

# An Empirical Comparison of SPM Preprocessing Parameters to the Analysis of fMRI Data

Valeria Della-Maggiore, Wilkin Chau, Pedro R. Peres-Neto,\* and Anthony R. McIntosh

Rotman Research Institute of Baycrest Centre, Toronto, Ontario M6A 2E1, Canada; and

\*Department of Zoology, University of Toronto, Toronto, Ontario M5S 3G5, Canada

Received June 27, 2001

**We present the results from two sets of Monte Carlo simulations aimed at evaluating the robustness of some preprocessing parameters of SPM99 for the analysis of functional magnetic resonance imaging (fMRI). Statistical robustness was estimated by implementing parametric and nonparametric simulation approaches based on the images obtained from an event-related fMRI experiment. Simulated datasets were tested for combinations of the following parameters: basis function, global scaling, low-pass filter, high-pass filter and autoregressive modeling of serial autocorrelation. Based on single-subject SPM analysis, we derived the following conclusions that may serve as a guide for initial analysis of fMRI data using SPM99: (1) The canonical hemodynamic response function is a more reliable basis function to model the fMRI time series than HRF with time derivative. (2) Global scaling should be avoided since it may significantly decrease the power depending on the experimental design. (3) The use of a high-pass filter may be beneficial for event-related designs with fixed interstimulus intervals. (4) When dealing with fMRI time series with short interstimulus intervals (<8 s), the use of first-order autoregressive model is recommended over a low-pass filter (HRF) because it reduces the risk of inferential bias while providing a relatively good power. For datasets with interstimulus intervals longer than 8 seconds, temporal smoothing is not recommended since it decreases power. While the generalizability of our results may be limited, the methods we employed can be easily implemented by other scientists to determine the best parameter combination to analyze their data.** © 2002 Elsevier Science (USA)

One of the primary statistical tools to examine changes in brain activity from fMRI datasets is Statistical Parametric Mapping (SPM) (Friston *et al.*, 1994, 1995a,b). SPM utilizes a general linear model (GLM) to assess task-specific, voxel-based differences in the magnitude of the blood-oxygenation-level-dependent (BOLD) signal. The fMRI time series is modeled at

each voxel with a linear combination of explanatory functions plus a residual error term (Friston *et al.*, 1995).

As with other parametric approaches, the application of GLM is contingent on two statistical assumptions of the data: normal distribution and independence of the error term. It is common to apply temporal and spatial smoothing to insure the time series conforms to a Gaussian Random Field (GRF), which validates the application of parametric statistical assessment (Friston *et al.*, 1995b; Worsley and Friston, 1995).

In addition to smoothing, other preprocessing steps have been implemented to further enhance statistical power while respecting the GLM's assumptions. Several factors contribute to image intensity changes in fMRI; these include the physiological component generated by alterations in the BOLD signal which vary in the order of 1 to 5% (Jezzard and Song, 1996; Turner *et al.*, 1998), and irrelevant noise of physiological and nonphysiological origin. Nonphysiological noise can be instrumental (e.g., thermal noise) or due to head movements, whereas physiological noise originates mainly from cardiac and respiratory cycles. SPM offers several options to model changes in the BOLD signal, including a canonical hemodynamic response function (HRF) and an HRF with time derivative (HRF/TD) aimed at adjusting for delays in the onset of the hemodynamic response. To eliminate random components of low frequency noise mentioned above, a high pass filter can be selected (Holmes *et al.*, 1997). Temporally correlated noise can be smoothed by convolving the time series with a known HRF function (low-pass filter), and/or by eliminating the temporal autocorrelation using a first-order autoregressive model (AR1) (Friston *et al.*, 2000).

To date, the basis for selecting the appropriate parameters to preprocess a fMRI dataset has been theoretical, a necessary but not sufficient criterion to ensure the appropriate statistical treatment of the data. There is little available information on the actual behavior of these parameters. The "sensitivity" of SPM parameters has been estimated empirically by evaluating the effect of preprocessing a real fMRI dataset

with different parameter combinations (Hopfinger *et al.*, 2000). However, statistical inferences based on this type of empirical testing may not be valid: on one hand, the estimation of power may be highly biased by the small number of sample tests ( $n$  = number of subjects); on the other hand, given that the magnitude of the signal and its spatial localization remain unknown to the experimenter, true activations cannot be distinguished from false positives. To our knowledge, no systematic study has been conducted to evaluate the robustness of these parameters.

In this paper we present the results derived from two sets of Monte Carlo simulations generated to evaluate the robustness of some of the SPM preprocessing parameters. Both power (i.e., the probability of detecting an activation if it exists) and type I error (i.e., the probability of detecting an activation if it does not exist) were estimated for five hundred datasets generated based on fMRI images obtained from an event-related study. Parametric and nonparametric approaches were used to generate the two sets of simulations. The parametric simulation entailed the generation of a white-noise baseline (plus AR1 correlated noise), and the addition of an HRF signal (Cohen, 1997). Given that we generated the signal, we could test the impact of varying the experimental design on the generality of the SPM results. The nonparametric simulation consisted in using the original fMRI data as the population from which simulated datasets were sampled. Although the statistical distribution of the original data remained unknown, this approach presented the advantage of preserving the spatial and temporal structure of real fMRI data. Power and false positive rate were assessed for combinations of the following parameters: basis (modeling) function, global scaling, low pass filter, high pass filter and autoregressive modeling of temporal correlations.

