Available online at www.sciencedirect.com
ScienceDirect
Computerized Medical Imaging and Graphics 31 (2007) 436–446
Computerized Medical Imaging and Graphics
www.elsevier.com/locate/compmedimag

Comparison of fMRI statistical software packages and strategies for analysis of images containing random and stimulus-correlated motion

Victoria L. Morgan [a,*], Benoit M. Dawant [a,b], Yong Li [b], David R. Pickens [a]

[a] *Vanderbilt University Institute for Imaging Science, Department of Radiology and Radiological Sciences, Vanderbilt University, Nashville, TN, USA*
[b] *Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, USA*

Received 14 September 2006; accepted 11 April 2007

Abstract

The objectives of this study were to use computer-generated phantoms containing real subject motion to: (1) compare the sensitivity of four commonly used fMRI software packages and (2) compare the sensitivity of three statistical analysis strategies with respect to motion correction. The results suggest that all four packages perform similarly in fMRI statistical analysis with SPM2 having slightly higher sensitivity. The most sensitive analysis technique was to perform motion correction and include the realignment parameters as regressors in the general linear model. This approach applies to all four packages examined and can be most beneficial when stimulus-correlated motion is present.

Keywords: Functional MRI; Data analysis; Imaging phantoms; ROC analysis; Image processing

1. Introduction

The objective of most functional MRI (fMRI) experiments is to determine the degree to which voxels in the brain respond to a given stimulus. The methods used to accomplish this task are difficult to evaluate and compare because there is no absolute ground truth for the results. Some comparisons between fMRI software packages have been limited to qualitative descriptions of documentation, functions, and ease of use [1]. Others have used multiple repetitions of the same experiment to increase reliability in the results [2–4]. Many others have found the use of various simulations or computer-generated phantoms valuable in examining fMRI analysis techniques. In these simulations, periods of signal increase and noise were added to null data sets to evaluate imaging and post-processing [5–7] and voxel-wise thresholding methods [8,9]. In all of these phantoms, motion was excluded or ignored.

Ardekani et al. added random rigid-body motion and noise to a data set without activation to compare the performance of four software packages based on realignment error [10]. Others have added regions of activation to the phantom with random motion [11] or real subject rigid body motion [12] and noise to evaluate the effects of motion correction on activation detection. In a similar study, we added real subject rigid and non-rigid body motion and varying levels of activation and noise to a null dataset in order to compare commonly used fMRI software packages based on the effect of their motion correction algorithms on detection of activations [13]. These three studies illustrated that the most accurate motion correction algorithms were able to increase specificity through accurate realignment while maintaining sensitivity through effective reslicing techniques. We also looked at the effect of motion correction algorithms on phantom data with no motion added and found no loss of sensitivity with motion correction.

More recent modifications to the phantom [14] employed updated registration methods for more accurate simulation of real subject motion. Also, a denoising approach was used to improve the signal-to-noise ratios of the base images from which the phantoms were constructed. A single-template method was used to create all the phantoms as a single subject, resulting in uniform creation of simulated activation regions. Finally, a collection of phantoms was generated with varying levels and types of real subject random and stimulus-correlated motion. These improvements have produced phantoms that can be used for more accurate evaluation of fMRI analysis of more realistic data with respect to activation sensitivity than previous phantoms.

* Corresponding author at: AA1105 MCN, Department of Radiology, Vanderbilt University, 1161 21st Avenue South, Nashville, TN 37232-2310, USA. Tel.: +1 615 343 5720; fax: +1 615 322 0734.
E-mail address: victoria.morgan@vanderbilt.edu (V.L. Morgan).

0895-6111/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compmedimag.2007.04.002

Evaluation of the utilization of motion correction algorithms and their effects on activation are important in the comparison of fMRI software packages because head motion is a primary source of error in processing of fMRI studies [15,16]. One main cause of error is that the same voxel location imaged in two different scans may contain MRI image information from two different locations within the brain. Also, signal variations occur when the head is in different positions within the magnetic field of the scanner in different scans. Finally, motion can influence signal intensity because the signal intensity is a function of the history of the position of the voxel in the magnetic field (spin history), especially in voxels whose relaxation time, T1, is much larger than the repetition time [17]. The rigid and non-rigid body motion taken from a real subject included in our current phantom incorporates the first error and some of the effects of the second, but does not include the third type of motion error.

Another concern in correcting for motion in fMRI data is the presence of stimulus-correlated motion. Stimulus-correlated motion can increase the number of false-positive activations when signal changes at anatomic boundaries appear to be due to the stimulus, when in reality they are due to motion [18,19]. After datasets containing stimulus-correlated motion are realigned, sensitivity to both false-positive and true-positive activations can be significantly reduced.

Motion correction may be further utilized for detection of activation by including the signal changes due to motion as confounding effects of no interest in the general linear model (GLM) [20]. In an investigation to determine the optimum design for fMRI studies involving overt speech, Birn et al. simulated datasets including signal increases due to activation and/or signal increases due to motion in block and event-related paradigms [21]. The data were analyzed using three different strategies: no correction for motion, ignoring all data at time points acquired during motion, and incorporating a model of the motion-induced signal as a regressor in the GLM. Their results showed that when activation related changes occur in the same voxels with motion related changes, modeling the motion as an additional regressor can improve the detection of the activation. In the same study, *in vivo* data were also analyzed using the same three strategies with similar results. One interesting note is that in all three strategies, datasets were not corrected for motion in the traditional sense of co-registration. In practice, most data are co-registered and resliced before statistical analysis. Also, the stimulus timing was used as the added motion regressor. This approach assumes that the subject moved in the same way in response to all of the stimuli. It is more common to use the actual motion translations and/or rotations as added regressors in the GLM. Our computer-generated phantom can be used to further evaluate the use of adding motion regressors to the GLM when the more common practice of motion correction (referring to co-registration and reslicing) is also employed.

In another study, the effects of task-correlated motion on effective connectivity were examined using three different strategies [22]. Specifically the investigators compared motion correction with SPM2 (http://www.fil.ion.ucl.ac.uk/SPM/SPM2.html), motion correction with SPM2 and including the rigid body motion parameters in the statistical model, and the FLIRT motion correction (http://www.fmrib.ox.ac.uk/fsl) [23] followed by an independent component analysis (ICA) to identify and remove the motion-related components. When implementing each of these preprocessing strategies on twelve healthy volunteers performing a word generation language task containing stimulus-correlated motion, they found no significant differences between activation maps. The effective connectivity, on the other hand, was greatly influenced by the preprocessing strategy.

The first objective of this study was to use our collection of computer-generated phantoms, which include a wide range of realistic subject motions and levels of signal change, to compare the fMRI analysis of four commonly used software packages. Specifically, quantitative comparisons were made based on activation detection (sensitivity) in 10 phantoms containing varying degrees of random and stimulus-correlated motion. The second objective of this study was to compare the sensitivity of three types of statistical analysis strategies with respect to motion correction: no motion correction, with motion correction, and with motion correction and including the motion parameters as regressors in the GLM. The completion of these two objectives will result in a framework upon which other quantitative comparisons of fMRI algorithms can be based.

