ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 41 (2008) 1242–1252

Evaluation and comparison of GLM- and CVA-based fMRI processing pipelines with Java-based fMRI processing pipeline evaluation system

Jing Zhang,[a,e,*] Lichen Liang,[b] Jon R. Anderson,[c] Lael Gatewood,[a] David A. Rottenberg,[c] and Stephen C. Strother[a,c,d]

[a]*Health Informatics Graduate Program, University of Minnesota, Minneapolis, MN 55455, USA*
[b]*Department of Electrical Engineering, University of Minnesota, Minneapolis, MN 55455, USA*
[c]*Department of Neurology, University of Minnesota, Minneapolis, MN 55455, USA*
[d]*Rotman Research Institute, Baycrest and Medical BioPhysics, University of Toronto, ON, Canada M6A2E1*
[e]*Mt. Sinai Medical Center, New York, NY 10029, USA*

Received 16 August 2007; revised 11 March 2008; accepted 17 March 2008
Available online 3 April 2008

Activation patterns identified by fMRI processing pipelines or fMRI software packages are usually determined by the preprocessing options, parameters, and statistical models used. Previous studies that evaluated options of GLM (general linear model)-based fMRI processing pipelines are mainly based on simulated data with receiver operating characteristics (ROC) analysis, but evaluation of such fMRI processing pipelines on real fMRI data is rare. To understand the effect of processing options on performance of GLM-based fMRI processing pipelines with real fMRI data, we investigated the impact of commonly-used fMRI preprocessing steps; optimized the associated GLM-based single-subject processing pipelines; and quantitatively compared univariate GLM (in FSL.FEAT and NPAIRS.GLM) and multivariate CVA (canonical variates analysis) (in NPAIRS.CVA)-based analytic models in single-subject analysis with a recently developed fMRI processing pipeline evaluation system based on prediction accuracy (classification accuracy) and reproducibility performance metrics. For block-design data, we found that with GLM analysis (1) slice timing correction and global intensity normalization have little consistent impact on fMRI processing pipelines, spatial smoothing and high-pass filtering or temporal detrending significantly increases pipeline performance and thus are essential for robust fMRI statistical analysis; (2) combined optimization of spatial smoothing and temporal detrending improves pipeline performance; and (3) in general, the prediction performance of multivariate CVA is higher than that of the univariate GLM, while univariate GLM is more reproducible than multivariate CVA. Because of the different bias–variance trade-offs of univariate and multivariate models, it may be necessary to consider a consensus approach to obtain more accurate activation patterns in fMRI data.

Keywords: fMRI; GLM; CVA; Prediction accuracy; Classification accuracy; Reproducibility

Introduction

Over the past one and a half decades, functional MRI (fMRI) has emerged as a powerful neuroimaging tool to study brain functions. In recent years, the potential of fMRI in diagnosing brain diseases has been identified (Sabatini et al., 2000; Bookheimer et al., 2000; Haslinger et al., 2001; Lipton et al., 2003; Machulda et al., 2003; Muller et al., 2003; Rombouts and Scheltens, 2005). As a non-invasive procedure, fMRI has become a critical step in preoperative surgical planning (Fernández et al., 2003; Stippich et al., 2007; Bookheimer, 2007; Tharin and Golby, 2007), and fMRI analysis for individual patient has been used for presurgical mapping which assists neurosurgeons and neuroradiologists in maximizing surgical outcomes while minimizing surgical risks. In addition, fMRI has become a potential tool for central nervous system drug development (FitzGerald et al., 1997; Stein, 2001; Borsook et al., 2006; Becerra and Borsook, 2006).

With the advancement of fMRI technology, various statistical models and methods have been developed to analyze fMRI data in order to meet the demands of growing fMRI applications. The statistical methods in fMRI analysis can be classified into two categories: (1) univariate statistical methods such as the univariate general linear model (GLM) (Friston et al., 1994, 1995a,b; Worsley

* Corresponding author. Neuroscience PET Laboratory, Mt. Sinai Medical Center. New York, NY 10029, USA.
E-mail address: jzhang0000@gmail.com (J. Zhang).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2008.03.034

and Friston et al., 1995) which characterizes region specific responses at each voxel based on assumptions; and (2) multivariate methods such as principal component analysis (PCA) (Andersen et al., 1999; Friston et al., 1999, 2000; Hansen et al., 1999; Laiand and Fang, 1999), canonical variates analysis (CVA) (Bullmore et al., 1996; Friston et al., 1995c; Worsley et al., 1997) and independent component analysis (ICA) (McKeown et al., 1998; Biswal and Ulmer, 1999; McKeown, 2000) which are often exploratory and data-driven, and have the potential to identify activation patterns that may reveal neural networks and functional connectivity of the brain.

A number of software tools such as Statistical Parametric Mapping (SPM) (Friston, 1996), Analysis of Functional NeuroImages (AFNI) (Cox, 1996), FMRIB Software Library (FSL) (Smith et al., 2004) have been developed and widely used for functional neuroimaging data analysis. However, these methods and software implementations often lack rigorous evaluation, and the fMRI analysis results generated by these software packages often lack careful validation. Various analysis models and/or software tools with the same functional neuroimaging dataset and similar parameter settings may identify different activation regions or spatial patterns in the brain and generate different statistical parametric images (SPIs) (Poline et al., 2006). There is little consensus on which fMRI processing pipeline, including a series of preprocessing steps and statistical analysis, or software package best detects brain activations and should be used in fMRI analysis. Consequently, the problem of lack of fMRI processing pipeline evaluation and result validation hinders the further development of optimal fMRI applications. Further, the incompatibility of various fMRI software tools has made it more difficult to compare numerous fMRI results (Fissell et al., 2003) and has become a large obstacle to collaborative efforts in fMRI studies (Skudlarski et al., 1999; Rex et al., 2003).

From clinical point of view, fMRI is not a fully established diagnostic neuroimaging method today. This is mainly due to a lack of standardization and guidelines, and the lack of licensing of important hardware and software components (Stippich, 2007). In order to solve the bottleneck problem in fMRI applications, it is crucial to explore methods to evaluate, compare, optimize and standardize heterogeneous fMRI processing pipelines. The most widely used approach to fMRI result validation and pipeline performance evaluation is the receiver operating characteristic (ROC) method. Since there is no easily measurable ground truth in real fMRI data, ROC analysis often requires simulation. Considerable work has been done to evaluate fMRI preprocessing steps and statistical methods with ROC approach on simulated data (Skudlarski et al., 1999; Gavrilescu et al., 2002; Lange et al., 1999; Lukic et al., 2002; Della-Maggiore et al., 2002; Beckmann and Smith, 2004). In particular, Skudlarski et al. (1999) found that the removal of intensity drifts by temporal detrending and high-pass filtering is beneficial to fMRI analysis, but temporal normalization of the global image intensity and low-pass filtering do not improve analytical power. Moreover, Lukic et al. (2002) and Beckmann and Smith (2004) reported evidence that multivariate data analytical approaches may outperform the widely used univariate GLM technique. However, the effectiveness of standard ROC analysis depends on how well the simulated data approximates the real data. To try to overcome simulation-dependent biases, several modified ROC methods were developed to work with real fMRI data and approximate true positive ratios (TPR) and false positive ratios (FPR) in various ways (Le and Hu, 1997; Genovese et al., 1997; Maitra et al., 2002; Nandy and Cordes, 2003; Liou et al., 2006). For example, the modified ROC curves that Nandy and Cordes, (2003) proposed depend on the proportion of active voxels for TPR and the fraction of voxels detected to be active in a separate rest-state data set for a FPR. Liou et al., using a mixed multinomial distribution approach introduced by Genovese et al. (1997), focus on the "reproducibility of a voxel defined as the degree to which the active status of the voxel, in responding to stimuli, remains the same across replicates implemented under the same conditions" (Liou et al., 2006). These approaches produce ROC curves that contain an unknown model bias characteristic of all attempts to measure reproducibility by our group and others. The fundamental problem of using approximated TPR and FPR, or reproducibility, as ground truth in ROC analysis for real fMRI data (where ground truth and model bias are generally not knowable) ultimately hinders the accuracy of all these modified ROC methods. We have attempted to overcome some of these limitations by focusing on the joint optimization of two metrics: (1) a global reproducibility measure based on comparing independent SPIs (Strother et al., 1997), and (2) a potentially unbiased predictive learning, or generalization measure, based on the accuracy of the estimated modeling parameters for predicting the experimentally defined brain states in a cross-validation framework (Stone, 1974; Larsen and Hansen, 1997).