## MATERIALS AND METHODS

### *Monte Carlo Simulations*

The power of a statistical test can be estimated using analytical or empirical methods. Analytical methods are based on the same probability theory and assumptions that are used to identify the appropriate statistical distribution for any traditional statistical method. Several assembled tables (e.g., Cohen, 1988) and computer software packages (see Thomas and Krebs, 1997) based on numerical solutions are available for estimating the power of most commonly used statistical tests. However, when analytical formulae for estimating power have not been derived, or when there is interest in assessing the power of a test in which statistical assumptions have been violated, power tables can be generated using a Monte Carlo approach (e.g., Stephens, 1974). In this case, one simulates statistical

populations and manipulates them in order to introduce a desirable effect size (e.g., difference between means) or sample variability (e.g., variance). Following this, a large number of samples are taken and the test statistic is calculated each time (Oden, 1991). If the effect size is manipulated to be zero (i.e., the null hypothesis is true), the probability of committing a type I error is estimated as the proportion of tests that erroneously rejected the null hypothesis. If the effect size is set to be different from zero, the proportion of cases in which the null hypothesis was correctly rejected is used as an estimate of statistical power. A comprehensive simulation study of this kind can provide a basis for understanding the behavior of any particular test and for comparing different tests (Peres-Neto and Olden, 2001). This aids in identifying the most appropriate statistical test for a particular scenario (i.e., combinations of factors that can influence the statistical test).

In the present study we used Monte Carlo simulations to compare the robustness of 16 different SPM models. These models were defined by several combinations of four preprocessing parameters (see Table 1): global scaling (remove global effects or not), low pass filter (HRF or none), high pass filter (default cutoff or none), modeling of temporal autocorrelation (AR1 or none). To enhance the reliability of the study, datasets were generated using two different simulation approaches: parametric and nonparametric. All simulated data was spatially smoothed with a 10-mm full-width half-maximum (FWHM) filter.

Two basis functions were initially evaluated: HRF and HRF/TD. HRF/TD is aimed at correcting for occasional delays in the onset of the hemodynamic response. However, modeling 500 datasets from the nonparametric simulations with HRF/TD drastically reduced the power compared to HRF alone. Further investigation of the efficiency of HRF/TD by generating datasets with 0-, 1-, or 2-s delay using the parametric approach, indicated that in fact, the effect of HRF/TD varied with the duration of the delay. Altogether, these results suggest that HRF/TD may be detrimental in estimating changes in brain activity when applying to the whole brain (see results for details). Thus, HRF was the only basis function tested for all models.

### *Parametric Simulation*

To study the performance of each parameter setting, a total of 500 datasets were simulated based on T2\*-weighted Echo-Planar Images (EPI) obtained from five subjects scanned with a GE 1.5T scanner during a visual attention study (100 datasets per subject) (Giesbrecht *et al.*, 2000). Each real dataset consisted of 24 axial slices ( $64 \times 64$  mm) of 180 vol; voxel size =  $3.8 \times 3.8 \times 5.0$  mm. For practical reasons, only 5 of the 24 slices were used to generate the simulated datasets.

**TABLE 1**  
Model Specification

SPM Parameter	Model type															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Basis function																
hrf alone	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Remove global effects																
Yes		x		x		x		x		x		x		x		x
No	x		x		x		x		x		x		x		x	
High pass filter																
Yes			x	x			x	x			x	x			x	x
No	x	x			x	x			x	x			x	x		
Low pass filter (HRF)																
Yes					x	x	x	x					x	x	x	x
No	x	x	x	x					x	x	x	x				
AR1																
Yes									x	x	x	x	x	x	x	x
No	x	x	x	x	x	x	x	x								

Note. x indicates the parameter settings that defined each of the 16 models tested for all simulated scenarios.

We first generated the baseline activity of the simulated datasets by using a first-order autoregressive plus white-noise model derived empirically by Purdon and Weisskoff (1998).

The model with additive white noise can be expressed as a recursive filter:

$$x[n] = ((1 - q) \times w[n] - q \times x[n - 1]) + v[n],$$

where  $w[n]$  and  $v[n]$  constitute the white noise, and  $q$  represents the degree of correlation between adjacent samples of the AR1 process. The AR1 component represents physiological and non-physiological low frequency noise characteristic of fMRI time series, while the white noise represents nonphysiological, scanner noise. The value of  $q$  was set to 0.82, and the variance of  $w[n]$  and  $v[n]$  were set to 1.16 and 3.52, respectively. These values were chosen so that the temporal autocorrelation of the simulated baseline was similar to that found in a real fMRI time series with a TR of 3 s. Since the time series of each voxel was generated independently, the spatial autocorrelation present in the real dataset was lost (Pettersson *et al.*, 1999). To solve this problem, a gaussian low-pass spatial filter with a kernel size of  $3 \times 3 \times 3$  voxels was applied to all simulated datasets. The kernel weight was determined empirically, so that the standard deviation of the voxel's time series was similar to that of the original dataset.

Five "active" regions were defined for each subject (Fig. 1a). Each region consisted of  $3 \times 3 \times 2$  voxels (i.e.,  $11.4 \times 11.4 \times 10$  mm). To simulate the signal, a hemodynamic response function derived from Cohen and

collaborators (Cohen, 1997) was added to the baseline time series:

$$h(t) = t^{8.6} e^{(-t/0.547)},$$

where  $t$  is time. Except for those datasets used to examine the effect of HRF/TD, the response latency of each time series was 1 s.

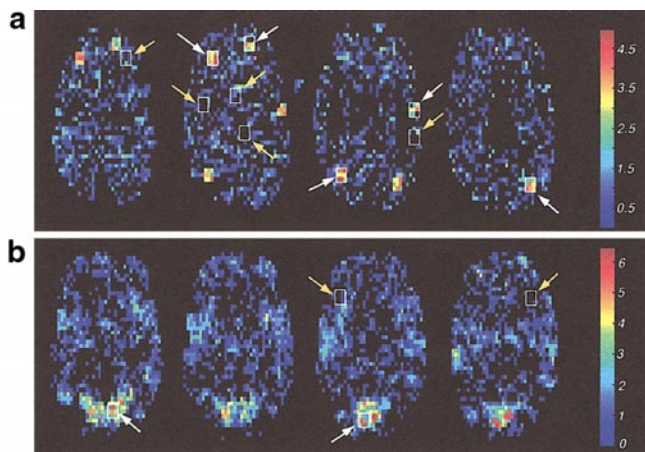
Simulated datasets were not spatially normalized to avoid the introduction of additional artifacts from non-linear warping. Instead, to maintain the same anatomical coordinates of the "active" regions across the 500 datasets we followed these steps: We first defined the five active regions from one of the subject's anatomical image. Using AFNI (Medical College of Wisconsin, Cox, 1996) we determined the coordinates of the five regions in Talairach coordinate space. The Talairach coordinates were then used to identify the corresponding locations in the native coordinate space of the other four subject's brains.