2. Materials and methods

2.1. Phantom

The computer-generated phantom, we created and utilized in this work is described in detail in Pickens et al. [14] and is available via website (http://www.vuiis.vanderbilt.edu/fmriphantoms). The phantom consists of a single gradient-echo, echo-planar image volume (TE = 35 ms, TR = 2000 ms, 90° flip angle, 64 × 64 matrix, 7 mm thickness, 0 mm gap, 19 slices per volume) denoised [24] and copied 99 times to create a 100 volume fMRI acquisition. Ten regions of activation of $3 \times 3 \times 2$ ($X \times Y \times Z$) voxels were added using a block design paradigm of 10 volumes per block convolved with a hemodynamic response function derived from the sum of two gamma functions (delay of response = 6 s, delay of undershoot = 16 s, dispersion of response = 1 s, dispersion of undershoot = 1 s, ratio of response to undershoot = 6, length of kernel = 32 s) [25]. The activation levels were 0.5, 1, 2, 4, and 6% signal change in two regions each. Rigid body motion estimated from a normal volunteer using mutual information [26] and non-rigid body motion from the same dataset, measured using an adaptive bases method we have developed [27], were then added. The non-rigid body motions include signal changes from physiological sources such as cardiac and respiratory motion and are small relative to the rigid-body motion. Last, a Rician noise level [28] comparable to the original image noise was added independently at each time point using trilinear interpolation. This phantom allows evaluation of fMRI analysis algorithms as performed in this study.

It should be noted that the design of the phantoms used here was optimized for comparison of fMRI statistical algorithms with respect to activation sensitivity. However, this method pro-

hibits absolute comparisons of motion correction algorithms in two ways. First, we use a single template, so the motion applied to the template is a combination of the motion from the real subject and the coregistration between the real subject and the template. This extra step of coregistration includes non-rigid body deformations. Second, the motion measured from the real subject is also a combination of rigid and non-rigid body motion at each voxel. Because of its non-rigid body motion component, the motion applied to our phantom cannot be directly compared to the estimate of the rigid body motion determined by the different packages. A set of phantoms optimized for the purpose of comparing motion correction approaches can be created with modifications to the design used here and may be pursued in a future study.

2.2. Motion models

Four categories of real subject head motion were incorporated into the phantom and compared in this study. These categories were chosen to be realistic, but to represent standard and somewhat extreme capabilities of the fMRI analysis software packages. The criteria used here for categorization only utilize the translational motion of the center of mass because we had previous threshold measurements of these data for comparison [29]. We also used correlation of the motion of the center of mass and the simulated task block design to determine task correlated motion. The details of this classification process are given Pickens et al. [14]. However, once a subject's motion was classified, the non-rigid and rigid body translations and rotations along the x, y, and z axes of every voxel were added to the phantom. The motion in Model 1 (low, random motion) was chosen to simulate a cooperative subject with very little motion. Model 2 (high, random motion) was chosen to simulate a subject that moves during the acquisition in a random manner. Models 3 (low, correlated motion) and 4 (high, correlated motion) were intended to simulate the same type of subjects as Models 1 and 2, respectively, except that their motion is correlated to the simulated task block paradigm. These correlated motions are most likely to occur in motor mapping experiments and create unique difficulties in the fMRI statistical analysis [18,19,22]. Datasets containing the types of motion in Models 2 and 4 may or may not be discarded in a typical analysis due to excessive motion. However, we wanted to evaluate how these different packages would be able to process these types of data if discarding these subjects were impractical.

Using data from approximately 20 normal control subjects, we were able to fit 3 subjects' motion each into Model 1, Model 2, and Model 4. Only one subject's motion met the criteria for Model 3. This resulted in 10 phantoms used in the following analyses of four commonly used fMRI software packages and three statistical methods. The characteristics of each phantom are given in Table 1 along with the limits defining low, high, random and task-correlated motion.

2.3. Analysis packages

Each phantom was analyzed by each of the fMRI analysis packages in order to quantify its ability to detect activation in the presence of various types of motion. All analyses were carried out using the GLM capabilities of that package with the intent to employ the most uniform analysis across packages as possible using the default options of the given package. The default options were chosen because these are recommended by the developers of the software and are most likely those used by the majority of the users. Other options customized for specific needs can also be evaluated using the methods presented. The primary regressor used in the GLM was the boxcar paradigm interleaving 10 volumes of rest and 10 volumes of stimulus convolved with the hemodynamic response function (hrf). This paradigm was the same used to determine correlation in the stimulus-correlated motion datasets and to create the simulated activation in the phantom. No slice timing correction was implemented prior to the statistical analyses because no slice timing artifact was included in

Table 1
Motion characteristics of 10 phantoms

Phantom	Translation of center of mass (mm)			Correlation to activation time course		
	X	Y	Z	X_{cc}	Y_{cc}	Z_{cc}
Low, random motion						
1	0.12	0.34	0.13	0.05	0.09	0.09
2	0.10	0.28	0.20	0.04	0.02	0.04
3	0.14	0.24	0.15	0.08	0.12	0.09
High, random motion						
4	0.25	0.35	0.51[a]	0.06	0.00	0.02
5	0.40	0.30	0.79[a]	0.19	0.16	0.06
6	0.14	0.74	0.68[a]	0.18	0.08	0.13
Low, correlated motion						
7	0.13	0.47	0.17	0.43[a]	0.03	0.13
High, correlated motion						
8	0.90[a]	0.50	0.94[a]	0.41[a]	0.29	0.26
9	2.16[a]	0.47	0.40	0.65[a]	0.40[a]	0.05
10	1.57[a]	0.54	0.39	0.49[a]	0.07	0.43[a]

[a] Values exceeded the threshold for high motion or correlation for that measurement ($X > 0.50$ mm, $Y > 1.60$ mm, $Z > 0.40$ mm, $X_{cc} > 0.19$, $Y_{cc} > 0.19$, $Z_{cc} > 0.19$).

the phantom. Similarly, no spatial smoothing was implemented during the preprocessing because the activation in this phantom was added in a constant magnitude across all voxels in the region of activation. In future modifications of the phantom, the activation may be added in a more realistic heterogeneous fashion across the region of activation, which will require spatial smoothing. However, the method of adding activation used here made the specific levels of activation known for each voxel. Each dataset was analyzed individually, so no spatial normalization or co-registration was performed other than for motion correction. The details of each package are described below.

The second goal of this work was to compare three different types of analyses relating to motion correction on these phantoms containing various types of motion. These three types of analysis are (1) analysis with no motion correction (UNCORR), (2) analysis with motion correction (CORR), and (3) analysis with motion correction and using the motion parameters as regressors in the GLM (CORR WITH PARAMS). All other aspects of the analyses remain constant within the package. Therefore, an fMRI statistical analysis was performed on 10 phantoms using four packages with three motion correction analyses for a total of 120 trials performed in this study.

