NeuroImage **18,** 10–27 (2003)
doi:10.1006/nimg.2002.1300

The Evaluation of Preprocessing Choices in Single-Subject BOLD fMRI Using NPAIRS Performance Metrics

Stephen LaConte,*,†,‡ Jon Anderson,†,§ Suraj Muley,†,§ James Ashe,§ Sally Frutiger,†,§ Kelly Rehm,¶,‖ Lars Kai Hansen,¶ Essa Yacoub,‡,‖ Xiaoping Hu,*,‡,‖ David Rottenberg,†,§,‖ and Stephen Strother*,†,§,‖

**Biomedical Engineering, ‡Center for Magnetic Resonance Research, §Neurology Department, and ‖Radiology Department, University of Minnesota, Minneapolis, Minnesota 55455; †PET Imaging Center, VA Medical Center, Minneapolis, Minnesota 55417; and ¶Department of Mathematical Modeling, Technical University of Denmark, Lyngby, Denmark*

Received January 30, 2002

This work proposes an alternative to simulation-based receiver operating characteristic (ROC) analysis for assessment of fMRI data analysis methodologies. Specifically, we apply the rapidly developing nonparametric prediction, activation, influence, and reproducibility resampling (NPAIRS) framework to obtain cross-validation-based model performance estimates of prediction accuracy and global reproducibility for various degrees of model complexity. We rely on the concept of an analysis chain meta-model in which all parameters of the preprocessing steps along with the final statistical model are treated as estimated model parameters. Our ROC analog, then, consists of plotting prediction vs. reproducibility results as curves of model complexity for competing meta-models. Two theoretical underpinnings are crucial to utilizing this new validation technique. First, we explore the relationship between global signal-to-noise and our reproducibility estimates as derived previously. Second, we submit our model complexity curves in the prediction versus reproducibility space as reflecting classic bias-variance tradeoffs. Among the particular analysis chains considered, we found little impact in performance metrics with alignment, some benefit with temporal detrending, and greatest improvement with spatial smoothing.

INTRODUCTION

Blood oxygenation level-dependent functional magnetic resonance imaging (BOLD fMRI) is a noninvasive method for imaging vascular responses to neural activity that was first reported in the early 1990s (Bandettini *et al.,* 1992; Kwong *et al.,* 1992; Ogawa *et al.,* 1990a,b; Turner *et al.,* 1991). During an fMRI experiment, a time series of brain volume images is acquired while the subject is presented with a stimulus intended to elicit a BOLD response. It is thus possible to assign class labels to each scan corresponding to the type of stimulus present during its acquisition (e.g., stimulus, control). We refer to these designations as brain-state class labels, which can be formalized as a covariate in an experimental design matrix in the general linear model (GLM) framework (Friston *et al.,* 1995c). The time the subject is in the MR scanner is a session, and each repeated fMRI experiment in the same session is an experimental run or an fMRI run. After acquisition, the data are preprocessed (which includes any transformation/filtering steps) and analyzed—most often with the goal of characterizing regions of the brain that changed their activity as a result of the stimulus paradigm. We define the term "analysis chain" as the sequence of preprocessing operations applied to the data and final statistical modeling step. The analysis chain ultimately results in an image of parameter values called an activation map or statistical parametric map (SPM).

The data analysis arena of fMRI research has long focused on finding alternative statistical methods for extracting functional signals or detecting regions of activation (Aguirre *et al.,* 1998a,b; Auffermann *et al.,* 2001; Bandettini *et al.,* 1993; Buchel *et al.,* 1998; Bullmore *et al.,* 1996; Constable *et al.,* 1995; Friston *et al.,* 1995c; LaConte *et al.,* 2000; Lange, 1996, 1997, 1999; McKeown *et al.,* 1998; Ngan and Hu, 1999; Ngan *et al.,* 2000; Petersson, 1998; Rabe-Hesketh *et al.,* 1997; Skudlarski *et al.,* 1999; Tegeler *et al.,* 1999; Worsley, 1997; Xiong *et al.,* 1996). This task has been complicated by the unknown temporal and spatial noise structure of the data and the inability of any one technique to fully describe all facets of the data (Skudlarski *et al.,* 1999). As an alternative to finding one all-encompassing analysis strategy, it has been proposed that multiple models be considered simultaneously (Hansen *et al.,* 2001; Lange *et al.,* 1999; Tegeler *et al.,* 1999). What is sometimes overlooked is the relative impact of the preprocessing components of the analysis chain. Much of the need for preprocessing arises from limitations of the fMRI data acquisition, which include (i) hardware effects such as electronic noise, finite precision of data collection/storage, and sensitivity to physical phenomena of interest (such as the BOLD effect) and (ii) physiologic effects such as patient movement, physiologically derived noise, and the intrinsic nonstationarity and nonlinearity of the brain itself. While much work continues to be performed to improve data acquisition, the complex and poorly understood nature of the data structure makes it difficult to evaluate optimal preprocessing

1053-8119/02 $35.00

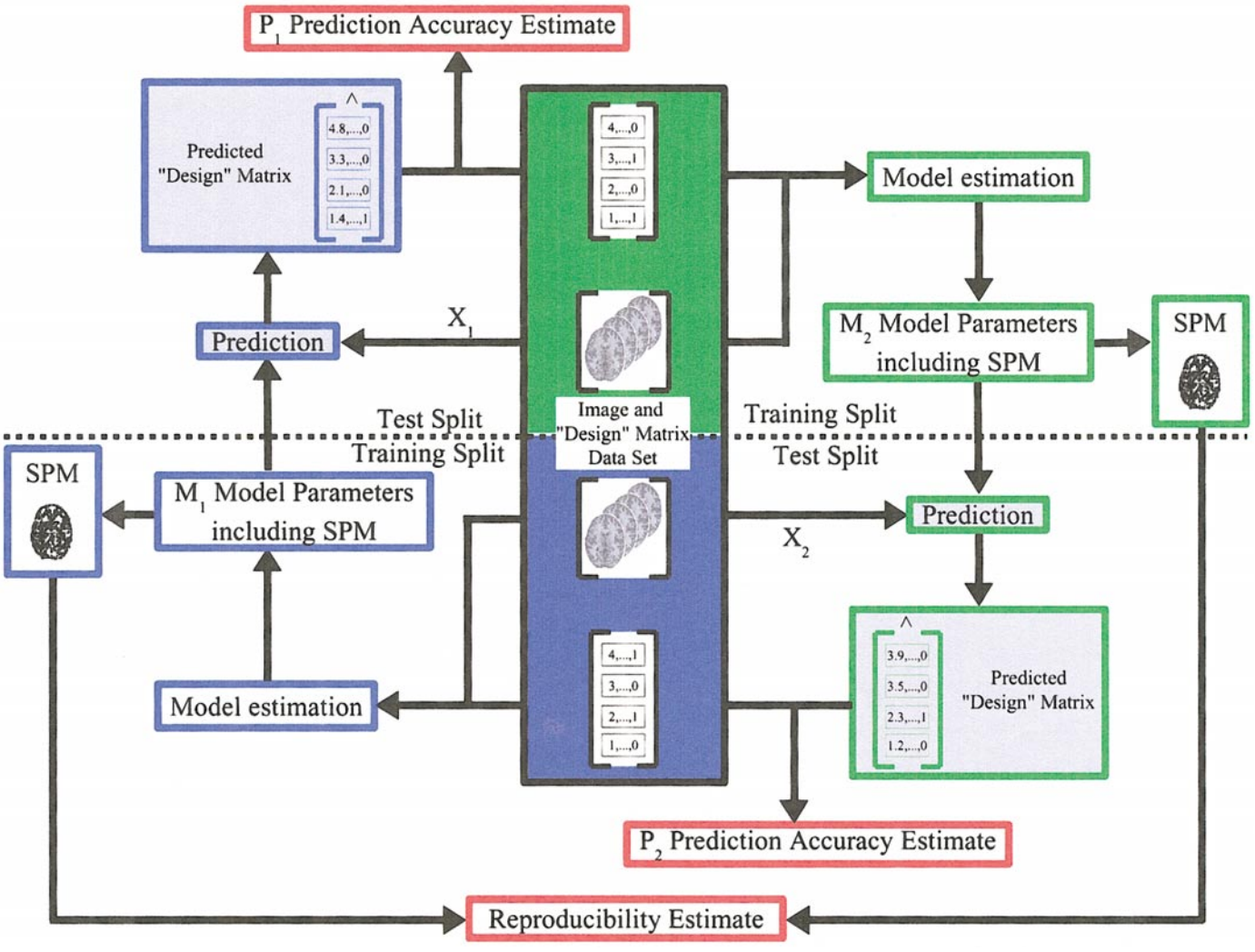


FIG. 1. Split-half resampling used to obtain reproducibility and prediction performance metrics. This figure illustrates the use of split-half resampling to obtain global SPM reproducibility and model prediction accuracy as provided by the NPAIRS framework (see text). A given data set consisting of fMRI image data and a corresponding design matrix are split symmetrically and termed "training" and "test." The training data are used to estimate parameters for a predetermined model. This model is then applied to the test fMRI images to estimate the design matrix for this split. Comparison of the predicted design matrix and the true design matrix led to an estimate of the training model's prediction accuracy. A completely symmetric process occurs by swapping the split designations of training and test, leading to a second model and a corresponding prediction accuracy estimate. Further, a subset of the model parameters from both training splits comprise the SPM. Thus a global pattern reproducibility metric is obtained by comparing the two SPMs.

and data analysis modeling within a theoretical framework derived from first principles.

Considering the difficulties and unknowns inherent in trying to appraise the analysis chain with a theoretical approach, empirical methods of evaluation are an appealing alternative. The most accepted tool is the receiver operating characteristic (ROC) analysis (Constable *et al.*, 1995; Hansen *et al.*, 2001; Le and Hu, 1997; Metz, 1978; Skudlarski *et al.*, 1999; Xiong *et al.*, 1996), measuring a method's accuracy by comparing the true-positive fraction of activated pixels against the false-positive fraction varied over some modeling parameter (e.g., significance level in the case of a *t* test data analysis model). Since this approach aims to discriminate between activated and nonactivated spatial regions, simulated data are required to assess what is "true" and what is "false." Unfortunately, this approach suffers from the same limitations that restrict our ability to start from first principles; it is currently impossible to simulate a completely comprehensive data set since the phenomena contributing to signal and noise components of the data are ill-characterized. In this study of preprocessing methodology, we illustrate an alternative to spatial ROC analysis that avoids simulations by making use of the nonparametric prediction, activation, influence, and reproducibility resampling (NPAIRS) framework (Strother *et al.*, 2002). Specifically we use the measures of SPM reproducibility and model prediction accuracy from known temporal information to evaluate the impact of preprocessing within the analysis chain.

Reproducibility is the ability to repeat an experiment or analysis and achieve consistent results. An important theoretical result demonstrated herein (and consistent with Strother *et al.,* 2002) is that the reproducibility of (unthresholded) activation maps as explored in Strother *et al.* (1997) and Tegeler *et al.* (1999) is monotonically related to the global signal-to-noise ratio (SNR) produced by the analysis chain. Prediction accuracy in neuroimaging has been described in Hansen *et al.* (1999), Kjems *et al.* (2002), Kustra and Strother (2001), Mørch *et al.* (1997), and Strother *et al.* (2002). The idea of formally using independent training and test sets to validate statistical models was introduced by Stone (1974) and is known as cross-validation. This has greatly influenced the rapidly evolving area of predictive learning in statistics (e.g., Cherkassky and Mulier, 1998; Ripley, 1996).

NPAIRS utilizes split-half resampling (a combination of twofold cross-validation and the delete-half jacknife) to estimate reproducibility and prediction by estimating model parameters on half of the data at a time and testing these parameters on the remaining half. To obtain prediction, we model the known temporal evolution of the experiment. If this model generates an SPM, then a comparison of model reproducibility is possible by, for example, correlating the SPMs across splits. Thus the prediction/reproducibility metrics provide an empirical means of methodologic validation that is specific to the data of interest and avoids dependence on simulation. Since we are using temporal classification labels to obtain prediction, it is also possible to perform ROC analysis temporally and substitute prediction for some detectability metric such as area under the ROC curve. Prediction, however, is a more general metric as it is more easily extended beyond the binary classification problem. With either prediction or another detection measure, reproducibility is vital to this framework because it allows us to account for the spatial patterns associated with the temporal model.