To enhance the reliability of the results, the robustness of the 16 SPM models was tested on two possible scenarios generated following an event-related experimental design with either fixed or variable interstimulus interval (ISI). The variable ISI, simulated random presentation of the stimuli.

Scenario 1: 1% signal change, constant ISI

Scenario 2: 1% signal change, variable ISI (mean ISI = 31 s)

These particular scenarios were chosen to represent common cases found in the fMRI literature (e.g., D'Esposito *et al.*, 1999; Hopfinger *et al.*, 2000). For a



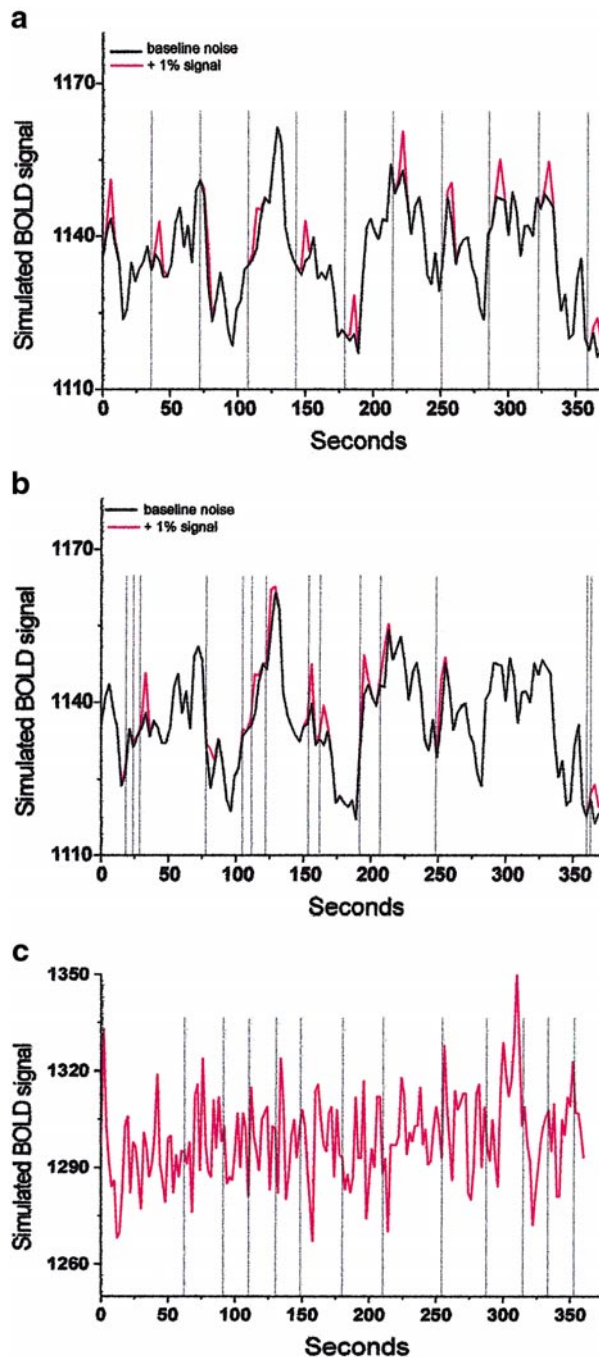
**FIG. 1.** Anatomical localization of “active” and “inactive” regions. Shown are the  $t$  values obtained from one dataset of the parametric (a), and nonparametric (b) simulations, overlay on four horizontal slices of a T1 image. The data displayed in the figure have not been spatially smoothed. Selected regions are signaled with a white box. “Active” regions are indicated with white arrows, whereas “inactive” regions are indicated with yellow arrows.

given scenario, the simulated percent signal change was constant within a voxel’s time series—thereby simulating only one experimental condition—and across spatial location (i.e., across the five “active” regions). Figures 2a and 2b illustrates scenarios 1 and 2, respectively, for one voxel of an “active” region. Both scenarios were simulated using the same  $TR = 3$  s.

### Nonparametric Simulations

Because the parametric approach described above may not maintain the spatial and temporal structure of real fMRI data, we designed an alternative nonparametric approach where these features were kept. This protocol was adapted from the parametric bootstrap (Efron and Tibshirani, 1993), where samples are drawn from a multivariate normal distribution and the variance/covariance structure obtained from empirical data. The original data used for this purpose was one session of the visual attention fMRI study described above. The experiment followed an event-related design with variable ISI (mean ISI = 26.4 s), where the stimuli were presented semi-randomly;  $TR = 2$  s, and the average signal change was 1%. The simulation protocol was as follows. Images were spatially normalized so that the between subjects’ standard deviation of each scan could be computed. From the 10 subjects of the original fMRI experiment, we selected those who exhibited task-related changes for the comparison of two conditions, “target” versus “cue” (for this purpose the data was analyzed with SPM using HRF alone). Five subjects who showed bilateral activation of the occipital cortex at a corrected alpha of 0.05 were chosen, and the data of each of them was designated as an

fMRI population. For each population, two areas (voxel size =  $3 \times 3 \times 2$ ) from the occipital cortex were defined as the “active” regions (Fig. 1b). The location of these areas varied slightly ( $\pm 3$  voxels) across the five sub-



**FIG. 2.** Simulated BOLD signal. The figure shows a portion of the simulated time series corresponding to one “active” voxel for a fixed (a) and a variable (b) interstimulus interval of the parametric approach and the nonparametric approach (c). Baseline noise is indicated in black (solid line); 1% signal is indicated in red. Vertical bars show the onset of the stimulus for each experimental design (for practical reasons, only the onsets for the cue—but not for the target—are illustrated in Fig. 2c).