2.3.1. SPM2

The first commonly used fMRI software package evaluated in this study was SPM2 (http://www.fil.ion.ucl.ac.uk/SPM/SPM2.html), which is freely distributed, but requires Matlab (The MathWorks, Inc., Natick, MA). The 10 phantoms were originally designed in ANALYZE© format (http://www.mrc-cbu.cam.ac.uk/Imaging/Common/analyze_fmt.shtml, Mayo Clinic) making them inherently compatible with this program. The volumes were motion corrected (CORR and CORR WITH PARAMS analyses only) using the *realign* and *reslice* function for creating corrected volumes. This process realigns all the volumes to the first volume using a least squares approach and a six parameter spatial transformation. The reslicing interpolation is done using B-splines.

The GLM was created using the following inputs: interscan interval = 2 s, scans per session = 100, specify design in scans, hrf basis set with no Volterra interactions, 1 condition, vector of onsets = 11, 31, 51, 71, 91, duration = 10, no parametric modulation. For the UNCORR and CORR analyses, no user-defined regressors were included. For the CORR WITH PARAMS analyses, six user-defined regressors were added to the model using the six columns of motion information (x, y, and z translations and rotations) from the rp_*.txt file created during the motion correction process. The last regressor was automatically supplied by SPM2 as a constant.

The GLM was implemented using no global scaling, the default high-pass cutoff period of 128 s and AR(1) correction for serial temporal correlations. The SPM2 t-map was created using the contrast of corresponding to the simulated task and the constant. The SPM2 software automatically outputs this t-map in ANALYZE format. This t-map was then used as the input to the sensitivity analysis.

2.3.2. AFNI

The second commonly used fMRI analysis software package evaluated in this study was Analyses of Functional Neuroimages (AFNI) version 2.56e (http://afni.nimh.nih.gov/afni) [30], which is a freely distributed package of C programs. AFNI can be run from command lines. The commands for the analyses performed in this study are given in the Appendix A.

For the CORR and CORR WITH PARAMS analyses, each dataset was motion corrected using the default iterated linearized weighted least squares approach with Fourier interpolation and the motion parameters were saved. The GLM without the six motion regressors was implemented for the UNCORR and CORR analyses using the 3dDeconvolve function with the option num_stimts = 1. The 6 motion parameters (CORR WITH PARAMS analysis) were incorporated using the option num_stimts = 7. The stimulus regressor file consisted of the time course of the hrf created by SPM2 and was identical to the one used to create the phantom activation time course. The output was the dataset that included the t-map. The t-map was converted to ANALYZE format using the *3dAFNItoANALYZE* function for use in the sensitivity analysis program.

2.3.3. BV

The third fMRI software package evaluated in this study was Brain Voyager QX (BV) version 1.2 (http://www.brainvoyager.de). This package is commercially available from Brain Innovation B.V. (Maastricht, The Netherlands) and is written in C++ to run on all major computing platforms. The data were entered directly into BV in ANALYZE format and converted to BV format. For the UNCORR analysis, preprocessing included the default options of temporal smoothing (low pass filter), linear trend removal and high pass filter with 3 cycles in the time course. The GLM was performed as a single study analysis using the same block-paradigm regressor as in the other packages and implementing the correction for serial correlations using the remove AR(1) and refit GLM option. The .glm file was then saved using the *Overlay GLM* function.

For the CORR analysis the data were motion corrected using the default parameters with trilinear interpolation and the GLM was performed in the same way. For the CORR WITH PARAMS analysis, the motion parameters (in a .rtc file) were added as regressors in GLM. A Matlab program was written to convert the GLM results to t-maps for each .glm file for use in the sensitivity analysis.

2.3.4. FSL

The last commonly used fMRI analysis software package evaluated in this study was FSL version 3.2β (http://www.fmrib.ox.ac.uk/fsl) [31], which is freely distributed. To convert from ANALYZE© format to a format recognized by FSL, we used the FSL tool *avwmerge*, which merged the 100 individual volume .img and .hdr file pairs into a single 4D file pair. To perform the analysis we used the FMRI Expert Analysis Tool v5.4 (FEAT). The data set was entered with a total of 100 volumes, TR = 2 s and the option of the default high pass filter cutoff. The default McFLIRT [23] realignment uses the middle time volume as the template and

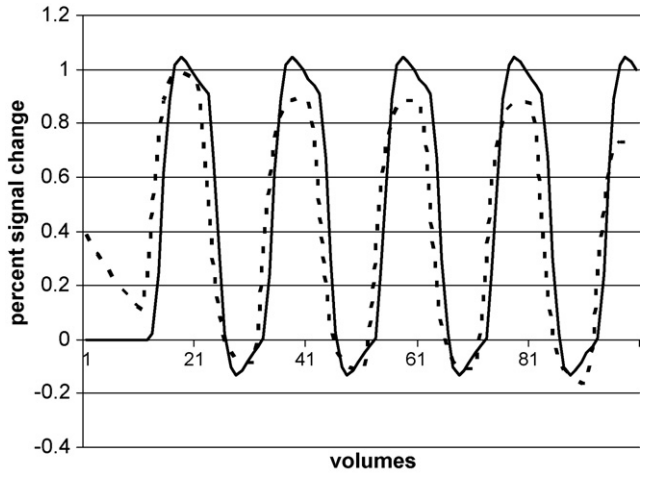

Fig. 1. Comparison of task regressors used in the GLM. The solid line represents the regressor used in SPM2, AFNI, and BV, as well as the time course used to add simulated activation in the phantoms. The dashed line represents the regressor created by using the defaults in FSL for the 10 off/ 10 on block design.

performs a coarse search and two finer searches to minimize the cost function and trilinear interpolation. This was performed for CORR and CORR WITH PARAMS analyses only. Preprocessing involved mean-based intensity normalization of all volumes by the same factor and high pass temporal filtering (Gaussian-weighted least-squares straight line fitting, with sigma = 30.0 s). These options are the default choices for this package.

For the UNCORR and CORR analyses, the design matrix was defined using the simple model setup entering 20 s blocks. This resulted in a design matrix consisting of the block design. For the CORR WITH PARAMS analysis, the design matrix was defined using the full model setup. A total of 7 regressors were added. The first was built using the following parameters: basic shape = square, skip = 0 s, off = 20 s, on = 20s, phase shift = 0 s, skip after = −1 s, convolution = gamma, phase = 0 s, S.D. = 3 s, and mean lag = 6 s. No temporal derivatives or filtering were used. These default parameters created a regressor very similar, but not identical, to the regressor used to create the phantom and used in the SPM2, AFNI and BV analyses. These two regressors are shown in Fig. 1. The other six regressors were taken from the file prefiltered_func_data_mcf.par from the mc directory created during the motion correction step. Each of the six columns of the file was used as each of the regressors without convolution, temporal derivatives or temporal filtering. Time-series statistical analysis was carried out using the FMRIB's Improved Linear Modeling (FILM) using pre-whitening with local temporal autocorrelation correction [32].

