NeuroImage **17,** 532–542 (2002)
doi:10.1006/nimg.2002.1226

Simulation of the Effects of Global Normalization Procedures in Functional MRI

Maria Gavrilescu,*,†,‡ Marnie E. Shaw,*,‡,§ Geoffrey W. Stuart,‡,¶ Peter Eckersley,*
Imants D. Svalbe,† and Gary F. Egan*,‡

**Howard Florey Institute, University of Melbourne, Australia; †School of Physics and Materials Engineering, Monash University, Melbourne, Australia; §School of Physics, University of Melbourne, Australia; ¶Mental Health Research Institute of Victoria, Melbourne, Australia; and ‡Centre for Neuroscience, Melbourne, Australia*

Received November 27, 2001

We report on differences in sensitivity and false-positive rate across five methods of global normalization using resting-state fMRI data embedded with simulated activation. These methods were grand mean session scaling, proportional scaling, ANCOVA, a masking method, and an orthogonalization method. We found that global normalization by proportional scaling and ANCOVA decreased the sensitivity of the statistical analysis and induced artifactual deactivation even when the correlation between the global signal and the experimental paradigm was relatively low. The masking method and the orthogonalization method performed better from this perspective but are both restricted to certain experimental conditions. Based on the results of these simulations, we offer practical guidelines for the choice of global normalization method least likely to bias the experimental results.

INTRODUCTION

In a typical fMRI experiment, there is extra variation in the mean signal level between individual scans in addition to the one due to alterations in blood flow caused by the activation paradigm. Potential sources of such variation include physical processes (head movements, instabilities of the scanner baseline) and underlying physiological processes: pulsation, breathing, and swallowing, as well as complex interactions between the activation signal and other processes (e.g., induced vasodilation—see Petersson *et al.*, 1999 for a review). Modeling the variation in the signal intensity across a large number of voxels is termed *global normalization*.

In the context of the General Linear Model (GLM) this variation can be accounted for in two ways: as an additive term in the model (ANCOVA), representing a covariate of no interest (Friston *et al.*, 1990),

$$Y_i(t) = \mu_i + \alpha_i h(t) + \beta_i (c(t) + G(t)) + \varepsilon_i(t), \quad (1)$$

or as multiplicative or ratio scaling (Fox *et al.*, 1984),

$$Y_i(t) = G(t)(\mu_i + \alpha_i h(t) + \beta_i c(t) + \varepsilon_i(t)), \quad (2)$$

where Y_i is the signal intensity for voxel i, μ_i represents the baseline intensity, $h(t)$ denotes the vector of covariates of interest reflecting task induced effects, $c(t)$ the vector of covariates of no interest, α_i and β_i are the constant vectors of parameter estimates, $G(t)$ is the term representing the global variation estimated as the average intensity of the intracranial voxels (global signal, Desjardins *et al.*, 2001), and $\varepsilon_i(t)$ represents the error term. One major difference between the two approaches is that proportional scaling scales the error variance (ε_i (t)) whereas ANCOVA does not. While no general consensus has been reached regarding the most appropriate method, the variance stabilizing properties provide an advantage for proportional scaling over the ANCOVA approach (Holmes *et al.*, 1997). The proportional scaling method is widely used for fMRI data (ANCOVA is not directly implemented in the fMRI module of spm99). However, we acknowledge that this preference is based on expert opinion[1] rather than published research. For this reason we have included ANCOVA in our study.

The widely used SPM99 package (http://www.fil.ion.ucl.ac.uk/spm) offers two ways of carrying out global normalization of fMRI data: grand mean session scaling and proportional scaling. Grand mean session scaling is a simple adjustment designed to remove only intersession variance in the global signal due mainly to changes in the gain of scanner amplifiers. In global normalization by proportional scaling, the intensity for each voxel in a scan is effectively divided by the global signal for that scan. This method attempts to remove both intersession (session-to-session) and intrasession

[1] This statement is based on a search performed on the spm discussion list at http://www.jiscmail.ac.uk/archives/spm.html: J. B. Poline (25/09/98); K. J. Friston (22/12/98).

1053-8119/02 $35.00

(scan-to-scan) variance in the global signal. The central assumption of this method is that global variation is not due to the changes produced by the activation paradigm. The global variation can be considered an orthogonal nuisance variable ($G(t)$ and $h(t)$ are not correlated) that can affect only the error term in the model and which reduces the sensitivity of the subsequent statistical analysis if not accounted for. Thus the global signal is used as the baseline against which to measure activation (or deactivation). These assumptions seem reasonable, given that changes in cerebral blood flow produced by activation tasks are subtle and are frequently confined to a small number of focal regions. This probably accounts for the popularity of the technique. However, when the assumptions are violated, the global signal becomes a confound; i.e., inclusion or exclusion of the global signal in the model will affect the relationship between the data and the variables of interest. To include this possible situation in the GLM we expanded $G(t)$ as follows:

$$G(t) = G + G_v(t) + A(t) + B(t), \qquad (3)$$

where G represents the mean global signal, $G_v(t)$ represents the variation around the mean that is not correlated with the covariates of interest ($h(t)$), $A(t)$ is the variation around the mean correlated with $h(t)$ due to task induced activation, and $B(t)$ represents the variation around the mean correlated with $h(t)$ due to other underlying physical or physiological processes. In practice, is not possible to estimate the separate contribution of different terms in $B(t)$; however, while some sources of variance are best removed from the global variation (e.g., task correlated motion artifacts; Freire *et al.*, 2001), part of this contribution is likely to express genuine global variation that should be preserved in the estimated global signal. Possible mechanisms able to generate correlation between the global signal and the paradigm are described by Aguirre *et al.* (1998).

$G_v(t)$, $A(t)$, and $B(t)$ each can be further expressed as a sum of two terms: one representing variation around the mean expressed in the activated voxels (denoted by $G_{Va}(t)$, $A_a(t)$, and $B_a(t)$, respectively) and the other one representing variation around the mean in the nonactivated voxels (denoted by $G_{Vn}(t)$, $A_n(t)$, and $B_n(t)$, respectively). The inclusion of $A_n(t)$ may seem inappropriate. Its inclusion is justified by the observation that the identification of activated voxels pertains to the method used to analyse the data. When certain voxels are labeled as nonactivated via a particular analysis method, the possibility that these voxels can express task-related variations in the intensity values is not excluded. Significant correlations between global and local activity may arise as a result of relatively strong and/or widespread activation (Aguirre *et al.*, 1998; Desjardins *et al.*, 2001). This can result in two unwanted effects when global normalization is employed: underestimation of the true levels of focal activity and the production of artifactual deactivations due solely to the overcorrection of global variation.

