ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 23 (2004) S196–S207

Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis

Stephen Strother,[a,b,c,*] Stephen La Conte,[b,d,1] Lars Kai Hansen,[e] Jon Anderson,[c] Jin Zhang,[c] Sujit Pulapura,[c] and David Rottenberg[a,c]

[a] *Radiology Department, University of Minnesota, United States*
[b] *Biomedical Engineering, University of Minnesota, United States*
[c] *Neurology Departments, University of Minnesota and VA Medical Center, United States*
[d] *Center for Magnetic Resonance Research, University of Minnesota, United States*
[e] *Informatics and Mathematical Modeling, Technical University of Denmark, Denmark*

Available online 11 September 2004

We argue that published results demonstrate that new insights into human brain function may be obscured by poor and/or limited choices in the data-processing pipeline, and review the work on performance metrics for optimizing pipelines: prediction, reproducibility, and related empirical Receiver Operating Characteristic (ROC) curve metrics. Using the NPAIRS split-half resampling framework for estimating prediction/reproducibility metrics (Strother et al., 2002), we illustrate its use by testing the relative importance of selected pipeline components (interpolation, in-plane spatial smoothing, temporal detrending, and between-subject alignment) in a group analysis of BOLD-fMRI scans from 16 subjects performing a block-design, parametric-static-force task. Large-scale brain networks were detected using a multivariate linear discriminant analysis (canonical variates analysis, CVA) that was tuned to fit the data. We found that tuning the CVA model and spatial smoothing were the most important processing parameters. Temporal detrending was essential to remove low-frequency, reproducing time trends; the number of cosine basis functions for detrending was optimized by assuming that separate epochs of baseline scans have constant, equal means, and this assumption was assessed with prediction metrics. Higher-order polynomial warps compared to affine alignment had only a minor impact on the performance metrics. We found that both prediction and reproducibility metrics were required for optimizing the pipeline and give somewhat different results. Moreover, the parameter settings of components in the pipeline interact so that the current practice of reporting the optimization of components tested in relative isolation is unlikely to lead to fully optimized processing pipelines.

Keywords: fMRI; Prediction; Reproducibility

Introduction

Neuroimaging researchers typically focus on extracting “neuroscientifically relevant” results from their data sets. Almost always this is done without attempting to optimize and/or understand the relative influence of the pipeline processing choices that were made in analyzing the data. Moreover, the generation of a “plausible result” that can be linked to the neuroscientific literature is often taken as justification of the pipeline choices made, providing a systematic bias in the field towards prevailing neuroscientific expectations and away from unexpected, new results (Skudlarski et al., 1999; Strother et al., 1995a,b, 2002). In addition, there is accumulating evidence in the literature that by applying a new processing pipeline to a raw data set, significantly modified spatial activation patterns may be obtained as a result of changing/optimizing preprocessing techniques (Della-Maggiore et al., 2002; Friston et al., 2000; LaConte et al., 2003a; Shaw et al., 2003a; Tanabe et al., 2002) and/or the data analysis approach (Beckmann and Smith, 2004; Friston et al., 1996; Kherif et al., 2002; Liou et al., 2003; Muley et al., 2001; Nandy and Cordes, 2003; Shaw et al., 2002; Strother et al., 1995a; Tegeler et al., 1999). These real-data results are supported by several simulation studies, which indicated that significant differences in signal detection performance should be expected for different preprocessing (Gavrilescu et al., 2002; Skudlarski et al., 1999) and data analysis (Beckmann and Smith, 2004; Lange et al., 1999; Lukic et al., 2002, 2004; Tzikas et al., 2004) approaches. These published results demonstrate the likelihood that new insights into human brain function may be obscured by poor and/or limited choices in the image processing pipeline (McIntosh, Private communication).

Simulations in which the true activation signal is known allow different pipeline choices to be ranked using standard signal detection metrics based on receiver operating characteristic (ROC) curves (Swets, 1988). However, for fMRI, this is problem-

* Corresponding author. Radiology Department and University of Minnesota, United States. Fax: +1 612 725 2068.

E-mail address: steve@neurovia.umn.edu (S. Strother).

[1] Now at Biomedical Engineering Department, Georgia Institute of Technology/Emory University.

Available online on ScienceDirect (www.sciencedirect.com.)

1053-8119/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2004.07.022

atic because the vascular, blood oxygenation level dependent (BOLD) signal and noise structure are not well understood, and it is generally unknown if a particular set of simulation results are relevant for any given fMRI data set, a problem that is compounded if we are interested in the BOLD fMRI signal and noise structure as a function of age and/or disease (D'Esposito et al., 2003).

In an attempt to avoid the need for simulations, researchers have proposed data-driven techniques that estimate performance metrics from the available data. Le and Hu (1997) suggested estimating the true distribution based on highly averaged results. However, the large number of repeat scanning runs required makes this approach impractical, even if it is not biased by the requirement for the mean to tend towards the true signal.

Other researchers have focused on the reproducibility, or reliability, of activation patterns based on the recognition that smaller p values do not imply a stronger likelihood of getting the same result in another replication of the same experiment, and the historical importance of replication as a fundamental criterion for a result to be considered scientific (Carver, 1993; Genovese et al., 1997; Kiehl and Liddle, 2003; Liou et al., 2003; Maitra et al., 2002; Moeller et al., 1999; Strother et al., 1997, 1998; Tegeler et al., 1999). This is one reason why minimizing p values as a quantitative performance measure for pipeline optimization is a poor choice, although it has been used repeatedly in the literature (e.g., Hopfinger et al., 2000; Tanabe et al., 2002).

Provided at least three repeat runs are available, an empirical-ROC curve may be estimated from the data (Genovese et al., 1997), and by incorporating local spatial correlation into the same framework a minimum of two runs is sufficient (Maitra et al., 2002). An interesting application of this empirical-ROC generation framework together with a technique for selecting the optimal operating point on the resulting ROC curve has been recently published by Liou et al. (2003). An alternative procedure for generating empirical ROC curves that requires a "control state" run to estimate false-positive rates together with a standard experimental run has been proposed by Nandy and Cordes (2003).

Strother et al. (1997, 1998) proposed an alternative reproducibility metric based on a principal components analysis (PCA) of two or more independently replicated statistical parametric images (SPIs). This approach was further developed in Kjems et al. (2002), LaConte et al. (2003a), Shaw et al. (2002, 2003a), Strother et al. (2002), and Tegeler et al. (1999). A correlation coefficient summarizes the reproducibility of two independent SPIs as reflected in their scatter plot. This reproducibility correlation coefficient also directly measures the overall signal-to-noise level of the single, reproducible, Z-scored, activation SPI that is extracted from the principal PCA axis of the scatter plot (Strother et al., 2002). However, this reproducibility metric is a biased measure because it inherits any data-analysis model biases that exist when measuring SPIs. It seems likely that the empirical-ROC metrics share this bias and that, like Strother's reproducibility metric, they should not be considered measures of true signal detection performance.

Simultaneously, Hansen and Strother, guided by the field of predictive learning in statistics (Hastie et al., 2001; Larsen and Hansen, 1997; Mjolsness and DeCoste, 2001), introduced the idea of using potentially unbiased cross-validation-based prediction metrics to measure data-analytic performance in functional neuroimaging (Hansen et al., 1999; Kjems et al., 2002; Kustra and Strother, 2001; Lautrup et al., 1995; Morch et al., 1997). Similar prediction metrics have recently been used by others (McKeown, 2000; Ngan et al., 2000). In addition, prediction metrics have been used to gain new insight into the debate over the spatially modular versus spatially distributed nature of human brain processing (Cox and Savoy, 2003; Haxby et al., 2001). We expect both prediction and reproducibility metrics to play an increasingly important role in the future optimization and interpretation of fMRI studies.

With this in mind, Strother et al. (2002) proposed the unique approach of simultaneously measuring and combining data-driven prediction and reproducibility metrics for pipeline and data analysis optimization using split-half resampling (a combination of two-fold cross-validation and delete-d jackknife resampling) to produce a ROC-like plot. They developed the NPAIRS (Nonparametric Prediction, Activation, Influence and Reproducibility reSampling) software package to implement and test this idea (Kjems et al., 2002; LaConte et al., 2003a; Shaw et al., 2003a; Web distribution and documentation at http://neurovia.umn.edu/incweb/npairs_info.html). In preliminary comparisons using simulations, Shaw et al. (2003b) have shown that prediction-reproducibility plots seem to perform at least as well as standard ROC curves.