It should be stressed that our proposed performance metric framework is quite flexible. To measure reproducibility, any statistical model generating an SPM is sufficient. Prediction accuracy estimates require some assumed truth (e.g., brain-state class labels) that may be used to define a prediction error metric. Canonical variates analysis (CVA), the multivariate extension of Fisher's linear discriminant analysis, satisfies both of these requirements. As has been previously presented (Bullmore *et al.,* 1996; Fletcher *et al.,* 1996; Friston *et al.,* 1995a; Kjems *et al.,* 2002; Kustra and Strother, 2001; Muley *et al.,* 2001; Strother *et al.,* 1996, 2002; Tegeler *et al.,* 1999; Worsley *et al.,* 1997), we apply CVA to fMRI images with brain-state class labels to obtain model parameters, including a SPM. To obtain reproducibility estimates, SPMs from each data split are compared. Prediction measures are estimated by classifying the test data based on the model parameters obtained from independent training data. Our perspective is that the data-driven performance metrics measure the interaction of the final statistical modeling step with all manipulations in the fMRI experiment and the analysis chain. For the purposes of this study, we define an analysis meta-model as including all parameters in the analysis chain defined by all preprocessing parameters and the final statistical model parameters. Note that this approach could also be extended to include all experimental, imaging, and image reconstruction parameters if desired. As our goal is to demonstrate NPAIRS for evaluating the impact of preprocessing decisions within our analysis chain, we perform CVA classification on differently preprocessed versions of the data, obtaining many analysis meta-models to evaluate. To explore and summarize the performance metric results from these meta-models, we utilize a second CVA discriminant analysis of the performance metrics themselves to characterize variations across models.

The targeted preprocessing choices for this study are (1) spatial smoothing, (2) alignment of whole-brain fMRI scans, and (3) temporal detrending. Here, spatial smoothing is used to increase the SNR of the data via spatial averaging, but other reasons for smoothing include allowing for more reliable intersubject averaging and stabilizing results from Gaussian random field analysis (Friston *et al.,* 1996; Poline *et al.,* 1997; Worsley *et al.,* 1992, 1996a,b). The disadvantage of liberal smoothing, of course, is the loss of spatial resolution. Postacquisition alignment techniques have been proposed to mitigate the effect of subject motion artifacts (Woods *et al.,* 1999). Some researchers, however, are concerned that these procedures introduce artifacts of their own [e.g. increasing the strength of autocorrelation structure (Lowe *et al.,* 1998)]. Temporal detrending is used to remove low-frequency drifts and is equivalent to high-pass filtering; however, this also changes the temporal autocorrelation structure of the data (Friston *et al.,* 1995b; Skudlarski *et al.,* 1999; Worsley and Friston, 1995).

Previous studies have examined optimal processing of fMRI data, relying upon ROC analysis. We must be careful to point out that in the following analysis (as in an ROC analysis), we do not claim to have discovered the optimal analysis chain for the data at hand. Instead, we outline a rational means of evaluating and comparing analysis methodologies without reliance upon simulation, and our results suggest several fruitful directions for future study of analysis methodology in our data.

THEORETICAL BACKGROUND

Our application of the NPAIRS framework for obtaining reproducibility and prediction performance metrics for a given meta-model and data set is illustrated in Fig. 1. The data set consists of the preprocessed fMRI image data as well as the corresponding design matrix, which accounts for any known experimental parameters (e.g., the brain-state class labels of each scan). The cross-validation resampling approach generates two sets of final statistical model parameters by alternately designating half of the data as "training." Two prediction accuracy estimates are obtained by applying both training models to the corresponding "test" image data, producing predicted design matrices that are then compared to the test design matrices. One meta-model reproducibility estimate is obtained by comparing the similarity of the two training set SPMs. The final statistical model that we have chosen to illustrate the NPAIRS framework is CVA. Important details of CVA in the context of neuroimaging within the split-half resampling framework, as well as concepts of pre-

diction accuracy and SPM reproducibility, are outlined below.

Canonical Variates Analysis with Principal Components Analysis

Principal component analysis (PCA) for reducing data dimensionality and controlling model complexity as well as CVA for producing linear, multivariate discriminant functions for separating brain-state class labels such as stimulus or baseline scans have been described in previous functional imaging contexts (Bullmore *et al.,* 1996; Kjems *et al.,* 1999, 2002; Kustra and Strother, 2001; Lange *et al.,* 1999; Strother *et al.,* 2002; Sychra *et al.,* 1994; Tegeler *et al.,* 1999). The following illustrates PCA/CVA in relation to the fMRI data space using linear algebra concepts [see Strother *et al.* (2002) and Kjems *et al.* (2002) for a probabilistic treatment of CVA as well as a multivariate statistics text such as Mardia *et al.* (1979) for a general development].

We define our data matrix, $\mathbf{X}$, to have each column correspond to a BOLD image volume at a specific time and each row to the time course of a specific voxel at a specific brain location. In neuroimaging contexts, the number of voxels (M rows) is typically much larger than the number of time scans (N columns), which can be represented by $\mathbf{X}_{\mathrm{MxN}}$. Without loss of generality, we constrain our row time series to be zero mean by removing the mean image volume across each fMRI procedure. In addition we normalize each column brain volume by its mean as in Moeller and Strother (1991). PCA is a convenient means of reducing the dimensionality of the data by producing a square matrix, $\mathbf{Q}_{\mathrm{NxN}}$. We obtain $\mathbf{Q}$ through a singular value decomposition (SVD) of $\mathbf{X}$.

$$\mathbf{U}^{\mathrm{T}}\mathbf{X} = \mathbf{\Lambda}\mathbf{V}^{\mathrm{T}} = \mathbf{Q}. \tag{1}$$

By convention the eigen-time series (the principal components comprising the rows of $\mathbf{Q}$) are ordered by the amount of variance they account for. It is common practice to truncate the latter, small-variance components for complexity control, resulting in $\mathbf{Q}^*_{\mathrm{N^*xN}}$. The standard problem then becomes that of how many components should be removed (Hansen *et al.,* 1999). Keeping too many components is analogous to overfitting, or fitting to the noise, which leads to increased model variance. Keeping too few components, however, corresponds to having a model that is too simplistic to be accurate, which manifests itself as bias. In either extreme, the resulting model does not adequately describe the observed data and is not optimized for describing future observations [two recent examples of these considerations are found in Mørch (1998) and Kjems *et al.* (2002), treatment of sample-size-dependent learning curves].

To perform a CVA of $\mathbf{Q}^*$ we calculate the canonical vector matrix, $\mathbf{L}$, from the eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$ (where $\mathbf{W}$ is the within-class variance and $\mathbf{B}$ is the between-class variance). Thus, $\mathbf{L}$ defines directions that maximize $\mathbf{B}$ while minimizing $\mathbf{W}$. Now calculate

$$\mathbf{C} = \mathbf{L}^{\mathrm{T}}\mathbf{Q}^* = \mathbf{L}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}*}\mathbf{X}, \tag{2}$$

where each row of $\mathbf{C}$ ($\mathbf{c}_i$, $i \in [1,N^*]$) holds the canonical score (or canonical variate) for the ith CVA dimension and the jth column ($j \in [1,N]$) represents the class-labeled observations in the canonical space for the jth image volume. Class membership can be defined by defining threshold boundaries for each canonical score, resulting in separating hyperplanes within the row space of $\mathbf{C}$. We have chosen an alternative classification that lends itself to Bayesian interpretation; each class is viewed as belonging to a multivariate Gaussian distribution in the canonical space (Strother *et al.,* 2002). The columns of the matrix $\mathbf{L}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}^*}$ are termed canonical eigenimages and are the SPMs obtained from PCA/CVA.

Prediction Accuracy and Split-Half Resampling

Resampling methods such as cross-validation are a nonparametric approach used to estimate prediction risk. They do not rely on assumptions about the statistical distribution that generated the data at the cost of being more computationally expensive than derived analytical models [e.g., the final prediction error of Akaike (1970); Ripley, 1998)]. The NPAIRS framework as described in Strother *et al.* (2002) relies on "split-half resampling," defined as twofold cross-validation applied to every possible combination of data splits. For the reproducibility estimates, it is convenient to have symmetric splits; in Strother *et al.* (2002), small but significant reductions in r for 5–3 versus 4–4 spits were observed. In this work, we use two repeated fMRI procedures resulting in only one possible split. In this case, the split-half resampling reduces to twofold cross-validation, the description of which follows from the more general treatment of k-fold cross-validation found in Cherkassky and Mulier (1998) and Efron and Tibshirani (1993).

Step 1. Divide the data, $\mathbf{X}$, into two disjoint samples of similar size. $\mathbf{X} = \{\mathbf{X}_1,\mathbf{X}_2\}$.

Step 2A. Estimate one model (M_1) from $\mathbf{X}_2$ and the other (M_2) from $\mathbf{X}_1$.

Step 2B. Estimate the first prediction accuracy value ($\mathbf{P}_1$) by applying M_1 to $\mathbf{X}_1$, and the second ($\mathbf{P}_2$), by applying M_2 to $\mathbf{X}_2$.

Step 3. Calculate average prediction accuracy by averaging $\mathbf{P}_1$ and $\mathbf{P}_2$.

In terms of the PCA/CVA model in the previous section, applying M_1 corresponds to using the canonical eigenimages ($\mathbf{L}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}^*}$ in Eq. [2] obtained from $\mathbf{X}_2$) to $\mathbf{X}_1$ and using the corresponding separating hyperplanes to classify each class-labeled brain volume. We define prediction accuracy as the posterior probability for each scan's true class membership, using Bayes formula

$$\begin{aligned} &P[\text{true class membership}|\text{test data; training model}] \\ &\quad = (1/K)P[\text{test data}|\text{true class membership; training model}] \\ &\qquad\qquad \times P[\text{true class membership}], \end{aligned} \tag{3}$$

where K is chosen such that the posterior probabilities for each class sum to 1. The likelihood term

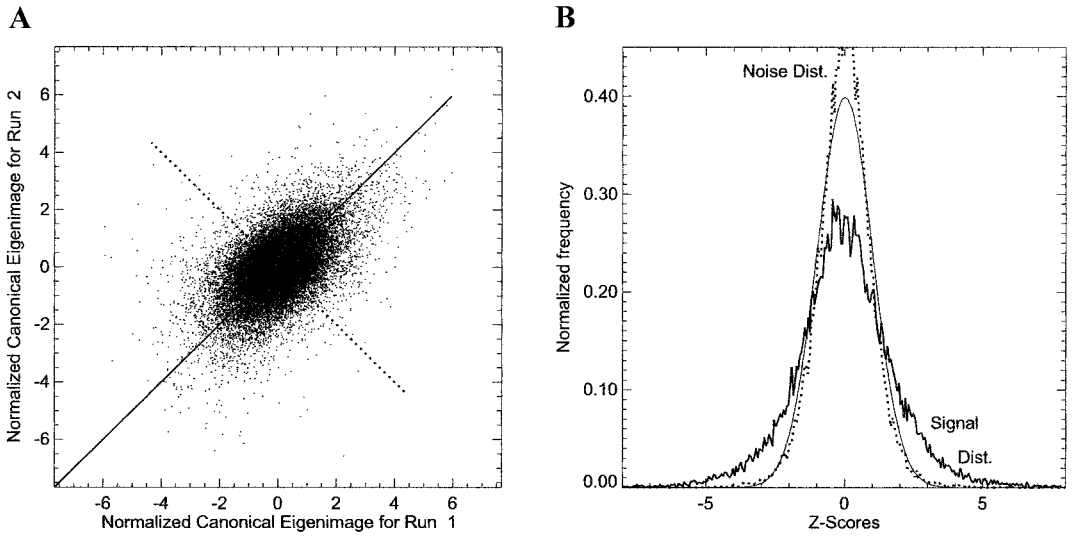


FIG. 2. Global SNR of an analysis. (A) Scatter plot of first canonical eigenimages (SPMs) in runs 1 and 2 for a two-class (force, baseline) CVA. Each data point represents a single voxel. The solid line represents the signal axis and the dotted line represents the noise axis from the major and minor PCA axes of the scatter plot after each axis was normalized by its standard deviation. (B) The signal (solid line) and noise (dotted line) histograms obtained by projecting the scatter plot data onto each corresponding PCA axis and normalizing by the noise axis standard deviation. The thin solid line is the theoretical $N\sim(0,1)$ distribution.

$$P[\text{test data}|\text{true class membership; training model}] = \exp[-\tfrac{1}{2}\| L^{\mathrm{T}}U^{\mathrm{T}*}(x_{\mathrm{te}} - \bar{x}^{\mathrm{c}}_{\mathrm{tr}})\|^2] \quad (4)$$

uses the perspective of each class belonging to a multivariate Gaussian distribution and is dependent on the Euclidean distance between the mean training-set scan for the class, $\bar{x}^{\mathrm{c}}_{\mathrm{tr}}$, and the test set scan x_{te}. The prior probability, P[true class membership], is assigned by the relative frequency of each class in the training data. We scale our prediction accuracy measurements to range from 0 to 1, producing normalized predictions $\mathbf{P}_{\mathrm{n1}}$, $\mathbf{P}_{\mathrm{n2}}$, and their average, $\bar{\mathbf{P}}_{\mathrm{n}}$ (Strother *et al.*, 2002).