Other global normalization methods were recently introduced to account for the correlation between the global signal and the paradigm. Andersson (1997) has proposed a more complex model for global normalization whereby the global variations are calculated independently of local changes in blood flow. This method, which calculates the global signal while "masking" out focal activations, was successfully applied to PET data from an experiment involving a large visual activation. The masking method attempts to remove $G_{Va}(t) + A_a(t) + B_a(t)$ from the calculation of the global signal by assuming that for spatially localized activations $A_a(t) + B_a(t) \ll A_n(t) + B_n(t)$. For the nonactivated voxels, any possible correlation between the global signal and the paradigm induced by processes other than the task induced activation is accounted for. The procedure is applied iteratively: at each step the activation pattern captured in the statistical parametric map is masked out and the global signal is calculated as the average intensity of the remaining voxels. This new global signal is then used for proportional scaling (or ANCOVA) and the number of activated/deactivated voxels is calculated ($P_{\text{uncorrected}} < 0.05$). A saturation effect for this number is then determined. A potential problem for this method arises for data sets with spatially extended activation, where the number of voxels remaining after the masking process may be too small to estimate global changes in the signal intensity. To our knowledge this method has not previously been tested for fMRI data.

The orthogonalization method as proposed by Desjardins *et al.* (2001) performs an adjustment of the global signal so that it becomes orthogonal to any nonconstant column of the design matrix. The correlation coefficient between the global signal and the nonconstant columns in the design matrix is effectively forced to be zero. The method corrects for $A(t)$ in all voxels, but attributes all the shared variance expressed by $B(t)$ to the covariates of interest. As we pointed out above, not all this variance should be discarded from the global signal. The underlying assumption of this method is that the major source of correlation between the global signal and the experimental paradigm is the activation signal itself, and the effects of other processes are either negligible or spatially incoherent. When this assumption does not hold (i.e., $A(t)$ and $B(t)$ are comparable), the orthogonalization method may increase the number of false positives. The major advantage of this method is that it can be applied to spatially extended activations that preclude the use of masking techniques.

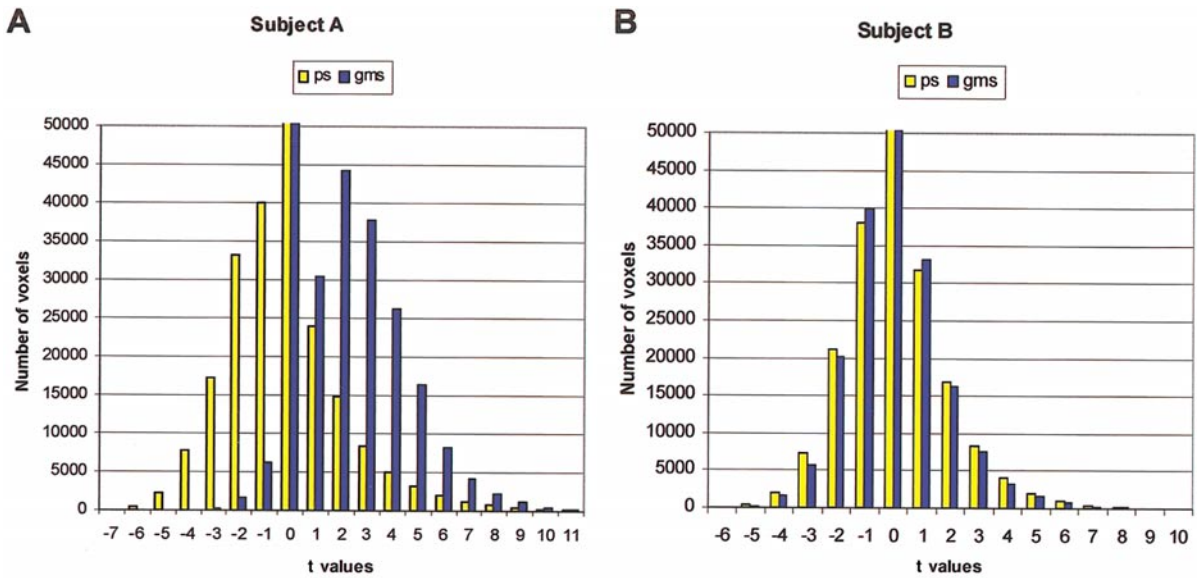

FIG. 1. (A) The histogram of the t maps with two global normalization methods: grand mean scaling and proportional scaling for subject A performing a visual attention paradigm. The t bins were defined as −0.5 to 0.5 around the values on the *x*-axis (e.g., 1 represents the bin 0.5 to 1.5). (B) Same comparison for subject B involved in a modified task. The vertical scale was restricted to values less than 50,000. This restriction affects only the voxels with $-0.5 < t < 0.5$ where both methods indicate a similar number of voxels for both subjects.

One disadvantage of both the masking and the orthogonalisation methods is that they optimize the estimation of the global signal against a proposed model and thus can not be used with a multivariate data-driven approach (such as principal components analysis). A completely new approach to global signal estimation, independent of both local intensity changes and the model imposed on the data, is presented by Andersson *et al.* (2001) based on multivariate characteristics of the PET data. This new method provides a robust way to preprocess the PET data. However, further work is needed to prove the method can be used in the context of very complex spatio-temporal structure characteristic of fMRI data.

In the examples presented in the literature to illustrate the negative impact of the correlation between the global signal and the experimental paradigm, the correlation is highly significant (t = 9.5, Aguirre *et al.*, 1998; $Z = 7.88$, $Z = 5.31$ Desjardins *et al.*, 2001). In practice we found that even for much lower correlation levels, global normalization by proportional scaling can sometimes produce conflicting results when compared with grand mean scaling. To illustrate this, Fig. 1A represents the histograms of the t values for proportional scaling versus grand mean scaling applied to one subject (subject A) performing a visual attention paradigm. A complete description of the experimental task and results can be found elsewhere (Chapman *et al.*, 2002); however, a short task description is provided in the methods section.