2.4. Sensitivity analysis

The *t*-maps resulting from each of the analyses described above were used to perform a receiver–operator-characteristic (ROC) type of analysis to objectively compare the sensitivity of each package for each type of motion and analysis. The standard ROC curve used in medical imaging applications is a plot of sensitivity (true-positives) versus specificity (false-positives) with each point on the curve derived from a different "cut-point" in defining a positive or negative result [5,6,33]. As the cut-point varies, the sensitivity will change inversely to the specificity. In order to simplify the large number of comparisons possible with these data, we have modified the standard ROC analysis to generate a point of information instead of a curve. This was accomplished by determining a fixed false-positive rate assumed to be most useful in the context of these data and then determining the sensitivity at this specificity value [33]. Although this does not allow comparisons at varying levels of specificity, it makes direct visual comparisons of these data at a reasonable false-positive rate possible. It also makes comparisons of the sensitivities of each region of interest with differing levels of signal change possible.

We implemented this sensitivity analysis using programs developed in IDL (ITT Visual Information Solutions, Inc., Boulder, CO) for this purpose. By ignoring all the voxels in the regions of simulated activations (ROIs) and outside the brain, a *t* threshold was determined that only 1% of the voxels exceeded. This 1% false-positive rate was selected as a reasonable threshold of error for fMRI, and therefore, a reasonable cut point for our comparisons. Since the location of each added activation region was known, it was possible to evaluate the number of true-positive activations in each ROI as the number of voxels above the determined *t* threshold. The percent true-positive rate (sensitivity) in an ROI was calculated as the number of true-positive activations divided by the known number of voxels in the ROI multiplied by 100%.

The percent true-positive rate of the two regions of simulated activation at each signal change level were averaged together in each phantom. For graphical purposes, the averaged true-positive rate at each signal change level was then averaged across the three phantoms in each motion category for Models 1, 2 and 4. Plots were made of percent true-positive activations versus percent signal change in an ROI for the selected 1% false-positive rate for each of the four packages and three analysis types. One plot was made for each of the four motion models for qualitative comparison.

To further quantify these results, another parameter was defined to describe the sensitivity of the trial over all the levels of percent signal change studied. This value, total sensitivity or TS, was computed as the sum of the percent true-positive activations across all signal levels divided by the total possible 500% true-positive rate (5 signal levels × 100% true-positives detected) for each phantom trial. This value has a range of 0 (no simulated activations detected) to 1 (all simulated activations detected).

2.5. Comparison of analysis strategies

All statistical analyses were performed using SPSS (SPSS, Inc., Chicago, IL). To determine the most sensitive analysis type for each of the four packages, a non-parametric equivalent of the one-sample repeated-measures ANOVA test, the Friedman Test, was performed on the TS values of the three analysis strategies for the particular package. In this analysis, the 10 phantoms were treated as 10 independent measures with no regard to motion cat-

egories and the TS for the three types of analysis were compared (UNCORR versus CORR versus CORR WITH PARAMS). This analysis ranks the TS values across the three types of analysis for each phantom (lower rank number = lower TS value) and then performs the statistical analysis on the ranks to test the null hypothesis that the ranks are random across phantoms. If this test determined a significant difference, then the non-parametric Wilcoxon Signed Ranks Test was used to evaluate which of the three analyses resulted in the greatest TS for each package. If the Friedman Test did not yield significance, then the analyses were not significantly different, but the analysis with the highest mean rank was considered as the most sensitive for that package.

It is hypothesized that the models with correlated motion would benefit most from the CORR WITH PARAMS analysis. To evaluate this, the difference between the TS of CORR and the TS of CORR WITH PARAMS was calculated for each phantom for each package. This yielded 16 values of this difference for correlated motion Models 3 and 4 (4 phantoms × 4 packages) and 24 values for this difference for the uncorrelated motion Models 1 and 2 (6 phantoms × 4 packages). A Mann–Whitney test, a non-parametric test using rankings, was used to determine whether the 16 correlated motion difference values were significantly higher than the 24 uncorrelated motion difference values.

2.6. Comparison of analysis packages

The Friedman Test was implemented across all 10 phantoms using the TS value of the one most sensitive analysis type (UNCORR, CORR, or CORR WITH PARAMS) of each of the four packages to determine if any one package is most sensitive across all motion types. In this analysis the ranks described the order of the TS values of the four packages for each phantom (SPM2 versus AFNI versus BV versus FSL). Again, all 10 phantoms were treated independently. As above, if the Friedman Test showed significant differences, then the Wilcoxon Signed Ranks Test was used for pair-wise comparisons to determine the most sensitive package. Otherwise, the highest average ranking package was considered the most sensitive.

3. Results

Fig. 2 shows three voxel time courses from a Model 4 (high, correlated motion) phantom without motion correction. Activation was determined by $t > 4.67$ ($p < 0.05$ corrected for multiple comparisons). The time course on top is from a voxel with simulated activation of 6% ($t = 55.7$). The middle shows a time course from a voxel with simulated activation of 2% ($t = 4.72$). On the bottom, a time course from a voxel with false-positive activation due to motion ($t = 4.97$) is given. Fig. 3a–d shows the results of the sensitivity analysis for each of the four motion models. Each plot contains a separate line for each package and for each type of analysis. Two regions were averaged at each signal change level.

3.1. Comparison of analysis strategies

The statistical rankings of each strategy for each package are given in Table 2 (higher rankings meaning higher TS values).

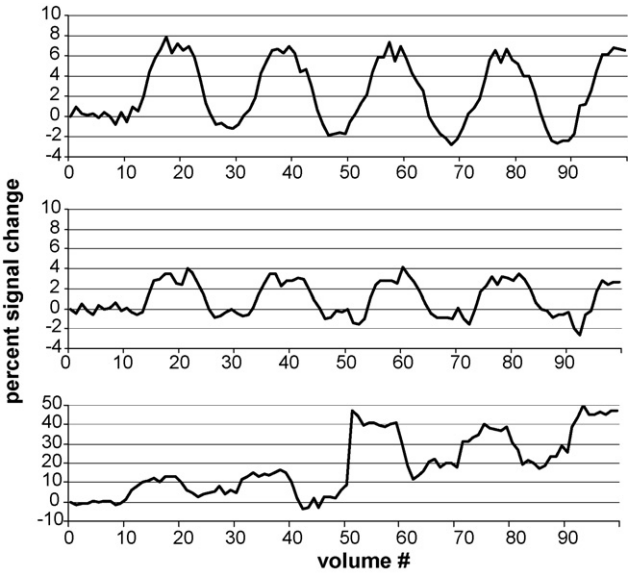

Fig. 2. Three voxel time courses from a Model 4 (high, correlated motion) phantom without motion correction. (Top) Voxel with simulated activation of 6% signal change. (Middle) Voxel with simulated activation of 4% signal change. (Bottom) Voxel with false-positive activation caused by motion.