Our CVA procedure relies on the number of PCs used to control model complexity. Moreover, we are applying this approach for a wide variety of preprocessing strategies. It has been noted previously (Cherkassky and Mulier, 1998; Friedman, 1994) that a single resampling in the case of complexity control and methodologic comparisons results in an optimistic prediction accuracy estimate. We have not followed the prescribed double resampling here because of constraints in the amount of data and because our aim is to demonstrate the ability to evaluate relative performance for different methodologies rather than focus on the true predictive ability of our models.

Reproducibility and SNR of an Analysis Model

Here we are measuring reproducibility as the correlation between two SPMs. Since the PCA/CVA procedure is only defined up to an arbitrary sign, we use the reference set filtering described in Strother *et al.* (2002), which results in positive values of r (small negative values are possible in cases of low SNR). Strother *et al.* (2002) derived the relationship between SPM reproducibility and the SNR of the reproducible SPM (rSPM). The relationship highlights the fact that the parameters of a given meta-model, including all data analysis model parameters, are subject to uncertainty and gives us some notion of the power of the modeling procedure. The rSPM is obtained from two SPMs (each normalized by its respective SD) whose similarity is in question. When plotted against each other, they produce a scatter plot with each common voxel represented as a data point. Figure 2A demonstrates this scatter-plot concept with results from an individual subject two-class CVA from both run 1 and run 2 as described later under Methods. The rSPM is the projection onto the direction of maximal signal within the scatter plot (the solid line in Fig. 2A). The uncorrelated noise image (nSPM) is defined by the direction perpendicular to the rSPM (the dotted line in Fig. 2A). The signal and noise directions of the scatter plot are found through PCA of the correlation matrix,

$$\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 1+r & 0 \\ 0 & 1-r \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix} \quad (5)$$

and correspond to 45° and 135° with variance $(1 + r)$ and $(1 - r)$, respectively. Note that r is the correlation coefficient of the two SPMs. The utility of this SNR representation is that the rSPM can be interpreted as a z score pattern, denoted rSPM(z), when scaled by the noise axis SD ($\sqrt{1-r}$) under the assumption that the noise distribution is Gaussian.

Moreover, Strother *et al.* (2002) proposed that, if we assume a Gaussian signal distribution, the spread of the tails of the normalized signal histogram could be summarized with the familiar concept of the confidence interval.

$$\mathrm{CI}(z) = (z_{1-\alpha/2} - z_{\alpha/2})\left(\frac{1+r}{1-r}\right)^{1/2} \quad (6)$$

and the approximation

$$\log(\mathrm{CI}(z)_{1-\alpha}) \approx \log(2z_{1-\alpha/2}) + (\log e)\left(r + \frac{r^3}{3} + \frac{r^5}{5}\right) \quad (7)$$

Equation (7) demonstrates that the reproducible Gaussian signal distribution may be thought of as being made up of a fixed noise distribution (with $r = 0$) and a signal that scales approximately linearly with r.

Figure 2B summarizes the scatter plot in Fig. 2A on a z score scale after the major and minor axes have been normalized by the minor axis SD. The thin solid line is the theoretical $N(0,1)$ distribution. The dotted and thick solid lines represent the noise and signal histograms, respectively, both normalized by the noise standard deviation.

METHODS

Data Acquisition

Behavioral protocol (the static force paradigm). The paradigm used for this study was a block design with each run similar to the PET static force protocol 2 described in Muley *et al.* (2001). 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*g*, 400*g*, 600*g*, 800*g*, 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 stimulus lasted 45 s and consisted of two red lines with a static yellow line in between. This was followed by a brief, 4-s transition period indicated by a "GET READY" message, prior to a 45-s force stimulus consisting of high and low boundary lines and a moving white trace line displaying the force applied to the transducer. The force stimulus boundary lines were constant across force level, requiring the subject to quickly adjust to a randomized force by trying to maintain the white trace line within the boundaries. The 45-s force period was ended with a sudden transition back to the static yellow line baseline stimulus. In all, each force level was presented once per procedure and was preceded and followed by a baseline period for a total of six baseline periods and five transition and force periods per procedure. This task was practiced prior to fMRI data collection outside (and briefly inside) the scanner until the subject could reliably stay within the boundary lines at each force level.

MRI. The data for this study were collected on a 1.5-T clinical scanner (Siemens Medical Systems, Iselin, NJ) with a standard quadrature head coil. An initial high resolution T1-weighted anatomical scan was taken using a 3D FLASH sequence [TR = 35 ms; TE = 6 ms; FA = 45°; NEX = 1; FOV = 165 × 220 mm; matrix, 192 × 256; slab thickness, 180 mm; number of slices, 180; voxel dimensions, 0.86 × 0.86 × 1.0 mm; orientation, oblique transverse (axial), 20°; shift mean, 6.4 mm (center of slice relative to magnet isocenter); imaging time, 20 min]. In all but the first four volunteers in this study, a second anatomic scan was acquired after the fMRI runs. This second anatomic MRI was identical to the first except the voxel dimension in the slice direction was doubled (number of slices, 90; voxel dimensions, 0.86 × 0.86 × 2.0 mm; imaging time, 10 min).

The fMRI runs were acquired using an EPI BOLD sequence [TR = 3986 ms; TE = 60 ms; FA = 90°; NEX = 1; FOV = 220 × 220 mm; matrix, 64 × 64; slab thickness, 150 mm; number of slices, 30; number of time points, 135; voxel dimensions, 3.44 × 3.44 × 5 mm; orientation, oblique transverse (axial), 20°; shift mean, 6.4 mm (center of slice relative to magnet isocenter); imaging time per procedure, 9 min].

Data acquisitions for the first four volunteer subjects consisted of the anatomic scan followed by three fMRI runs. Of these, the best two (based on assessment of motion—see "Preliminary Data Analysis") were used. All other subjects had a first anatomic scan followed by two fMRI procedures.

Subjects. Seventeen subjects were recruited from the community surrounding the University of Minnesota Twin Cities campus. Sixteen of the seventeen 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 year) 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 Analysis

The software used for this work was written in IDL. The NPAIRS algorithm is part of the VAST software library (http://neurovia.umn.edu/incweb/npairs_info.html) at the VA Medical Center, Minneapolis, Minnesota.

Preprocessing. As the relative impact of preprocessing on the analysis chain is the focus of this investigation, we outline our generic methodology and its variations. The approach taken here was to (1) align each fMRI volume and resample it into a Talairach reference space (Talairach and Tournoux, 1988), (2) spatially smooth these volumes, and (3) remove confounds by performing volume mean normalization and then removing temporal trends and experimental block effects within a GLM framework.

FMRI scan alignment was implemented with the automated image registration (AIR 3.08) program (Woods *et al.*, 1998). The anatomic and fMRI data were first stripped of scalp, eyeballs, fat, and other structures, providing a mask of brain voxels. After stripping, AIR was used to obtain a six-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 procedure. As an alternative, the case of no fMRI scan alignment was also considered (effectively corresponding to the identity transformation for each individual scan).

Talairach resampling was ultimately affected by applying

a single interpolation step to each fMRI scan. This transformation was derived from the fMRI scan alignment transformation (the identity transformation for the case of no alignment), a mean fMRI-to-structural MRI transformation, and a structural-to-Talairach transformation. The mean fMRI-to-structural MRI (6-parameter) transformation also used AIR 3.08. Applying the fMRI alignment transformations and simply averaging the scans calculated the mean fMRI volume. For the case of no alignment, a separate mean volume for each experimental procedure was obtained. The structural MRI-to-Talairach transformations used 12 parameters 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. The amount of smoothing applied was dependent upon the full-width at half-maximum (FWHM) of the smoothing kernel, which took pixel values {0, 1.5, 6.0} multiplied by the in-plane pixel size (3.44 × 3.44 mm). For simplicity, we refer to these smoothing levels as {no, low, and high} smoothing, respectively.

After volume mean normalization, temporal detrending was performed by using a linear combination of cosine basis functions within the GLM framework as suggested by Holmes *et al.* (1997); cosine and constant terms constituted the covariates within a design matrix and the residuals of the GLM model were retained as the detrended data. The number of cycles used per procedure was {0, 0.5, 2.0 cycles}. In all cases, the DC term (run mean) was also subtracted from each time course. Thus, we referred to the detrending levels as {dc, low, and high} detrending. Note that this is a modification from the procedure reported in LaConte *et al.* (2001); there, run means were only removed in the case of 0.5 and 2.0 cycle detrending. For the "no detrending" case in LaConte *et al.* (2001), run means were removed before the PCA/CVA step for the training data, and this training mean was also removed from the test data (rather than the actual test-data mean).

In total, 10 preprocessing combinations were studied; 1 was no preprocessing (i.e., dc detrending, no smoothing, and no alignment) and the other 9 were combinations of the three detrending and three smoothing levels with alignment.

Preliminary data analysis. As standard practice, we advocate an initial screening of data preceding a full-blown analysis. In many cases this step is as simple as screening the data for motion or surveying a handful of scans for the presence of distortions. In this case, the initial investigation was more thorough and was used as a guiding step for proceeding with the analysis. Our general philosophy was to explore the data set for inherent structure (without imposing a priori knowledge) and then to examine flexible models before committing to any particular model (Bullmore *et al.*, 1996, 2000; Rabe-Hesketh *et al.*, 1997; Strother *et al.*, 1995).

An initial PCA study was used to explore the data of individual subjects by applying it to Talairach aligned volumes with no smoothing or detrending. Possible structures of interest were PCs that appeared to correspond to the experimental stimulus design and aberrant components corresponding to undesirable phenomenon. As part of the initial PCA study, the initial scans before T1 relaxation reached equilibrium were identified and removed. After removing the initial scans, the AIR3 alignment calculations were performed with respect to the new "first" time point, and then new fMRI-to-fMRI and fMRI-to-structural transformations were calculated as previously described. This second alignment was used to estimate maximum and mean pixel displacement to screen for subject motion.