The spatial location of activated/deactivated areas is not relevant in this context. As can be seen in Fig. 1A, proportional scaling adjustment produced a greater number of significantly deactivated voxels while grand mean scaling produced mainly significant activation. Figure 1B represents another subject (subject B) performing a modified visual attention task where the two global normalization methods give similar results. The correlation between the global signal and the paradigm converted to a Z score was $Z = 1.470$ ($df = 248$) for subject A and $Z = 0.350$ ($df = 256$) for subject B. Both

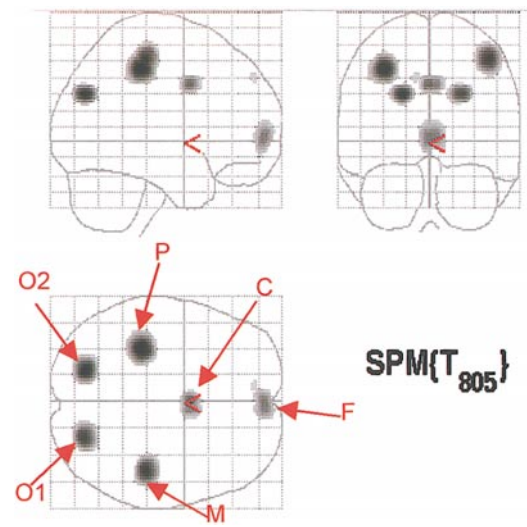

FIG. 2. Spatial distribution of the simulated activation for 1.5% amplitude in smaller clusters data sets: O1, O2 clusters in occipital lobe, P cluster in parietal lobe, M cluster in motor cortex, C cluster in cingulate cortex, F cluster in frontal lobe.

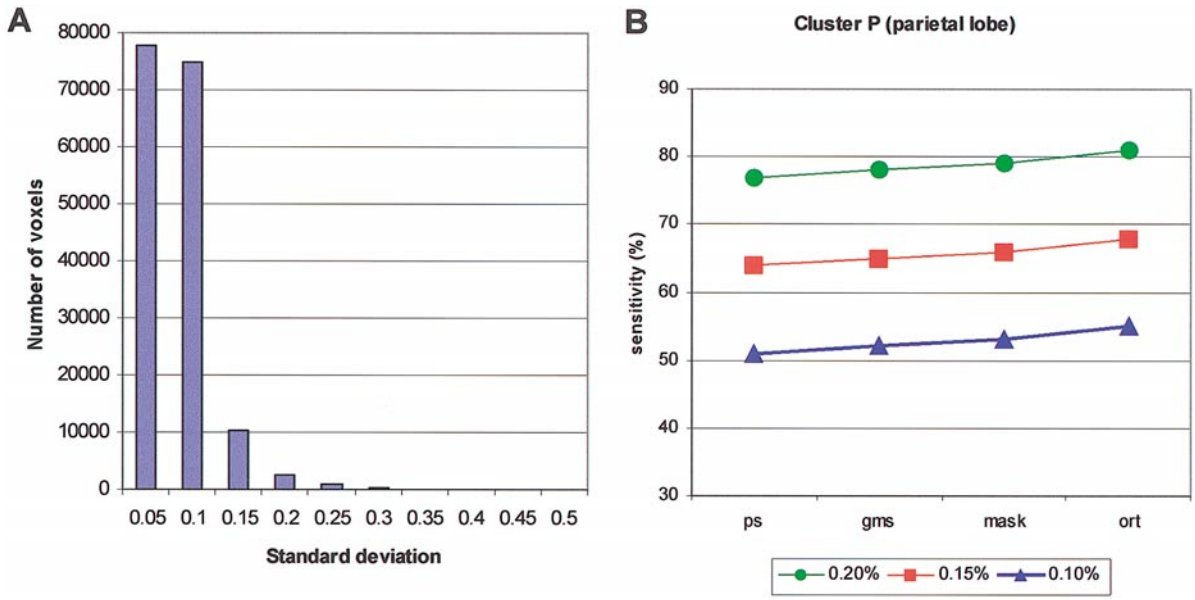

FIG. 3. (A) The histogram of the standard deviations around the mean represented as percentage from the mean for one subject in the resting state. (B) The effect of cutoff value for the definition of the cluster of interest in sensitivity measures for cluster P in parietal cortex in 1.5% amplitude small clusters data set. Three cutoff values were compared: 0.2, 0.15, and 0.1% of the mean signal intensity.

values are smaller than the values reported in the literature. These examples illustrate the need for a more informed and objective choice of the global normalization adjustment. There is also a need to determine the "boundary conditions" that make the use of each particular technique appropriate and that avoid bias in the statistical inference. This outcome can be achieved using known activation patterns incorporated in a background of real fMRI noise.

In this study we used fMRI data, acquired in the resting state, with added simulated focal activity, and compared the sensitivity and false-positive rate across different methods of global normalization. Previous studies have not addressed whether global normalization by proportional scaling introduces a confound or indeed a nuisance variable for data sets where the correlation between the global signal and the paradigm is relatively low. Therefore, in this study we examined a range of correlation values that, although not large, are similar to the calculated values for our real data sets.

METHODS

Brain images for seven healthy subjects were acquired at rest (eyes closed) using an EPI sequence (GE 1.5 T, TR = 3500 ms, TE = 40 ms, FA = 60, 64 × 64 × 15 matrix, 4 × 4 × 7-mm spatial resolution, 160 images/subject, in five sessions of 112 s each). Prior to embedding the simulated signal, the images were aligned and spatially normalised to the EPI template using SPM99 software, in order to place the activation signal in equivalent brain locations rather than in the same image position.

Six activation clusters were inserted in different brain areas, shaped as Gaussian spheres (Fig. 2). For the resting state data, the mean intensity varied from 765.8 ± 5.9 to 1093.9 ± 6.1 across subjects. Five data sets were created by combining three levels of signal amplitude (0.75, 1.5, and 2.5% of mean image intensity) with two levels of spatial extent (smaller clusters with standard deviation σ from 4 to 7 mm and larger clusters with σ from 6 to 13 mm). The Gaussian distributions were truncated at 3.8σ. The centroids of O1, O2, P, C, and M clusters were placed in slightly different voxel locations (by one to two voxels) across data sets. For cluster F the centroid was placed in exactly the same voxel for all data sets.

The correlation values between the simulated paradigm ($h(t)$) and the global signal ($G(t)$) are given in Table 1. The correlation between the global signal in the resting state data and the simulated paradigm was different from zero (r_0). The embedded activation induced an increase in the correlation value from r_0 to r (via $A(t)$).

The simulated activations were modeled in time by convolving the haemodynamic response function (hrf) from SPM99 with a square wave (21-s period). For the cluster in the frontal area F (Fig. 2), we introduced a 3-s spread in the response delay relative to onset across the seven subjects.

Given the way we constructed the clusters, the signal intensity for the voxels at the edge of the clusters was similar to the intensity of the underlying noise and

TABLE 1

Correlation Induced by the Simulated Activation across Data Sets

Signal amplitude (%)	Spatial extent σ (mm)	Activated voxels (%)	Correlation[a] (r/r_o)	Correlation[b] (Z)	Artifactual correlation[a] (r_a/r_o)
0.75	4–7	0.8	1	$-0.3359 \rightarrow 0.4273$	1.2
1.5	4–7	1.4	2	$-0.3128 \rightarrow 0.4748$	−0.4
0.75	6–13	3.9	3.5	$-0.2689 \rightarrow 0.5736$	−2.2
1.5	6–13	7	7.5	$-0.1156 \rightarrow 0.9248$	−0.1
2.5	6–13	10.3	12	$0.0876 \rightarrow 1.3501$	1.8

Note. r/r_0, the ratio of induced correlation and the correlation in the resting state data; Z, the absolute size of the correlation induced; r_a, artifactual correlation calculated as the mean value of the correlation across 1000 random permutations of the experimental paradigm.