The Friedman Test showed that for SPM2, the CORR WITH PARAMS analysis has significantly higher TS than the other two analyses. For AFNI, the results were similar with significantly higher TS with the CORR WITH PARAMS analysis. For BV, there was no significant difference between the three analysis types across all phantoms. However, the mean ranks showed the same trends as SPM2 and AFNI with CORR WITH PARAMS having the highest mean rank. The results for FSL were very similar to BV.

The Mann–Whitney test showed that the difference between the TS of the CORR WITH PARAMS and the CORR analysis was greater in the phantoms with correlated motion than in the phantoms with uncorrelated motion ($p < 0.001$, mean difference in correlated motion phantoms = 0.12 ± 0.12, mean difference in uncorrelated motion models = 0.01 ± 0.03). This test was performed across all packages.

Table 2
Comparison of analysis types across all 10 phantoms

	Friedman *P*	UNCORR mean rank	CORR mean rank	CORR WITH PARAMS mean rank
SPM2	0.002	1.40	2.00[a]	2.60[b]
AFNI	0.014	1.45	1.95[a]	2.60[b]
BV	0.393	1.70	2.10	2.20
FSL	0.368	1.75	1.95	2.30

Note: UNCORR is analysis without motion correction, CORR is with motion correction, and CORR WITH PARAMS is motion correction with inclusion of rigid body motion parameters in the statistical analysis.

[a] CORR ranks were greater than UNCORR ranks ($p < 0.05$).

[b] CORR WITH PARAMS ranks were greater than CORR ranks ($p < 0.05$).

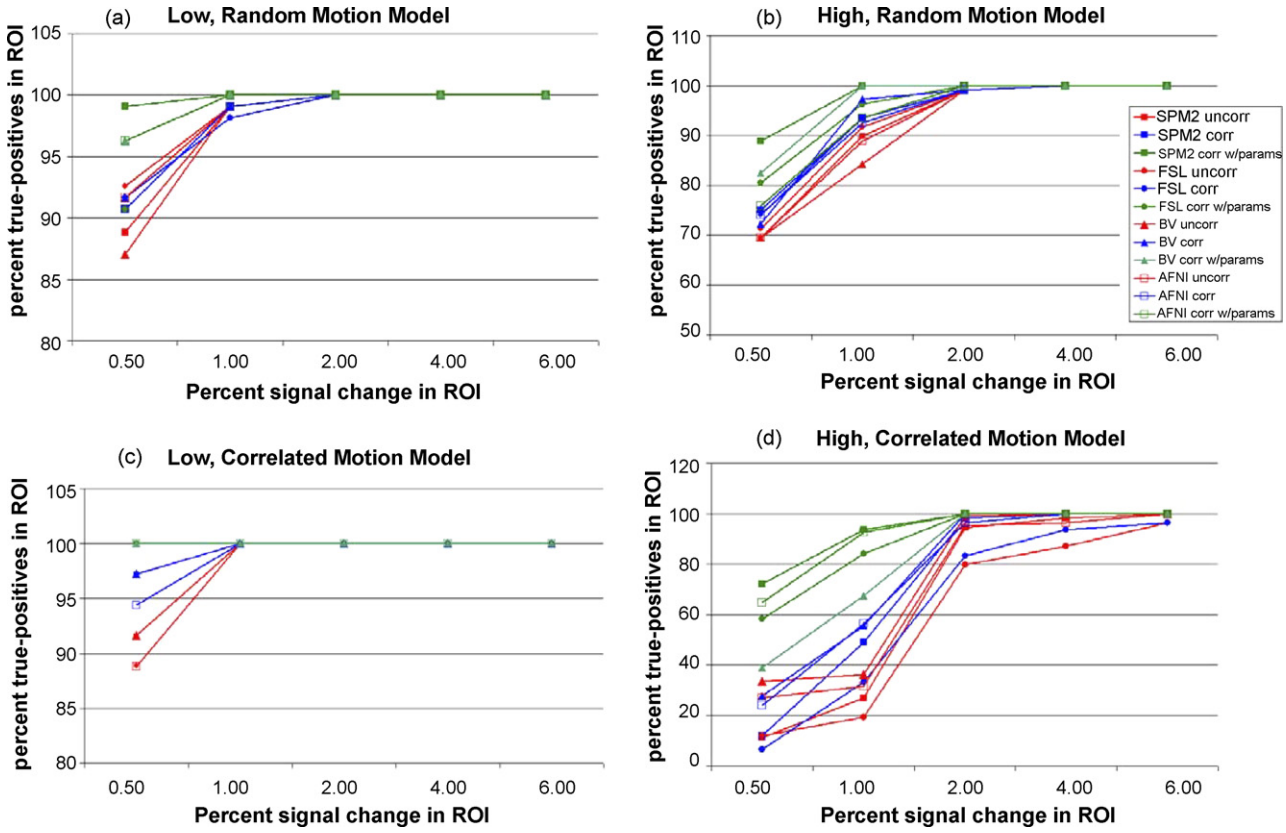

Fig. 3. Sensitivity analysis at 1% false-positive rate in (a) Model 1 (low, random motion), (b) Model 2 (high, random motion), (c) Model 3 (low, correlated motion), and (d) Model 4 (high, correlated motion). Each line represents the average of three phantoms analyzed with the specific package with the specific analysis type. At each level of percent signal change, two regions of activation are averaged. The UNCORR results are shown in red lines, the CORR results in blue lines and the CORR WITH PARAMS in green lines. The SPM2 results are shown in closed squares (■), the FSL in closed circles (●), the BV in closed triangles (▲) and the AFNI in open squares (□).

3.2. Comparison of analysis packages

From these results the CORR WITH PARAMS analysis was chosen as the most sensitive statistical strategy for each package and used in the analysis to compare the four packages. The Friedman Test showed that the four packages were significantly different in their TS across all 10 phantoms (Friedman: $p = 0.007$, SPM2 mean rank = 3.50, AFNI mean rank = 2.30, BV mean rank = 2.15, FSL mean rank = 2.05). The Wilcoxon Signed Ranks test showed that SPM2 was more sensitive than AFNI ($p = 0.028$), BV ($p = 0.018$) and FSL ($p = 0.018$).

4. Discussion

In this study we compared four commonly used fMRI software packages and three types of fMRI analyses relating to motion correction in single-session studies. Our computer-generated phantom with rigid and non-rigid body, low and high, random and stimulus-correlated motion from real subjects allowed extensive evaluation of the motion related statistical capabilities of each package.

In general, at 0.5% and 1% signal change, the results of the trials varied greatly across phantoms, packages and analyses. In Model 1 (low, random motion), the true-positive rates were in the range of approximately 87% and above, depending on the analysis (Fig. 3a). The lowest true-positive rates were seen in Model 4 (high, correlated motion) with rates as low as 6% (Fig. 3d). At approximately 2% signal change and greater, most trials yielded 80% true-positives in the ROI at the 1% false-positive rate. Fig. 3a–d shows that the general trend was that the UNCORR analysis yielded the least sensitive results (red lines), while the CORR WITH PARAMS yielded the most sensitive results (green lines) across signal change levels.