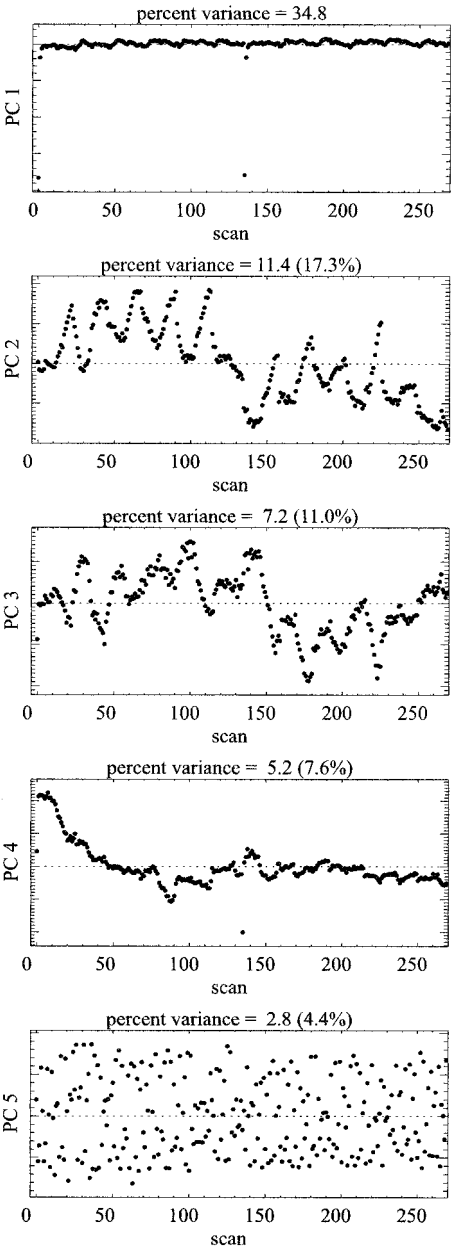


FIG. 3. Exploratory PCA of an AIR3 aligned dataset (no smoothing or detrending). The initial scans before T1 relaxation has reached equilibrium in both run drive the greatest portion of the variance (PC1) and are also apparent in subsequent components. After removal of the first three scans in each run, the components shown here were "promoted" (PC2 became PC1, etc.) with original-to-promoted correlation coefficients of 1.00, 0.99, 0.95, and 0.99, respectively. The percentages of variance explained by promoted components are shown in parentheses.

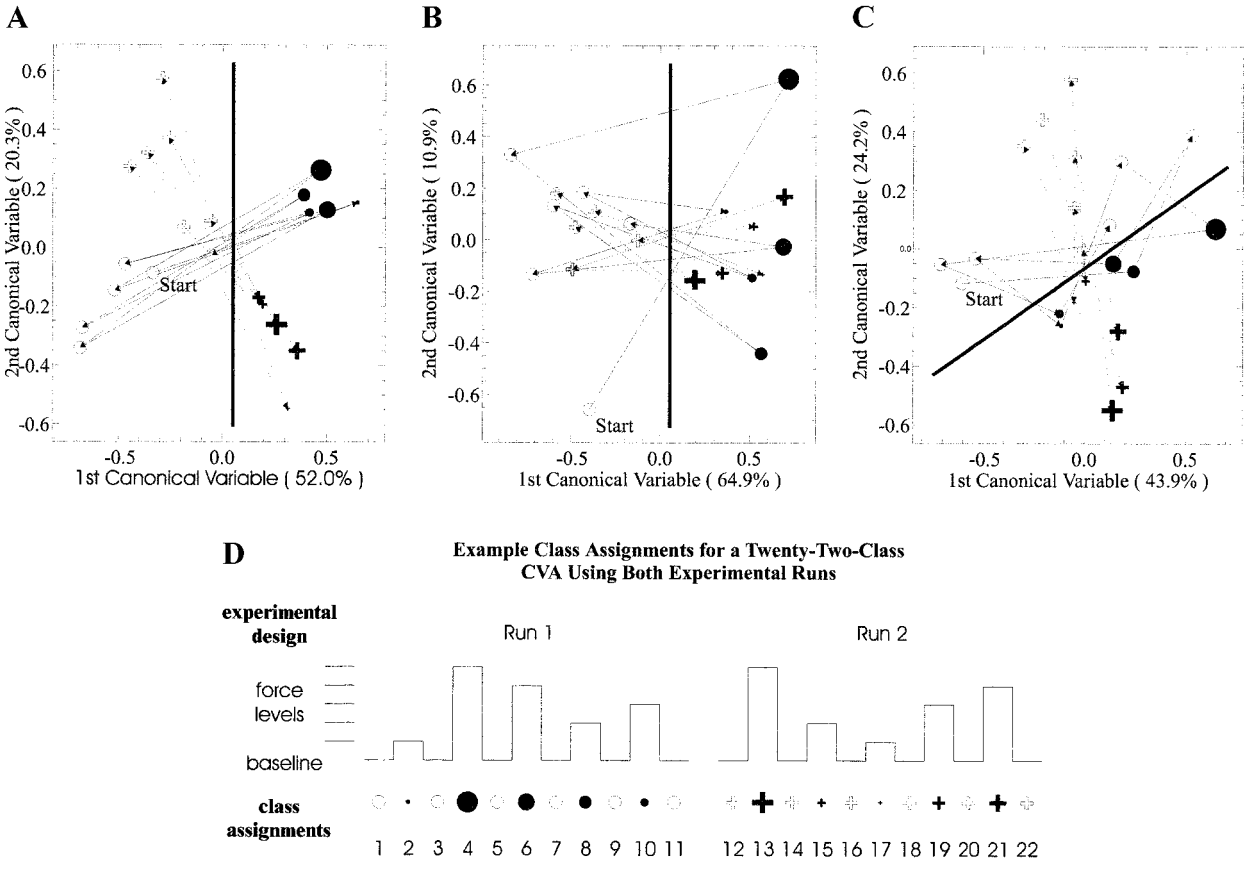


FIG. 4. Twenty-two class CVA of three subjects (A–C) preprocessed with high detrending, low smoothing, and AIR3 alignment. Each block of baseline and force scans in both run was designated by a unique class label (for visualization, force levels are illustrated with their relative symbol size) as in (D). For each procedure, open symbols represent the mean canonical variable values for the scans in each baseline block; closed symbols and their sizes represent the mean force levels. Circles and crosses represent the first and second procedures, respectively. (A–C) Arrows represent the temporal order of block-to-block transitions, and the bold lines each represent one possible discriminant boundary between baseline and force classes.

After removal of the nonequilibrium scans, flexible CVA models were explored. Both data procedures were used to build one 22-class and two 11-class models for each subject. The data were preprocessed with high detrending, low smoothing, and AIR3 alignment. For the 22-class CVA, each control and force period in each run had a unique class label. Class labels for the 11-class CVA consisted of the 6 class labels based on the temporal order of the control periods and 5 class labels for each of the force levels that was randomized in time for a single procedure. Only scans acquired entirely within the 45-s control and 45-s force states (neglecting the 4-s "ready" effect) were considered. Therefore, three to four scans acquired during the transition from control to force (between scans 11 and 14, 33 and 37, 56 and 60, 79 and 82, and 102 and 105), as well as two transition scans from force to control (between scans 23 and 25, 46 and 48, 69 and 71, 92 and 94, and 115 and 117) were excluded from the analysis. The variability of excluded scans arose from slight variations of timing between the fMRI stimulus task control and the scanner acquisition TR of 4 s. On average, 30 time points (initial scans plus transition scans) were excluded from the total 135 scans in each procedure. Both the 22-class and the 11-class CVA models were built on the first 50 PCs (of the average possible 210 and 105, respectively).

Study of the analysis chain. The focus of this article is on the evaluation of preprocessing decisions within an analysis chain, with each analysis chain resulting in a meta-model that includes the parameters for the preprocessing operations as well as the final statistical analysis. In our specific case, an analysis chain is composed of the Talairach resampling, smoothing, and detrending operations as well as the PCA and CVA steps. For each of the 16 subjects, two meta-models (one for each run) were derived for each of the 10 combinations of preprocessing methods described above using five levels of model complexity {10, 25, 50, 75, and 100 PCs}. Thus, the procedure depicted in Fig. 1 was applied 800 times. As in our initial data exploration, transition scans (those not exclusively acquired during only control or force periods) were removed from the PCA/CVA step, leaving approximately 105 scans in each run. Based on results of our initial data exploration (described below) and to avoid the

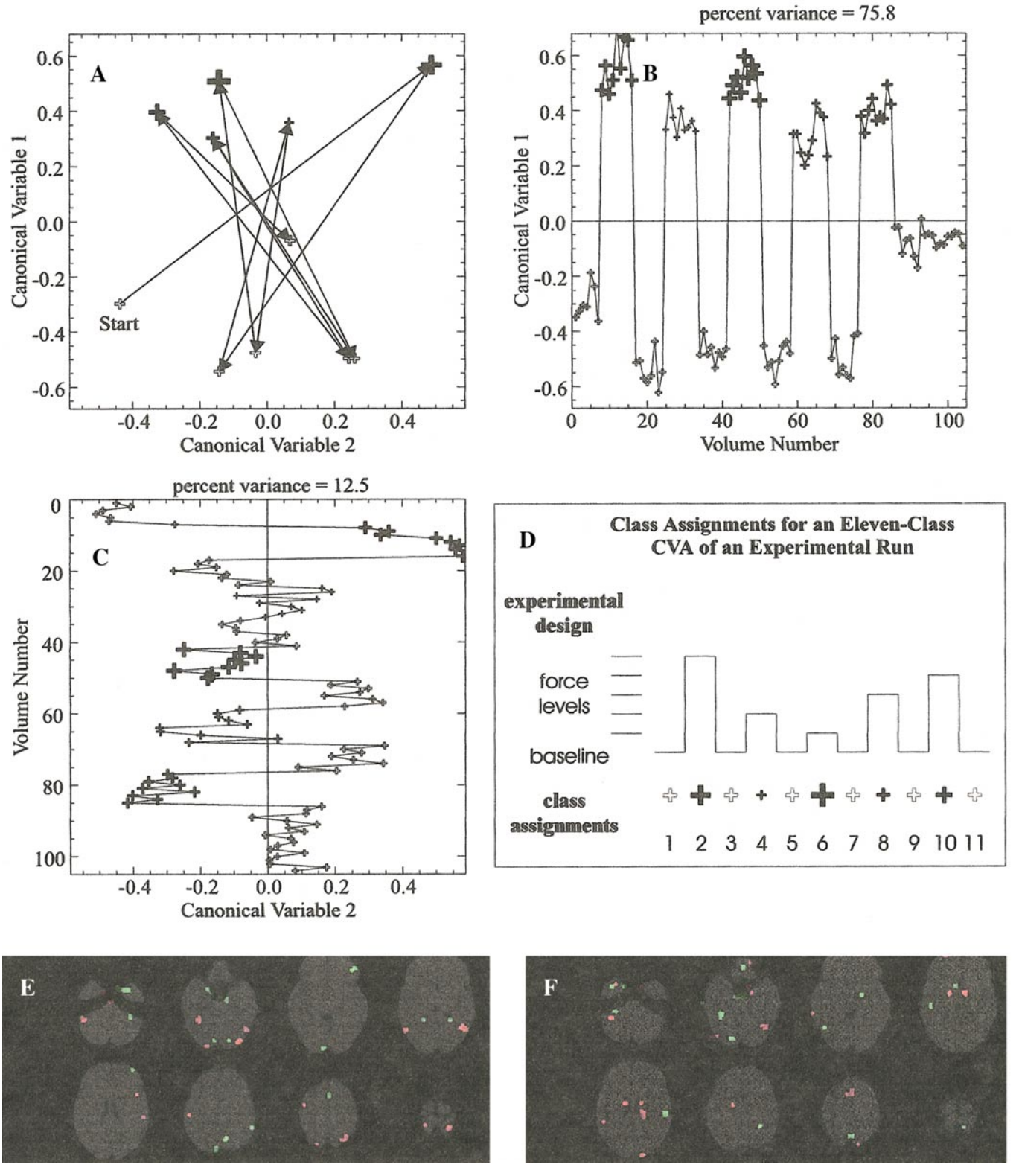


FIG. 5. Eleven-class CVA of run 2 of an individual subject preprocessed with high detrending, low smoothing, and AIR3 alignment. Each baseline segment and force block was designated by a unique class label as shown in (D). (A–D) Open symbols represent baseline, closed symbols represent force (with relative size indicating force level). The connecting arrows in (A) show the temporal evolution of the experiment. Canonical variables one and two are shown in (B) and (C), respectively, and are arranged to illustrate their relation to (A). The top 1% of values from the canonical eigenimages corresponding to canonical variables 1 and 2 are superimposed on anatomic data and shown in (E) and (F), respectively.