**TABLE 2**  
Model Robustness

Parametric		Model type															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Fixed ISI	Power	0.728	0.740	0.853	0.864	0.208	0.227	0.405	0.445	0.437	0.462	0.556	0.580	0.150	0.161	0.260	0.296
	Cluster size	5.200	5.400	7.200	7.200	2.800	3.000	4.000	4.000	3.400	3.400	4.000	4.000	2.000	2.600	3.000	3.200
	False positives (of 500)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Variable ISI	Power	0.918	0.862	0.932	0.887	0.635	0.513	0.651	0.560	0.720	0.592	0.684	0.592	0.489	0.374	0.464	0.377
	Cluster size	10.00	8.200	9.200	8.200	5.400	4.600	5.400	4.800	6.000	4.600	5.200	4.400	4.200	3.400	4.000	3.800
	False positives (of 500)	3.400	2.000	2.200	0.600	0.600	0.600	1.200	0.400	0.400	0.000	0.000	0.000	0.400	0.000	0.200	0.200
Non parametric	Power	0.916	0.704	0.923	0.728	0.454	0.231	0.567	0.334	0.616	0.328	0.639	0.344	0.432	0.216	0.476	0.280
	Cluster size	15.00	11.00	15.00	11.50	12.00	10.00	13.00	13.00	14.00	13.00	14.50	15.00	11.50	9.500	12.00	14.50
	False positives (of 500)	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

*Note.* Shown are the power and number of false positives for the parametric and nonparametric simulations corresponding to the models defined in Table 1. Power and false positives were obtained from averaging across all datasets and across all active and inactive regions, respectively. Cluster size indicates the average number of voxels whose corrected  $P$  value was smaller than the familywise alpha level of 0.05 for all active regions. The first column indicates the simulation approach used to estimate power and number of false positives.

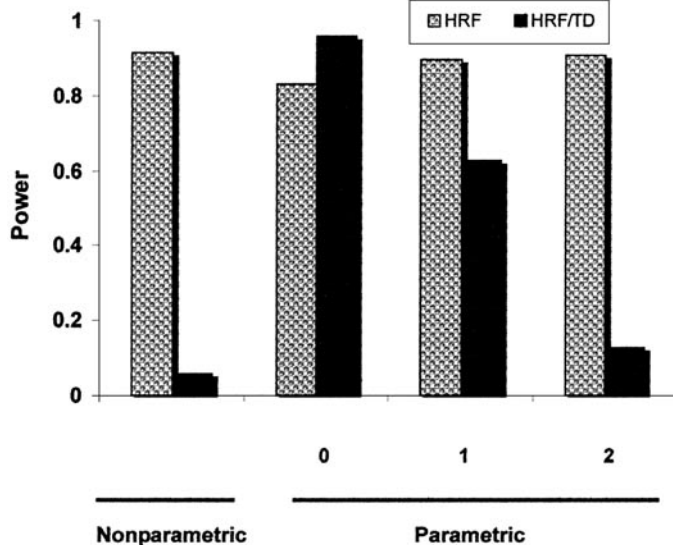
jects according to individual anatomical differences. The standard deviation was computed for each scan of the brain across the 10 subjects. One hundred datasets were generated per population (total = 500) by adding to each scan a normally distributed random error based on the standard deviation of the corresponding volume. The Pearson correlation between the time series of the simulated datasets and the population was around 0.75. Although we used normally distributed errors, this approach was considered as nonparametric in the sense that signal and experimental design were

not manipulated. Our rationale for using the standard deviation for the 10 initial subjects, instead of the one based on the 5 subjects from whom the spatiotemporal series were generated, was that they provided a better estimate of the between-subjects error.

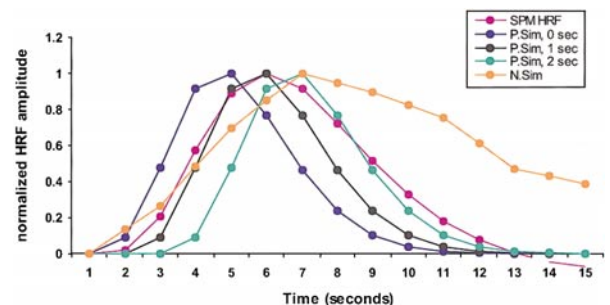
#### Estimation of Power and Type I Error

Once all simulated datasets were generated following the two approaches, they were statistically analyzed using the SPM correction for multiple comparisons based on the theory of GRF (Adler, 1981; Worsley *et al.*, 1992). Corrected  $P$  values were obtained for all voxels, but only the peak voxel of each “active” region was kept for the estimation of power. The same criterion was used to estimate the false positive rate (see below).

Power and false positive rate were assessed at a familywise alpha level (i.e., the alpha level obtained



**FIG. 3.** Effect of using the time derivative of HRF on power estimation. Average power estimates for models 1 and 2 (without time derivative and with time derivative, respectively) are displayed for the nonparametric and the parametric simulations. The latter includes the results from using datasets with response latencies of 0, 1, or 2 s.



**FIG. 4.** Hemodynamic response functions (HRF). Shown are the HRF generated for the parametric simulation (Cohen *et al.*, 1997) (clear blue), the same function with 1-s delay (black) and 2-s delay (green), and the ideal HRF from SPM (pink). The HRF from the nonparametric simulation (orange) is the average HRF estimated from one “active” region of the occipital lobe.

after correcting for multiple comparisons) of 0.05. Power was estimated as the ratio between the number of times that a model yielded a significant outcome for an “active” area and the total number of samples (true positive rate) ( $n = 500$ ). Because the familywise alpha level obtained after adjusting for multiple corrections was extremely low ( $\alpha \approx 0.00005$ , as estimated by interpolating the “threshold”  $t$  value and the degrees of freedom reported by SPM into the  $t$  probability distribution) we were not able to determine type I error with our sample size ( $n = 500$  datasets). In fact, around 20000 simulated datasets would have been needed to compute the type I error correctly. However, given that the number of false positives (the number of times that a model reported a significant outcome for an “inactive” area) of 500 datasets ranged from 0 to 3, we were not too concerned about underestimating type I error rates by using our current sample size. Thus, instead of reporting the type I error, we showed the number of false positives out of 500 datasets (Table 2). To assess the number of false positives from the parametric simulations, five  $3 \times 3 \times 2$  “inactive” areas, i.e., areas composed of voxels where no activation was added to the baseline noise, were selected (Fig. 1a). Two “inactive” regions of the same dimensions were designated for the nonparametric datasets, from areas of the brain where the  $t$  statistic was close to zero (Fig. 1b). Both power and false positives were assessed from different regions of the same datasets. The average number of voxels that reached statistical significance for “active” regions was quantified and is displayed on Table 2.