[a] Group level correlation.

[b] Min and max value across subjects.

thus we cannot expect to detect those voxels as activated. For this reason, in the sensitivity measurements we defined a *cluster of interest* as all of the voxels inside the simulated clusters with intensity values greater than 0.2% of the mean image intensity. This cutoff value was chosen based on the histogram of the standard deviation in the signal intensity for the resting data represented as a percentage of the mean intensity (Fig. 3A). Increasing this cutoff value will restrict the sensitivity measure to the top of each cluster, where all methods are expected to perform equally well. Decreasing this value will only reduce the absolute values of sensitivity without changing the relative variations across methods as illustrated in Fig. 3B (for cluster P in 1.5% amplitude, smaller clusters data set).

While the sensitivity was measured inside the clusters of interest, the false-positive rate was measured outside the simulated clusters in this way avoiding the contamination of the results by the weakly activated voxels at the edge of the clusters. Because the area outside the simulated clusters varied for each data set, we expect the false-positive rate to be different for the same method across data sets. In data sets with larger clusters, the tails of the spatial Gaussian distributions overlapped. At the cluster of interest cutoff value of 0.2% of the mean image intensity, some of the clusters of interest overlapped and in these cases we could not measure an individual cluster sensitivity value. This was the case for cluster *P* at the 2.5% amplitude in the larger clusters data set. Since the amplitude of the simulated signal varied within the cluster, the probability of detecting a voxel as activated diminished at the cluster boundary, resulting in reduced absolute sensitivity. However, we are interested only in *relative* variations in sensitivity across the global normalization methods described herein.

For each data set we performed five SPM99 fixed-effects analyses, maintaining all parameters constant except for the global normalization procedure. A high-pass filter (49 s) was applied in all analyses. For one data set (1.5% amplitude in larger clusters) we also performed a random-effects analysis. The global normalization methods compared were grand mean session scaling global normalization (gms), global normalization by proportional scaling (ps), ANCOVA (ANC), masking method (mask), and orthogonalization method (ort). ANCOVA analysis where the mean centered global signal was included in the design matrix as a session specific user defined regressor (session specific ANCOVA - ANC) was performed for all simulated and real data sets. A subject-specific ANCOVA (ANCs) was also performed for 2.5% amplitude, and 1.5% amplitude in the larger clusters data sets (the mean centered global signal was introduced as a subject specific regressor in the design matrix). For the masking method the activation was defined based on the t map thresholded at $P_{\text{corrected}} < 0.05$. For this method we also studied how the probability threshold value affects the sensitivity. To achieve this we defined the activation to be excluded at four threshold levels: $P_{\text{uncorrected}} < 0.1$ (mask_1); $P_{\text{uncorrected}} < 0.01$ (mask_2); $P_{\text{uncorrected}} < 0.001$ (mask_3); $P_{\text{corrected}} < 0.05$ (mask_4). The number of voxels excluded by each mask is different for each data set and increased from mask_4 to mask_1. As the number of excluded voxels increased, the estimation of the global signal was less reliable. For mask_3 and mask_4 the estimation of the global signal was based on at least 90% of the intracranial voxels in all data sets.

For all simulated data sets the t map was thresholded at $P_{\text{corrected}} < 0.05$ for fixed effects and $P_{\text{uncorrected}} < 0.001$ for random effects to measure the sensitivity (percentage of identified activated voxels from the total number of activated voxels inside the clusters of interest) and the false-positive rate (the proportion of voxels identified as activated from the total number of voxels outside the simulated clusters). The effect of the global normalization techniques on induced deactivation was measured by counting all the voxels in the deactivation map (voxels with intensity values smaller than the baseline intensities) thresholded at $P_{\text{uncorrected}} < 0.001$.

The data set for subject A (Fig.1A) was obtained from a comparison between two conditions in a selective visual attention experiment using a reach and grasp task. The experimental condition required subject A to reach out and grasp a target stimulus flanked by nontargets prior to movement initiation. This condition was compared to a "view only" condition where the subject observed three stationary targets for the duration of the condition period. Data for subject B (Fig. 1B) was also obtained from a selective visual attention experiment but the conditions differed slightly. In the condition of interest the subject was cued to the target with a fiber-optic light and the nontarget objects remained on the display during the reach-to-grasp response. This condition was compared to a "view only" condition and once again the baseline condition for subject B differed slightly to that used for subject A: the three nontargets moved synchronously in and out of the display during the condition period. The two subjects were preprocessed in the same way. A high-pass filter with a time constant of 105 s was included for both subjects. The motion parameters and/or their derivatives were not included in the design matrix.

The session-specific ANCOVA, masking, and the orthogonalization methods were applied to subjects A and B. For the masking method five iterations were performed using the t map of the contrast thresholded at $P_{\text{corrected}} < 0.05$ and the F map of the effects of interest thresholded at $P_{\text{uncorrected}} < 0.05$, as suggested by Andersson (1997). Since for real data the "ground truth" is unknown, we compared the results in terms of the number of voxels reported in the activation/deactivation statistical maps thresholded at $P_{\text{corrected}} < 0.05$.

RESULTS

An initial analysis was carried out on resting state data prior to inclusion of simulated clusters at the induced paradigm frequency (0.048 Hz) and no significant activation or deactivation emerged ($P_{\text{corrected}} < 0.05$). At a lower threshold ($P_{\text{uncorrected}} < 0.001$) 201 voxels were present in the deactivation map with grand mean scaling, 200 with ANC, 440 with orthogonalization, and zero with proportional scaling. The masking method could not be applied to the null data since no activation was detected.

Sensitivity. Global normalization by proportional scaling decreased the sensitivity values for simulated data sets with $r/r_0 \gg 1$ (Figs. 4A and 4B). The orthogonalization method resulted in slightly higher sensitivity (1–2% relative increase) than grand mean scaling, which in turn outperformed the masking technique (1–2% relative increase). The sensitivity values obtained with ANCOVA were placed between proportional scaling and grand mean scaling values. Subjects-specific and session-specific ANCOVA gave very similar results for the two data sets were ANCs was tested.