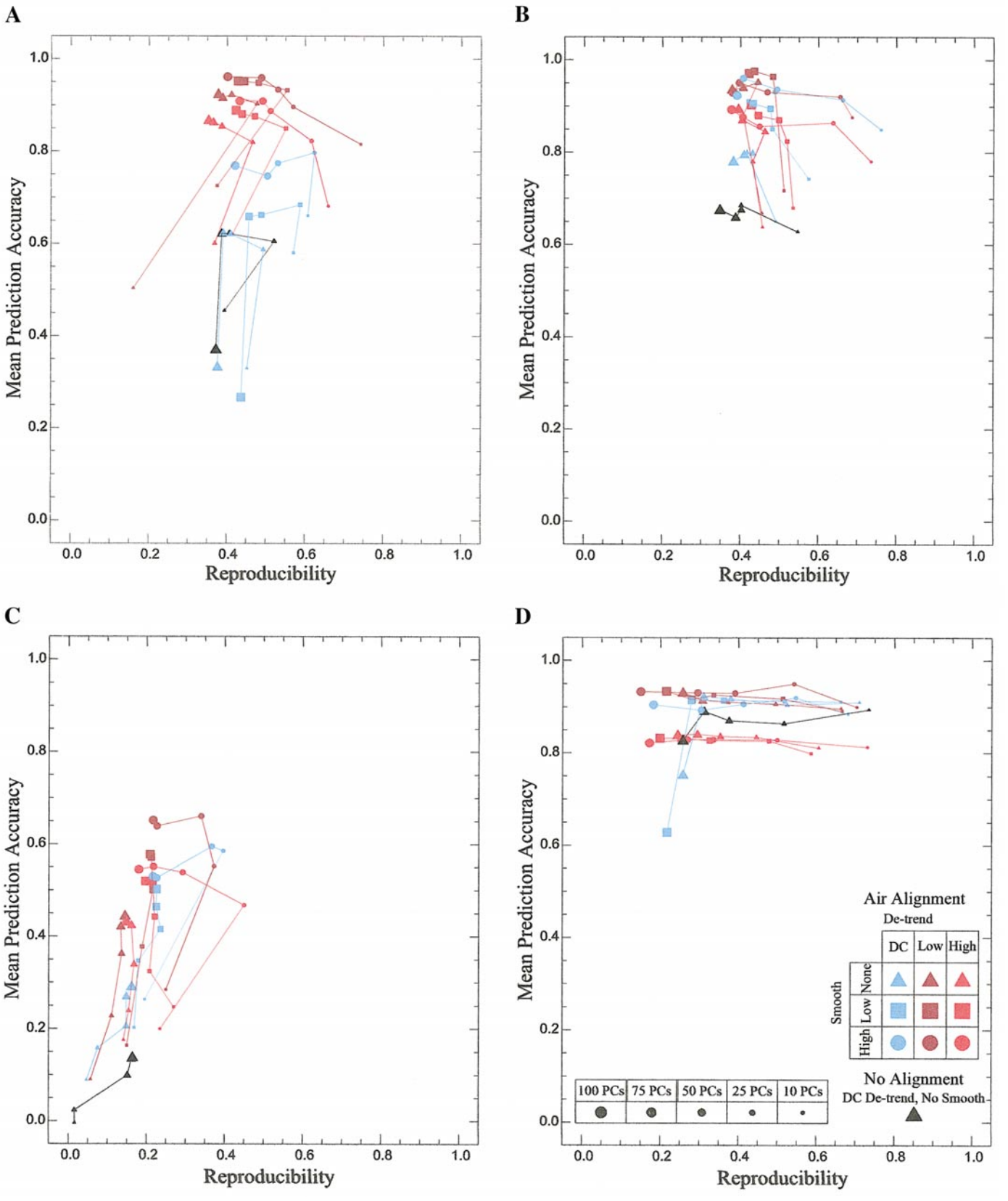


FIG. 6. Prediction accuracy versus SPM reproducibility for four individual subjects (A–D). These plots are the NPAIRS alternative to ROC analysis. The optimal graph location, representing the ideal data set and analysis should provide perfect prediction and reproducibility $(p,r) = (1,1)$. Each curve represents a combination of smoothing, alignment, and temporal detrending analyzed with a range of simple to complex models (i.e., the number of PCs). See key in lower portion of (D). These curves depict a tradeoff between reproducibility, the match to the experimental design structure (prediction), and model complexity (degrees of freedom).

additional complexities of a multidimensional result per meta-model we used a two-class (task and control) CVA statistical model. As provided by the NPAIRS framework, reproducibility and prediction metrics were used to evaluate the meta-model choices. For each subject, the meta-model for each run produced a SPM and the correlation coefficient between the two SPMs was calculated. Similarly, the prediction accuracy for both models were calculated.

RESULTS

Preliminary Data Analysis

Figure 3 illustrates the utility of PCA for separating the fMRI time series into interpretable variance components in the two repeated runs from a single subject. The data are AIR3 aligned without smoothing or detrending. Shown are the first five principal components before removal of the initial scans (those before T_1 relaxation has reached equilibrium). These first three preequilibrium scans in both procedures are clearly outliers and drive the greatest portion of the variance in the data as indicated in the first component and are observable, although subtle in effect, in PC 3 and PC 4. Upon removal of these initial scans in both procedures, the new PCs largely represent a promotion of the originals (PC 2 correlates almost perfectly with the new first PC after the initial scan removal—see the legend to Fig. 3 for precise correlation values). These remaining PCs suggest stimulus-coupled effects (PC 2 and PC 3) and procedure differences (PC 2, PC 3, and PC 4), coupled with equilibration of low-frequency fluctuations such as movement (PC 4) and other higher frequency (perhaps physiologically derived) components (PC 5).

Three individual subject results typifying our findings from the 22-class CVA are shown in Fig. 4 (see Fig. 4D for a graphical reminder of the class structure). Displayed is the $\mathbf{c}_1$, $\mathbf{c}_2$ subspace (the first two rows of $\mathbf{c}$). Recall that these are the two canonical variables that explain the most variance in the $\mathbf{W}^{-1}\mathbf{B}$ matrix and thus define the directions that give the best separation between the chosen class structure while simultaneously minimizing the pooled, within-class dispersion. Each symbol represents the class-mean canonical variable values for the scans in each baseline or force block. Most striking is the nonstationary baseline-force response from run to run. For example, the subject in Fig. 4A shows similar baseline-force pairs within each run, but run 1 is nearly perpendicular to run 2 in the $\mathbf{c}_1$, $\mathbf{c}_2$ plane. In Fig. 4B, the baseline-force response is nearly the same for both procedures other than the first few baseline-force pairs. The subject in Fig. 4C has a run 2 response much like that in Fig. 4A. The first run in Fig. 4C, however, is unusual in that it has three distinct baseline-force pair directions: (i) horizontal (e.g., the baseline one to force one pair), (ii) positive diagonal (e.g., force one to baseline two), and (iii) negative diagonal (e.g., baseline three to force three). The bold lines in Figs. 4A–4C illustrate that the mean baseline-force effect in both runs is linearly separable for each subject. Such a discriminant boundary existed for all but one subject (not shown). Beyond the baseline-force effect, no other structure (such as temporal order of the experimental blocks or separation of the individual force levels) was consistently observed across subjects.

Figure 5 demonstrates the results of the 11-class CVAs (see Fig. 5D) applied to run 2 of an individual subject and illustrates the temporal and spatial information provided by a CVA. Figures 5B and 5C represent the first and second canonical scores. Figure 5A shows the mean class locations in the $\mathbf{c}_1$, $\mathbf{c}_2$ space as derived from Figs. 5B and 5C. Figures 5E and 5F represent the first and second canonical eigenimages obtained from the columns of the $\mathbf{L}^{T}\mathbf{U}^{T^*}$ matrix [Eq. (2)], respectively, superimposed on the anatomic data. Shown in Figs. 5E and 5F are the 1% extreme values (top 0.5% positive values in red and bottom 0.5% negative values in green, which are both shown since sign is arbitrary in CVA). The SPM in Fig. 5E, which corresponds to the time course in Fig. 5B, shows a different pattern from Fig. 5F (corresponding to the time course in Fig. 5C). To qualitatively summarize the results for the other subjects (data not shown), most 11-class CVA results were able to clearly discriminate force and baseline, but showed little other consistently discernible structure.

Study of the Analysis Chain

Evaluation of prediction versus reproducibility. Figure 6 demonstrates the relationship of prediction and reproducibility metrics for four individual subjects. Each curve represents a combination of smoothing, alignment, and temporal detrending analyzed with a range of simple to complex models (i.e., number of PCs). As in an ROC analysis, the prediction versus reproducibility plots have a clear optimal graph location: the ideal data set and analysis chain should provide perfect prediction and reproducibility $(p,r) = (1,1)$. This result, however, is impossible to obtain in practice, as perfect reproducibility ($r = 1$) requires infinite SNR. Thus decisions within the (p,r) space should take into account that each curve depicts a tradeoff between reproducibility (SPM SNR), the match to the experimental design structure (prediction), and model complexity (degrees of freedom). The most striking feature for these single-subject plots is the differences across subjects. In Fig. 6A, sensitivity to the various analysis chains seems to be primarily in the direction of prediction accuracy, with detrending having the largest impact (in the order of dc, high, low). Moreover, within each of these detrending levels, there is an ordering with degree of spatial smoothing. Reproducibility in Fig. 6B tends to be highest with low model complexity, while prediction tends to favor high model complexity. Figure 6C illustrates a clear progression from left-to-right (increasing global reproducibility) with smoothing, as well as a tendency for improved prediction accuracy. Within these trends, higher complexity seems to correspond with prediction accuracy while intermediate complexity (around 25 PCs) optimizes reproducibility. Figure 6D highlights an inverse relationship with complexity and reproducibility that is prevalent for all for all but Fig. 6C, which has relatively low levels of prediction and reproducibility. This relationship indicates a loss of SNR with the more flexible models. Also for Fig. 6D, other components of the analysis chain tend to have relatively little impact on either performance metric.

Figure 7 summarizes the average prediction versus reproducibility results of the 16 subjects. On average, the different preprocessing combinations have a striking effect on both performance metrics, reinforcing our notion that these choices should be optimized. We do not see evidence of any advantage to just alignment, comparing detrending and no spatial smoothing (black triangle and blue triangle curves). For the curves without temporal detrending (blue), there appears to be an optimal model complexity for prediction (that is, a tendency for an intermediate level of complexity to result in a maximum prediction value). For the curves with some detrending, more complex models tend to converge toward better prediction with a large drop in reproducibility. At the same time, simple models sacrifice prediction for reproducibility. This is a classic illustration of a bias-variance tradeoff; high bias (from simple models) tends to favor reproducibility at the cost of prediction, and increased variance (from more flexible models) has the reverse effect. Finally, it is interesting to note that the best performance in terms of optimizing either metric is obtained with heavy smoothing, which results in a general trend upward and to the right in the (p,r) space. This may be unacceptable for many neuroscientific questions, indicating that optimization using these metrics must be performed as a function of spatial scale (i.e., smoothing kernel size). Based on the evidence for converging performance curves (e.g., brown circle-square and red circle-square) we expect different preprocessing and model choices to perform best at different spatial scales.