This paper is concerned with the combined use of prediction and reproducibility metrics to test the relative importance of different processing pipeline choices in the detection of large-scale brain networks from the combined BOLD-fMRI scans of 16 subjects performing a block-design, parametric-static-force task. We have investigated the impact and interaction of interpolation, within-plane spatial smoothing, temporal detrending, between-subject alignment using affine and nonlinear polynomial registration (i.e., warps), and "tuning" the data analysis approach. The large-scale brain networks were detected for separate, uncorrelated OFF–ON and parametric force responses using canonical variates analysis (CVA), a flexible multivariate form of linear discriminant analysis that may be tuned to fit the data. The prediction and reproducibility metrics were measured using split-half resampling of the 16-subject group within the NPAIRS framework. We found that the metrics could easily detect the smoothing difference between sinc and trilinear-based interpolation, and that even a small amount of smoothing together with tuning the CVA model were by far the most important processing parameters. Detrending was found to be essential to remove low-frequency time trends, and to allow a reliable parametric force response to emerge despite pseudo-randomization of the force levels across two runs per session. In contrast, using an affine registration compared with 3rd to 7th order polynomial warps had only a minor impact on the performance metrics. However, our results make it clear that both prediction and reproducibility metrics are required for optimization as they individually select different optimal pipeline parameter settings that are associated with somewhat different activation patterns. In addition, the parameter settings of components in the pipeline interact so that the current practice of reporting the optimization of components tested in relative isolation is unlikely to lead to optimized processing pipelines.

Methods

Data acquisition

For a detailed description of data acquisition protocols, see La Conte et al. (2003a,b).

Behavioral protocol (the static force paradigm)

Volunteers were visually cued to alternate between resting quietly while passively viewing the visual feedback screen (control state) and applying a randomly presented force level with the right thumb and forefinger to a force transducer (force state). The force levels used were 200, 400, 600, 800, and 1000 g, and the visual stimulus was back-projected onto the bottom one third of a screen at the foot of the scanner couch. Each baseline and stimulus epoch lasted 45 s. Each force level was presented once per fMRI run and was preceded and followed by a baseline period for a total of six baseline periods and five transition and force periods per fMRI run, during which 124 scans were acquired in 8.25 min; two runs were acquired per scanning session, which lasted for less than 1 h. The task was practiced before fMRI data collection outside (and briefly inside) the scanner.

MRI

We used a Siemens 1.5 T clinical scanner with the following acquisition parameters: fMRI EPI BOLD, TR/TE = 3986/60 ms, FOV = 22 × 22 × 15 cm, slices = 30, voxel = 3.44 × 3.44 × 5 mm; MRI: T1-weighted 3D FLASH.

Subjects

Sixteen volunteer subjects were included in this study after screening for motion (maximum pixel movement <0.5 cm), performance of the task, and general image quality. The 16 subjects were composed of 8 men (ranging in age from 25 to 44 years with a mean of 31 years) and 8 women (ages 19 to 44 years, mean 25 years). All subjects tested right-handed with the Edinburgh handedness inventory (Oldfield, 1971) and underwent a neurologic examination as in Muley et al. (2001).

Data processing

Software

The NPAIRS software used for this work is written in IDL™ (Research Systems Inc., Boulder, CO). The NPAIRS algorithm is part of the VAST software library from the VA Medical Center, Minneapolis, Minnesota, and the distributed NPAIRS module may now be run without an IDL license (see http://neurovia.umn.edu/incweb/npairs_info.html).

Preprocessing

After removal of the initial nonequilibrium scans per run, we (1) aligned each fMRI volume and resampled it into a Talairach reference space using either sinc or trilinear interpolation, (2) spatially smoothed these volumes, and (3) removed temporal trends and experimental block effects within a GLM framework. fMRI scan alignment was implemented with the Automated Image Registration program (AIR 5.03, Woods et al., 1998a,b). The anatomic and fMRI data were first stripped to provide a mask of brain-only voxels. After stripping, AIR was used to obtain a 6-parameter alignment transformation for each masked 3D fMRI volume (from both experimental runs), bringing that volume into alignment with the first scan of the first run. Applying the fMRI alignment transformations and averaging the aligned scans per session provided a mean fMRI volume. Talairach resampling was ultimately affected by applying a single sinc or trilinear interpolation step to each fMRI scan derived from the fMRI scan alignment transformation, a mean fMRI-to-structural MRI transformation (6 parameter, AIR 5.03), and a structural-to-Talairach transformation. The structural MRI-to-Talairach transformations were performed with four different increasingly nonlinear transformations (AIR1 = 12 parameter affine, and AIR 3, 5 and 7 = 3rd, 5th and 7th order polynomial warps) to map the structural volume for each subject to a Talairach reference volume. Smoothing was achieved by convolving each axial slice of each volume with a 2D Gaussian kernel with a full-width at half-maximum (FWHM), which took pixel values {0, 1.0, 1.5, 2.0, 3.0, 4.0, 6.0, 8.0} multiplied by the in-plane pixel size (3.44 × 3.44 mm). Temporal detrending was performed, after principal component analysis (PCA; see below), on the PCA-denoised subspace passed to the CVA model by using a linear combination of cosine basis functions within the GLM framework (Holmes et al., 1997). Cosine basis functions and run means constituted the unwanted covariates within a design matrix, and results from the first six columns representing baseline and static force effects, together with the residuals of the GLM model, were retained as the detrended data (see Fig. 1). The number of cycles used per procedure included all half and full cycles up to the following cutoff values {0, 0.5, 1.0, 1.5, 2.0, 3.0 cycles}, where one cycle has a period of 69 s. These high-pass cutoffs should be compared with the 5.5 cycles of baseline-force epochs per run (see Fig. 2). In total, 168 preprocessing combinations were studied (four fMRI to Talairach space transformations with sinc-based interpolation, seven in-plane smoothing levels, and six detrending levels) for each of ten different parameterizations of the CVA model (a total of 1680 different processing pipelines). Seven additional pipelines were studied for trilinear interpolation with an affine between-subject registration and seven smoothing levels, together with the optimal detrending and CVA parameterizations identified in the earlier sinc-based studies.

Resampling and data analysis

Each of the preprocessed data sets described above had transition scans excluded from subsequent analysis so that only steady-state scans within the 45-s control and 45-s force states (neglecting the 4-s “ready” period before each force epoch) were considered; see LaConte et al. (2003a) for details. We did this to increase the maximum CVA cost function, based on the ratio of between-group to within-group covariance, by removing the highly variable transition scans from the within-group covariance. Thirty time points (initial nonequilibrium scans plus transition scans) were excluded from the total 124 scans per run leaving an average of 187 scans/session with 93 or 94 scans/run.

After dropping the transition scans, the remaining scans were each partly preprocessed (i.e., masked, aligned and smoothed), and normalized by their scan means. Only voxels that existed in the AND of the individual subjects’ aligned, brain-only masks were retained for analysis. A PCA was performed on the 2992 scan (16 subjects × 187 scans/session) × 23,389 masked brain voxels’ data matrix, and a “denoised” subspace of 748 principal components (PCs), 25% of the total of 2992 PCs generated by the 2992 scans, was passed on for subsequent resampling, detrending and CVA analysis. For computational efficiency, we computed a single PCA of each partly preprocessed large data matrix and then performed smaller second-level PCA operations on the training and test split-half partitions of the denoised subspace of 748 PCs, as described by Kjems et al. (2002). Note that without the

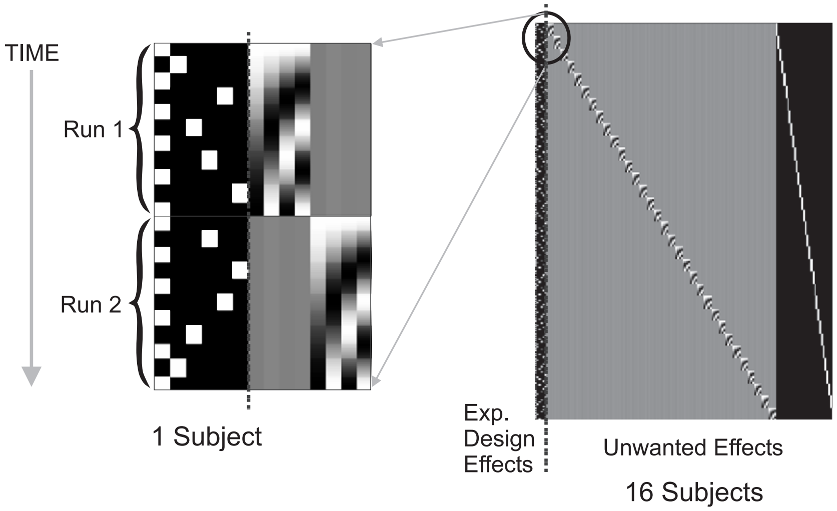

Fig. 1. Illustration of the design matrix used for removal, by regression within the general linear model, of unwanted voxel-based components: (1) low frequency temporal effects removed using cosine basis functions of 0.5, 1.0, 1.5 and 2.0 cycles and (2) mean effects per run. The first six columns illustrate the wanted effects due to the six baseline and five parametric force (columns two to six) epochs (see Fig. 2).