Guided by statistical learning, the nonparametric prediction, activation, influence and reproducibility resampling (NPAIRS) approach was proposed (Strother et al., 2002) to evaluate fMRI processing pipelines on real fMRI data based on prediction accuracy and SPI reproducibility. A detailed description can be found in Strother et al. (2002) and LaConte et al. (2003) regarding the NPAIRS approach and its metrics. Briefly, in the NPAIRS approach, fMRI data is split into 2 independent halves (on subjects, sessions, or runs): the training set and the test set. For example, in this study we use a within-subject split defined by the two available runs per subject. Prediction accuracy is obtained by applying model parameters estimated in the training set (e.g., run 1) to the test set (e.g., run 2), cross-validating (in this study by swapping the definition of the training set to run 2 and the test set to run 1), and averaging the prediction accuracy estimates (in this study two per subject). Prediction accuracy (p) is measured as the average posterior probability of each fMRI volume's true class membership (i.e., predicted baseline or activation brain state) in the test set based on the training set parameters and Bayes formula (Mardia et al., 1979; Strother et al., 2002). For each independent pair of split-half data sets the resulting SPIs' reproducibility is defined as the correlation (r) between all pairs of spatially aligned voxels in the brain. In general the average, or median, of the distribution of such correlation values is obtained from the independent SPIs of many split-half resamplings of the fMRI data (e.g., in this study there is one split and one correlation value per subject). Reproducible SPIs can be obtained from arbitrary data analysis approaches on a Z-score scale (Strother et al., 2002) in the NPARIS approach.

The NPAIRS approach was implemented in the NPAIRS software package (http://neurovia.umn.edu/incweb/download_home.html), which also provides models such as GLM and PCA/CVA for statistical analysis. Utility of the NPAIRS framework has been demonstrated by a number of single-subject and group analyses in functional neuroimaging (Strother et al., 2002, 2004; Kjems et al., 2002; Shaw et al., 2003; LaConte et al., 2003; Chen et al., 2006). However, there are some limitations to the existing software implementation of the NPAIRS approach, an Interactive Data Language (IDL)-based NPAIRS package. The main limitations of the NPAIRS package are (1) lack of system interoperability, which makes it difficult to evaluate modules in other software packages, and (2) lack of a GLM prediction measure, which hinders the evaluation of GLM-

based pipelines. Hence, a Java-based pipeline evaluation system has been developed to allow the evaluation of GLM and CVA-based heterogeneous fMRI processing pipelines with the NPAIRS approach (Zhang et al., in press).

Since in the NPAIRS package a prediction measure was not available to GLM-based pipelines but only to CVA-based pipelines, the scope of our previous research using (p, r) plots was largely restricted to the evaluation of CVA-based pipelines (LaConte et al., 2003; Shaw et al., 2003; Strother et al., 2004; Chen et al., 2006; Zhang, 2005). In this study, with the Java-based fMRI processing pipeline evaluation system, we (1) evaluated the impact of a series of preprocessing steps: slice timing correction, motion correction, spatial smoothing, temporal filtering (including two forms of high-pass filtering: Gaussian-weighted running line smoother of FSL and cosine basis set detrending which is close to the SPM approach to high-pass filtering) and global intensity normalization on GLM-based single-subject processing pipelines; (2) optimized the associated GLM-based single-subject processing pipelines (3) quantitatively compared the performance of fMRI processing pipelines with univariate GLM (in FSL.FEAT and NPAIRS.GLM) and multivariate CVA (in NPAIRS.CVA) in single-subject analysis.

Methods

fMRI and MRI data

This study used a BOLD fMRI dataset, in which a block-design parametric static-force task was applied to 16 normal subjects who were scanned on a 1.5-T Siemens scanner. Two fMRI runs per scan session were acquired with an EPI BOLD sequence [TR = 3986 ms; TE = 60 ms; FA = 90°; Matrix, 64 × 64; FOV = 220 × 220 mm; number of slices, 30; number of time points, 135; voxel dimensions, 3.44 × 3.44 × 5 mm; slab thickness, 150 mm; orientation, oblique transverse (axial), 20°; shift mean, 6.4 mm (center of slice relative to magnet isocenter); imaging time per procedure, 9 min]. In each run, there were six baseline periods, which alternated with five activation periods during which a static force was applied to a force transducer held by the subject between the right thumb and forefinger with randomly assigned force levels (200 g, 400 g, 600 g, 800 g, 1000 g) monitored via a visual feedback loop. Each baseline and activation epoch lasted for 45 s and the fundamental frequency of the block design was 0.011 Hz. More details of this data set are available from LaConte et al. (2003).

Data analysis environment

We employed a Java-based fMRI processing pipeline evaluation system (Zhang et al., in press) which integrates YALE (or RapidMiner), a machine learning environment, into Fiswidgets, a fMRI pipeline environment (Fissell et al., 2003). In this environment we implemented a GLM prediction measure by applying the GLM prediction algorithm (Kjems et al., 2002) to evaluate heterogeneous fMRI processing pipelines. The preprocessing software used was Visualization and Analysis Software Tools (VAST*), an IDL-based software library developed at the VA Medical Center in Minneapolis. The preprocessing software and the IDL-based NPAIRS package were integrated into Fiswidgets through the Java wrappers that Fiswidgets provides. FMRI processing pipelines were built and run on the Fiswidgets GlobalDesktop. Some adaptations were made for modules that were not completely incorporated into Fiswidgets, e.g., FSL.FEAT (Image Analysis Group, FMRIB at Oxford). FSL.FEAT was run in batch mode with parameter files through the Unix command widget that Fiswidgets provides.

Preprocessing

Slice timing correction was performed by the FSL.slicetimer. High-pass temporal filtering was realized by applying the high-pass filter in FSL.FEAT. Temporal detrending was achieved by specifying a linear combination of cosine basis functions in the GLM design matrix and retaining the residuals and desired effects of the GLM model as the detrended data (Strother et al., 2004). Spatial smoothing was implemented through convolution with a 2D (within-slice) Gaussian kernel. Global intensity normalization was performed by dividing the intensities of each fMRI volume by its volume mean.

Motion correction was carried out with AIR.alignlinear applying a 6-parameter rigid body transformation to align each fMRI volume with the first volume of the first run in order to remove head motion. AIR.alignlinear was also used in the mean fMRI-to-structural MRI transformation (6-parameter). The intra-subject alignment from individual fMRI space to structural MRI space was derived by multiplying the fMRI motion correction transformation and the mean fMRI-to-structural MRI transformation. The fMRI volumes were then resampled to the individual MRI space by applying the derived transformation to each fMRI volume and projecting it into the subject's structural MRI space. Inter-subject alignment was performed for pipeline optimization and evaluation of analytic models. The inter-subject alignment transformation was derived by combining the intra-subject alignment transformation with a structural MRI-to-MNI152 (Montréal Neurological Institute template) transformation using a 7th order polynomial warp in AIR5.03 (Woods et al., 1998). The fMRI volumes were then aligned to the MNI template brain through inter-subject alignment.

Statistical analysis

For the processing pipelines tested, univariate GLM analysis was carried out with FSL.FEAT and NPAIRS.GLM, both applied to split-half runs within the Java-based NPAIRS pipeline evaluation framework. In FSL.FEAT, the default square waveform with default options was used to convolve with the default (single) gamma hemodynamic response function (HRF) and no temporal derivative was added. In NPAIRS.GLM, since convolution was not available in the NPAIRS package, no HRF convolution (nor temporal derivative) was applied.