The subject variability demonstrated in Fig. 6 is so great that the mean curves in Fig. 7 may not provide a meaningful summary. Figure 8 provides a direct multivariate test of the mean differences while allowing for random subject effects (Kustra, 2000). In Fig. 8, preprocessing is used as the class structure for this summary CVA of our individual subject prediction–reproducibility curves. Thus, each data point in Fig. 8A represents a preprocessing curve for an individual subject as described by the matrix of model results versus preprocessing in Fig. 8D. In this matrix, each subject formed a block of preprocessing data vectors consisting of the one r and two p values for all five levels of model complexity. The mean vector for each subject block was removed, and a CVA was applied with results shown in Figs. 8A–8C with the first two canonical eigenvectors accounting for 91% of the variance. The preprocessing class means and 95% confidence circles illustrate that the 2D (p,r) curve shapes seen in Fig. 7 reflect statistically meaningful differences after removal of random subject effects. Unlike the Fig. 5 canonical score time courses, the plots in Figs. 8B and 8C represent classification across subjects (not time) and stack each subject's preprocessing class labels for visualization of the spread about each class mean. The canonical score in Fig. 8B is largely influenced by the three levels of smoothing, creating a ramp for each level of detrending. There is also a mild suggestion of an upward trend with increased degree of detrending (the "smoothing ramp" for dc detrending is lower than the low detrending ramp which is lower than the high detrending ramp). The main effect seen in the canonical score in Fig. 8C is separation of dc detrending from the other detrending levels. Within the higher detrending levels, there is also some influence from smoothing. Some smoothing and/or detrending seems to reduce spread about the preprocessing class means, as we see a greater spread for dc detrending with no smoothing (black and blue triangles) in both Fig. 8B and Fig. 8C, than for the other preprocessing combinations.

The arrows in Fig. 8A indicate the preprocessing choices used to generate the SPMs in Fig. 9; shown are dc detrending, no smoothing, no alignment (Fig. 9A), low detrending, low smoothing, alignment (Fig. 9B), high detrending, high smoothing, and alignment (Fig. 9C). These SPMs also correspond to the appropriate average preprocessing lines in Fig. 7. The SPMs are the average of the rSPM(z)s (the normalized reproducible SPMs) across all subjects and all five levels of model complexity. We are not advocating statistical inference on averaged z scores, but rather we wanted to display meaningful images that would provide some intuition about the relationship of our reproducibility and prediction performance metrics with the resulting SPMs. The three patterns displayed are very similar, and differences seem to largely arise from smoothing. Contributions of detrending and alignment may produce more subtle effects, but it is not possible to claim this from inspecting these average maps.

Global SPM SNR and reproducibility. The densities shown in Fig. 2B are typical of those for the other 15 subjects' two-class CVA results, where we have noticed a consistency in the noise distribution being slightly peaked and having extended tails compared to the $N(0,1)$ distribution. One possible explanation for this phenomenon is that the fMRI noise properties are not spatially stationary. In other words, different regions of the acquired brain volume are noisier than others owing to nonhomogeneous vascular signal contributions or sensitivity to imaging parameters. We intend to study this effect further in later studies.

Figure 10 shows the 16 subjects' reproducibility results versus their rSPM and nSPM confidence intervals for: dc detrending, no smoothing, no alignment, and medium model complexity (75 PCs); low detrending, high smoothing, alignment, and high model complexity (100 PCs); and low detrending, high smoothing, alignment, and low model complexity (25 PCs). These plots allow us to summarize the information in Fig. 2 arising from many subjects compared with the solid lines that represent theoretical values for a Gaussian distribution from Eq. (7). Values pertaining to the spread of the noise histograms are on the x axis of these plots since the nSPMs, by definition, have reproducibility values of zero. Figure 10A demonstrates the least Gaussian noise distribution with the longest tail (experimental 99% values $>$ Gaussian) and the most peaked center (experimental 90% values $<$ Gaussian). In Fig. 10B, the preprocessing and increased model complexity have generated a more consistent, more Gaussian-like noise distribution, with increased prediction (Fig. 7) and no change in global SNR as reflected in the reproducibility values of Fig. 10B versus Fig. 10A. Compared to Fig. 10B, reducing model complexity in Fig. 10C generates a less consistent noise distribution with a longer tail, similar prediction (Fig. 7) but much higher reproducibility and hence global SNR values. Overall, the offset from the theoretical Gaussian results seems dependent on the analysis chain and

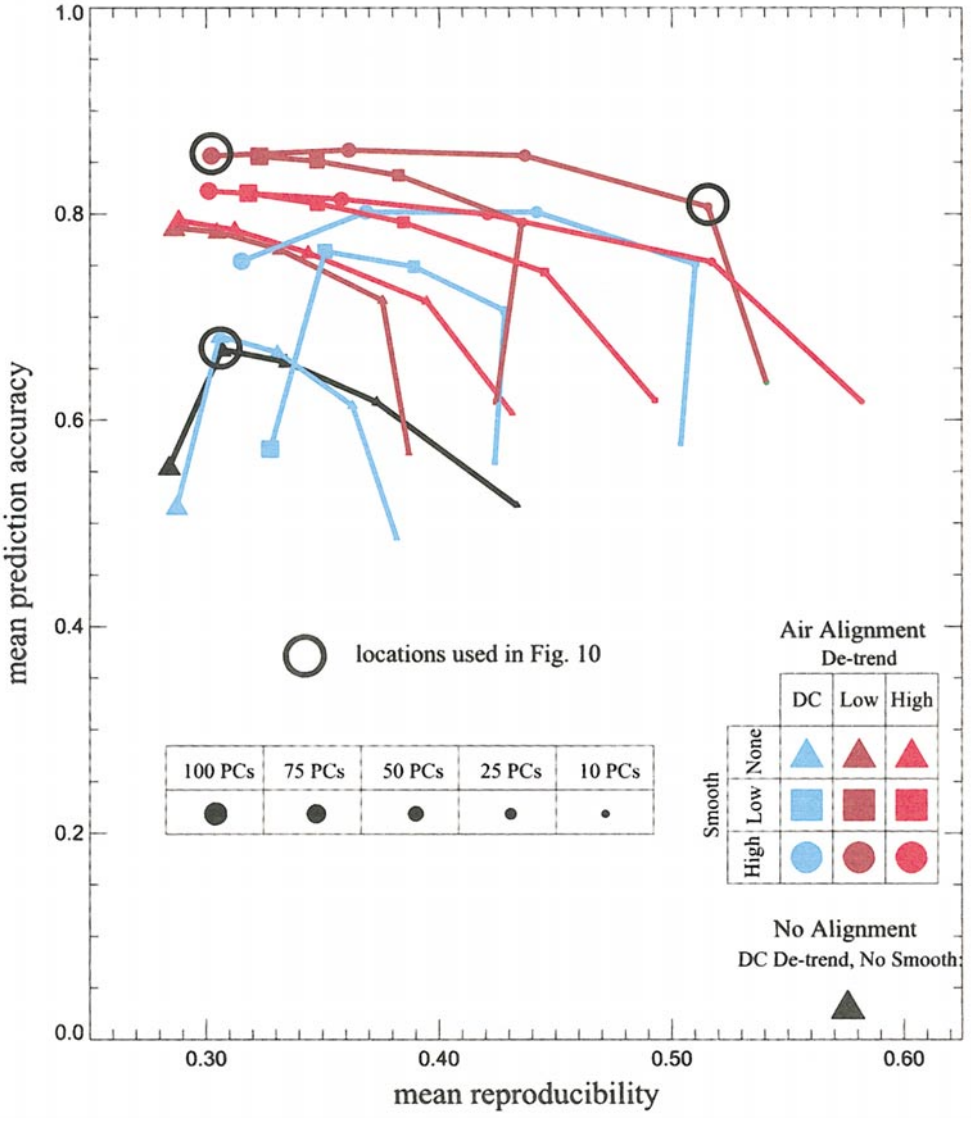


FIG. 7. Prediction accuracy versus SPM reproducibility averaged across all 16 subjects. See key in lower portion.

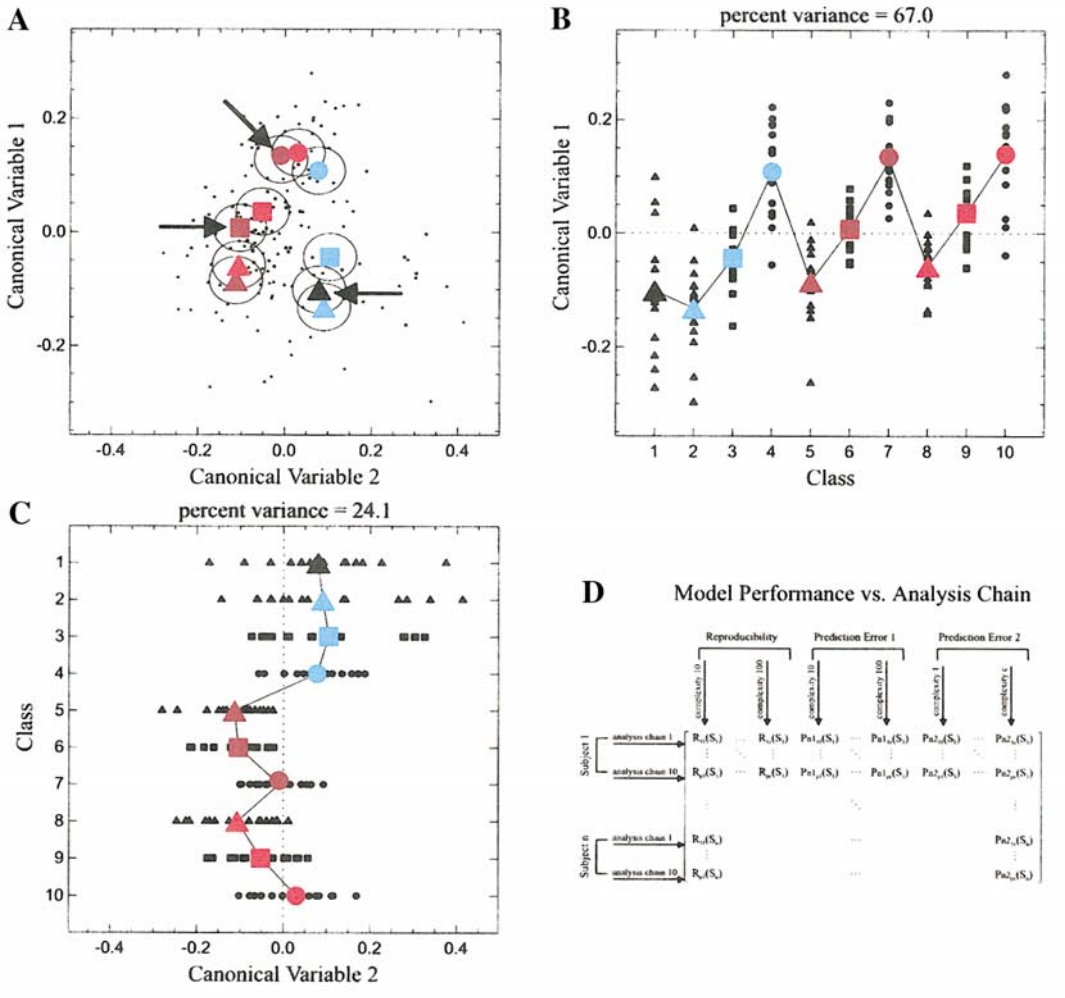
is remarkably consistent across subjects for a broad range of reproducibility values.