With the masking method, we observed no improvement in sensitivity after the first iteration. Also, the sensitivity value was unaffected by the probability threshold (Fig. 4C) after the significant activation was masked out (mask_3 and mask_4). This result was obtained for all data sets included in the analysis. The correlation coefficient between the global signal and the paradigm decreased toward the resting state value (Fig. 4D). Similar variations in sensitivity were obtained in the random effects analysis.

For subjects A and B (real data) the number of activated voxels ($P_{\text{corrected}} < 0.05$, $t > 4.71$) were 24,148 (gms), 5966 (ps), 10,324 (ANC), 45,152 (ort) for subject A and 2021 (gms), 2581 (ps), 2484 (ANC), 2619 (ort) for subject B. With the masking method the use of the t map ($P_{\text{corrected}} < 0.05$) or F map ($P_{\text{uncorrected}} < 0.05$, results reported in Fig. 5) rendered very similar results. The mask based on the t map left 87.5% of the intracranial voxels for subject A and 98.8% for subject B to estimate the global signal while the mask based on the F map left only 75.5% for subject A and 86.1% for subject B. For both subjects the number of activated/deactivated voxels stabilized after three iterations: 6505 activated voxels for subject A and 2826 activated voxels for subject B. After five iterations, the correlation coefficient between the global signal and the paradigm dropped to $Z = 1.370$ for subject A and to $Z = 0.320$ for subject B.

False-positive rate. The false-positive rate varied by less than 1% across methods for all data sets ($P_{\text{corrected}} < 0.05$, Fig. 4E). However, the trend in false-positive rate variations across methods is opposite to the trend observed for sensitivity: proportional scaling had the smallest proportion of false positives (0.32% for 1.5% amplitude in larger clusters data set) with ANCOVA, masking, grand mean scaling, and orthogonalization methods being higher (0.39% for ANC, 0.40% for ANCs, 0.42% for masking, 0.46% for grand mean scaling, and 0.55% for orthogonalization method when 1.5% activation amplitude in larger clusters data was used for comparison).

Deactivation. Proportional scaling and ANCOVA induced a dramatic increase in the number of deactivated voxels: 10,951 voxels for proportional scaling (Fig. 6B) and 8031 voxels for ANC (8521 for ANCs), compared with 780 voxels for the masking method (Fig. 6C), 469 voxels for grand mean scaling (Fig. 6A), and 240 voxels for the orthogonalization technique (Fig. 6D) using the 2.5% amplitude, larger clusters data set. The effect was consistent for all data sets with $r/r_0 \gg 1$, for both fixed and random-effects analysis. For the data sets with $r/r_0 \approx 1$ a similar trend was observed, but the differences across methods were much smaller (462 voxels for proportional scaling, 601 voxels for ANC, 587 voxels for the masking method, 389 voxels