For multivariate NPAIRS.CVA, the baseline-activation transition volumes were dropped to improve the model cost function as described in LaConte et al. (2003). NPAIRS.CVA is based on PCA which reduces the input data dimension, controls model complexity and avoids singular covariance matrices. In NPAIRS.CVA, fMRI data was first decomposed using PCA, and a reduced number of principle components (PCs) were then passed to CVA where the within- (W) and between-class (B) covariance matrices were constructed. We used a two-class CVA, or Fisher Linear Discriminant, with baseline and activation volumes assigned to the two separate classes (Mardia et al., 1979). The matrix $W^{-1}B$ was decomposed with a further PCA where the single PC obtained maximizes the ratio of between-class mean variance to the pooled within-class variance. The number of PCs can be optimized to control model complexity and trade-off prediction with reproducibility as outlined in LaConte et al. (2003) and Strother et al. (2002, 2004).

Table 1
Pipeline options and parameters for the preprocessing steps tested

Preprocessing steps in pipelines tested		Options and parameters of processing pipelines																	
		1			2			3			4						5		
		Slice timing correction (FSL)			Motion correction (AIR)			Spatial smoothing (VAST) (FWHM in pixels)			Temporal filtering						Global intensity normalization (VAST)		
											High-pass filter (s) (FSL)			Detrending (VAST) (≤# cosine cycles/run)					
P1	Slice timing correction (FSL)	x	F		A			0, 2			x			0, 2			x		
P2	Motion correction (AIR)	x			x	A		0, 2			x			0, 2			x		
P3	Spatial smoothing (VAST)	x			A			0	1.5, 2, 4, 6		x			0, 2			x		
P4	Temporal filtering P4.1 high-pass filter (FSL)	x			A			0, 2			0	128, 176		0			x		
	P4.2 detrending (VAST)	x			A			0, 2			x			0	1, 1.5, 2, 3		x		
P5	Global intensity normalization (VAST)	x			A			0, 2			x			0, 2			x		V

P*i* [*i*=1, 2, 3, 4 (4.1, 4.2), 5]—Pipelines in category *i* that tested the *i*th preprocessing step; FWHM—full-width-half-maximum; F—FSL.slicetimer; A—AIR.alignlinear; x—no operation performed; V—VAST.

Evaluating the impact of preprocessing steps

The impact of preprocessing steps including slice timing correction, motion correction, spatial smoothing, temporal detrending and global intensity normalization was evaluated for NPAIRS.GLM-based pipelines. Only the impact of temporal filtering was evaluated for FSL.FEAT-based pipelines.

To avoid testing preprocessing steps in isolation, a set of preprocessing options in a range of settings (Table 1) were set to evaluate each preprocessing step tested. For example, as shown in Table 1, the impact of slice timing correction was tested by turning it on and off in pipelines with motion correction, 0 and 2-pixel spatial smoothing, and cosine detrending settings of all half and full cycles up to the following cutoff cycles {1.0, 1.5, 2.0, 3.0}, where one cycle has a period of 69 s. However, to limit the total number of combinations, tested parameters were dropped from further testing when no significant effect was found.

We used a prediction accuracy (p) vs. reproducibility (r) plot to evaluate the performance of the functional neuroimaging data processing pipelines outlined above. As it is unclear whether one metric is to be preferred over another we gave an equal weighting (1:1) to p and r measures in this study to calculate the Euclidean distance between the (p, r) pair of the pipeline tested and the optimal values of (1, 1) for perfect prediction accuracy and reproducibility.

The mean distance change (delta$M$) across the 16 individually processed and tested subjects was used to measure the impact of each preprocessing step tested. This was defined as the difference of the mean distance across all subjects to perfect prediction and reproducibility (1, 1) calculated by subtracting the mean distance with the step from that without the step. It can be expressed as:

$$\begin{aligned}\text{delta}M &= \overline{D}_0 - \overline{D} \\ &= \frac{1}{N}\left\{\sum_{i=1}^{N}\sqrt{(1-p_{i0})^2+(1-r_{i0})^2} - \sum_{i=1}^{N}\sqrt{(1-p_i)^2+(1-r_i)^2}\right\}\end{aligned} \quad (1)$$

where $\overline{D}$ is the mean distance between the (p, r) performance of the pipeline tested and (1, 1); p_{i0} and r_{i0} are the prediction accuracy and reproducibility without the preprocessing step tested for the ith subject, p_i and r_i are the ones with the preprocessing step for the ith subject; and N is the total number of subjects in the dataset. Note that improved pipeline performance implies either $p_i > p_{i0}$ and/or $r_i > r_{i0}$, and that delta$M = \overline{D}_0 - \overline{D} > 0$ To compare the relative impact of the preprocessing steps tested, relative variation was further computed through dividing mean distance change (delta$M$) by its standard deviation.

Optimizing single-subject preprocessing steps

For NPAIRS.GLM and FSL.FEAT-based, single-subject pipelines the optimization of preprocessing steps based on the spatial smoothing and temporal filtering results from the evaluation of the impact of the different steps was performed on inter-subject aligned data. For pipelines with NPAIRS.GLM the parameters were: (1) spatial smoothing with in-plane Gaussian full-width-half-maximum (FWHM)=0, 1.5, 2, 4, 6 pixels multiplied by the in-plane pixel size (3.44 mm^2), and (2) temporal detrending, cosine cycle of 0 and ≤1, 1.5, 2, 3. For FEAT, the spatial smoothing options were FWHM=2, 4, 6 pixels and high-pass filtering cutoffs were 176 s (similar to 2-cosine cycles in a run) and 128 s (similar to 3-cosine cycles in a run). The impact of such optimization on GLM-based pipelines was examined with both NPAIRS performance metrics, and between-subject reproducibility (BSR).

Using NPAIRS performance metrics and BSR to assess the impact of pipeline optimization is described in Zhang (2005). Briefly, the optimized pipeline was compared with the best performing non-optimized (penultimate) pipeline in order to examine the impact of pipeline optimization. In the BSR approach, the number of activated voxels ($Z>3$) common to each pair of subjects relative to the average number of activated voxels between both subjects was measured and this procedure was repeated for all possible pairs of subjects to obtain a conjunction matrix. The BSR for all 16 subjects was measured as the average of the conjunction matrix values for all possible pairs (Shaw

et al., 2003). Based on the pipeline optimization results of the 16 subjects, an optimized BSR matrix (16×16) was formed. The non-optimized BSR matrices were calculated using the SPIs generated by the non-optimized pipelines and they were ranked by mean BSR across all subject pairs to obtain the best performing non-optimized pipeline (or the penultimate pipeline). The distribution of pairwise BSR values for the penultimate pipeline was then compared with that from the optimized BSR matrix using a Wilcoxon matched-pair rank sum test to see whether average group homogeneity improved after optimization.

Evaluating heterogeneous pipelines

In this study, the evaluation of the heterogeneous pipelines across four models (NPAIRS.GLM, NPAIRS.CVA with #PCs=5, NPAIRS.CVA with optimized #PCs (#PCs tested=2, 5, 10 and 25), and FSL.FEAT) was performed at 2, 4, 6-pixel smoothing levels, ≤2-cosine detrending (for NPAIRS.CVA and NPAIRS.GLM) or 176 s high-pass filtering (for FSL.FEAT), with intra- and inter-subject alignment. The combination of the pipeline choices together with 4 types of statistical models formed 24 pipelines in total.