DISCUSSION

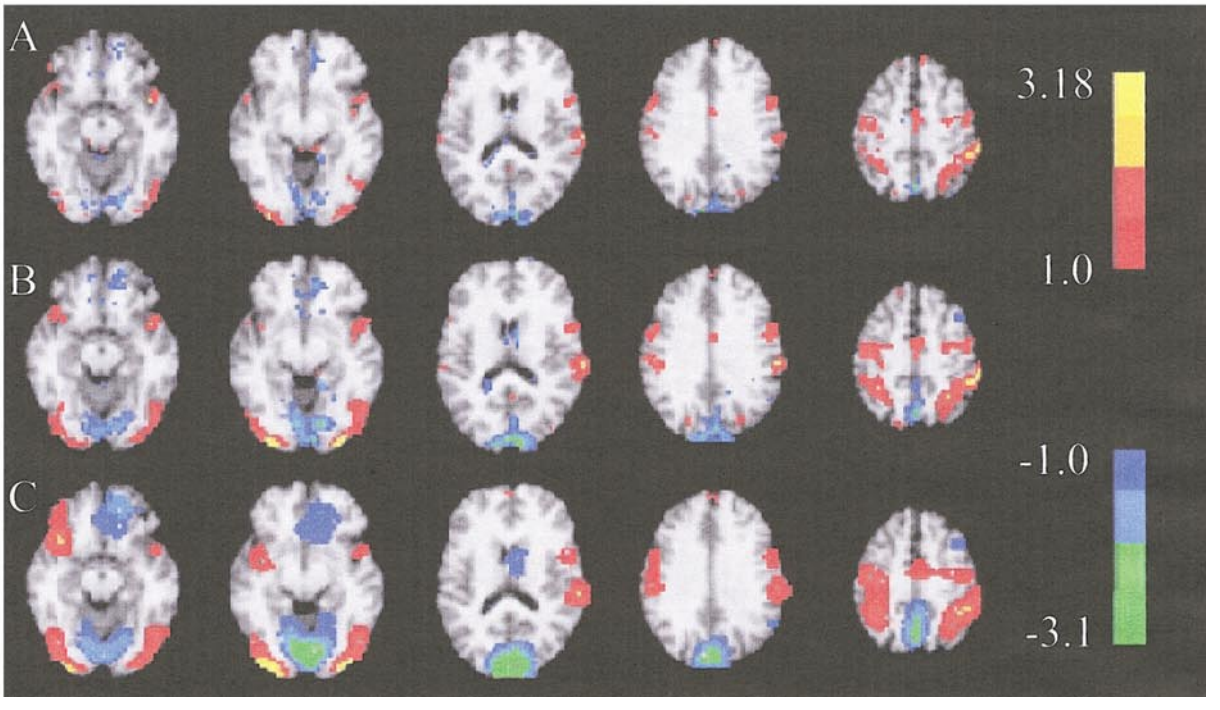
We have demonstrated a flexible data analysis framework for appraising various analysis chains for individual subjects with repeated procedures given the NPAIRS performance metrics, namely, prediction accuracy and reproducibility. These performance metrics provide complementary information about the quality of a given meta-model by making use of test set validation. The NPAIRS framework can easily be broadened to compare experimental variations across functional tasks as well as across multiple subjects (Strother *et al.*, 2002).

FIG. 8. CVA summary of model performance versus preprocessing. A CVA was performed on our performance metric result data using the data matrix defined in (D). The 10-class structure consisted of the 10 preprocessing combinations used, and each class had 16 members (the 16 subjects). The variable space consisted of the reproducibility measure and two prediction accuracy estimates obtained for each level of model complexity. (A–C) Large symbols represent mean locations, while small symbols represent actual data points. The symbol shapes themselves distinguish the 10 preprocessing classes and correspond to the plots in Figs. 6 and 7. Arrows demark analysis chains displayed in Fig. 9.

FIG. 9. Average of the normalized reproducible SPMs (rSPM(*z*)s) for the 16 subjects and five levels of model complexity. Shown are dc detrending, no smoothing, no alignment (A); low detrending, low smoothing, and alignment (B); and high detrending, high smoothing, and alignment (C). Average of the normalized reproducible SPMs (rSPM(*z*)s) for the 16 subjects and five levels of model complexity. Shown are dc detrending, no smoothing, no alignment (A); low detrending, low smoothing, and alignment (B); and high detrending, high smoothing, and alignment (C).


A
Canonical Variable 1
Canonical Variable 2
B
percent variance = 67.0
Canonical Variable 1
Class
C
percent variance = 24.1
Class
Canonical Variable 2
D
Model Performance vs. Analysis Chain
Reproducibility
Prediction Error 1
Prediction Error 2
Subject 1
Subject 6

8


A
B
C
3.18
1.0
-1.0
-3.1

9

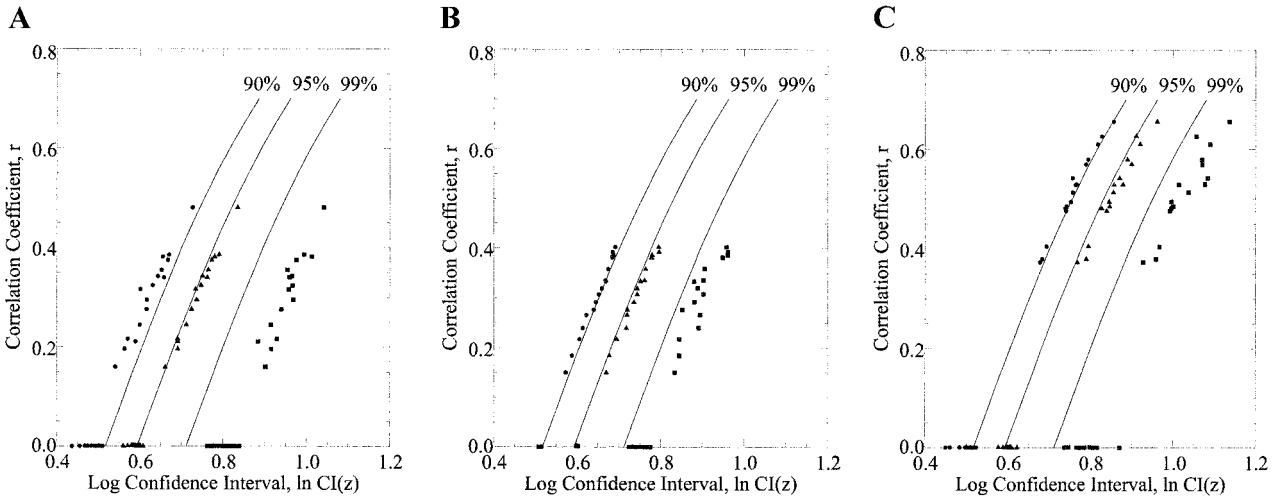


FIG. 10. SPM reproducibility versus log confidence interval. Two correlation coefficients were calculated from scatter plots as in Fig. 2 of the first canonical eigenimages in both runs for a two-class (force, baseline) CVA and correspond to the rSPM (major axis) and the nSPM (which are defined by the direction of the minor axis and have correlation coefficients of zero by definition). Confidence intervals (width of histogram) were calculated using histograms as illustrated in Fig. 2 (those corresponding to zero correlation were obtained from the noise histograms). Circles, triangles, and squares correspond to an individual subject's confidence intervals of 90, 95, and 99%, respectively. (A) A preprocessing of no detrending, no smoothing, and no alignment with model complexity of 75 PCs. (B) Low detrending, high smoothing, AIR3 alignment with 100 PCs; (C) Low detrending, high smoothing, and AIR3 alignment with 25 PCs.

Our general methodologic development was intended to effectively demonstrate the flexibility of NPAIRS. The preliminary multivariate data analysis allowed us to screen the quality of our volunteer data for motion, image quality, and performance of the experimental task. We also very quickly arrived at an appreciation for the variability of the data across runs and across subjects. In addition, this analysis step allowed us to evaluate the type of model best suited to demonstrate the NPAIRS methodology. During our study of the analysis chain, the choice of the CVA class structure could have been parameterized within the resampling exercise along with the rest of the analysis chain. The decision to evaluate preprocessing using a two-class model arose from several factors. The initial 22-class and 11-class model results demonstrated a baseline force effect that was robust for virtually every subject. In addition, the primary goal of this work was to demonstrate our performance metric-based approach for evaluating preprocessing. We therefore focused our energy on finding a suitable model for demonstrating this framework rather than the ideal model from the perspective of a neuroscience interpretation.

We acknowledge that our estimates of prediction accuracy are biased from a pure machine-learning point of view since we have resampled for model complexity without an additional resampling for prediction accuracy [a situation that is known to lead to optimistic estimates (Cherkassky and Mulier, 1998; Friedman, 1994)]. We feel, however, that this procedure makes sense for relative comparisons in this neuroimaging setting—it is very natural to treat repeated runs or individual subjects as independent units. For the case treated here, more experimental runs than the two we collected would be necessary for a second resampling estimate of the true prediction accuracy value. For complex functional tasks, it is very difficult to obtain several runs of high quality (in terms of motion and independent behavioral measurements), and long scanning session times introduce additional concerns over stationarity issues of both the scanner and the weary volunteer. Further, the true prediction accuracy results are only of secondary importance—what is necessary for these studies is the relative impact of prediction accuracy for each methodologic decision.

It is also interesting to note that, unlike all other machine-learning settings we are aware of, our model selection is not solely based on prediction. For our data, the global SPM reproducibility metric often acts as an additional penalization against complex models. Most cases reported in Figs. 6 and 7 illustrate that complex models tend to sacrifice reproducibility and global SNR, even if prediction is improved. In a few instances, however, we saw the opposite effect (e.g., Fig. 6C). Within the NPAIRS framework the prediction versus reproducibility curves of Figs. 6 and 7 represent a viable, data-driven alternative to ROC analysis for evaluating methodologies. As with ROC, there is one optimal graph location; the ideal data set and analysis should provide perfect prediction and reproducibility $(p,r) = (1,1)$. Barring the ideal case of both perfect prediction and reproducibility, choosing one analysis chain at the exclusion of several others requires careful consideration. Is the point (0.6,0.6) better or worse than (0.55,0.85)? It is not clear that Euclidean distances are appropriate within this space, especially since points close together in the p–r space can originate from vastly different models as is most easily appreciated by viewing canonical eigenimages arising from different levels of smoothing. Ultimately, choosing a model from these curves represents a bias-variance tradeoff, with simple models tending toward high bias (lacking the degrees of freedom to adequately describe the data) and complex models tending toward in-

creased variance (having the flexibility to incorporate spurious features). One solution may be to use consensus methods (Hansen *et al.,* 2001) by combining a subset of competing meta-models that are closest to the ideal.

Figures 6–8 attempt to summarize the relative performance of our candidate analysis chains for this single-subject study. The results for Fig. 6 indicate that competing analysis chains impact our performance metrics differently for each subject. Also some subjects tend to have better performance metrics than others, independent of our experimental quality control. This has also been recently reported in (Shaw, 2002). From this, a strong argument could be made that the analysis chain should be optimized for each individual. While this approach may, indeed, be beneficial in some cases, it is important to realize that the prediction and reproducibility results are resampled estimates and therefore subject to uncertainty. Thus the average results over all subjects (Fig. 7) may be more indicative of the relative impact of each analysis chain. In Fig. 8, we tested whether or not the mean curves in Fig. 7 provide a statistically meaningful summary of the impact of preprocessing choices for all sixteen subjects. Mean results in Figs. 7 and 8 do not show an impact in performance metrics with alignment (black and blue triangles); however, the scope of our analysis chains did not cover the case of no alignment combined with other preprocessing operations. Detrending made some impact, but spatial smoothing provides the greatest benefits. We believe that the optimal preprocessing in general is highly dependent on the analysis chain as well as all other experimental parameters. Thus, as in ROC studies, our specific findings may or may not apply directly to other data sets.

At present, we observe slight but systematic deviations in our noise estimates from our assumed Gaussian distribution, as illustrated by the noise histogram shapes (such as in Fig. 2B) as well as the systematic offsets in the reproducibility versus confidence interval results in Fig. 10. We attribute this to the fact that we are globally characterizing spatially varying noise. Taken together, the results in Figs. 6–8 and 10 demonstrate that while there is large variation across subjects (Figs. 6, 8, and 10) the reproducible signal and noise distributions resulting from different analysis chains vary systematically across subjects in ways that may be characterized within the NPAIRS framework. We are also exploring extensions of Eqs. (6) and (7) to other distribution assumptions. Our analysis of the SNR of the reproducible activation patterns is important because this development provides standardized SPMs, which can be compared to the results of other models. As was pointed out by an anonymous reviewer, it is likely that combining the normalized split-half SPMs with smoothing would allow us to account for spatially varying noise. This comes from recognizing that our rSPMs are random effects SPMs with a pooled variance as noted in Strother *et al.* (2002) [see Worsley *et al.,* 2002, for details of a related method]. In addition, CVA analysis by itself provides an approximate random effects correction depending on the chosen class structure (Kustra, 2000). These important issues are a key focus of our ongoing research and will be address in detail in a subsequent paper.