cascaded PCAs to produce invertible split-half data matrices of 1496 scans (2992/2 observations) × 748 PCs (variables) instead of the noninvertible matrix of 1496 scans × 23,389 voxels, the CVA could not be performed. Each 1496 × 748 split-half data matrix was detrended as described above and a flexible 11-class CVA model was applied. As illustrated in Fig. 2, the 11 class labels consisted of six class labels for the temporal order of the control/baseline periods, and five class labels for each of the force levels that were randomized in time for each run. This “agnostic” class structure was used to detect any unknown but consistent parametric force response that existed across runs and subjects. One of the advantages of this data analysis approach for group studies is that it provides an approximate random effects model (Kustra, 2000) with further random effects adjustments for inter-subject noise applied as a result of the Z score normalization within the split-half resampling procedure (Strother et al., 2002). For each split-half group the 11-class CVA analyses were performed using the first 5, 10, 25, 50, 75, 100, 150, 200, 300, and 500 principal components of the possible total of 748 PCs from the second level PCA.

Study of the processing pipeline

Each processing pipeline results in a meta-model that includes the parameters for the preprocessing operations as well as those of the final data analysis stage. In our specific case, an analysis pipeline is composed of the masking, Talairach resampling, smoothing, and detrending operations as well as the PCA and CVA steps. For the denoised subspace from each of the 168 preprocessed sets, NPAIRS was run with 50 split-half resamplings that randomly separated the 16 subjects into two independent 8-subject groups (1496 scans/group). Using CVA parameter estimates from each pair of 8-subject groups, we generated two predictions and one reproducibility metric value. The prediction value generated per group was the median of all of the individual test-scan prediction values obtained when a CVA model built on one 8-subject group (training set) was used to predict the class of each of the 1496 scans in the independent 8-subject group (test set). Test and training sets were then swapped to get the second median prediction value for a given split-half sample. A reproducibility metric was generated for each of the 10 canonical eigenimages (11 classes provide 10 dimensions) from the two 8-subject groups. The box-whisker plots reported in the results below are distribution summaries of the 100 or 50, prediction and reproducibility metric values, respectively, from the 50 split-half resamplings. Curves are plotted through the median values of these distribution summaries to minimize the effect of outlying performance values from particular split-half groups. These distribution summaries do not provide error bars per se as they are made up of correlated estimates from the split-half groups. However, the relative range of the distributions reflects the relative homogeneity of the subjects compared between the split-half groups, as discussed below.

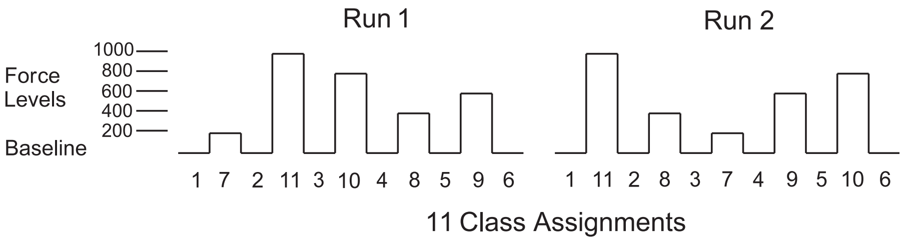

Fig. 2. Graphical depiction of experimental design for two runs per subject of five pseudo-randomized parametric force epochs, with levels of 200 to 1000 g, alternating with baseline epochs. The 11 linear discriminant classes for the canonical variable analysis approach are assigned under the assumption that baseline and force effects are stationary across runs, but follow an unknown, common temporal course within runs.

Results

Figs. 3 and 4 demonstrate the basic behavior of the NPAIRS prediction and reproducibility metrics for the 11-class CVA model as a function of polynomial warp order and within-slice smoothing. Figs. 3A and 4A illustrate the median of 50 split-half prediction medians for an 11-class CVA model built on the first 100 principal components and detrended with 0 and 1.5 cycle cosine-basis-function cut-offs, respectively. In panels B and C, model performance is split into the underlying, uncorrelated canonical dimensions, which are summarized by the NPAIRS reproducibility values of the underlying canonical eigenimages (Figs. 3B and 4B), and for 4.0 pixel FWHM (13.8 mm) smoothing, the canonical variate scores (represented as one dot per subject about the 11-class means that are connected by a solid line, Figs. 3C and 4C), for baselines (scans 1–6) as a function of time, and static force scans as a function of force (scans 7–11, pseudo-randomized in time).

In Figs. 3 and 4, the most striking feature is the rapid rise in both prediction and reproducibility (see dimension one) for small amounts of smoothing from 1 to 2 pixel FWHM (3.4 to 6.9 mm) with broad prediction and reproducibility maxima being attained at 4 and 6 pixel FWHM, respectively (13.8 mm for Figs. 3A and 4A, and 20.6 mm for Dimension 1, Figs. 3B and 4B). Comparing Fig. 4A to 3A, the overall prediction level drops with detrending, and in Figs. 4A and B, there is a tendency for the 5th (blue line) and 7th (red line) order polynomial warps to provide the optimal prediction and reproducibility performance metrics, respectively. Moreover, the box-whisker distribution ranges are generally larger in Figs. 3A, and B (dimension four), compared with Figs. 4A and B (dimension two). This reflects reduced subject heterogeneity as a result of sufficient detrending to remove large low-frequency temporal variations allowing the subtle benefits of the 5th and 7th order warps to be reflected in both the median and range of the performance metrics.

Fig. 4A also illustrates the impact of trilinear interpolation compared to sinc interpolation; by extrapolating horizontally from the 0 pixel FWHM value, we see that trilinear interpolation is equivalent to a little less than 1.5 pixel FWHM Gaussian smoothing, which has a significant impact on model performance. This finding is also mirrored by the reproducibility values and further reinforces the importance of subtle smoothing operations on the data.

In panels B and C, the first dimension clearly represents a strong, reproducible OFF–ON force response, but the detection of a much weaker parametric force response in the higher dimensions is highly dependent on detrending. With no detrending (0 cosine) in Fig. 3C, the 2nd and 3rd dimensions represent reliable baseline temporal trends that appear to interact with the parametric force response despite the pseudo-randomization of the force levels with time across the runs. The marginally reproducible 4th dimension might represent a reliable parametric force response with mean baseline canonical variates that are approximately constant. We found that cosine basis functions with a 1.5 cycle cut-off were required to remove the effects of dimensions two and three (Fig. 3C), as illustrated in Fig. 4C. Dimension one in Fig. 4B has slightly reduced reproducibility, and hence overall Z-score SNRs, compared to Fig. 3B. However, compared to dimension four of Fig. 3, dimension two in Fig. 4 is more reproducible with a clearly linear parametric force response, and mean baseline canonical variates that are constant with value