In order to compare relative pipeline performance across heterogeneous models, classification accuracy was employed as a measure of prediction performance (Stone, 1974; Bullmore et al., 1995; Lautrup et al., 1994). Classification accuracy is defined as: $\frac{\text{number correctly classified scans}}{\text{total number of scans}}$. The threshold of posterior probability was set as 0.5, which is used to determine an fMRI volume's class membership based on posterior probability (i.e., if posterior probability ≥0.5, the fMRI volume belongs to the class; otherwise,

Table 2
The impact of preprocessing steps tested with GLM-based pipelines

Detrending (cosine)	0-pixel smoothing			2-pixel smoothing		
	delta*M*	Std. Dev.	Sig.	delta*M*	Std. Dev.	W. S. Sig.
(P1) The impact of slice timing correction (tested with FSL.slicetimer)						
0	−0.03	0.178	0.45	−0.10	0.160	**0.04**
2	0.04	0.214	0.37	−0.04	0.195	0.78
(P2) The impact of motion correction (tested with AIR.alignlinear)						
0	−0.08	0.105	**0.01**	−0.07	0.149	**0.03**
2	−0.07	0.174	0.10	−0.07	0.208	0.24

(P3) The impact of spatial smoothing (tested with VAST spatial smoothing module)

Smoothing (pixel)	0-cosine detrending			2-cosine detrending		
	delta*M*	Std. Dev.	Sig.	delta*M*	Std. Dev.	W. S. Sig.
1.5	0.09	0.044	**0.00**	0.09	0.041	**0.00**
2	0.14	0.062	**0.00**	0.12	0.059	**0.00**
4	0.16	0.128	**0.00**	0.14	0.130	**0.00**
6	0.17	0.093	**0.00**	0.14	0.087	**0.00**

(P4.1) The impact of high-pass filtering (tested with FSL high-pass filter)

Filter Cutoff(s)	2-pixel smoothing			4-pixel smoothing		
	delta*M*	Std. Dev.	Sig.	delta*M*	Std. Dev.	W. S. Sig.
176	0.04	0.049	**0.00**	0.04	0.046	**0.00**
128	0.03	0.047	**0.00**	0.04	0.049	**0.00**

(P4.2) The impact of cosine basis detrending (tested with VAST detrending module)

≤ # cosine cycles [a]	0-pixel smoothing			2-pixel smoothing		
	delta*M*	Std. Dev.	Sig.	delta*M*	Std. Dev.	W. S. Sig.
1	0.06	0.054	**0.00**	0.05	0.041	**0.00**
1.5	0.07	0.056	**0.00**	0.05	0.048	**0.00**
2	0.07	0.058	**0.00**	0.06	0.051	**0.00**
3	0.07	0.061	**0.00**	0.05	0.052	**0.00**

(P5) The impact of global intensity normalization (tested with VAST global intensity normalization module)

Detrending (cosine)	0-pixel smoothing			2-pixel smoothing		
	delta*M*	Std. Dev.	Sig.	delta*M*	Std. Dev.	W. S. Sig.
0	−0.01	0.025	**0.03**	−0.01	0.021	0.27
2	−0.00	0.017	0.21	−0.00	0.020	0.55

Std. Dev.—Standard deviation; W. S. Sig.—Significance tested by Wilcoxon matched-pair rank sum test across subjects.

[a] This means that all cosine basis function were applied simultaneously as unwanted covariates in steps of 0.5 cycles up to and including the number of cycles listed.

Table 3
Summary of the impact of the preprocessing steps with 2-pixel smoothing, and 2-cosine detrending or temporal filtering, tested with GLM-based pipelines

Preprocessing steps	Statistical model used	deltaM	Std.	W. S. Sig.	Rela. Vari.
1 Slice timing correction	GLM_{NPAIRS}	−0.04	0.195	0.78	−0.205
2 Motion correction	GLM_{NPAIRS}	−0.07	0.208	0.24	−0.337
3 **Spatial smoothing**	**GLM_{NPAIRS}**	**0.12**	**0.059**	**0.00**	**2.034**
4 **Temporal detrending**	**GLM_{NPAIRS}**	**0.06**	**0.051**	**0.00**	**1.176**
High-pass filtering	**$FEAT_{FSL}$**	**0.04**	**0.049**	**0.00**	**0.816**
5 Global normalization	GLM_{NPAIRS}	−0.00	0.020	0.55	−0.000

deltaM.—Mean distance change (mean distance without the tested preprocessing step − mean distance with the tested preprocessing step); Std.—Standard deviation; W. S. Sig.—Significance tested by Wilcoxon matched-pair rank sum test across subjects; Rela. Vari.—Relative variation (deltaM divided by standard deviation; significance $p<0.05$ is highlighted).

not). Mean classification accuracy (defined as the average classification accuracy across all the subjects in the dataset) and mean reproducibility (i.e., the average SPI reproducibility across all the subjects) was calculated to make the mean classification accuracy vs. mean reproducibility plot.

Results

Evaluating the impact of preprocessing steps

The impact of the preprocessing steps tested using NPAIRS.GLM- or FSL.FEAT-based pipelines with intra-subject or inter-subject alignment is presented in Table 2. As defined in the Methods, the mean distance change is a result of turning on and off a preprocessing step and delta$M>0$ implies improved performance. The bold numbers in Table 2 represent statistical significance better than $p<0.05$, calculated using a Wilcoxon matched-pair rank sum test across subjects.

Table 2 indicates that slice timing correction, global intensity normalization and motion correction all have significantly negative impacts across subjects but only when no detrending is applied. This indicates the overall importance of detrending in the processing pipeline and the potential for interactions between steps. Spatial smoothing, temporal detrending and high-pass filtering were found to significantly improve the performance of all GLM-based pipelines.

Table 3 summarizes the impact of GLM-based evaluation on inter-subject aligned data, for 2-pixel smoothing, and ≤2-cosine detrending or 128 s temporal filtering. The relative variation results demonstrate that spatial smoothing has the largest impact among preprocessing steps, followed by temporal detrending and/or high-pass filtering. The much higher standard deviation of slice timing and motion correction across the 16 subjects demonstrates that the impact of these steps may be quite heterogeneous across subjects compared with the more homogeneous response to smoothing and temporal filtering. Compared with results of CVA-based pipelines (Zhang, 2005), the significant positive impact of preprocessing steps: spatial smoothing, temporal detrending and high-pass filtering, and the insignificant average impact of slice timing correction and global intensity normalization are consistent across the univariate GLM and multivariate CVA models, which suggests that such effects are model independent.

Optimizing single-subject preprocessing steps

Table 4(A) summarizes the pipeline optimization results from NPAIRS.GLM and FSL.FEAT-based pipelines. In Table 4(A) GLM_{NPAIRS} outperforms $FEAT_{FSL}$ (i.e., smaller D_{min}) for 13/16 subjects, and the optimized preprocessing options vary considerably from subject to subject. For example, the optimized FWHM values for spatial smoothing across the 16 subjects vary from 2 to 6 voxels. Table 4(B) demonstrates that pipeline optimization with NPAIRS.GLM and FSL.FEAT improves pipeline performance significantly compared with the next-best-performing, non-optimized pipeline. Table 4(C) shows that on average pipeline optimization improves the BSR significantly compared with the average of the best and worst

Table 4(A)
Optimized pipeline options and performance for 16 single subjects

Subj.	Smoothing (pixel)		Detrending		D_{min} [a]	
	GLM_{NPAIRS}	$FEAT_{FSL}$	GLM_{NPAIRS}	$FEAT_{FSL}$	GLM_{NPAIRS}	$FEAT_{FSL}$
S1	6	6	1.5	176	0.41	0.43
S2	6	6	1	176	0.44	0.52
S3	4	4	3	176	0.29	0.32
S4	2	4	2	128	0.11	0.11
S5	2	4	1.5	176	0.33	0.30
S6	6	6	1.5	176	0.30	0.36
S7	4	4	0	176	0.17	0.19
S8	4	6	2	128	0.26	0.30
S9	4	4	2	176	0.44	0.52
S10	4	4	1.5	176	0.23	0.29
S11	4	6	2	176	0.47	0.49
S12	4	4	1.5	128	0.25	0.25
S13	4	6	2	176	0.55	0.64
S14	2	4	2	176	0.45	0.52
S15	6	2	1.5	176	0.18	0.25
S16	6	6	3	176	0.26	0.27

Note: Subj.—subject; Smoothing—spatial smoothing; Detrending—temporal detrending (GLM_{NPAIRS}: ≤# cosine basis function; $FEAT_{FSL}$: high-pass filter cutoff seconds.

[a] D_{min}—the minimum distance between the (prediction, reproducibility) pair of the tested pipeline to (1, 1) perfect prediction and reproducibility pair; GLM—NPAIRS.GLM; FEAT—FSL.GLM.