CONCLUSION

We have demonstrated a flexible data analysis framework for evaluating preprocessing decisions in fMRI analysis using prediction and reproducibility metrics provided by the NPAIRS framework. Using reproducibility we were able to characterize the global SNR properties of our analysis and generate *z* score images useful for direct comparison with other analysis approaches. Finally, we have demonstrated cross-validation-derived prediction versus reproducibility curves as an alternative to simulation-based ROC analysis.

ACKNOWLEDGMENTS

The authors acknowledge the thoughtful comments from our anonymous reviewers; the practical discussions with Professor Vladimir Cherkassky; the helpful comments of Dr. Shing-Chung Ngan, Kirt Shaper, and Craig Benson; and the technical assistance from James Arnold. This work was partly supported by the NIH Human Brain Project P20 Grant MN57180.

REFERENCES

Aguirre, G. K., Zarahn, E., and D'Esposito, M. 1998a. A critique of the use of the Kolmogorov-Smirnov (KS) statistic for the analysis of BOLD fMRI data. *Magn. Reson. Med.* **39:** 500–505.

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

Akaike, H. 1970. Statistical predictor identification. *Ann. Inst. Stat. Math.* **22:** 203–217.

Auffermann, W. F., Ngan, S.-C., Sarkar, S., Yacoub, E., and Hu, X. 2001. Nonadditive two-way ANOVA for event-related fMRI data analysis. *NeuroImage* **14:** 406–416.

Bandettini, P. A., Wong, E. C., Hinks, R. S., Tikofsky, R. S., and Hyde, J. S. 1992. Time course EPI of human brain function during task activation. *Magn. Reson. Med.* **25:** 390–398.

Bandettini, P. A., Jesmanowicz, A., Wong, E. C., and Hyde, J. S. 1993. Processing strategies for time-course data sets in functional MRI of the human brain. *Magn. Reson. Med.* **30:** 161–173.

Buchel, C., Holmes, A. P., Rees, G., and Friston, K. J. 1998. Characterizing stimulus-response functions using nonlinear regressors in parametric fMRI experiments. *NeuroImage* **8:** 140–148.

Bullmore, E. T., Horwitz, B., Honey, G., Brammer, M., Williams, S., and Sharma, T. 2000. How good is good enough in path analysis of fMRI data? *NeuroImage* **11:** 289–301.

Bullmore, E. T., Rabehesketh, S., Morris, R. G., Williams, S. C. R., Gregory, L., Gray, J. A., and Brammer, M. J. 1996. Functional magnetic resonance image analysis of a large-scale neurocognitive network. *NeuroImage* **4:** 16–33.

Cherkassky, V., and Mulier, F. 1998. *Learning from Data: Concepts, Theory, and Methods.* Wiley, New York.

Constable, T. R., Skudlarski, P., and Gore, J. C. 1995. An ROC approach for evaluating functional brain MR imaging and postprocessing protocols. *Magn. Res. Med.* **34:** 57–64.

Efron, B., and Tibshirani, R. J. 1993. *An Introduction To the Bootstrap.* Academic Press, San Diego.

Fletcher, P. C., Dolan, R. J., Shallice, T., Frith, C. D., Frackowiak, R. S. J., and Friston, K. J. 1996. Is multivariate analysis of PET data more revealing than the univariate approach? Evidence from a study of episodic memory retrieval. *NeuroImage* **3:** 209–215.

Friedman, J. H. 1994. An overview of predictive learning and function approximation. In *From Statistics to Neural Networks: Theory*

and Pattern Recognition Applications (V. Cherkassky, J. H. Friedman, and H. Wechsler, Eds.). Springer-Verlag, Berlin.

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

Friston, K. J., Holmes, A., Poline, J-B., Price, C. J., and Frith, C. D. 1996. Detecting activations in PET and fMRI: Levels of inference and power. *NeuroImage* **40:** 223–235.

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

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J-B., Firth, C. D., and Frackowiak, R. S. J. 1995c. Statistical parametric maps in functional neuroimaging: A general linear approach. *Hum. Brain Map.* **2:** 189–210.

Hansen, L. K., Larsen, J., Nielsen, F. A., Strother, S. C., Rostrup, E., Savoy, R., Lange, N., Sidtis, J., Svarer, C., and 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., and Lange, N. 2001. Consensus inference in neuroimaging. *NeuroImage* **13:** 2001.

Holmes, A. P., Josephs, O., Buchel, C., and Friston, K. J. 1997. Statistical modeling of low-frequency confounds in fMRI. *NeuroImage* **5:** S480.

Kjems, U., Hansen, L. K., and Strother, S. C. 2002. The quantitative evaluation of functional neuroimaging experiments: Generalization error and learning curves. *NeuroImage* **15:** 772–786.

Kjems, U., Strother, S. C., Anderson, J. A., Law, I., and 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. Img.* **18:** 306–319.

Kustra, K. 2000. Statistical Analysis of Medical Images with Applications to Neuroimaging. PhD Thesis, University of Toronto. (http://www.utstat.utoronto.cal~rafal/thesis.ps.gz.)

Kustra, R., and 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. Img.* **20:** 376–387.

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., Kennedy, D. N., Hoppel, B. E., Cohen, M. S., Turner, R., Cheng, H-M., Brady, T. J., and Rosen, B. R. 1992. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc. Natl. Acad. Sci. USA* **89:** 5675–5679.

LaConte, S., Strother, S. C., Anderson, J., Muley, S., Frutiger, S., Hansen, L. K., Yacoub, E., Hu, X., and Rottenberg, D. A. 2001. Evaluating pre-processing choices in single-subject BOLD-fMRI studies using data-driven performance metrics. *NeuroImage* **13**(Part 2): S179.

LaConte, S. M., Ngan, S-C., and Hu, X. 2000. Wavelet transform based Wiener filtering of event-related fMRI data. *Magn. Reson. Med.* **44:** 746–757.

Lange, N. 1996. Statistical approaches to human brain mapping by functional magnetic resonance imaging. *Stat. Med.* **15:** 389–428.

Lange, N. 1997. Empirical and substantive models, the Bayesian paradigm and meta-analysis in functional brain imaging. *Hum. Brain Map.* **5:** 259–263.

Lange, N. 1999. Statistical procedures for functional MRI. In *Medical Radiology-Diagnostic Imaging and Radiation Oncology: Functional MRI* (P. Bandettini and C. Moonen, Eds.). Springer Verlag, New York.

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

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

Lowe, M. J., Mock, B. J., and Sorenson, J. A. 1998. Functional connectivity in single and multislice echoplanar imaging using resting-state fluctuations. *NeuroImage* **7:** 119–132.

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

McKeown, M. J., Makeig, S., Brown, G. G., Jung, T.-P., Kinderman, S. S., Bell, A. J., and Sejnowski, T. J. 1998. Analysis of fMRI data by blind separation into independent spatial components. *Hum. Brain Map.* **6:** 160–188.

Metz, C. E. 1978. Basic principles of ROC analysis. *Semin. Nuclear Med.* **8:** 283–298.

Moeller, J. R., and Strother, S. C. 1991. A regional covariance approach to the analysis of functional patterns in positron emission tomographic data. *J. Cereb. Blood. Flow Metab.* **11:** A121–A135.

Mørch, N. 1998. *A Multivariate Approach to Functional Neuromodeling,* Ph.D. thesis. Danish Technical University.

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

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

Ngan, S-C., and Hu, X. 1999. Analysis of functional magnetic resonance imaging data using self-organizing mapping with spatial connectivity. *Magn. Reson. Med.* **41:** 939–946.

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

Ogawa, S., Lee, T.-M., Kay, A. R., and Tank, D. W. 1990a. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc. Natl. Acad. Sci. USA* **87:** 9868–9872.

Ogawa, S., Lee, T.-M., Nayak, A. S., and Glynn, P. 1990b. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magn. Reson. Med.* **14:** 68–78.

Oldfield, R. C. 1971. The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia* **9:** 97–113.

Petersson, K. 1998. Comments on a Monte Carlo approach to the analysis of functional neuroimaging data. *NeuroImage* **8:** 108–112.

Poline, J-B., Worsley, K. J., Evans, A. C., and Friston, K. J. 1997. Combining spatial extent and peak intensity to test for activations in functional imaging. *NeuroImage* **5:** 83–96.

Rabe-Hesketh, S., Bullmore, E. T., and Brammer, M. J. 1997. The analysis of functional magnetic resonance images. *Stat. Methods Med. Res.* **6:** 215–237.

Ripley, B. D. 1996. *Pattern Recognition and Neural Networks.* Cambridge Univ. Press, Cambridge/New York.

Ripley, B. D. 1998. Statistical theories of model fitting. In *Neural Networks and Machine Learning* (C. M. Bishop, Ed.). Springer-Verlag, Berlin.

Shaw, M., Strother, S. C., Podzebenko, K., Anderson J., Gavrilescu, M., Egan, G., and Watson, J. 2002. Optimized pre-processing for improved signal detection in fMRI. [abstract]. Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2–6, 2002, Sendai, Japan. Available on CD-Rom in *NeuroImage*, Vol. 16, No. 2.

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

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

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

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

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

Strother, S. C., Lange, N., Savoy, R. L., Anderson, J. R., Sidtis, J. J., Hansen, L. K., Bandettini, P. A., O'Craven, K., Rezza, M., Rosen, B. R., and Rottenberg, D. A. 1996. Multidimensional state-spaces for fMRI and PET activation studies. *NeuroImage* **3**(Pt 2): S98.

Sychra, J. J., Bandettini, P. A., Bhattacharya, N., and Lin, Q. 1994. Synthetic imags by subspace transforms. I. principal components images and related filters. *Med. Phys.* **21**(2): 193–201.

Talairach, P., and Tournoux, J. 1988. *A Stereotactic Coplanar Atlas of the Human Brain.* Thieme, Stuttgart.

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

Turner, R., Le Bihan, D., Moonen, C. T., Despres, D., and Frank, J. 1991. Echo-planar time course MRI of cat brain oxygenation changes. *Magn. Reson. Med.* **22:** 159–166.

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

Woods, R. P., Dapretto, M., Sicotte, N. L., Toga, A. W., and Mazziotta, J. C. 1999. Creation and use of a Talairach-compatible atlas for accurate, automated, nonlinear intersubject registration and analysis of functional imaging data. *Hum. Brain Map.* **8:** 73–79.

Worsley, K. J. 1997. An overview and some new developments in the statistical analysis of PET and fMRI data. *Hum. Brain Map.* **5:** 254–258.

Worsley, K. J., Evans, A. C., Marrett, S., and Neelin, P. 1992. A three-dimensional statistical analysis of CBF activation studies in human brain. *J. Cereb. Blood Flow Metab.* **12:** 900–918.

Worsley, K. J., and Friston, K. J. 1995. Analysis of fMRI time-series revisited—Again. *NeuroImage* **2:** 173–181.

Worsley, K. J., Liao, C. H., Aston, J., Petre, V. Duncan, G. H., Morales, F., and Evans, A. C. 2002. A general statistical analysis for fMRI data. *NeuroImage* **15:** 1–15.

Worsley, K. J., Marrett, S., Neelin, P., and Evans, A. C. 1996a. Searching scale space for activation in PET images. *Hum. Brain Map.* **4:** 74–90.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., and Evans, A. C. 1996b. A unified statistical approach for determining significant signals in images of cerebral activation. *Hum. Brain Map.* **4:** 58–73.

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

Xiong, J., Gao, J-H., Lancaster, J. L., and Fox, P. T. 1996. Assessment and optimization of functional MRI analysis. *Hum. Brain Map.* **4:** 153–167.