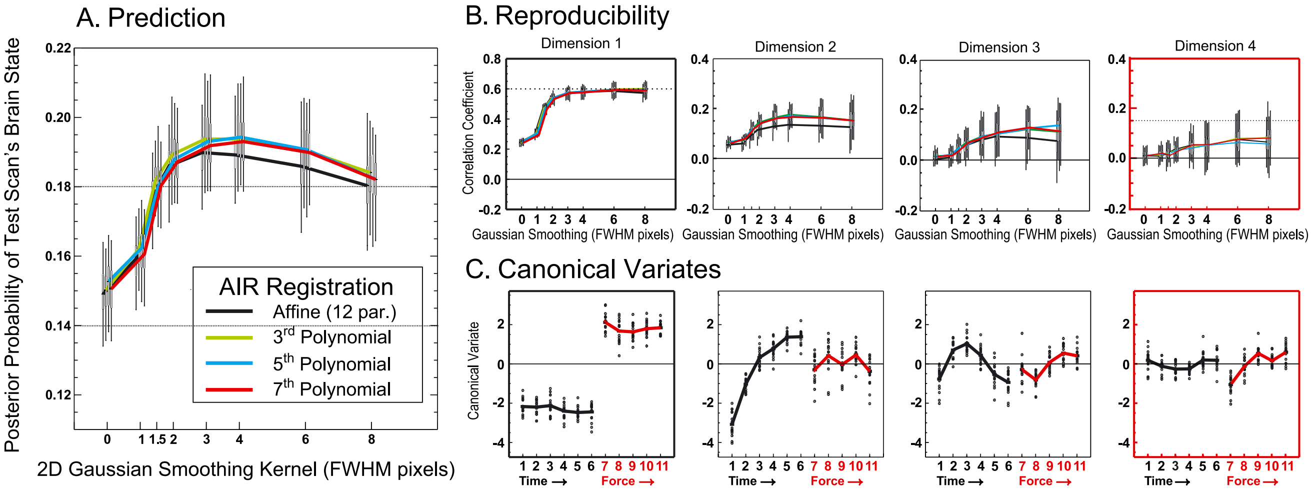

Fig. 3. For no temporal detrending, plots of posterior probability prediction (A), and correlation coefficient reproducibility for the first three discriminant dimensions (B), as a function of within-plane spatial smoothing for Gaussian FWHMs of 0 to 8 pixels (1 pixel = 3.44 × 3.44 mm^2), and four between-subject alignment techniques from AIR 5.03: 12 parameter affine (black line), and 3rd (green line), 5th (blue line), and 7th (red line) order polynomial warps with within-plane sinc function interpolation. The split-half distributions of median prediction, and reproducibility correlation coefficients for the 16-subject group are plotted as thin vertical lines for each combination of parameters. The lines at prediction equals 0.18 (A), and reproducibility equals 0.6 and 0.15 for dimensions one and four (B), respectively, are included for comparison with Fig. 4. Note that in (B) dimensions two to four are plotted with an expanded vertical scale compared to dimension one. (C) For the first four discriminant dimensions, plots of canonical variate class means for each subject (black circle), and their grand means for baseline time courses (black lines), and parametric force responses (red lines), with a Gaussian FWHM smoothing kernel of 4.0 pixels (13.8 mm), no temporal detrending, a 7th-order polynomial warp and within-plane sinc interpolation. The percent variance accounted for by dimensions one to four in (C) is 64.4%, 16.6%, 5.6%, and 3.9%, respectively. In B and C, the panels illustrating (1) the OFF–ON force response of dimension one are highlighted by a thick black outline, and (2) a possible parametric force response without strong baseline-time interactions in dimension four are highlighted by a thick red outline.

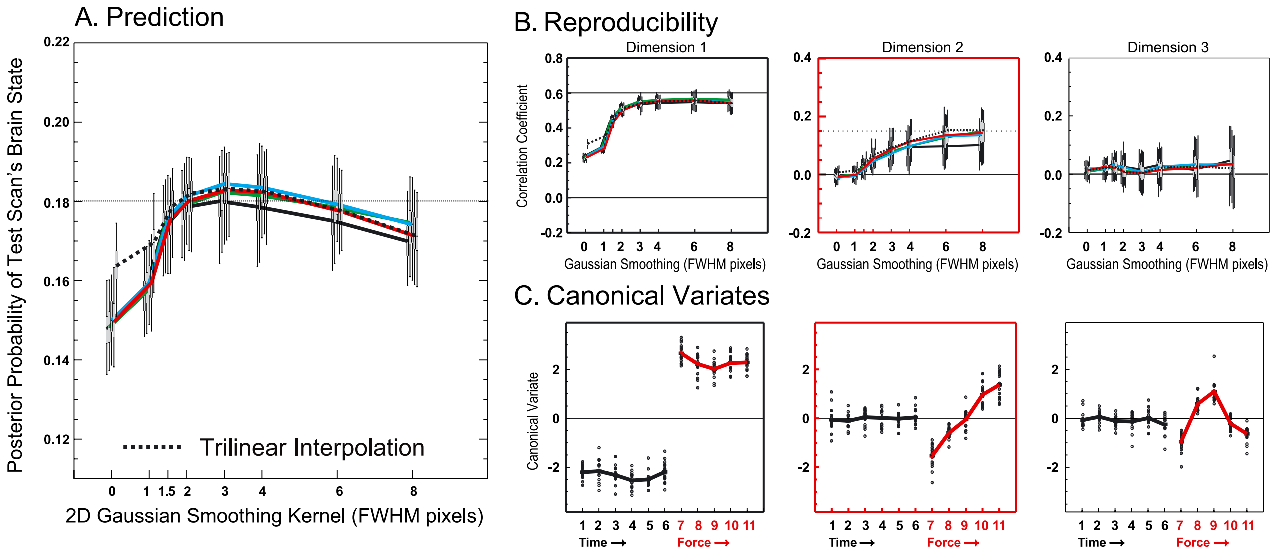

Fig. 4. For temporal detrending with a 1.5 cycle cosine-basis-function cut-off, plots of posterior probability prediction (A), and correlation coefficient reproducibility for the first three discriminant dimensions (B), as a function of within-plane spatial smoothing for Gaussian FWHMs of 0 to 8 pixels (1 pixel = 3.44 × 3.44 mm^2), and four between-subject alignment techniques from AIR 5.03: 12 parameter affine (black line), and 3rd (green line), 5th (blue line) and 7th (red line) order polynomial warps. In A and B, solid lines represent within-plane sinc function interpolation and the dotted line represents trilinear interpolation with an AIR7 warp. The split-half distributions of median prediction, and reproducibility correlation coefficients for the 16-subject group are plotted as thin vertical lines for each combination of parameters. The lines at prediction equals 0.18 (A) and reproducibility equals 0.6 and 0.15 for dimensions one and two (B), respectively, are included for comparison with Fig. 3. Note that in (B) dimensions two to three have an expanded vertical scale compared to dimension one. (C) For the first three discriminant dimensions, plots of canonical variate class means for each subject (black circle), and their grand means for baseline time courses (black lines), and parametric force responses (red lines) with a Gaussian FWHM smoothing kernel of 4.0 pixels (13.8 mm), detrending with a 1.5 cycle cosine-basis-function cut-off, a 7th order polynomial warp and within-plane sinc interpolation. The percent variance accounted for by dimensions one to three in (C) is 81.5%, 6.1%, and 3.4%, respectively. In B and C, the panels illustrating (1) the OFF–ON force response of dimension one are highlighted by a thick black outline, and (2) a possible parametric force response without baseline–time interactions in dimension two are highlighted by a thick red outline.

zero. Furthermore, the approximately quadratic force response seen in dimension three of Fig. 4C is not associated with a reproducible spatial pattern at any smoothing scale (Fig. 4B) and therefore was ruled out as a reliable group response. We have repeatedly seen reproducibility results such as those in Fig. 4B as a function of dimension that clearly and unambiguously indicate the dimensionality of the result; in this case two, with subsequent reproducibility/dimension equal to zero.

Fig. 5A demonstrates the impact of the number of PCs passed to the CVA after the second-level PCA with 4-pixel FWHM smoothing and a fixed cosine detrending cut-off of 1.5 cycles. For 4-pixel FWHM smoothing Fig. 5B shows the effect of the number of cosine basis functions used for a fixed 100 PC subspace. Distributions of test-scan prediction medians for all scans are plotted, together with distributions of the force and baseline scans taken separately to illustrate their quite different behavior. The baseline prediction medians tend to a little below the value 1/6 = 0.167, the value expected if the model can always tell a baseline scan from a force scan, but is completely confused as to which baseline class (1–6) a particular baseline scan comes from, that is, the model performs no better than random guessing for allocating baseline scans to classes 1–6. The small bias below the 0.167 value for truly random baseline scans is probably due to outlying scans with lower probabilities of being a baseline than would be expected for the multivariate Gaussian distributions assumed in the CVA model. More detailed study of subsets of baseline prediction values is required to confirm this. Nevertheless, the prediction medians for baseline classes reach a very shallow minimum for a 1.5 cycle cosine detrending cut-off. This baseline prediction minimum coupled with the canonical variates observation of elimination of components with non-constant baseline means was the reason we chose a 1.5 cycle cut-off as the optimal detrending setting.

