of fMRI Activation's Using Blank Runs and the Randomization of Image Order.* Human Brain Mapp. **2:**.

Sorenson, J. A. 1995. *ROC Method for Evaluation of fMRI Techniques*. 3rd SMR 1995, Nice.

Sorenson, J. A., Wang, X. 1996. ROC methods for evaluation of fMRI techniques. *Magn. Reson. Med.* **36:**737–744.

Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. *Science* **240:**1285–1293.

Weiskoff, R. M., Baker, J., *et al.* 1993. *Power Spectrum Analysis of Functionally-Weighted MR Data: What's in the Noise.* SMRM 12th annual meeting, New York.

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

Xiong, J., Gao, J., *et al.* 1996. Assessment and optimization of functional MRI analyses. *Human Brain Mapp.* **4:**153–167.

Xiong, J., Jia-Hong Gao, J. L. L., Fox, P. T. 1995. Clustered pixels analysis for functional MRI activation studies of the human brain. *Human Brain Mapp.* **3**(4):287–301.