The statistics showed that for SPM2 and AFNI, the CORR WITH PARAMS analysis was most sensitive across phantoms (all types of motion). For BV and FSL, the analyses were not significantly different ($p > 0.05$), but the trend showed CORR WITH PARAMS as the most sensitive technique. These results are consistent with the belief that when using the GLM, including more regressors to describe your data is beneficial [15]. When comparing the four packages using the CORR WITH PARAMS analysis results for each phantom, the statistics showed that there was a significant difference between them and that SPM2 was significantly more sensitive than the other packages across the 10 phantoms.

These statistics are based on all the phantoms. We were not able to produce enough phantom variations in each motion category to analyze each of these separately. However, Fig. 3a–d provides some insight into the differences between motion types.

When using the GLM it is assumed that the variance of the dataset can be divided into two orthogonal or independent sources: those due to motion and those due to signal changes of interest [19]. When the motion is correlated with the stimulus, these two are not independent. Bullmore et al. [19] suggests that a way to examine the effect of non-orthogonal regressors is to compare the analyses CORR and CORR WITH PARAMS. We found that the change in sensitivity using the CORR WITH PARAMS rather than the CORR analysis is greater in the phantoms with stimulus-correlated motion than random motion. This was found across all four packages. It might be expected that sensitivity would be decreased due to decreasing the power of the experimental effect when using the motion parameters as regressors. However, our results show increases in TS at the set false-positive rates using the CORR WITH PARAMS analysis.

It should be noted that one primary difference between the four packages studied is in the way each deals with the temporal autocorrelation of the noise of the data [32]. In general, the versions of BV, FSL and SPM2 used in this study analyze the time course of each voxel twice. First, the GLM is fitted without considering the serial correlations. Then, the residuals are analyzed by computing an autocorrelation. The serial correlations are then removed (prewhitening) and the GLM is fit again. The differences in these packages lie in the estimation of the autocorrelation coefficients. FSL uses a local prewhitening approach involving Tukey tapering, dividing the data into overlapping subsets that are Fourier transformed [32]. SPM2 uses restricted maximum likelihood estimates of the variance components [34]. BV uses pseudogeneralized least squares to estimate the coefficients [35]. AFNI does not include any correction for the temporal autocorrelations in their 3dDeconvolve function. The sources of these autocorrelations include low frequency scanner drifts [16,36] and cardiac and respiratory susceptibility changes [37,38]. These characteristics are not included in our current phantom; therefore, the results of this study may have been different had these sources been present.

However, motion is a possible source of temporal autocorrelations which is included in the phantom [15,16]. This effect is illustrated in Fig. 4a and d, which shows the average signal of the brain through time without simulated activation for a phantom from Model 2 (high, random motion) and Model 4 (high, correlated motion), respectively. Fig. 4b and e shows the frequency spectrum of the time courses. The autocorrelations are shown in Fig. 4c and f. Ideal autocorrelations of white noise would be an impulse at zero lag and zero at all other lag times. The autocorrelations resulting from the data with random motion (Fig. 4c) fall mostly within the horizontal lines delineating the 95% confidence limits, indicating near random data. The autocorrelations of the data with stimulus-correlated motion (Fig. 4f) fall outside the 95% confidence limits indicating a higher degree of autocorrelation due to motion. Therefore, this implies that there are minor temporal correlations present in our stimulus-correlated motion phantoms due to the periodicity of the motion. We looked at this issue by analyzing all 10 phantoms with and without the correction for temporal correlations implemented in SPM2 using the CORR WITH PARAMS analysis. With the temporal correlations, TS increased slightly (approximately 1.5%) in only 2 phantoms, both of which were in Model 4 (high, correlated motion). No others were changed. Additional sources of temporal correlations will need to be incorporated into the phantoms to adequately compare the packages with respect to this issue.

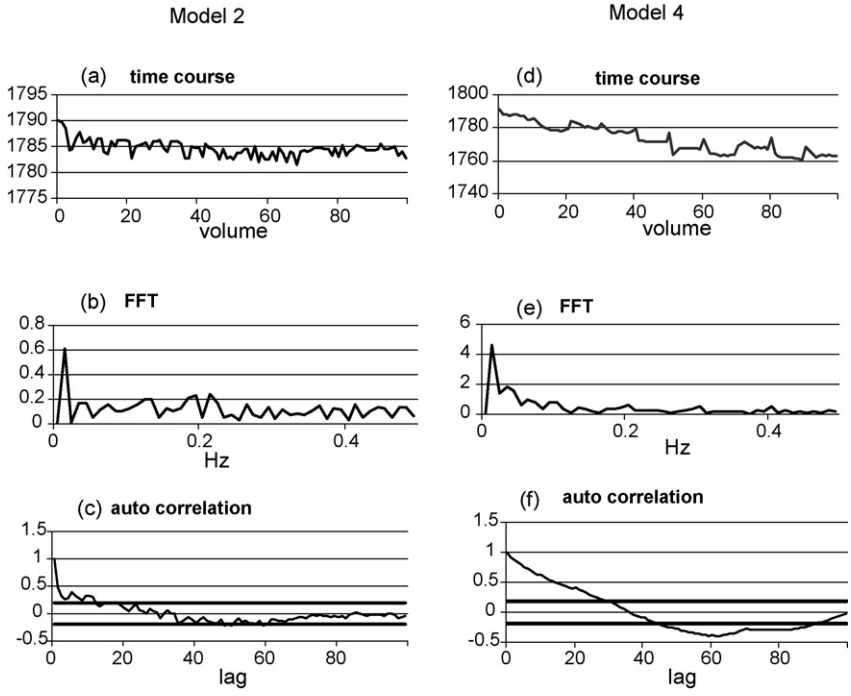

Fig. 4. Analysis of temporal autocorrelations of two phantoms without simulated activation. (a–c) Phantom from Model 2 (high, random motion). (d–f) Phantom from Model 4 (high, correlated motion). (a and d) time course of average signal over whole brain. (b and e) FFT of time course. (c and f) Autocorrelation of time course as a function of lag. The horizontal thick lines represent the 95% confidence limits.

There are several advantages of the computer-generated phantom used in this study over *in vivo* subject data for the type of comparisons performed here. Most importantly, the computer-generated phantom allows for measurement of true-positive and false-positive activations for direct measurement of activation sensitivity for varying levels of activation and noise. The phantom also contains head motion taken from *in vivo* studies, so that specific types of motion that are of interest can be utilized (i.e. stimulus-correlated motion).

However, the disadvantage of these phantoms are that they do not include all of the characteristics of data that would be found in analyzing *in vivo* subject data including varied hemodynamic responses [25], slice timing artifact [39], as well as susceptibility artifacts due to motion and physiological noise [15,17,37,38] already mentioned. Therefore, the current generation of phantom cannot be used to evaluate components of the software packages that relate to these issues. Also, there is the possibility of systematic interpolation errors in adding the motion to the phantoms. We did not examine this issue or any other types of interpolation in this study. Although we did not specifically add a low frequency baseline drift [36], the nature of our stimulus with a rest at the beginning and a task at the end, will introduce a slight linear drift with a positive slope.