Similarly, the force prediction medians start to rise above 1/5 = 0.2, the value expected if the model can always tell a force from a baseline scan but is completely confused as to the true force level of a particular force scan. Even the best prediction values for force scans in Fig. 5 indicate that the spit-half models are confused and unable to reliably distinguish between different parametric force levels. These observations may be generally summarized for all possible pairs of true-class and associated predicted-class labels using confusion matrices as described in Kjems et al. (2002). Despite the low force prediction values, there is a slight peak for a 1.5 cycle detrending cut-off, reinforcing this choice for optimal modeling. Fig. 5B demonstrates why the overall prediction levels fell with detrending in Fig. 4A compared to Fig. 3A. The drop is caused by removing temporal-baseline trends that the model fits and uses to improve overall prediction values (e.g., dimension 2, Fig. 3C). These results indicate the dangers of relying on prediction values alone to judge meta-model performance. In this data set, the additional constraint of obtaining constant baselines means, achieved with a 1.5 cycle detrending cut-off, is required to select an optimal meta-model. After selecting approximately optimal smoothing, detrending, and alignment values of 4.0 pixels FWHM, 1.5 cycle cosine basis cut-off, and a 7th-order warp, the CVA model must still be optimized as a function of the number of PCs used.

Fig. 6 is a prediction–reproducibility plot of dimensions one and two for the prediction medians of the force scans alone as a

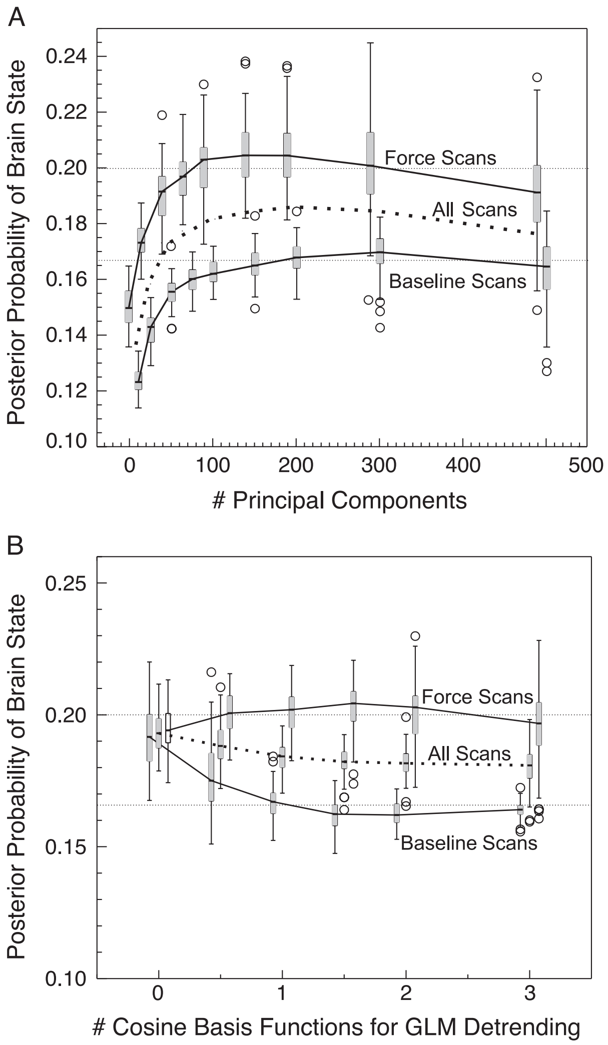

Fig. 5. For a Gaussian FWHM smoothing kernel of 4.0 pixels (13.8 mm), a 7th-order polynomial warp and sinc function interpolation, plots of median posterior probability prediction values for all scans, and force and baseline scans taken alone, (A) as a function of the number of principal components passed to the canonical variates discriminant model with detrending by a fixed 1.5 cycle cosine-basis-function cut-off, and (B) as a function of the cosine-basis-functions cut-off used with 100 principal components passed to the discriminant model.

function of the number of second-level PCs passed to each split-half CVA with the following pipeline parameters: sinc interpolation, 4 pixel FWHM (13.8 mm), 1.5 and 2 cosine basis function cut-offs, 7th order polynomial warp. The prediction values are dimensionless and summarize the complete model, while reproducibility values are dimension specific. The tendency for reproducibility to fall with increasing numbers of PCs (i.e., increasing model parameterization) is clearly seen once a stable subspace has been found at about 50–100 PCs. In addition, as noticed by LaConte et al. (2003a) prediction tends to peak at much higher numbers of PCs than does reproducibility, indicating that reproducibility measures tend to pick different optimal model parameters and their resulting activation patterns from those chosen by prediction measures.

Fig. 7 illustrates the differences in reproducible SPI Z-score patterns for optimal reproducibility compared to optimal prediction. In canonical eigenimage one (Fig. 7A), there are pronounced changes in the activation patterns between 50 PCs (optimal reproducibility) and 200 PCs (optimal prediction), for example, see outlined regions in slices 18 and 25. In canonical eigenimage two (Fig. 7B), note the lack of primary motor response in slices 26–28 with 100 PCs (optimal reproducibility, Fig. 6) compared to the much stronger primary motor response with 200 PCs (optimal prediction, Fig. 6; see outlined regions in slice 26–28). These results provide an example of the pattern differences that may result from focusing on maximal signal-to-noise (i.e., minimizing *p* values), as reflected in the reproducibility correlation coefficient, versus optimal predictive modeling.

Discussion

Our choices for the pipeline components to manipulate in this study were based on some preliminary testing, computational expediency and standard practice in our laboratory. We acknowledge that we have not exhaustively optimized even the components tested, which would require further testing of the preprocessing components (interpolation, smoothing, detrending and warps) for 150 and 200 PCs passed to the CVA to cover the parameterization between optimal reproducibility and optimal prediction performance. Our goal was to illustrate the issues

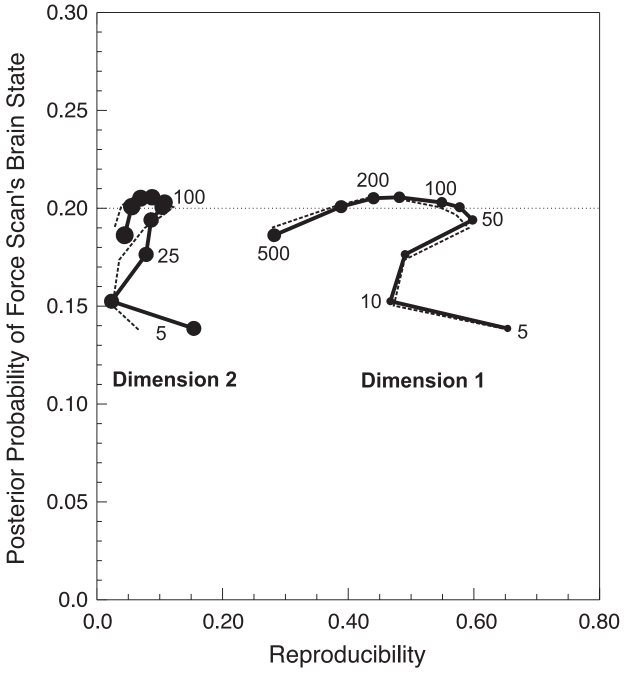

Fig. 6. Plots of the median posterior probability prediction for force scans alone versus correlation coefficient reproducibility values from dimensions one and two of the 11-class discriminant model as a function of the number of principal components (PCs) passed to the model for: a Gaussian FWHM smoothing kernel of 4.0 pixels (13.8 mm), detrending with 1.5 cycle (thick black line), and 2.0 cycle (thin dashed line) cosine-basis-function cut-offs, a 7th order polynomial warp and within-plane sinc interpolation. The line at 0.2 represents the performance expected when randomly assigning scans to the five force classes. Optimal performance is represented by the point (1, 1) with perfect prediction and infinite signal to noise, that is, a correlation coefficient of 1.0. The closest linear distances to (1,1) are 50 and 100 PCs for dimensions one and two, respectively.

involved in pipeline optimization using prediction and reproducibility rather than to produce the final, optimized parametric static force result. Fortunately, the ability to rapidly (i.e., overnight computing on multi-processor arrays) and flexibly set up and test different processing pipelines across multiple software packages is being developed within the Fiswidget (Fissell et al., 2003; Strother, 2003) and LONI pipeline (Rex et al., 2003) software environments. Both tools provide Java-based software environments for incorporating different components from heterogeneous software packages into an fMRI processing pipeline. We have "wrapped" NPAIRS within Fiswidgets with the goal of conducting future optimization studies within this framework.