Table 4(B)
The impact of pipeline optimization (with NPAIRS performance metrics)

Pipeline model	ΔD_{min} [a]	Std. Dev.	Paired t Sig.	W. S. Sig. [b]
NPAIRS.GLM	0.020	0.012	**0.00**	**0.00**
FSL.FEAT	0.006	0.011	**0.05**	**0.05**

[a] ΔD_{min}—Mean difference across 16 subjects of (best performing non-optimized distance — optimized distance).
[b] W. S. Sig.—Significance of Wilcoxon matched-pair rank sum test.

non-optimized conjunction matrices (Avg. Mean Diff. BSR<0 implies BSR improvement), but not for the best non-optimized BSR conjunction matrix (indicated by Mean Diff. BSR). This is consistent with the corresponding results on NPAIRS.CVA-based pipelines in (Zhang et al., in press). This may suggest a minor gain in group homogeneity (i.e., the improved common activation detection across all subjects) with individual pipeline optimization, in agreement with the results of Shaw et al. (2003).

Evaluation of heterogeneous pipelines

The mean classification accuracy vs. mean reproducibility plot in Fig. 1 illustrates that the pipelines with NPAIRS.GLM (blue) and NPAIRS.CVA at optimized #PCs (green) outperform those with NPAIRS.CVA at fixed 5 PCs (red) and FSL.FEAT (orange). Further, we see from Fig. 1 that, in general, CVA-based pipelines ranging from CVA with fixed 5 PCs (red) to those with optimized #PCs (green) have higher classification accuracy than GLM-based pipelines (blue for NPAIRS.GLM, orange for FSL.FEAT), but are less reproducible than GLM-based pipelines for the cases tested.

Discussion

In previous studies of fMRI pipelines with the NPAIRS approach (Laconte et al., 2003; Shaw et al., 2003; Strother et al., 2004; Chen et al., 2006; Zhang, 2005), due to the limitations of the IDL-based NPAIRS package used, the research scope was largely restricted to CVA-based pipelines with modules within the NPAIRS package. The present study extended these previous studies by obtaining prediction accuracy for GLM-based, single-subject fMRI processing pipelines across software packages with a Java-based processing pipeline evaluation system (Zhang et al., in press). The findings of this study, based on a single block design, static-force dataset should be generalized to other datasets with care, but provide a view of how various fMRI processing pipeline options influence fMRI processing

Table 4(C)
The impact of pipeline optimization with between-subject reproducibility (BSR)

Pipeline model	Measure type	BSR	Std. Dev.	Paired t Sig.	W. S. Sig. [a]
NPAIRS.GLM	Mean Diff. [b]	0.113	0.129	**0.00**	**0.00**
	Avg. Mean Diff. [c]	−0.158	0.117	**0.00**	**0.00**
FSL.FEAT	Mean Diff. [b]	−0.002	0.072	0.77	0.26
	Avg. Mean Diff. [c]	−0.030	0.065	**0.00**	**0.00**

[a] Mean Diff. BSR—Mean difference (best performing non-optimized BSR conjunction matrix − optimized BSR conjunction matrix).
[b] Avg. Mean Diff. BSR—Mean difference (the average of the best performing and worse performing non-optimized BSR conjunction matrix − optimized BSR conjunction matrix).
[c] W. S. Sig.—Significance of Wilcoxon matched-pair rank sum test.

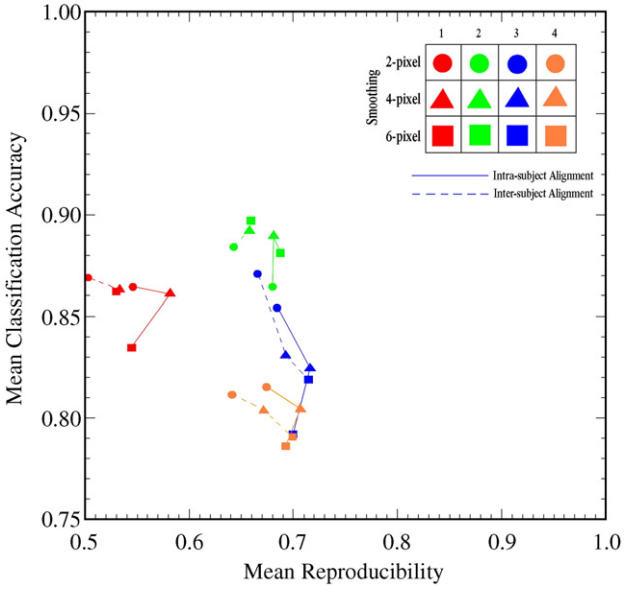

Fig. 1. Mean classification accuracy plotted vs. reproducibility of pipelines comparison at 2, 4, 6-pixel smoothing. Pipelines in comparison: 1. NPAIRS.CVA at 5 #PCs (red); 2. NPAIRS.CVA at optimized #PCs (green); 3. NPAIRS.GLM (blue); 4. FSL.FEAT (orange).

performance (in prediction accuracy and SPI reproducibility metrics) across heterogeneous analytic models and software packages.

Evaluating the impact of preprocessing steps

The finding of a significant positive impact of spatial smoothing, high-pass filtering and temporal detrending on GLM-based pipelines in this study are consistent with what was observed in our previous studies (LaConte et al., 2003; Zhang, 2005) with CVA-based single-subject fMRI processing pipelines, and in previous ROC analyses on simulated block-designed and event-related data (Skudlarski et al., 1999; Della-Maggiore et al., 2002). The nonsignificant impact of slice timing correction and global intensity normalization on GLM-based pipelines is consistent with the observations for CVA-based pipelines (Zhang, 2005), and ROC findings on simulated block-designed data (Skudlarski et al., 1999). These results suggest that the relative importance of many preprocessing steps may be ranked using NPAIRS results from real data, and are similar for both univariate GLM and multivariate CVA approaches.

However, we have demonstrated that motion correction can significantly improve the performance of CVA-based pipelines (Zhang, 2005), but its impact on the same data set is significantly negative without any temporal detrending or not significant for GLM-based pipelines (Table 2). Previous ROC analysis findings (Skudlarski et al., 1999) indicate that motion correction may not change the relative efficiency of the steps in GLM-based fMRI data analysis. Jezzard et al. (2001) further pointed out that when subject motion is stimulus uncorrelated, motion correction makes the activation analysis more sensitive; but when the motion is stimulus correlated, motion correction reduces the statistical power and the apparent level of activation. While it is unclear exactly what is causing this differential sensitivity to motion correction we suspect significant stimulus coupled motion negatively impacts the GLM results, but may be removed by the PC denoising associated with CVA on a PCA basis set.

Pipeline optimization

The individually optimized spatial smoothing and temporal detrending choices for GLM-based, single-subject fMRI processing pipelines optimization in this study are only a small set of possible pipeline options and parameter choices, chosen based on previous research and our initial impact evaluation (Laconte et al., 2003; Shaw et al., 2003; Strother et al., 2004; Zhang, 2005). Compared with optimization results of CVA-based pipelines (Zhang, 2005) based on spatial smoothing, temporal detrending and #PCs for PCA/CVA, these results demonstrate that such pipeline optimization significantly improves pipeline performance regardless of the univariate-or-multivariate model (GLM or CVA) and the software package used (FSL or NPAIRS). This result reflects what is already a common practice in the field that it is important to include spatial smoothing and temporal detrending/filtering in fMRI processing pipelines. Furthermore, the prediction-reproducibility framework provides a quantitative means of evaluating the relative importance of such processing steps. These results demonstrate that the level of spatial smoothing chosen is the most important processing choice made, and that the significant positive impact of such pipeline optimization is somewhat model independent.

Further, the positive impact on the averaged between-subject reproducibility (BSR) across GLM- and CVA-based pipelines suggests a minor gain in group homogeneity (i.e., the improved common activation detection across all subjects) independent of univariate GLM and multivariate CVA models. This supports previous results showing that aggregating the individually optimized data in a random effect group analysis may result in improved group results (Shaw et al., 2003).