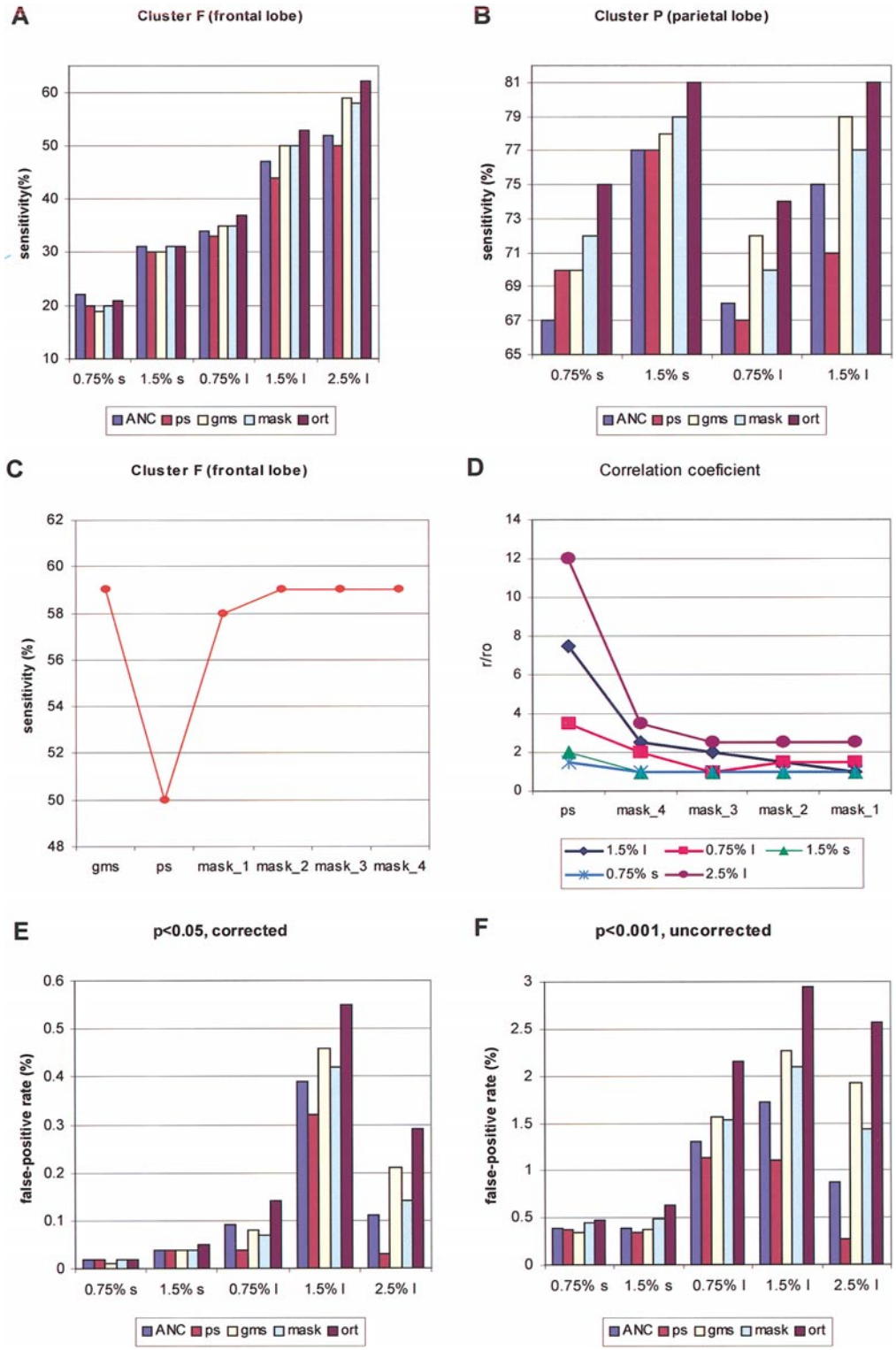

FIG. 4. (A) Sensitivity values across global normalization methods and data sets for cluster *F* ($P_{corrected} < 0.05$); s, smaller clusters; l, larger clusters. (B) Sensitivity values across global normalization methods and data sets for cluster P ($P_{corrected} < 0.05$); s, smaller clusters; l, larger clusters. These values are representative for all clusters except F. For the 2.5% amplitude, larger clusters data set an individual sensitivity value for the P cluster could not be calculated since O1, O2, P, C, and M clusters were overlapped. (C) The effect of the probability threshold on the sensitivity values with the masking method for 2.5% amplitude, larger cluster data set, cluster F (frontal). $P_{uncorrected} < 0.1$ (mask_1); $P_{uncorrected} < 0.01$ (mask_2); $P_{uncorrected} < 0.001$ (mask_3); $P_{corrected} < 0.05$ (mask_4). (D) The reduction of correlation coefficient between the global signal and the paradigm (r) toward the resting state value (r_0) with the masking method. (E) False-positive rate values across data sets for $P_{corrected} < 0.05$; s, smaller clusters; l, larger clusters. (F) False-positive rate values across data sets for $P_{uncorrected} < 0.001$; s, smaller clusters; l, larger clusters.

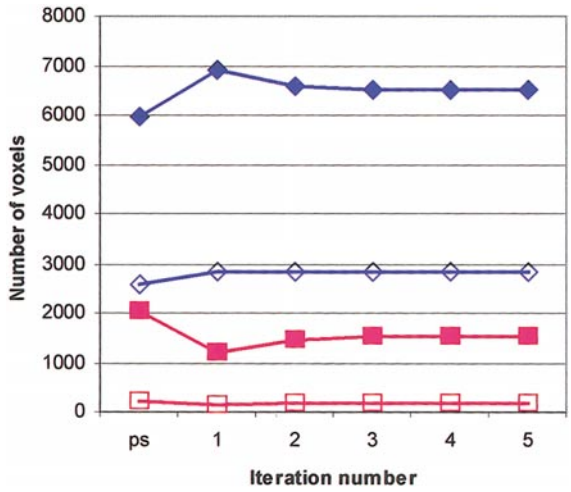

FIG. 5. The effects of five successive applications of the masking method for subject A (filled symbols) and for subject B (open symbols). ◆ represents the activated voxels and ■ represents the deactivated voxels. The mask was defined by the *F* map of the effects of interest ($P_{\text{uncorrected}} < 0.05$); ps, proportional scaling.

for grand mean scaling, and 291 voxels for the orthogonalization technique, when the 0.75% amplitude, smaller clusters data set was used for comparison). At $P_{\text{uncorrected}} < 0.001$, 167 voxels were expected to exceed the threshold in the deactivation map. With a more stringent threshold ($P_{\text{corrected}} < 0.05$) no deactivation was present with grand mean scaling, masking, and orthogonalization, while 120 voxels were reported with proportional scaling and 212 voxels with ANC (227 voxels with ANCs).

In the real data, for subject A no voxels were reported in the deactivation map ($P_{\text{corr}} < 0.05$) with grand mean scaling and orthogonalization, only 9 voxels with ANC while 2036 voxels were reported with proportional scaling and 1522 voxels with the masking method after three iterations. Subject B had 88 deactivated voxels after grand mean scaling, 222 voxels after proportional scaling, 169 voxels after ANC, 246 voxels after orthogonalization and 169 voxels after masking (after five iterations).

DISCUSSION

We used resting-state fMRI data with added simulated activation in order to assess the performance of different global normalization methods using objective measurements of sensitivity and false-positive rate. As in other studies based on simulated data (Skudlarski *et al.*, 1999), we are unable to make any inference about the absolute sensitivity values or absolute significance

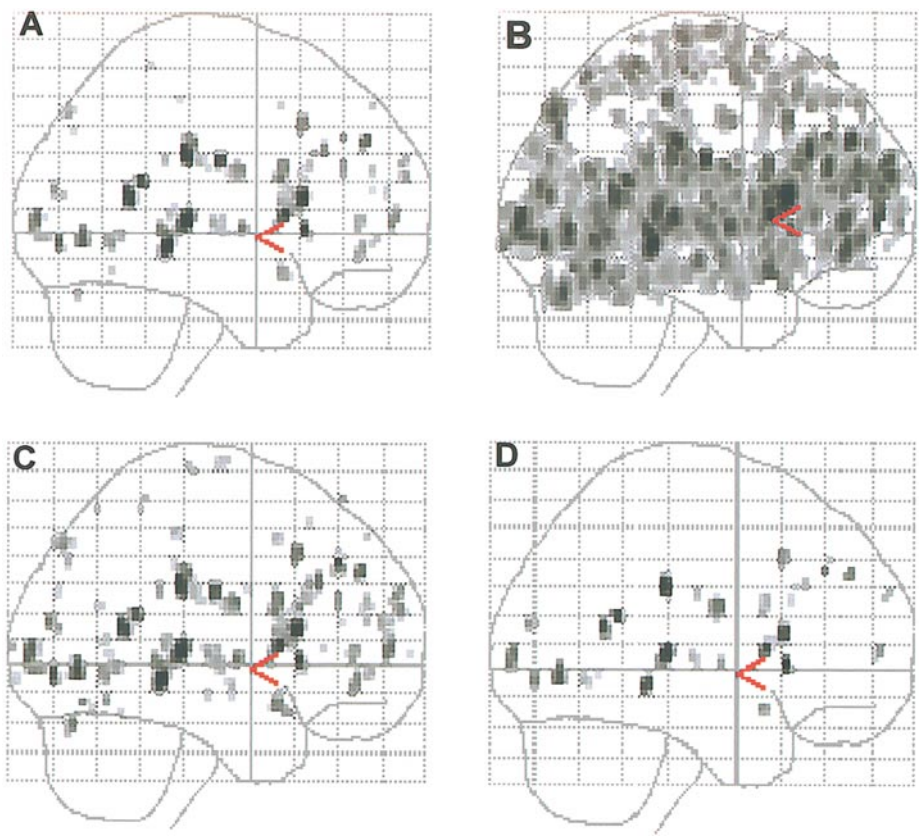

FIG. 6. Sagittal view of the deactivation *t* maps ($P_{\text{uncorrected}} < 0.001$). (A) grand mean scaling; (B) proportional scaling; (C) masking method (mask_4); (D) orthogonalization method.

of the statistical results, since they are expected to be somewhat higher than real activations, whose properties are unknown.