The current phantom also does not allow for multi-session analyses without some estimation and modeling of inter-session variability. One study was found that compared SPM99 and FSL v1.3 [4]. In this single subject, mutli-session study design, a single subject was scanned in 99 separate sessions performing motor, visual and cognitive tasks. Inter-session variability was estimated as the difference between fixed effects variability (within-session) and simple mixed effects variability (within-session and inter-session) for different combinations of SPM99 and FSL for motion correction and statistics. Their findings were that FSL induced less inter-session error than SPM99, thus implying that FSL was more efficient in performing these higher-level analyses. Both the SPM99 and FSL packages used are earlier versions than those used in this study.

In a study by Oakes et al. [12] real human subjects and phantoms containing varying levels of real subject rigid body motion were used to compare the motion correction tools of five fMRI analysis packages including the four studied here and automated image registration (AIR) [40]. They found the most accurate motion correction results in the phantoms using AFNI followed by SPM2 with AIR being the poorest. The GLM statistical results showed the highest recovery of activation after motion correction with AFNI while BV had the least. This was true for both block-design and event-related designs. In the human subjects, however, they did not find significant differences between packages in the activation results, but the results from Brain Voyager were slightly lower than the others. This study complements the present work, by quantifying the motion correction, examining event-related designs, utilizing human data, and comparing speed and usability of each package which we have not done. Our study utilizes phantoms with non-rigid body motion, incorporates task-correlated motion, compares different analysis strategies, evaluates different levels of signal intensity changes and examines the differences in each package's statistical processing of the GLM not addressed by Oakes et al. Both studies agree that there are only minimal differences between all of these packages with BV being slightly less accurate than the others in most cases.

In this paper we describe the results of a study to compare four commonly used fMRI software packages and three analyses relative to motion correction in computer-generated phantoms containing four different models of subject motion. In general, we have provided a framework for comparative analyses of various fMRI analysis techniques using activation sensitivity as a parameter of accuracy in single-session data. Our results suggest that the most sensitive analysis technique we studied is to perform motion correction and then include these realignment parameters as regressors in the general linear model. This applies to all four packages examined and can be most beneficial when stimulus-correlated motion is present. Our results also suggest that all four packages studied perform similarly in fMRI statistical analysis, however, SPM2 resulted in slightly higher sensitivities in these single session datasets. Therefore, we conclude, like Oakes et al. [12], that selecting an fMRI processing package based on strong local support and usability may be most beneficial.

5. Summary

In previous work a set of computer-generated fMRI phantoms containing simulated activation coupled with random and stimulus-correlated head motion taken from real subject datasets was created. The objective of the current study was to use these phantoms to create a framework for quantitative comparison of fMRI analysis of single subject data based on activation detection (sensitivity). To demonstrate this we performed two investigations: (1) comparison of the sensitivity of four commonly used fMRI software packages: SPM2, Brain Voyager, AFNI and FSL and (2) comparison of the sensitivity of three statistical analysis strategies with respect to motion correction: no motion correction, with motion correction, and with motion correction and including the rigid body motion parameters as regressors in the general linear model. The sensitivity was defined as the percent of true-positives at a 1% false-positive rate. The results suggest that all four packages perform similarly in fMRI statistical analysis with SPM2 having slightly higher sensitivity, and that the most sensitive analysis technique is to perform motion correction and include the realignment parameters as regressors in the general linear model. This approach applies to all four packages examined and can be most beneficial when stimulus-correlated motion is present.

Acknowledgements

Supported in part by grants R01 NS046077 and M01 RR-00095, from the National Institutes of Health.

Appendix A. AFNI command lines

- To motion correct and to output the new motion corrected dataset and the 6 time courses of motion parameters similar

to those from SPM2:

```
3dvolreg -prefix rPh1 -Fourier -verbose -1Dfile Ph1_motion.txt Ph1 + orig
```

- To perform the GLM without the 6 motion regressors:

```
3dDeconvolve -input Ph1 + orig -num_stimts 1 -stim_file 1 block10hrf.1D\
-stim_label 1 task -tout -glt 1 contrast1.txt -glt_label 1 taskvrest\
-bucket rPh1_glm
```

- To incorporate the 6 motion parameters (CORR WITH PARAMS analysis):

```
3dDeconvolve -input rPh1 + orig -num_stimts 7\
-stim_file 1 block10hrf.1D -stim_label 1 task\
-stim_file 2 Ph1_motion.txt[0] -stim_base 2\
-stim_file 3 Ph1_motion.txt[1] -stim_base 3\
-stim_file 4 Ph1_motion.txt[2] -stim_base 4\
-stim_file 5 Ph1_motion.txt[3] -stim_base 5\
-stim_file 6 Ph1_motion.txt[4] -stim_base 6\
-stim_file 7 Ph1_motion.txt[5] -stim_base 7 -tout –glt -bucket rPh1_glm.
```

- To convert *t*-map to ANALYZE format:

```
3dAFNItoANALYZE Ph1T rPh1_glm + orig[5]
```

References

[1] Gold S, Christian B, Arndt S, Zeien G, Cizadlo T, Johnson DL, et al. Functional MRI statistical software packages: a comparative analysis. Hum Brain Mapp 1998;6:73–84.

[2] Le TH, Hu X. Methods for assessing accuracy and reliability in functional MRI. NMR Biomed 1997;10:160–4.

[3] Maitra R, Roys SR, Gullapalli RP. Test-retest reliability estimation of functional MRI data. Magn Reson Med 2002;48:62–70.

[4] Smith SM, Beckmann CF, Ramnani N, Woolrich MW, Bannister PR, Jenkinson M, et al. Variability in fMRI: a re-examination of inter-session differences. Hum Brain Mapp 2005;24:248–57.

[5] Constable RT, Skudlarski P, Gore JC. An ROC approach for evaluating functional brain MR imaging and postprocessing protocols. Magn Reson Med 1995;34:57–64.

[6] Sorenson JA, Wang X. ROC methods for evaluation of fMRI techniques. Magn Reson Med 1996;36:737–44.

[7] Della-Maggiore V, Chau W, Peres-Neto PR, McIntosh AR. An empirical comparison of SPM preprocessing parameters to the analysis of fMRI data. Neuroimage 2002;17:19–28.

[8] Logan BR, Rowe DB. An evaluation of thresholding techniques in fMRI analysis. Neuroimage 2004;22:95–108.

[9] Marchini J, Presanis A. Comparing methods of analyzing fMRI statistical parametric maps. Neuroimage 2004;22:1203–13.

[10] Ardekani BA, Bachman AH, Helpern JA. A quantitative comparison of motion detection algorithms in fMRI. Magn Reson Imaging 2001;19:959–63.

[11] Jiang A, Kennedy D, Baker J, Weiskoff R, Tootell R, Woodds R, et al. Motion detection and coorection in functional MR imaging. Hum Brain Mapp 1995;3:224–35.