Heterogeneous pipeline evaluation

Since GLM is the most widely used univariate method in fMRI analysis and CVA is a typical multivariate fMRI analysis method, the evaluation and comparison of these two methods (or models) across fMRI software packages has practical meaning. As indicated in Fig. 1 univariate GLM is not perfect—it has relatively similar SPI reproducibility, but lower prediction accuracy than CVA. A possible reason for the slightly better SPI reproducibility in GLM (in Fig. 1) is that it fits the data at each voxel using a fixed, nonadaptive design matrix that avoids the model overfitting problem and generates slightly biased, but lower spatial variance results and hence somewhat higher reproducibility. In addition, the GLM prediction measure used in this study is based on the algorithm described in Kjems et al. (2002) and Zhang et al. (in press). A limitation of the algorithm is that it uses a product of individual voxel prediction estimates based on the assumption that the voxels in fMRI data are independent. However, voxels are not independent due to local spatial autocorrelation and long-range network interactions. Univariate GLM ignores these spatio-temporal covariance structures, compared with the adaptive covariances estimated in CVA, leading to lower prediction accuracy. An improved algorithm measuring prediction performance of GLM that avoids such an assumption remains to be investigated. In addition, Friston and Penny (2003) proposed the posterior probability map, which provides another way to obtain GLM prediction measures.

In contrast, multivariate CVA is built on a PCA basis and adapted to the spatio-temporal covariance structure in the data to find a linear combination of variables that maximally discriminates the two groups (baseline and activation) identified by prior knowledge of the brain states of the image volumes. Therefore, it is not surprising that prediction for CVA is somewhat higher than for GLM given that CVA is designed to maximize prediction. The fact that CVA has a slightly lower reproducibility than GLM-based models suggests that adapting the covariance structures to the data may somewhat overfit producing higher model variance and lower reproducibility. This trade-off between prediction and reproducibility as a function of #PCs used to build the CVA model is clearly seen in LaConte et al. (2003). These differential performance results probably reflect different bias–variance trade-offs and neither is obviously better than the other. If the goal is prediction of image-volume brain states, an emerging experimental approach in fMRI, then CVA or some other predictive modeling technique (e.g., Support Vector Machines) may be preferred (LaConte et al., 2005). On the other hand, if the goal is local regional signal detection then the higher reproducibility across runs provided by GLM-based pipelines may be preferred.

In Fig. 1, it is interesting to note that changing the time-course model had little impact on the SPI reproducibility for a given smoothing level compared with moderate changes in prediction performance: (1) In NPAIRS.CVA, baseline-activation transition volumes were dropped, which improves the predicted label accuracy by removing the transition image volumes that are difficult to classify; (2) In FSL.FEAT, the transition volumes were kept and convolution with the HRF was performed (that blurred and delayed the hemodynamic response), which mismatches the predicted GLM labels with the experimental volume labels that were not time-shifted producing lower prediction values; (3) In NPAIRS.GLM, the transition volumes were kept and convolution with the HRF was not performed leaving a closer, but still imperfect match between the experimental volume labels and predicted GLM response. By only calculating prediction on non-transition image volumes or by shifting the baseline-activation label structure these values can be improved, but this requires an additional modeling assumption to estimate the extent of the shift, particularly for TRs shorter than the 4 s of this study. These results suggest that across these different models with their different temporal assumptions, the two NPAIRS performance metrics may be relatively uncoupled, reinforcing the use of the prediction vs. reproducibility performance plot. Our data also suggests that statistical reliability as measured by reproducibility is not a strong function of the details of transition block modeling, an issue we return to below.

Why NPAIRS

The NPAIRS approach was proposed, as an alternative to simulations with the ROC approach, for evaluating the performance of fMRI processing pipelines with real fMRI data (Strother et al., 2002). NPAIRS, like an ROC, uses measures from two conceptually orthogonal domains (prediction in the time domain and reproducibility in the space domain), providing two partially related metrics that are compared in a 2D plot such that a single corner of the 2D graph represents ideal modeling characteristics; for NPAIRS perfect prediction ($p=1$) and infinite global SNR ($r=1$); for ROCs the optimal corner of the plot is where TP=1 and FR=0.

Given the smoothly varying nature of fMRI regressors typically employed in GLM models with delayed peaks and post-stimulus undershoot, the notion of classifying fMRI volumes according to their brain states (such as baseline and activation) may seem quaint (see Discussion above). However, classifying fMRI volumes makes

sense within the NPAIRS-CVA framework because classification provides a constraint that helps to control potentially increased bias through simply trying to maximize reproducibility by manipulating design matrix regressors or preprocessing parameters, e.g., spatial smoothing. When testing a GLM model in this framework the typical smoothly varying GLM regressors with HRFs may still be used to model each split-half data set (e.g., our FSL model). In terms of SNR represented by the GLM_{NPAIRS} and $FEAT_{FSL}$ reproducibility *Z*-scores little seems to be lost by treating the transition volumes in different ways. This is particularly likely to hold true when the transition volumes represent only a small fraction of the total active or baseline block length, e.g., in this case there are 2–3 transition volumes (dropped for $NPAIRS_{CVA}$) from blocks of 10–11 images each (TR=4 s; see LaConte et al. (2003) for details). This result is supported by unpublished simulations of a block-design experiment of this type where dropping transition volumes compared with results using exact HRF-models reduces signal detection by <6% based on ROC measurements. A simple two-class classification metric with dropped transition volumes may be less useful with shorter TRs and shorter blocks, or certainly for single-event studies. It will then be necessary to adopt alternative prediction measures such as the mean-square error on the complete HRF time-course in a cross-validation framework, as used by Kay et al. (2007).

We recognize that the reproducibility metric in NPAIRS is imperfect and may be improved, perhaps by restricting split-half comparisons to only gray matter voxels or by augmenting it with a voxel-based measure, such as that developed by Genovese et al. (1997). However, neither the approach of Nandy and Cordes (2003) nor that of Genovese or Liou et al. (2006) can be applied to this data set because they need, respectively, a separate baseline run per subject or at least three independent runs. Our reproducibility measure, like the others in the literature, contains an unknown bias. For example, the reproducibility correlation coefficient tends to increase with spatial smoothing as the intervoxel dependence increases because we are likely to be introducing bias into out modeled SPI output. A model that outputs close to a constant spatial pattern that is relatively independent of the input data (e.g., such as might be produced by a very large spatial smoothing filter) will provide highly reproducible results that are useless because they largely ignore the data. This is an extreme example of a bias–variance trade-off with very large bias and very little variance in the output. As we do not know where any particular model should lie on its bias–variance curve, we used an equal weighting between prediction accuracy and SPI reproducibility to calculate the distance to ($p=1$, $r=1$). More work is needed to explore other evaluation metrics and alternative weight ratios.

It is anticipated that once the NPAIRS approach is further refined and becomes more mature, it can be applied to other fMRI processing software packages and/or modules such as SPM, AFNI and FSL.MELODIC (ICA), which will allow quantitative evaluation of the performance of such fMRI processing pipelines.

Model bias-and-variance trade-off and consensus approach

As Breiman (1998) pointed out, there is a bias-and-variance trade-off involved in the method (or classifier) evaluation and comparison in machine learning. Prediction methods such as classification trees and neural networks are generally less biased, but unstable (with high variance); while methods such as linear discriminant analysis (LDA) and k-nearest neighbors (KNN) are stable (with low variance), but can be more biased. All function neuroimaging models lie somewhere on a data and model-dependent bias–variance curve and the NPAIRS framework is an initial attempt to start to compare these trade-offs for preprocessing pipelines and data analysis models. Since both univariate and multivariate models have their own limitations and strengths which are often complementary to each other, fMRI analysis might best proceed by using a combination of techniques. A univariate model can be used for activation detection and a multivariate model used for the detection of network, functional connectivity. Similar views are expressed in fMRI pipeline evaluation (Friston et al., 1995c; Strother et al., 2002; LaConte et al., 2003). To overcome model-dependent biases and limitations, Hansen et al. (2001) proposed a consensus map through model averaging and demonstrated that an enhanced ROC curve was obtained through model averaging in a simulated study. Such consensus methods may help to resolve the bias–variance trade-off problem in fMRI model comparison and selection.