This simulation directly models $A(t)$ (see Eq. (3)). By the way we measured the sensitivity both $A_a(t)$ (via the voxels included in the clusters of interest) and $A_n(t)$ (via the weakly activated voxels at the edge of the clusters not included as activated) are represented. The effect of $A_n(t)$ may explain why the correlation coefficient did not decrease to the resting state value with the masking method for the 2.5% activation in the larger clusters data set (Fig. 4D). By including real fMRI noise as the background for the activation we modeled, at least in part, $G_v(t) + B(t)$ (the correlation was different from zero in the resting state). Since in practice it is not possible to estimate the separate contributions to $G(t)$, which are expected to be task-dependent, it is hard to appreciate how well resting state data reflects the noise variance in real activated data. However, the presence of processes generating $B(t)$ in real data is substantiated by the residual correlation for subjects A and B after five successive applications of the masking method. We assume that $G_v(t) + B(t)$ was underestimated in this simulation as Skudlarski *et al.* (1999) suggested that the total variance in resting fMRI data is smaller than in data acquired during a task performance.

The size of correlation between the global signal and the experimental paradigm induced by the simulated signal was small compared to the values reported in the literature (Aguirre *et al.*, 1998; Desjardins *et al.*, 2001). However, the correlation was consistent with the range of values we have observed in our real data (Fig. 1). We advocate reporting the correlation value in future studies, to more completely describe the global normalization problem in fMRI data sets.

Both ANCOVA and proportional scaling provided lower sensitivity values and an increased number of deactivated voxels as compared to the other three methods. Although previous studies reported similar results for proportional scaling and ANCOVA (Aguirre *et al.*, 1998 for fMRI; Andersson, 1997 for PET) we found that the ANCOVA method produced slightly higher sensitivity values than proportional scaling associated with slightly higher false-positive rate values. The difference between the two methods is exacerbated for data sets with $r/r_0 \gg 1$. The results for subjects A and B confirmed this trend. Session-specific and subject-specific ANCOVA performed comparably. Lower false-positive rate values support the use of proportional scaling as opposed to ANCOVA. It should be noted that the global signal estimated via masking and orthogonalization methods can also be used with ANCOVA rather than with proportional scaling.

For the simulated data sets where the correlation values were close to the resting state ($r/r_0 \approx 1$), all global normalization methods performed equivalently. However, as the correlation between the global signal and the experimental paradigm increases, the proportional scaling normalization results deteriorate (Figs. 4A and 4B). Perhaps the most problematic effect is the induced deactivation (Fig. 6B). For the proportional scaling method, some of the deactivation clusters were still significant at higher threshold ($P_{\text{corrected}} < 0.05$). This result demonstrates the tendency of proportional scaling to produce false deactivations when the global signal is confounded with the experimental paradigm. The deactivation results for subject A reveal the complexity of this problem for real data. Three of the methods indicated significant ($P_{\text{corrected}} < 0.05$) deactivation (proportional scaling, ANC, and masking) while the other methods reported no deactivation (grand mean scaling and orthogonalization). This suggests that, in this range of correlations, any deactivations identified only by some global normalization methods should be interpreted with caution. One cannot automatically infer that such deactivation is not real. However, in such circumstances, the deactivation should be well justified by the integrated psychological and physiological theories about that particular task.

When the activation is confined to spatially restricted areas, as implemented in this simulation study, the masking method produced a reasonable trade-off between optimal sensitivity and false-positive rate values. Moreover, the number of deactivated voxels is much smaller than for proportional scaling and similar to the other normalization methods included in this study. The masking method aims to correct for $A_a(t)$ but in fact also corrects for $G_{Va}(t) + B_a(t)$ and does not take into account $A_n(t)$. This method will be a good approximation as long as the number of activated voxels is much smaller than the number of nonactivated voxels.

For all simulated data sets we observed no improvement in sensitivity after the first masking iteration. This result is likely to be due to the fact that the simulated activation was placed in the same position in all subjects and all significant activation is captured in the first iterative step. The situation may be different for a real activation as suggested by the results for subjects A and B. For the simple two-state activation we simulated, the F map of the effects of interest is equivalent to the t map ($h(t)$ has just one column). However, this is not in general the case with a complex, multitask design where the columns of $h(t)$ can be correlated in different degrees with the global signal. In this case the correlation coefficient should be estimated for each column and the correlations should be individually corrected for. Andersson (1997) proposed the use of the F map with $F > 1.96$ ($P_{\text{uncorrected}} < 0.05$) since for this threshold at least 95% of the nonactivated voxels were still used to estimate the adjusted global signal. For most fMRI data, this F threshold is probably too low. From our experience with real fMRI acti-

vations the number of voxels with $F > 1.96$ in the effects of interest map is too big for this map to define the mask. Since the F map captures all the variance in the data explained by a particular model (design matrix), it is expected that the number of voxels exceeding a certain threshold in this map will increase with the complexity of the model, provided that the model accounts for extra variance. The masking method is appropriate as long as the number of voxels used to estimate the global signal is large enough to determine *global* effects. Our results show that once the real activation is masked out (mask_3 and mask_4), the sensitivity values cannot be further improved by decreasing the threshold value (Fig. 4C). Therefore for a multitask design when the contrasts of interest can be correlated in different degrees with the global signal, two solutions can be adopted to implement the masking method: estimation of the global signal based on the t map for each contrast or the use of the F map of the effects of interest (maybe thresholded at a higher level). These solutions were tested for subjects A and B, both of whom performed a multitask paradigm, and produced similar results. The masking method had a smaller impact on the number of activated/deactivated voxels (Fig. 5) than for the PET data analyzed by Andersson (1997). This may be due to the intrinsic differences between PET and fMRI data and/or to the small correlation for the two subjects we reported here.

The very simple grand mean scaling normalization produced better results than we expected. This may be because the intrasession variance in the global signal for all subjects was much smaller than the intersession variance (2 to 30 times smaller across subjects), and also because of the short sessions (112 s). We anticipate that short time fluctuations with large amplitudes in the magnetic field strength or physiological parameters could negatively affect the grand mean scaling results in other contexts. We note that with the implementation of GLM in SPM99 where session effects are automatically included as covariates of no interest ($c(t)$) in the design matrix, grand mean scaling is practically equivalent to no global normalization when no filtering is employed. However, a high/low-pass filter included in the design matrix will account for low/high-frequency components of the global signal intrasession variance.

The differences in sensitivity between masking and grand mean scaling, although small for the simulated data, can be exacerbated in real data as the results for subject A demonstrate. The simulated data also indicated a slightly bigger false-positive rate for grand mean scaling compared with masking. Taking into account these observations, in practical situations, grand mean scaling can be used in studies where there are hypotheses about the activated areas, providing that the scan-to-scan variance in the global signal is not bigger than session-to-session variance. However, for studies where new tasks and/or brain mechanisms are explored, the masking method represents a safer alternative.