[12] Oakes TR, Johnstone T, Ores Walsh KS, Greischar LL, Alexander AL, Fox AS, et al. Comparison of fMRI motion correction software tools. Neuroimage 2005;28:529–43.

[13] Morgan VL, Pickens DR, Hartmann SL, Price RR. Comparison of functional MRI image realignment tools using a computer-generated phantom. Magn Reson Med 2001;46:510–4.

[14] Pickens DR, Li Y, Morgan VL, Dawant BM. Development of computer-generated phantoms for FMRI software evaluation. Magn Reson Imaging 2005;23:653–63.

[15] Friston KJ, Williams S, Howard R, Frackowiak RS, Turner R. Movement-related effects in fMRI time-series. Magn Reson Med 1996;35: 346–55.

[16] Jezzard P, Clare S. Sources of distortion in functional MRI data. Hum Brain Mapp 1999;8:80–5.

[17] Muresan L, Renken R, Roerdink JB, Duifhuis H. Automated correction of spin-history related motion artefacts in fMRI: simulated and phantom data. IEEE Trans Biomed Eng 2005;52:1450–60.

[18] Hajnal JV, Myers R, Oatridge A, Schwieso JE, Young IR, Bydder GM. Artifacts due to stimulus correlated motion in functional imaging of the brain. Magn Reson Med 1994;31:283–91.

[19] Bullmore ET, Brammer MJ, Rabe-Hesketh S, Curtis VA, Morris RG, Williams SC, et al. Methods for diagnosis and treatment of stimulus-correlated motion in generic brain activation studies using fMRI. Hum Brain Mapp 1999;7:38–48.

[20] Friston K, Holmes A, Worsley K, Poline J-B, Frith C, Frackowiak R. Statistical parametric maps in functional imaging. Hum Brain Mapp 1995;2:189–210.

[21] Birn RM, Cox RW, Bandettini PA. Experimental designs and processing strategies for fMRI studies involving overt verbal responses. Neuroimage 2004;23:1046–58.

[22] Gavrilescu M, Stuart GW, Waites A, Jackson G, Svalbe ID, Egan GF. Changes in effective connectivity models in the presence of task-correlated motion: an fMRI study. Hum Brain Mapp 2004;21:49–63.

[23] Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 2002;17:825–41.

[24] Nowak R. Wavelet-based Rician noise removal for magnetic resonance imaging. IEEE Trans Image Process 1999;8:1408–19.

[25] Handwerker DA, Ollinger JM, D'Esposito M. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. Neuroimage 2004;21:1639–51.

[26] Wells III WM, Viola P, Atsumi H, Nakajima S, Kikinis R. Multi-modal volume registration by maximization of mutual information. Med Image Anal 1996;1:35–51.

[27] Rohde GK, Aldroubi A, Dawant BM. The adaptive bases algorithm for intensity-based nonrigid image registration. IEEE Trans Med Imaging 2003;22:1470–9.

[28] Gudbjartsson H, Patz S. The Rician distribution of noisy MRI data. Magn Reson Med 1995;34:910–4.

[29] Pickens DR, Price RR, Morgan VL, Holburn MA, Parks M, Martin PD. Long-term repeatability of cerebellar fMRI correlated with an independently monitored motor task (abstract). Radiology 1998;209: 245.

[30] Cox RW, Hyde JS. Software tools for analysis and visualization of fMRI data. NMR Biomed 1997;10:171–8.

[31] Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, et al. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 2004;23(Suppl. 1):S208–19.

[32] Woolrich MW, Ripley BD, Brady M, Smith SM. Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage 2001;14:1370–86.

[33] Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology 2003;229:3–8.

[34] Friston KJ, Penny W, Phillips C, Kiebel S, Hinton G, Ashburner J. Classical Bayesian inference in neuroimaging: theory. Neuroimage 2002;16:465–83.

[35] Bullmore E, Brammer M, Williams SC, Rabe-Hesketh S, Janot N, David A, et al. Statistical methods of estimation and inference for functional MR image analysis. Magn Reson Med 1996;35:261–77.

[36] Smith AM, Lewis BK, Ruttimann UE, Ye FQ, Sinnwell TM, Yang Y, et al. Investigation of low frequency drift in fMRI signal. Neuroimage 1999;9:526–33.

[37] Raj D, Anderson AW, Gore JC. Respiratory effects in human functional magnetic resonance imaging due to bulk susceptibility changes. Phys Med Biol 2001;46:3331–40.

[38] Windischberger C, Langenberger H, Sycha T, Tschernko EM, Fuchsjager-Mayerl G, Schmetterer L, et al. On the origin of respiratory artifacts in BOLD-EPI of the human brain. Magn Reson Imaging 2002;20:575–82.

[39] Turner R, Howseman A, Rees GE, Josephs O, Friston K. Functional magnetic resonance imaging of the human brain: data acquisition and analysis. Exp Brain Res 1998;123:5–12.
[40] Woods RP, Cherry SR, Mazziotta JC. Rapid automated algorithm for aligning and reslicing PET images. J Comput Assist Tomogr 1992;16: 620–33.

Victoria L. Morgan received her Bachelor's degree in biomedical engineering from Wright State University in 1990, her Master's degree from Vanderbilt University in 1994, and her PhD from Vanderbilt University in 1996. Currently, she is an assistant professor of Radiology at Vanderbilt University performing research in computer-generated phantoms and functional MRI of the brain. Her clinical research interests include seizure localization in epilepsy and presurgical activity mapping in neurosurgical patients.

Benoit M. Dawant, PhD, received his Master's degree in electrical engineering from Catholic University of Louvain, Belgium in 1982 and his Doctorate in systems engineering from University of Houston, TX in 1987. Dr. Dawant is an associate editor of *IEEE Transactions on Biomedical Engineering*, was a guest editor for *IEEE Transactions on Information Technology in Biomedicine* and is a member of the steering committee for *IEEE Transactions on Medical Imaging*. He currently is a professor in the departments of Electrical and Computer Engineering and Radiology at Vanderbilt University investigating image registration techniques and their application in image-guided surgery.

Yong Li received his Bachelor's degree from Shanghai Jiao Tong Univerity, China, in 1997, and his MS from Tsinghua University, China in 2000. He is currently a PhD student in the Department of Electrical Engineering and Computer Science, Vanderbilt University, USA. His research interests include medical image registration, distortion correction, segmentation, and software phantom for fMRI analysis validation.

David R. Pickens received a BA from the University of the South at Sewanee, Tennessee, in 1969 majoring in biology, a BE (cum laude) in biomedical engineering, and MS and PhD in mechanical engineering from Vanderbilt University in 1971, 1978, and 1981, respectively. He is currently associate professor of Radiology and Radiological Sciences and Associate Professor of Biomedical Engineering at Vanderbilt University. He has worked in various aspects of medical imaging including functional magnetic resonance imaging (FMRI), publishing several papers and abstracts on this subject. For the last four years, he has participated in the development of software phantoms for evaluation of FMRI processing algorithms.