As with any other statistical measure, power estimates are subject to random variation and therefore confidence intervals are needed to compare differences between models. The most common way of constructing confidence intervals around power estimates generated through Monte Carlo simulations is by assuming a binomial distribution, because each individual sample test contributes with one out of two possible mutually exclusive outcomes (i.e., reject or accept the null hypothesis). However, in cases where the sample variation within simulation scenarios (i.e., models) is greater than random, a binomial approach would overestimate power differences across models. In the present study, the large variation observed within and between subjects suggests that using a binomial distribution would not be appropriate. An alternative approach, commonly used in the realms of robust estimation (e.g., Dryden and Walker, 1999), is to construct confidence intervals empirically by resampling the original data (i.e., sample test probability values) a large number of times and calculating the power for each subset. The confidence interval is then constructed based on the variation around the values for the subsets. These intervals will be influenced by the sampling variability due to subjects and regions, thus providing a more conservative approach for comparing

power between models. Our protocol for estimating confidence intervals was as follows: (1) sample with replacement 100 probability values out of the total values available per model (i.e., 100 sample tests  $\times$  5 subjects  $\times$  5 regions for the parametric simulation, and  $100 \times 5$  subjects  $\times$  2 regions for the nonparametric simulation), using this subset to calculate power; (2) repeat step 1, 1000 times; (3) based on 1000 values generated in step 2, construct a 95% percentile confidence interval. A sampling size of 100 probability values was chosen to estimate the confidence intervals because it represented the smallest sample unit where sampling variation was only due to chance (i.e., number of sample tests generated per subject). Because type I error rates were generally smaller than the specified alpha level for all scenarios, confidence intervals for these estimates were not constructed.

## RESULTS AND DISCUSSION

Overall, the results from our simulations indicate that, despite particular differences, the parameter combination yielding the most powerful results was consistent across the four scenarios of the parametric approach and the nonparametric approach. Specifically, the selection of HRF as the basis function and the high-pass filter (models 1 and 3) were more efficient than any of the other parameters in modeling the fMRI data. Moreover, it is worth emphasizing that the pattern of results obtained using the nonparametric approach resembled closely that from scenario 2 of the parametric approach. Interestingly, although the spatial and temporal structure for the two sets of simulations was very different, their experimental design followed a variable ISI. This finding is important as it reinforces the generalizability of our work. Finally, except for the effect of global scaling, the pattern of results obtained for scenario 1 of the parametric simulation was similar to that obtained for scenario 2 and the nonparametric simulation. However, regardless of the SPM model, datasets generated with a fixed ISI yielded less powerful results. This observation is consistent with the results of a recent simulation study indicating that event-related designs with fixed ISI are less efficient for power detection than those where the presentation of the stimulus is random or semi-random (Liu *et al.*, 2001). A detailed discussion concerning the impact of each SPM preprocessing parameter on power and type I error, follows below.

### *Basis Function: HRF Alone versus HRF/TD*

As mentioned in the methods section, the optimal basis function to model fMRI time series was determined before running all Monte Carlo simulations. To decide whether HRF/TD would be evaluated as another preprocessing parameter, the impact of HRF/TD

**TABLE 3**  
Efficiency of HRF/TD in Adjusting for Delays in the Onset of HRF

Approach	Delay	Model 1				Model 2			
		Effect variance	Residual variance	Fitting residuals	$t$ value	Effect variance	Residual variance	Fitting residuals	$t$ value
Parametric	0	0.138	0.025	16.486	5.421	0.219	0.034	15.368	6.493
	1	0.092	0.021	10.948	4.440	0.106	0.028	10.853	3.735
	2	0.093	0.018	8.725	5.065	0.064	0.025	8.567	2.544
Nonparametric		0.299	0.055	79.214	5.436	0.185	0.074	76.446	2.494

*Note.* Shown are the  $t$  values obtained by dividing the effect variance and the residual variance according to SPM's formula (for details see discussion)  $= T = cb/(c^2(G^{*T}G^*)^{-1}G^{*T}VG(G^{*T}G^*)^{-1}c^T)^{1/2}$ , where the nominator is the effect variance and the denominator, the residual variance. The fitting residuals were obtained from fitting the time series of one active voxel with HRF (Model 1) and HRF/TD (Model 2). The variables were measured based on one dataset derived from a representative subject of each simulation approach. Parametric simulations were generated with a response latency of either 0, 1, or 2 s.

in modeling fMRI data was assessed by testing 500 nonparametric datasets with and without HRF/TD. Figure 3 shows the power computed from averaging the number of true positives across all regions for models 1 (with HRF alone) and 2 (HRF/TD). The results indicate that power was drastically reduced when HRF/TD was selected. The efficiency of this parameter in correcting for differences in response latency was further assessed using the parametric approach by simulating 500 datasets with 0-, 1-, or 2-s delay in the onset of the hemodynamic response curve. To make these results comparable to those obtained using the nonparametric approach, the data was generated according to scenario 2 (variable ISI). The normalized shape and time course of the hemodynamic response curve corresponding to SPM, the parametric (with the three delay conditions) and nonparametric simulations are illustrated by Fig. 4. The outcomes, depicted in Fig. 3, indicate that including HRF/TD as an extra covariate to the GLM only increased the power for the 0 sec delay condition. However, it attenuated the power significantly for datasets with a response latency of 1 s and drastically for datasets with a response latency of 2 s. Together with the onsets displayed by Fig. 4, these results served to explain why the power was so low when nonparametric datasets were modeled using HRF/TD.