Although the orthogonalization method produced the highest signal recovery (sensitivity), there are potential problems with the assumptions of this method. The mechanisms generating fluctuations in the signal intensity for a large number of voxels (global effects) are not well understood. It is acknowledged that the observed correlation between the experimental paradigm and the estimate of the global effects (global signal) is not solely generated by the task induced activity (Aguirre *et al.,* 1998; Desjardins *et al.,* 2001). The contributions of the underlying physiological processes to the correlation values ($B(t)$) may be equivalent or even greater than those induced by the activation signal ($A(t)$). In such circumstances, the use of the orthogonalization method along with an increase in the probability of detecting weakly activated voxels (increase in sensitivity) can also artificially increase the chance of nonactivated voxels exceeding the level of statistical significance (increase in false-positive rate, i.e., decrease in specificity). In our simulation these kinds of processes may have been underestimated. However, the increase we observed in the false-positive rate (Figs. 4E and 4F) suggests that the orthogonalization method may overcorrect the correlation between the global signal and the paradigm even under our simulation conditions. In the above context, overcorrecting refers to the observation that while the masking method tends to reduce the correlation value to the baseline value (which is bigger than zero), the orthogonalization method reduces the correlation value to zero. In this case, the only potential source of correlation other than the activation signal itself resides in the spatio-temporal structure of the fMRI noise that can be, by chance, correlated with any experimentally imposed paradigm. In order to measure this effect, we created 1000 random permutations of the paradigm and measured for each occurrence the correlation with the global signal across all data sets. The ratio of the mean correlation values across all permutations to the correlation of the resting state data is reported in Table 1. By bounding to zero the correlation coefficient between the global signal and all the nonconstant columns of the design matrix, the orthogonalization method corrects for this artifactual correlation as well, while the masking method does not (Fig. 4D). This may explain the increase in sensitivity associated with the increase in false-positive rate observed with this method (Figs. 4A, 4B, 4E, 4F). It may be argued that appropriate spatial filtering and/or the use of extent thresholds can reduce the false-positive rate. However, the differences in false-positive rate values across methods may be exacerbated in real fMRI data sets. This observation is supported by the increase in relative variations of false-positive rate at a lower proba-

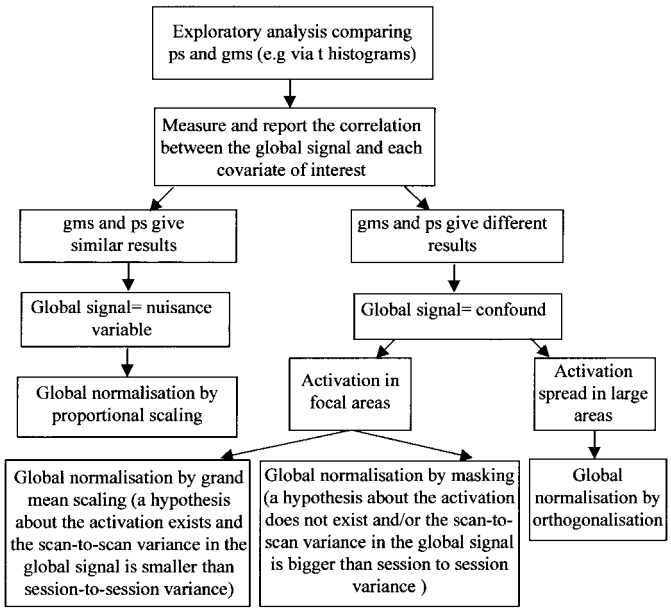

FIG. 7. Practical guide to global normalization.

bility threshold ($P_{\text{uncorrected}} < 0.001$, Fig. 4F) and also by the huge increase in the number of activated voxels with the orthogonalization method for subject A.

The masking method performs better from this perspective. When the estimate of the global signal is still based on a large number of voxels (at least 90% in these simulated data), all the underlying physiological processes are included for these voxels. However, when the task-induced activity is largely spread across the brain as in the studies described in Desjardins *et al.* (2001), the masking method cannot be used. In such a case the contribution of the activation signal to the correlation value is far bigger than any other contributions ($A(t) \gg B(t)$) and the orthogonalization method is a more appropriate choice.

Based on the results of this simulation study we offer a practical guide for global normalization of fMRI data (Fig. 7).

The results for the application of these five global normalization methods to subjects A and B (real data) support the main finding with the simulated data. Subject A falls in $r/r_0 \gg 1$ category while subject B behaves more like $r/r_0 \approx 0$ data sets.

ACKNOWLEDGMENTS

We express our gratitude to the two anonymous reviewers for their valuable suggestions that significantly improved the manuscript. We also thank Dr. Mark Jenkinson for very useful discussions and Dr. Kent A. Kiehl for kindly providing the software for the orthogonalization method.

REFERENCES

Aguirre, G. K., Zarahn, E., and D'Esposito, M. 1998. The inferential impact of global covariates in functional neuroimaging analyses. *NeuroImage* **8:** 302–306.

Andersson, J. L. 1997. How to estimate global activity independent of changes in local activity. *NeuroImage* **6:** 237–244.

Andersson, J. L., Ashburner, J., and Friston, K. 2001. A global estimator unbiased by local changes. *NeuroImage* **13:** 1193–1206.

Chapman, H., Gavrilescu, M., Wang, H., Kean, M., Egan, G., and Castiello, U. 2002. Posterior parietal cortex control of reach-to-grasp movement in humans. *Eur. J. Neurosci.* **15:** 2037–2042.

Desjardins, A. E., Kiehl, K. A., and Liddle, P. F. 2001. Removal of confounding effects of global signal in functional MRI analyses. *NeuroImage* **13:** 751–758.

Fox, P. T., and Raichle, M. E. 1984. Stimulus rate dependence of regional cerebral blood flow in human striate cortex, demonstrated with positron emission tomography. *J. Neurophysiol.* **51:** 1109–1121.

Freire, L., and Mangin, J.-F. 2001. Motion Correction Algorithms may create spurious activations in the absence of subject motion. *NeuroImage* **13:** 709–722

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

Holmes, A. P., and Friston, K. J. 1997. Statistical models and experimental design (page 17). SPM97 course notes. http://www.fil.ion.ucl.ac.uk/spm/course/notes.html#notes97: 1:47

Petersson, K. M., Nichols, T. E., Poline, J. B., and Holmes, A. P. 1999. Statistical limitations in functional neuroimaging. I. Non-inferential methods and statistical models. *Phil. Trans. R. Soc. London B* **354:** 1239–1260.

Skudlarski, P., Constable, R. T., and Gore, J. C. 1999. ROC analysis of statistical methods used in functional MRI: Individual subjects. *NeuroImage* **9:** 311–329.