Conclusion

In this study, we found that for block-design fMRI data, (1) slice timing correction and global intensity normalization have little consistent impact on GLM-based fMRI processing pipelines, but spatial smoothing and high-pass filtering or temporal detrending significantly increased pipeline performance; (2) combined optimization of spatial smoothing and temporal detrending processing steps improved performance of GLM-based pipelines; and (3) in general, the prediction performance of CVA models is higher than that of the GLM models, while GLM models are more reproducible than CVA models. In addition, our results indicate that it may be necessary to consider a consensus approach to obtain more accurate activation patterns in fMRI data due to the different bias–variance trade-offs of the univariate GLM and multivariate CVA models.

Non-invasive fMRI has fundamentally changed the way we study the brain and the future of fMRI looks bright in terms of a transition from basic research to clinical applications (Borsook et al., 2006). Looking ahead, we believe that the current status in fMRI analysis (i.e., the lack of fMRI software evaluation and result validation) will be changed gradually and advances in evaluating, optimizing, standardizing and validating fMRI processing pipelines will eventually lead to medical licensing of fMRI software, which will help fMRI to be fully established as a clinical neuroimaging method.

Acknowledgments

We thank James Ashe, M.D. and Suraj Muley, M.D. for providing the static-force data. We are also grateful to Kelly Rehm, Kirt Schaper and Kate Fissell for their technical assistance and support. This work was partly supported by the NIH Human Brain Project P20 Grant MN EB002013.

References

Andersen, A.H., Gash, D.M., Avison, M.J., 1999. Principal component analysis of the dynamic response measured by fMRI: a generalized linear systems framework. Magn. Reson. Imaging 17 (6), 795–815.

Beckmann, C.F., Smith, S.M., 2004. Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE Trans. MedImg. 23 (2), 137–152.

Becerra, L., Borsook, D., 2006. Insights into pain mechanisms through functional MRI. Drug Discov. Today 3 (3), 313–318.

Biswal, B.B., Ulmer, J.L., 1999. Blind source separation of multiple signal sources of fMRI data sets using independent component analysis. J. Comput. Assist. Tomogr. 23 (2), 265–271.

Bookheimer, S.Y., Strojwas, M.H., Cohen, M.S., Saunders, A.M., Pericak-Vance, M.A., Mazziotta, J.C., Small, G.W., 2000. Patterns of brain activation in people at risk for Alzheimer's disease. N. Engl. J. Med. 343 (7), 450–456.

Bookheimer, S., 2007. Pre-surgical language mapping with functional magnetic resonance imaging. Neuropsychol. Rev. (Electronic publication ahead of print).

Borsook, D., Becerra, L., Hargreaves, R., 2006. A role for fMRI in optimizing CNS drug development. Nat. Rev. Drug Discov. 7, 1–14.

Breiman, L., 1998. Arcing classifiers. Ann. Statist. 26, 801–849.

Bullmore, E.T., Brammer, M., Rouleau, G., Everitt, B., Simmons, A., Sharma, T., Frangou, S., Murray, R., Dunn, G., 1995. Computerized brain tissue classification of magnetic resonance images: a new approach to the problem of partial volume artifact. NeuroImage 2 (2), 133–147.

Bullmore, E.T., Rabe-Hesketh, S., Morris, R.G., Williams, S.C., Gregory, L., Gray, J.A., Brammer, M.J., 1996. Functional magnetic resonance image analysis of a large-scale neurocognitive network. Neuroimage 4 (1), 16–33.

Chen, X., Pereira, F., Lee, W., Strother, S., Mitchell, T., 2006. Exploring predictive and reproducible modeling with the single-subject FIAC dataset. Human Brain Mapping 27, 452–461.

Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162–173 [http://afni.nimh.nih.gov/afni/].

Della -Maggiore, V., Chau, W., Peres-Neto, P.R., McIntosh, A.R., 2002. An empirical comparison of SPM preprocessing parameters to the analysis of fMRI data. NeuroImage 17 (1), 19–28.

Fernández, G., Specht, K., Weis, S., Tendolkar, I., Reuber, M., Fell, J., Klaver, P., Ruhlmann, J., Reul, J., Elger, C.E., 2003. Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. Neurology 60 (6), 969–975.

Fissell, K., Tseytlin, E., Cunningham, D., Iyer, K., Carter, C.S., Schneider, W., Cohen, J.D., 2003. Fiswidgets: a graphical computing environment for neuroimaging analysis. Neuroinformatics 1 (1), 111–126 [http://grommit.lrdc.pitt.edu/fiswidgets].

FitzGerald, D.B., Cosgrove, G.R., Ronner, S., Jiang, H., Buchbinder, B.R., Belliveau, J.W., Rosen, B.R., Benson, R.R., 1997. Location of language in the cortex: a comparison between functional MR imaging and electrocortical stimulation. Am. J. Neuroradiol. 18, 1529–1539.

Friston, K.J., 1996. Statistical parametric mapping and other analysis of functional imaging data. Brain Mapping: the Methods. Academic Press, pp. 363–385.

Friston, K.J., Penny, W., 2003. Posterior probability maps and SPMs. Neuroimage 19 (3), 1240–1249.

Friston, K.J., Jezzard, P., Turner, R., 1994. Analysis of functional MRI time-series. Hum. Brain Mapp. 1, 153–171.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.J., 1995a. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 2, 189–210.

Friston, K.J., Holmes, A.P., Poline, J.B., Grasby, P.J., Williams, S.C.R., Frackowiak, R.S.J., Turner, R., 1995b. Analysis of fMRI time series revisited. Neuroimage 2, 45–53.

Friston, K.J., Frith, C.D., Frackowiak, R.S.J., Turner, R., 1995c. Characterizing dynamic brain responses with fMRI: a multivariate approach. NeuroImage 2 (2), 166–172.

Friston, K.J., Phillips, J., Chawla, D., Buchel, C., 1999. Revealing interactions among brain systems with nonlinear PCA. Hum Brain Mapp. 8 (2–3), 92–97.

Friston, K.J., Phillips, J., Chawla, D., Buchel, C., 2000. Nonlinear PCA: characterizing interactions between modes of brain activity. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 355 (1393), 135–146.

Gavrilescu, M., Shaw, M.E., Stuart, G.W., Eckersley, P., Svalbe, I.D., Egan, G.F., 2002. Simulation of the effects of global normalization procedures in functional MRI. NeuroImage 17 (2), 532–542.

Genovese, C.R., Noll, D.C., Eddy, W.F., 1997. Estimating test-retest reliability in fMRI I: statistical methodology. Magn. Reson. Med. 38, 497–507.

Hansen, L.K., Larsen, J., Nielsen, F.A., Strother, S.C., Rostrup, E., Savoy, R., Lange, N., Sidtis, J., Svarer, C., Paulson, O.B., 1999. Generalizable patterns in neuroimaging: how many principal components? Neuroimage 9 (5), 534–544.

Hansen, L.K., Nielsen, F.Å., Strother, S.C., Lange, N., 2001. Consensus inference in neuroimaging. NeuroImage 13 (6), 1212–1218.

Haslinger, B., Erhard, P., Kämpfe, N., Boecker, H., Rummeny, E., Schwaiger, M., Conrad, B., Ceballos-Baumann, A.O., 2001. Event-related functional magnetic resonance imaging in Parkinson's disease before and after levodopa. Brain 124 (3), 558–570.

Jezzard, P., Matthews, P.M., Smith, S.M., 2001. Functional MRI: an Introduction to Methods, 1st edition. Oxford University Press.

Kay, K.N., David, S.V., Prenger, R.J., Hansen, K.A., Gallant, J.L., 2007. Modeling low-frequency fluctuation and hemodynamic response timecourse in event-related fMRI. Hum. Brain. Map. 29 (2), 142–156.

Kjems, U., Hansen, L.K., Anderson, J., Frutiger, S., Muley, S., Sidtis, J., Rottenberg, D., Strother, S.C., 2002. The quantitative evaluation of functional neuroimaging experiments: mutual information learning curves. NeuroImage 15 (4), 772–786.