Further comparison of the parameter estimates (beta coefficients), the variance and the residuals for HRF and HRF/TD, using SPM's test of statistical inference<sup>1</sup> (Worsley and Friston, 1995), helped us understanding the nature of these results. Table 3 displays the results obtained for one active voxel of one dataset

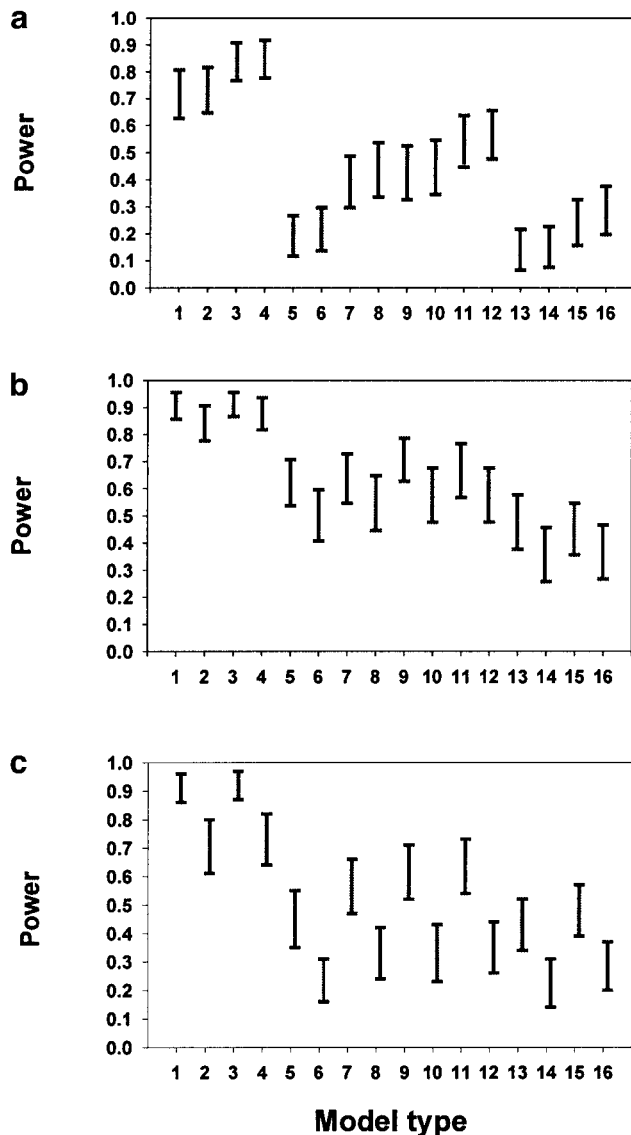
of the parametric simulation (scenario 2) and the nonparametric simulation. The addition of the time derivative decreased the curve fitting residuals, but also increased the residual variance used for calculation of the  $t$  statistic and hence, did not always yield higher  $t$  values. A higher  $t$  value was obtained for the 0-s delay condition, because the effect variance for the time series was significantly augmented by HRF/TD. Conversely, the slightly higher effect variance obtained for the 1-s delay condition was not enough to overcome the high residual variance associated with the inclusion of the derivative, resulting in a lower  $t$  value. Finally, the addition of the time derivative to model the 2-s delay condition and the nonparametric time series yielded to a lower effect variance, drastically reducing both  $t$  values.

These findings are consistent with the results of our simulations and support the observation that the efficacy of the time derivative in accounting for delays in the response onset depends on the magnitude of the delay. Moreover, they suggest that improving the model fitness does not always lead to higher power estimates. Based on one real fMRI dataset, Hopfinger and collaborators (2000) have reported similar sensitivity of HRF and HRF/TD. Given the interchangeability of the two basis functions, the authors suggested to use HRF/TD to account for occasional delays in HRF. Our results, however, indicate that depending on the response latency, HRF/TD may drastically reduce the power of the analysis. For these reasons, HRF/TD was not considered as a parameter set for further evaluation in this study.

#### *Global Intensity Normalization: Scale or None?*

Scaling the data by the global mean attenuated power for those datasets with variable ISI (i.e., scenario 2 of the parametric simulation and the nonparametric simulations) (Fig. 5). This effect was particu-

<sup>1</sup>  $t$  values were computed according to the formula  $T = cb/(ce^2(G^{*T}G^*)^{-1}G^{*T}VG(G^{*T}G^*)^{-1}c^T)^{1/2}$ , where  $\mathbf{c}$  represents the contrast of interest,  $\mathbf{b}$  is the parameter estimate,  $e^2$  is an unbiased estimator of the error variance  $\sigma^2$ ,  $\mathbf{V} = \mathbf{K}\mathbf{K}^T$ , where  $\mathbf{K}$  is a matrix whose rows represent the hemodynamic response function and  $\mathbf{G}^* = \mathbf{K}\mathbf{G}$ , where  $\mathbf{G}$  is the design matrix. <sup>T</sup> indicates the matrix transpose.



**FIG. 5.** Power estimates of SPM models. Shown are the confidence intervals for all 16 models, corresponding to scenarios 1 (a) and 2 (b) of the parametric approach, and the nonparametric approach (c) ( $n = 500$ ). Confidence intervals were obtained from spatially smoothed datasets with a 10-mm FWHM.

larly pronounced for all models of the nonparametric simulations.

The use of global scaling to process neuroimaging data remains controversial. Global signals, i.e., variations in signal that are common to the entire brain volume, were initially considered to represent the underlying background to regional changes in activity (Ramsay *et al.*, 1993). When the global mean is independent of the experimental condition, scaling by the grand mean can be beneficial because it reduces inter-subject variability thereby improving the sensitivity at the group level of analysis (McIntosh *et al.*, 1996; Aguirre *et al.*, 1998). However, there is likely little

benefit for the analysis of single subjects. This hypothesis is consistent with our results. The reason why global scaling was only detrimental to datasets with variable ISI is, however, unclear.