Our results demonstrate the overwhelming importance of spatial smoothing in fMRI signal detection with the importance of the local pixel neighborhood demonstrated by the sharp rise in prediction metrics for smoothing FWHM of 0 to 1.5 pixels (0–5.1 mm). Reproducibility also rises sharply in dimension one for FWHM from 0 to 2.0 pixels (0–6.9 mm), and continues rising to a shallow peak at 6 pixels (20.6 mm); dimension two does not appear to become reliable/reproducible until the smoothing FWHM reaches 1.5 to 2.0 pixels and gradually rises to a shallow peak at a FWHM of 6.0 pixels for the 7th-order warp (the significance of a particular dimension's reproducibility distribution may be tested using a computationally intensive second level, permutation resampling within each pair of split-half resampling groups, as described by Strother et al. (2002)). As a consequence of the sharp rise in prediction and reproducibility performance for smoothing in a small pixel neighborhood, we have shown that interpolation choices may significantly affect model performance and must be carefully considered. It is tempting to speculate that the 0 to 5 or 6 mm smoothing range identified above represents an approximate matched filter response to an intrinsic smoothing scale composed of BOLD data smearing resulting from reconstruction, physiological (Malonek and Grinvald, 1996), and anatomical–functional smearing. Note that Woods et al. (1998b) demonstrated that 80% of the structural landmarks tested lay within an average distance of 5 mm of each other across subjects for affine through 5th-order warps. However, any such assignment of cause and effect must await further analysis, particularly of single-subject results, which may need to be individually optimized (Shaw et al., 2003a), and

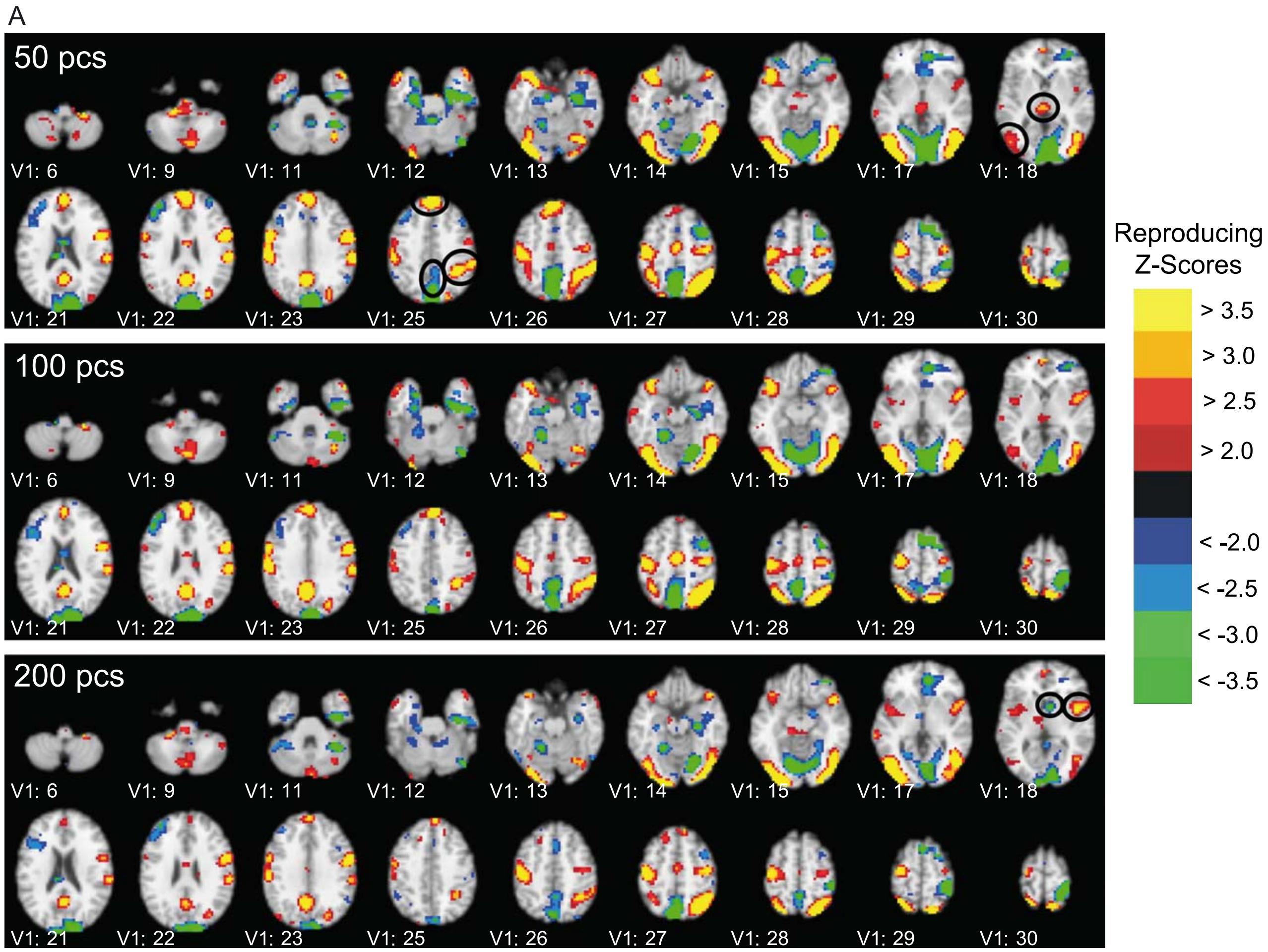

Fig. 7. Selected slices displaying the activation pattern of the first (A) and second (B) discriminant dimensions (i.e., canonical eigenimages) for 50, 100, and 200 principal components passed to the discriminant model with a Gaussian FWHM smoothing kernel of 4.0 pixels, detrending with a 1.5 cycle cosine-basis-functions cut-off, a 7th-order polynomial warp and within-plane sinc interpolation. The first and second dimensions, respectively, reflect the OFF–ON and linear force responses of the canonical variates in Fig. 4C. The regions highlighted in black outlines should be compared across the SPIs for 50, 100, and 200 PCs and are discussed in the text. Image left = brain left.

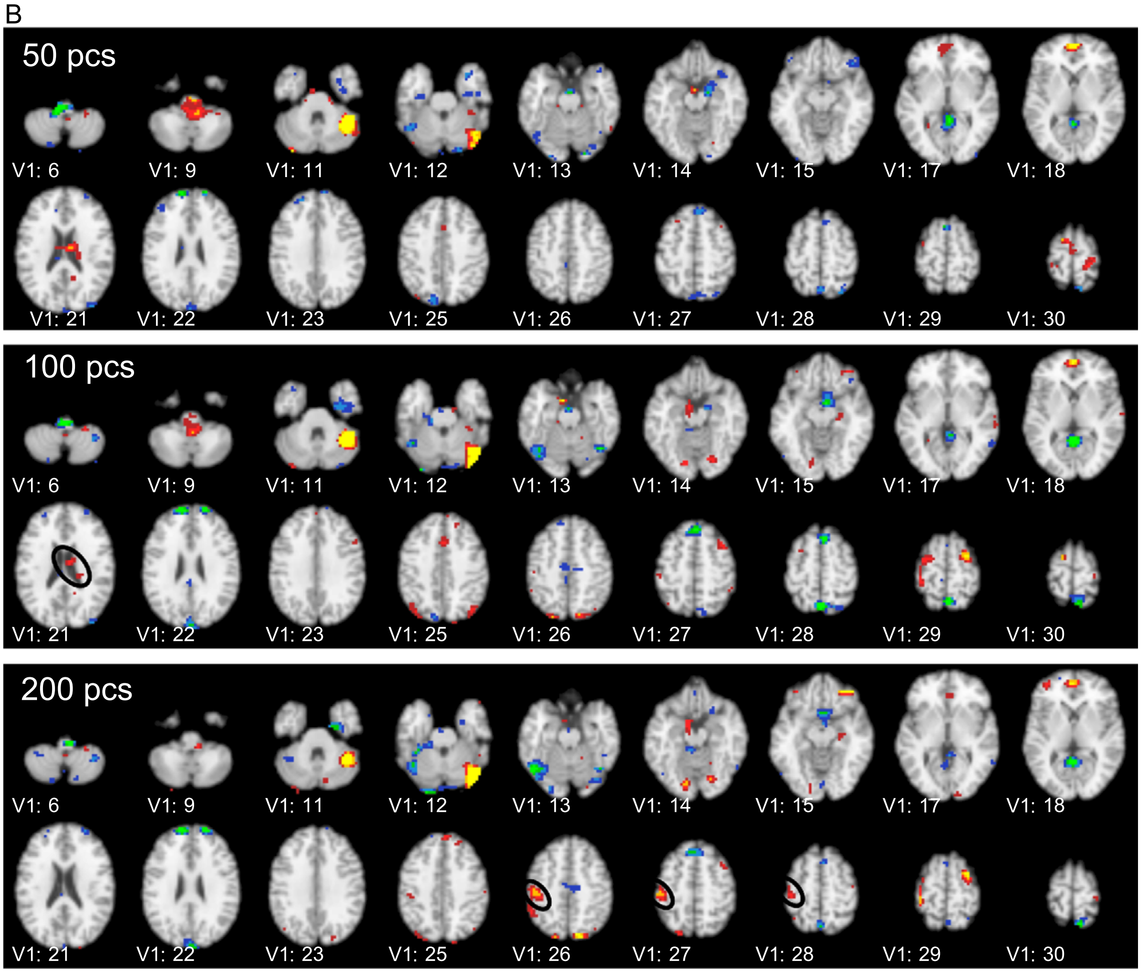

Fig. 7 (*continued*).