LaConte, S., Anderson, J., Muley, S., Ashe, J., Frutiger, S., Rehm, K., Hansen, L.K., Yacoub, E., Hu, X., Rottenberg, D., Strother, S., 2003. The evaluation of preprocessing choices in single-subject BOLD fMRI using NPAIRS performance metrics. NeuroImage 18 (1), 10–27.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X., 2005. Support vector machines for temporal classification of block design fMRI data. NeuroImage 26 (2), 317–329.

Laiand, S.H., Fang, M., 1999. A novel local PCA-based method for detecting activation signals in fMRI. Magn Reson Imaging. 17 (6), 827–836.

Lange, N., Strother, S.C., Anderson, J.R., Nielsen, F.Å., Holmes, A.P., Kolenda, T., Savoy, R., Hansen, L.K., 1999. Plurality and resemblance in fMRI data analysis. NeuroImage 10 (3), 282–303.

Larsen, J., Hansen, L.K., 1997. Generalization: the hidden agenda of learning. In: Hwang, J.-N., Kung, S.Y., Niranjan, M., Principe, J.C. (Eds.), The Past, Present, and Future of Neural Networks for Signal Processing. IEEE Signal Processing Magazine, 14(6), pp. 43–45.

Lautrup, B., Hansen, L.K., Law, I., Morch, N., Svarer, C., Strother, S.C., 1994. Massive weight sharing: a cure for extremely ill-posed problems. In: Hermann, H.J., Wolf, D.E., Poeppel, E. (Eds.), Proceedings of the Workshop on Supercomputing in Brain Research: from Tomography to Neural Networks. World Scientific, Ulich, Germany, pp. 137–144.

Le, T.H., Hu, X., 1997. Methods for assessing accuracy and reliability in functional MRI. NMR Biomed. NMR Biomedicine 10, 160–164.

Liou, M., Su, H.R., Lee, J.D., Aston, J.A., Tsai, A.C., Cheng, P.E., 2006. A method for generating reproducible evidence in fMRI studies. Neuroimage 29 (2), 383–395.

Lipton, A.M., McColl, R., Cullum, C.M., Allen, G., Ringe, W.K., Bonte, F.J., McDonald, E., Rubin, C.D., 2003. Differential activation on fMRI of monozygotic twins discordant for AD. Neurology 60 (10), 1713–1716.

Lukic, A.S., Wernick, M.N., Strother, S.C., 2002. An evaluation of methods for detecting brain activations from PET or fMRI images. Artif. Intell. Med. 25, 69–88.

Machulda, M.M., Ward, H.A., Borowski, B., Gunter, J.L., Cha, R.H., O'Brien, P.C., Petersen, R.C., Boeve, B.F., Knopman, D., Tang-Wai, D.F., Ivnik, R.J., Smith, G.E., Tangalos, E.G., Jack Jr., R., 2003. Comparison of memory fMRI response among normal, MCI, and Alzheimer's patients. Neurology 61 (4), 500–506.

Maitra, R., Roys, S.R., Gullapalli, R.P., 2002. Test-retest reliability estimation of functional MRI Data. Magn. Reson. Med. 48, 62–70.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press, San Diego.

McKeown, M.J., Makeig, S., Brown, G.G., Jung, T.P., Kindermann, S.S., Bell, A.J., Sejnowski, T.J., 1998. Analysis of fMRI data by blind separation into independent spatial components. Hum. Brain Mapp. 6 (3), 160–188.

McKeown, M.J., 2000. Detection of consistently task-related activations in fMRI data with hybrid independent component analysis. Neuroimage 11 (1), 24–35.

Muller, J.L., Euticke, C.D., Putzhammer, A., Roder, C.H., Hajak, G., Winker, J., 2003. Schizophrenia and Parkinson's disease lead to equal motor-related changes in cortical and subcortical brain activation: an fMRI fingertapping study. Psychiatry Clin. Neurosci. 57, 562–568.

Nandy, R.R., Cordes, D., 2003. Novel ROC-type method for testing the efficiency of multivariate statistical methods in fMRI. Magn. Reson. Med. 49 (6), 1152–1162.

Poline, J.B., Strother, S.C., Dehaene-Lambertz, G., Egan, G.F., Lancaster, J.L., 2006. Motivation and synthesis of the FIAC experiment: reproducibility of fMRI results across expert analyses. Hum. Brain Mapp. 27, 351–359.

Rex, D.E., Ma, J.Q., Toga, A.W., 2003. The LONI pipeline processing environment. NeuroImage 19 (3), 1033–1048.

Rombouts, S., Scheltens, P., 2005. Functional connectivity in elderly controls and AD patients using resting state fMRI: a pilot study. Curr. Alzheimer Res. 2 (2), 115–116.

Sabatini, U., Boulanouar, K., Fabre, N., Martin, F., Carel, C., Colonnese, C., Bozzao, L., Berry, I., Montastruc, J.L., Chollet, F., Rascol, O., 2000. Cortical motor reorganization in akinetic patients with Parkinson's disease, a functional MRI study. Brain 123 (2), 394–403.

Shaw, M.E., Strother, S.C., Gavrilescu, M., Podzebenko, K., Waites, A., Watson, J., Anderson, J., Jackson, G., Egan, G., 2003. Evaluating subject specific preprocessing choices in multisubject fMRI data sets using data-driven performance metrics. NeuroImage 19 (3), 988–1001.

Skudlarski, P., Constable, R.T., Gore, J.C., 1999. ROC analysis of statistical methods used in functional MRI: individual subjects. NeuroImage 9 (3), 311–329.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M., Matthews, P.M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23 (S1), 208–219 [http://www.fmrib.ox.ac.uk/fsl/].

Stein, E.A., 2001. fMRI: a new tool for the in vivo localization of drug actions in the brain. J. Anal. Toxicol. 25 (5), 419–424.

Stippich, C., 2007. Presurgical functional magnetic resonance imaging (fMRI). Clin. Neuroradiology 17 (2), 69–87.

Stippich, C., Rapps, N., Dreyhaupt, J., Durst, A., Kress, B., Nennig, E., Tronnier, V.M., Sartor, K., 2007. Localizing and lateralizing language in patients with brain tumors: feasibility of routine preoperative functional MR imaging in 81 consecutive patients. Radiology 243 (3), 828–836.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Statist. Soc. B 36, 111–147.

Strother, S.C., Lange, N., Anderson, J.R., Schaper, K.A., Rehm, K., Hansen, L.K., Rottenberg, D.A., 1997. Activation pattern reproducibility: measuring the effects of group size and data analysis models. Hum. Brain Mapp. 5, 312–316.

Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., Rottenberg, D., 2002. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. NeuroImage 15 (4), 747–771.

Strother, S.C., Laconte, S., Hansen, L.K., Anderson, J., Zhang, J., Pulapura, S., Rottenberg, D., 2004. Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics. NeuroImage 23 (supple.1), 196–207.

Tharin, S., Golby, A., 2007. Functional brain mapping and its applications to neurosurgery. Neurosurgery 60 (Supl 2), 185–201.

Woods, R.P., Grafton, S.T., Watson, J.D., Sicotte, N.L., Mazziotta, J.C., 1998. Automated image registration: II. Intersubject validation of linear and nonlinear models. J. Comput. Assist. Tomogr. 22 (1), 153–165.

Worsley, K.J., Friston, K.J., 1995. Analysis of fMRI time-series revisited — again. Neuroimage 2, 173–181.

Worsley, K.J., Poline, J.B., Friston, K.J., Evans, A.C., 1997. Characterizing the response of PET and fMRI data using multivariate linear models (MLM). NeuroImage 6, 305–319.

Zhang, J., 2005. Evaluating and optimizing fMRI processing pipelines with the NPAIRS approach for decision support in fMRI applications. PhD Thesis, University of Minnesota.

Zhang, J., Liang, L., Anderson, J., Gatewood, L., Rottenberg, D., Strother, S.C., in press. A Java-based fMRI processing pipeline evaluation system for the assessment of GLM and CVA-based heterogeneous pipelines. Neuroinformatics.