#### *Temporal Filtering: High Pass Filter, Low Pass Filter, and Autoregressive Models*

Due to serial dependency of physiological and non-physiological components of the noise, fMRI time series violate the independence of the error term, one of the assumptions of the general linear model. This colored noise represents a problem for inference based on time series regression parameters (Bullmore *et al.*, 2001), and thus should be controlled for. Uncorrelated low-frequency noise can be removed by using a high-pass filter. Colored high-frequency noise can be either treated with a low-pass filter that convolves the time series with a Gaussian filter of the width of HRF (Friston *et al.*, 1995b; Worsley and Friston, 1995; Zarahn *et al.*, 1997) or removed by using an autoregressive model (Friston *et al.*, 2000). In a recent paper, Friston and collaborators (Friston *et al.*, 2000) reported that modeling unwanted frequency components by a combination of a high and a low-pass filter (temporal smoothing) provided a good parameter estimation of the GLM, while protecting for inferential bias. Using a 1st order autoregressive model (AR1) was more efficient at parameter estimation but significantly enhanced inferential bias.

Our results comparing the confidence intervals for all simulations (Fig. 5) indicate that the use of a high-pass filter set at default cutoff, may be beneficial depending on the experimental design. Indeed, although no obvious improvement was observed for those simulations with variable ISI, those with fixed ISI showed higher power when the high pass filter was included (see Fig. 5a, models 3 and 4). Nevertheless, it is important to keep in mind that the use of a high pass filter will depend on the amount of low-frequency noise, which varies with the scanner.

The efficacy of using the low-pass filter or AR1 to model temporal autocorrelation should be evaluated in relation with the type I error rates, which varied with the experimental design (i.e., fixed or variable ISI) and the ISI. Although the number of false positives obtained for the parametric simulations with fixed ISI were 0 of 500 datasets for all models, those obtained for the parametric simulations with variable ISI were higher for the models with higher power. Although at first glance the average number of false positives for the first three models does not appear particularly high for a nominal alpha of 0.05 (Table 2: 3.4, 2, 2.2 of 500 for models 1, 2, and 3, respectively), they are certainly much higher than the familywise alpha resulting after correction for multiple comparisons, i.e., around  $5 \times 10^{-5}$ . Note that the implementation of the low-pass filter or

AR1 reduced the number of false positives (see models 5 to 16 from Table 2), suggesting that they were efficient in modeling serial autocorrelations. However, the number of false positives for the nonparametric simulation, where stimuli were also presented with a variable ISI, was not high (only 1 false positive was obtained for model 3, whereas the rest had no false positives). We think that this difference may have originated in the length of the ISI. Figure 2 indicates the ISI between some of the stimuli of the parametric simulation was very short (as short as 3 or 6 s in occasions), whereas the minimum ISI for the nonparametric simulation was 8 s (minimum ITI was also 8 s). Although the inactive areas lacked any activation, the regression model used to fit the fMRI time series for the inactive voxels was determined by the stimuli onsets of the active regions. We hypothesize that a regression model specified according to the onsets displayed by Fig. 2b (corresponding to the parametric simulation with variable ISI) would result in a better fit for the noise of the inactive voxels than that specified according to the onsets displayed by Fig. 2c (corresponding to the non-parametric simulation also with variable ISI). As a result, the number of false positives occurring in those areas would increase for the model specified by the parametric simulation with variable ISI but not as much for the nonparametric simulation with longer ISI. That was in fact, the pattern obtained for the type I errors (Table 2). We confirmed our hypothesis by running 500 additional parametric simulations with variable ISI, in which we increased the ISI to at least 8 s, which showed a reduction in type I error with no changes in the power estimates (data not shown).

Based on these findings, we conclude that the efficiency of the low-pass filter and AR1 in modeling serial autocorrelations depends on the ISI. In our case, the relatively low incidence of false positives associated with the most powerful models (1 and 3), suggest that the implementation of a low-pass filter or AR1 is not necessary for valid inference. However, when dealing with fMRI time series where stimuli are presented close together, such as in rapid presentation event-related designs, the risk of inferential bias would increase. In those cases, the use of AR1 is recommend over the low-pass filter as it appears to decrease the number of false positives while maintaining a relatively high power (see model 9 from Table 2).

## CONCLUSIONS

Our main goal was to conduct a simulation study where differences in performance of SPM preprocessing parameters could be contrasted and revealed. Several conclusions can be extracted from assessing the robustness of the SPM preprocessing parameters using simulated fMRI datasets. It is important to keep in

mind that these conclusions are based on the scenarios considered in this study for individual subject analysis, and thus may not apply to all fMRI experiments. However, we have adopted a framework that is sufficiently general to guide SPM users in assessing the robustness of other data sets or scenarios that may be more appropriate to their specific questions.

To begin, given the inconsistencies associated with the use of HRF/TD, it would seem wise to use HRF over HRF/TD to model fMRI data. The use of an HRF that is empirically derived for each voxel, rather than a canonical HRF, may prove to be the best overall solution to the discrepancy. The use of the high-pass filter is recommended when analyzing fMRI datasets with fixed ISI. No improvement in power, over the use of HRF alone, was evident when applied to datasets with variable ISI. The use of global scaling for individual analysis should be avoided since it can significantly reduce the power, in particular for datasets with variable ISI.

Finally, given that both the low pass filter and the first-order autoregressive function decrease power, it is only recommended to use them for fMRI datasets with short ISI (<8 s), which are more susceptible to inferential bias. In those cases, AR1 appears to be more efficient than HRF in that it controls for the incidence of false positives while maintaining a relatively high power. The effect of using the gaussian filter, alternative hemodynamic response functions, and a putative more efficient high-pass filter remains to be tested.