structure–function variability as described by White et al. (2001). In the mean time, we note that these results generally support the default spatial smoothing values suggested for the FSL (FWHM = 5 mm) and SPM99 (FWHM = 2–3 pixels) software packages. To obtain any reliable parametric force dimension, some smoothing is essential and up to FWHM = 20 mm may be optimal in terms of the overall activation SNR, although this will probably be unacceptable for some, and perhaps many neuroscience applications with spatially localized hypotheses.

Temporal detrending appears to play an important role in removing reproducing time–force interactions while allowing a reliable parametric force response to emerge, which can take optimal advantage of higher-order warps. Our data clearly show that the pseudo-randomization of the force levels with time across two runs is insufficient to eliminate such interactions, which remain particularly strong without explicit detrending (Fig. 3). By introducing the additional requirement that the baseline means are constant with time, we were able to overcome this problem, and we demonstrated two complementary ways of measuring the elimination of baseline time trends. Experimentally, we observed the elimination of the canonical dimensions with obvious time–force interactions as a function of the amount of detrending applied (Figs. 3 and 4), while simultaneously, a parametric force component with constant baseline-class means emerged. In addition, we noted that constant baseline-class means are equivalent to a maximally confused model that is unable to predict baseline-class membership any better than random guessing; this detrending point may be identified using estimated prediction values for baseline scans alone. Our choice of a detrending cutoff of 1.5 cosine basis cycles was further reinforced by also being the point at which the best predictive model (albeit a poor model) for parametric force values was attained (Fig. 5B). Without the baseline scans available to judge the time–force interactions, it would be very difficult to make any judgement about the meaning of the measured parametric force responses. Reproducibility itself is insufficient to demonstrate a reliable force-dependent response because canonical dimensions two and three in Fig. 3, with clear time–force interactions, are as reproducible as canonical dimension two in Fig. 4. Nevertheless, assumptions requiring the baseline-class means to be constant are potentially restrictive from the perspective of discovering new time-dependent brain responses. Future experimental designs should try to eliminate the need for such assumptions by avoiding spectral overlap between low frequency noise and the fundamental paradigm frequency.

While the reproducibility metric is insufficient to indicate a reliable component that reflects primarily experimental manipulation, it may be used to determine the number of canonical dimensions that must be considered for further interpretation; the final judgment of statistical significance per dimension can be performed by a nested permutation resampling within the split-half resampling (Strother et al., 2002). Determining the number of significant dimensions is a particularly difficult problem for multivariate models (e.g., Beckmann and Smith, 2004; Hansen et al., 1999) and we propose the reproducibility metric as a function of spatial smoothing scale as a means of making this determination. Testing as a function of the smoothing scale is important because weak effects may become considerably stronger with increased smoothing as illustrated for the parametric force response in Fig. 4, dimension two. While prediction metrics may be used to determine dimensionality (Hansen et al., 1999) we believe reproducibility may sometimes have advantages because a relatively stable and reproducible spatial network may be associated with variable subject and/or run-dependent experimental stimulus responses that do not allow for reliable predictive modeling across the group, at least with a linear discriminant model. This seems to be the situation for the parametric-static-force data set analyzed here with a stable, reproducible canonical eigenimage and linear parametric force weights (Figs. 4 and 7) associated with predicted canonical variates that are only marginally better than randomly guessing the parametric force levels. Such uncoupling between reproducibility and prediction, with reliable/reproducible spatial activation patterns associated with poorly predicting models, and perhaps even vice versa, as a function of different types of models (e.g., linear discriminant; support vectors machines, LaConte et al., 2003b) is an important area of future research.

Our results also illustrate the important tradeoff between reproducibility (i.e., Z-score SNRs) and prediction as a function of linear discriminant (i.e., CVA) parameterization; reproducibility tends to peak for models with lower parameterization, well before prediction reaches its maximum value, and the reproducibility of the first dimension peaks at a lower parameterization than the second dimension (Fig. 6). Our reported optimizations of smoothing and detrending were performed with 100 PCs passed to the CVA because this is the closest linear distance to the optimal performance point for dimension two, that is, prediction and reproducibility = (1, 1). However, Fig. 7B raises questions about this choice as canonical eigenimage two for 100 PCs contains a reproducing "artifact" crossing the lateral ventricles in slice 21 (see black ellipse) with a very weak left-sided motor response. One might expect the strong right-sided cerebellar response to be coupled with a clear left-sided primary motor response for this right-handed task. However, for the optimal prediction point at 200 PCs, compared to the pattern for 100 PCs, the artifact has disappeared and the coupled cerebellar-primary motor response is seen. Choosing the 200 PC activation pattern would amount to a selection based on our neuroscientific expectations, the very thing the metrics were introduced to avoid. Reconciling the different activation patterns that are obtained for optimal prediction and reproducibility is an important area for future research. One possibility is to take a consensus of the patterns between the two optimal points following the approach of Hansen et al. (2001). Ultimately, automated techniques relying on nonbiological mathematical constraints may need to be externally validated by carefully chosen, well-established neuroscientific results, but this must be done with great care to avoid the circularity described earlier that reinforces prevailing neuroscientific expectations.

Our results demonstrate the challenge involved in optimizing functional neuroimaging pipelines, even when considering only a subset of the parameters in the meta-model that constitutes the whole pipeline. The five components we considered (interpolation, spatial smoothing, temporal detrending, between-subject registration, and CVA model complexity) clearly interact so that it does not seem reasonable to try to optimize then individually. Nevertheless, when considering the whole pipeline and perhaps multiple techniques and their software implementations for each pipeline component, it currently seems necessary to utilize some form of greedy search optimization to keep the combinatorial explosion of pipeline options computationally tractable. Almost all of the previous processing literature in functional neuroimaging is concerned with optimization of one or two components in isolation, typically associated with introducing a new and "better" procedure for one component, for example, detrending (Tanabe et al., 2002) and registration (Jenkinson et al., 2002; Kjems et al., 1999). This prevailing approach represents the strongest greedy search assumption that there is no interaction between pipeline components other than those being tested, an assumption that our results demonstrate is false. We believe that the functional neuroimaging field should now enter a new phase of testing in which interactions of components and their software implementations are emphasized along with the testing of new procedures and their associated software tools. In this way, it will become possible to design better greedy search approaches that emphasize the key components and their interactions, based on a growing testing literature, as we move towards testing and optimizing complete processing pipelines.

Acknowledgments

Important contributions to this study were made by James Ashe, MD and Suraj Muley, MD of the University of Minnesota and VA Minneapolis Neurology Departments who, respectively, designed the behavioral static-force experiment and recruited and scanned the volunteer subjects. We also thank Xiaoping Hu, PhD and Essa Yacoub, PhD of the Center for Magnetic Resonance Research at the University of Minnesota for designing the fMRI/MRI acquisition strategies and running the MRI scanner, and the anonymous reviewers for helpful and insightful suggestions. This work was partly supported by the NIH Human Brain Project P20 Grant MN57180.

References

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

Carver, R.P., 1993. The case against statistical significance testing, revisited. J. Exp. Educ. 61, 287–292.

Cox, D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19 (2 Pt. 1), 261–270.

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, 19–28.

D'Esposito, M., Deouell, L.Y., Gazzaley, A., 2003. Alterations in BOLD fMRI signal with ageing and disease: a challenge for neuroimaging. Nat. Rev., Neurosci. 4, 863–872.

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 (The Fiswidgets URL is: http://neurocog.lrdc.pitt.edu/fiswidgets/).

Friston, K.J., Poline, J.B., Holmes, A.P., Frith, C.D., Frackowiak, R.S.J., 1996. A multivariate analysis of PET activation studies. Hum. Brain Mapp. 4, 140–151.