## ACKNOWLEDGMENTS

The first two authors of the paper have equally contributed to this work. We thank Barry Giesbrecht and George R. Mangun for providing us with the fMRI Data used to generate the Monte Carlo simulations. The computer code may be obtained by contacting Dr. Wilkin Chau at wchau@rotman-baycrest.on.ca. We are also grateful to Craig Easdon for his helpful comments on our manuscript. Funded by Natural Sciences and Engineering Research Council and Canadian Institutes of Health Research held by A. R. McIntosh.

## REFERENCES

- Adler, R. J. 1981. *The Geometry of Random Fields*. Wiley, New York.
- Aguirre, G. K., Zarahn, E., and D'Esposito, M. 1998. The inferential impact of global signal covariates in functional neuroimaging analyses. *Neuroimage* **8**(3): 302–306.
- Cox, R. W. 1996. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**(3): 162–173.
- Cohen, M. S. 1997. Parametric analysis of fMRI data using linear systems methods. *NeuroImage* **6**: 93–103.
- Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. L. Erlbaum, Hillsdale, NJ.
- D'Esposito, M., Zarahn, E., and Aguirre, G. K. 1999. Event-related functional MRI: Implications for cognitive psychology. *Psychol. Bull.* **125**(1): 155–164.
- Dryden, I. L., and Walker, G. 1999. Highly resistant regression and object matching. *Biometrics* **55**: 820–825.

- Efron, B., and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*. Chapman & Hall.
- Friston, K. J., Jezzard, P., and Turner, R. 1994. Analysis of functional MRI time series. *Hum. Brain Mapp.* **1**: 153–171.
- Friston, K. J., et al. 1995. Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.* **2**: 189–210.
- Friston, K. J., Frith, C. D., Turner, R., and Frackowiak, R. S. 1995a. Characterizing evoked hemodynamics with fMRI. *NeuroImage* **2**(2): 157–165.
- Friston, K. J., Holmes, A. P., Poline, J. B., Grasby, P. J., Williams, S. C., Frackowiak, R. S., and Turner, R. 1995b. Analysis of fMRI time-series revisited. *NeuroImage* **2**(1): 45–53.
- Friston, K. J., Josephs, O., Zarahn, E., Holmes, A. P., Rouquette, S., and Poline, J. 2000. To smooth or not to smooth? Bias and efficiency in fMRI time-series analysis. *NeuroImage* **12**(2): 196–208.
- Giesbrecht, B., Woldorff, M. G., Fichtenholtz, H. M., and Mangun, G. R. 2000. Isolating the neural mechanisms of spatial and non-spatial attentional control. 30th Annual Meeting of the Society for Neuroscience, New Orleans, LA.
- Holmes, A. P., Josephs, O., Büchel, C., and Friston, K. J. 1997. Statistical modeling of low-frequency confounds in fMRI. Proceeding of the 3rd International Conference of the Functional Mapping of the Human Brain, S480.
- Hopfinger, J. B., Buchel, C., Holmes, A. P., and Friston, K. J. 2000. A study of analysis parameters that influence the sensitivity of event-related fMRI analyses. *NeuroImage* **11**(4): 326–33.
- Liu, T. T., Frank, L. R., Wong, E. C., and Buxton, R. B. 2001. Detection power, estimation efficiency, and predictability in event-related fMRI. *NeuroImage* **13**: 759–773.
- Jezzard, P., and Song, A. W. 1996. Technical foundations and pitfalls of clinical fMRI. *NeuroImage* **4**(3 Pt 3): 63–75.
- Mcintosh, A. R., Grady, C. L., Haxby, J. V., Maisog, J. Ma, Horwitz, B., and Clark, C. M. 1996. Within subject transformations of PET regional cerebral blood flow data: ANCOVA, ratio, and Z score adjustments on empirical data. *Hum. Brain Mapp.* **4**: 93–102.
- Oden, N. L. 1991. Allocation of effort in Monte Carlo simulation for power of permutation tests. *J. Am. Stat. Assoc.* **86**: 1074–1076.
- Peres-Neto, P., and Olden, J. D. 2001. Assessing the robustness of randomization tests: Examples from behavioral studies. *Animal Behav.* **61**: 79–86.
- Petersson, K. M., Nichols, T. E., Poline, J.-B., and Holmes, A. P. 1999. Statistical limitations in functional neuroimaging II. Signal detection and statistical inference. *Philos. Trans. R. Soc. Lond. B* **354**: 1261–1281.
- Purdon, P. L., and Weisskoff, R. M. 1998. Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI. *Hum. Brain Mapp.* **6**(4): 239–249.
- Ramsay, S. C., Murphy, K., Shea, S. A., Friston, K. J., Lammertsma, A. A., Clark, J. C., Adams, L., Guz, A., and Frackowiak, R. S. 1993. Changes in global cerebral blood flow in humans: Effect on regional cerebral blood flow during a neural activation task. *J. Physiol.* **471**: 521–534.
- Stephens, M. A. 1974. EDF statistics for goodness of fit and some comparisons. *J. Am. Stat. Assoc.* **69**: 730–737.
- Thomas, L., and Krebs, C. 1997. A review of statistical power analysis software. *Bull. Ecol. Soc. Am.* **78**: 126–140.
- Turner, R., Howseman, A., Rees, G. E., Josephs, O., and Friston, K. 1998. Functional magnetic resonance imaging of the human brain: Data acquisition and analysis. *Exp. Brain Res.* **123**(1–2): 5–12.
- Worsley, K. J., Evans, A. C., Marrett, S., and Neelin, P. A. 1992. Three-dimensional statistical analysis for CBF activation studies in human brain. *J. Cereb. Blood Flow Metab.* **12**(6): 900–1180.
- Worsley, K. J., and Friston, K. J. 1995. Analysis of fMRI time-series revisited—Again. *NeuroImage* **2**(3): 173–181.
- Worsley, K. J., Marrett, S., Neelin, P., Vanal, 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 Mapp.* **4**: 58–73.
- Zarahn, E., Aguirre, G. K., and D'Esposito, M. 1997. Empirical analyses of BOLD fMRI statistics. I. Spatially unsmoothed data collected under null-hypothesis conditions. *NeuroImage* **5**(3): 179–197.