Friston, K.J., Josephs, O., Zarahn, E., Holmes, A.P., Rouquette, S., Poline, J.B., 2000. To smooth or not to smooth. NeuroImage 12, 196–208.

Gavrilescu, M., Shaw, M., Stuart, G., Eckersley, P., Svalbe, I., Egan, G., 2002. Simulation of the effects of global normalisation procedures in functional MRI. NeuroImage 17, 532–542.

Genovese, C.R., Noll, D.C., Eddy, W.F., 1997. Estimating test–retest reliability in functional MR imaging: 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, 534–544.

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

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning Theory. Springer-Verlag, New York.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Science 293, 2425–2430.

Holmes, A.P., Josephs, O., Büchel, C., Friston, K.J., 1997. Statistical modeling of low-frequency confounds in fMRI. NeuroImage 5 (Pt. 2 of 4), S480.

Hopfinger, J.B., Buchel, C., Holmes, A.P., Friston, K.J., 2000. A study of analysis parameters that influence the sensitivity of event related fMRI analyses. NeuroImage 11, 326–333.

Jenkinson, M., Bannister, P.R., Brady, J.M., Smith, S.M., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17, 825–841.

Kherif, F., Poline, J.B., Flandin, G., Benali, H., Simon, O., Dehaene, S., Worsley, K.J., 2002. Multivariate model specification for fMRI data. NeuroImage 16, 1068–1083.

Kiehl, K.A., Liddle, P.F., 2003. Reproducibility of the hemodynamic response to auditory oddball stimuli: a six-week test–retest study. Hum. Brain Mapp. 18 (1), 42–52.

Kjems, U., Strother, S.C., Anderson, J.A., Law, I., Hansen, L.K., 1999. Enhancing the multivariate signal of [^{15}O]water PET studies with a new non-linear neuroanatomical registration algorithm. IEEE Trans. Med. Imag. 18, 306–319.

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

Kustra, R., 2000. Statistical Analysis of Medical Images with Applications to Neuroimaging. PhD Thesis, University of Toronto. Available at http://www.utstat.utoronto.ca/~rafal/thesis.ps.gz.

Kustra, R., Strother, S.C., 2001. Penalized discriminant analysis of [^{15}O]water PET brain images with prediction error selection of smoothing and regularization hyperparameters. IEEE Trans. Med. Imag. 20, 376–387.

LaConte, S., Anderson, J., Muley, S., Frutiger, S., Hansen, L.K., Yacoub, E., Xiaoping, H., Rottenberg, D., Ashe, J., Strother, S.C., 2003a. Evaluating preprocessing choices in single-subject BOLD-fMRI studies using data-driven performance metrics. NeuroImage 18, 10–27.

LaConte, S., Strother, S.C., Cherkassky, V., Hu, X., 2003b. Predicting Motor Tasks in fMRI Data with Support Vector Machines. ISMRM, Toronto. May.

Lange, N., Strother, S.C., Anderson, J.R., Nielsen, F.A., Holmes, A.P., Kolenda, T., Savoy, R., Hansen, L.K., 1999. Plurality and resemblance in fMRI data analysis. NeuroImage 10, 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 Process. Mag., pp. 43–45.

Lautrup, B., Hansen, L.K., Law, I., Morch, N., Svarer, C., Strother, S.C., 1995. 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. 137–148.

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

Liou, M., Su, H.-R., Lee, J.-D., Cheng, P.E., Huang, C.-C., Tsai, C.-H., 2003. Bridging functional MR images and scientific inference: reproducibility maps. J. Cogn. Neurosci. 15, 935–945.

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.

Lukic, A.S., Wernick, M.N., Galatsanos, N.P., Yang, Y., Strother, S.C., 2004. Reversible jump Markov chain Monte Carlo signal detection in functional neuroimaging analysis. IEEE Symposium on Biomedical Imaging, Washington, April.

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

Malonek, D., Grinvald, A., 1996. Interactions between electrical activity and cortical microcirculation revealed by imaging pectroscopy implications for functional brain mapping". Science 272 (5261), 551–554.

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

Mjolsness, E., DeCoste, D., 2001. Machine learning for science: state of the art and future prospects. Science 293, 2051–2055.

Moeller, J.R., Nakamura, T., Mentis, M.J., Dhawan, V., Spetsieres, P., Antonini, A., Missimer, J., Leenders, K.L., Eidelberg, D., 1999. Reproducibility of regional metabolic covariance patterns: comparison of four populations. J. Nucl. Med. 40 (8), 1264–1269.

Morch, N., Hansen, L.K., Strother, S.C., Svarer, C., Rottenberg, D.A., Lautrup, B., Savoy, R., Paulson, O.B., 1997. Nonlinear versus linear models in functional neuroimaging: learning curves and generalization crossover. In: Duncan, J., Gindi, G. (Eds.), Lecture Notes in Computer Science 1230: Information Processing in Medical Imaging. Springer-Verlag, Berlin, pp. 259–270.

Muley, S.A., Strother, S.C., Ashe, J., Frutiger, S.A., Anderson, J.R., Sidtis, J.J., Rottenberg, D.A., 2001. Effects of changes in experimental design on PET studies of isometric force. NeuroImage 13, 185–195.

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

Ngan, S.-C., LaConte, S.M., Hu, X., 2000. Temporal filtering of event-related fMRI data using cross-validation. NeuroImage 11, 797–804.

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

Shaw, M.E., Strother, S.C., McFarlane, A.C., Morris, P., Anderson, J., Clark, C.R., Egan, G.F., 2002. Abnormal functional connectivity in post-traumatic stress disorder. NeuroImage 15, 661–674.

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

Shaw, M.E., Waites, A.B., Strother, S.C., 2003. A comparison of methods for evaluating functional neuroimaging results: ROC curves and performance metrics. Int. Conf. on Functional Mapping of the Human Brain, New York. June.

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.

Strother, S.C., 2003. Commentary: a developer's commentary on FisWidgets. Neuroinformatics 1, 131–134.

Strother, S.C., Anderson, J.R., Schaper, K.A., Sidtis, J.S., Liow, J.-S., Woods, R.P., Rottenberg, D.A., 1995a. Principal component analysis and the scaled subprofile model compared to intersubject averaging and statistical parametric mapping: I "Functional connectivity" of the human motor system studied with [^{15}O]water PET. J. Cereb. Blood Flow Metab. 15, 738–753.

Strother, S.C., Kanno, I., Rottenberg, D.A., 1995b. Principal component analysis, variance partitioning and "functional connectivity". J. Cereb. Blood Flow Metab. 15, 353–360.

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

Strother, S.C., Rehm, K., Lange, N., Anderson, J.R., Schaper, K.A., Hansen, L.K., Rottenberg, D.A., 1998. Measuring activation pattern reproducibility using resampling techniques. In: Carson, R.E., Daube-Witherspoon, M.E., Herscovitch, P. (Eds.), Quantitative Functional Brain Imaging with Positron Emission Tomography. Academic Press, San Diego, pp. 241–246.

Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Siditis, J., Frutiger, S., Muley, S., LaConte, S., Rottenberg, D., 2002. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. NeuroImage 15, 747–771 (The URL of the NPAIRS software package described in this paper is: http://neurovia.umn.edu/incweb/npairs_info.html).

Swets, J.A., 1988. Measuring the accuracy of diagnostic systems. Science 240 (4857), 1285–1293.

Tanabe, J., Miller, D., Tregellas, J., Freedman, R., Meyer, F.G., 2002. Comparison of detrending methods for optimal fMRI preprocessing. NeuroImage 15, 902–907.

Tegeler, C., Strother, S.C., Anderson, J.R., Kim, S.-G., 1999. Reproducibility of BOLD-based functional MRI obtained at 4T. Hum. Brain Mapp. 7, 267–283.

Tzikas, D.G., Likas, A., Galatsanos, N.P., Lukic, A.S., Wernick, M.N., 2004 (April). Relevance vector machine analysis of functional neuroimages. IEEE Symposium on Biomedical Computing, Washington.

White, T., O'Leary, D., Magnotta, V., Arndt, S., Flaum, M., Andreasen, N.C., 2001. Anatomic and functional variability: the effects of filter size in group fMRI data analysis. NeuroImage 13, 577–588.

Woods, R.P., Grafton, S.T., Holmes, C.J., Cherry, S.R., Mazziotta, J.C., 1998a. Automated image registration: I. General methods and intrasubject, intramodality validation. J. Comput. Assist. Tomogr. 22, 139–152.

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