◆ Human Brain Mapping 7:267–283(1999) ◆

Reproducibility of BOLD-Based Functional MRI Obtained at 4 T

Carola Tegeler,[1] Stephen C. Strother,[2–4]* Jon R. Anderson,[2] and Seong-Gi Kim[1]

[1]*Center for Magnetic Resonance Research, Medical School, University of Minnesota, Minneapolis, Minnesota*
[2]*PET Imaging Service, Veterans Affairs Medical Center, Minneapolis, Minnesota*
[3]*Department of Radiology, University of Minnesota, Minneapolis, Minnesota*
[4]*Department of Neurology, University of Minnesota, Minneapolis, Minnesota*

Abstract: The reproducibility of activation patterns in the whole brain obtained by functional magnetic resonance imaging (fMRI) experiments at 4 Tesla was studied with a simple finger-opposition task. Six subjects performed three runs in one session, and each run was analyzed separately with the t-test as a univariate method and Fisher's linear discriminant analysis as a multivariate method. Detrending with a first- and third-order polynomial as well as logarithmic transformation as preprocessing steps for the t-test were tested for their impact on reproducibility. Reproducibility across the whole brain was studied by using scatter plots of statistical values and calculating the correlation coefficient between pairs of activation maps. In order to compare reproducibility of ''activated'' voxels across runs, subjects and models, 2% of all voxels in the brain with the highest statistical values were classified as activated. The analysis of reproducible activated voxels was performed for the whole brain and within regions of interest. We found considerable variability in reproducibility across subjects, regions of interest, and analysis methods. The t-test on the linear detrended data yielded better reproducibility than Fisher's linear discriminant analysis, and therefore seems to be a robust although conservative method. Preliminary data indicate that these modeling results may be reversed by preprocessing to reduce respiratory and cardiac physiological noise effects. The reproducibility of both the position and number of activated voxels in the sensorimotor cortex was highest, while that of the supplementary motor area was much lower, with reproducibility of the cerebellum falling in between the other two areas. *Hum. Brain Mapping 7:267–283, 1999.*

Key words: reproducibility; fMRI; BOLD; brain mapping; high field; motor; statistical tests

INTRODUCTION

Functional imaging techniques are routinely used to locate activation sites related to mental tasks in humans. Recently developed functional magnetic resonance imaging (fMRI) techniques provide good spatial resolution and are completely noninvasive procedures. Thus, fMRI has extended the utility for cognitive neurosciences. It is well-accepted that higher fields (e.g., 4.0 Tesla) provide higher signal-to-noise and contrast-to-noise ratios in fMRI [Bandettini and Wong, 1995; Ogawa et al., 1993; Turner et al., 1993; Weiskoff et al., 1994]. Moreover, functional maps obtained at high fields contain fewer relative contributions from large venous vessels than those from low field systems [Gati et al., 1997; Ogawa et al., 1993; Weiskoff et al., 1994].

Grant sponsor: National Institutes of Health; Grant numbers: Human Brain Project MH57180, RR08079; Grant sponsor: Deutsche Forschungsgemeinschaft.

*Correspondence to: Stephen Strother, Ph.D., PET Imaging Service (11P), Veterans Administration Medical Center, One Veterans Drive, Minneapolis, MN 55417. E-mail: steve@pet.med.va.gov

Received for publication 30 June 1998; Revised 19 February 1999

Thus, many institutes have set up high field laboratories with field strengths greater than or equal to 3 T. Currently more than 20 systems with 3 T or 4 T magnets exist worldwide. While several laboratories have investigated the reproducibility of functional imaging with low field systems (1.5 T) [Mattay et al., 1996; Moser et al., 1996; Noll et al., 1997; Wexler et al., 1997; Yetkin et al., 1996], to our knowledge, reproducibility studies have not been performed at high fields. In this study, we used a 4 T system to examine reproducibility in the whole brain during a simple finger-opposition task, which was used previously to study reproducibility at 1.5 T [Mattay et al., 1996]. Reproducibility of statistical values, activation volume, and foci location has been investigated in the primary sensorimotor areas as well as associative motor areas including the supplementary motor area and the cerebellum.

Activation maps from functional imaging studies are very sensitive to the thresholds used to classify voxels as active or inactive. Various attempts have been made to determine ''optimal'' thresholds [Arndt et al., 1997; Forman et al., 1995; Genovese et al., 1997; Kleinschmidt et al., 1995], but no single data analysis model or thresholding procedure has become established as a gold-standard processing method in fMRI studies. The same holds for the preprocessing of data prior to statistical tests. While some laboratories detrend their data by subtracting individual slopes from the time courses of each voxel [Mattay et al., 1996; Bandettini et al., 1993], others do not. In previous reproducibility studies, model-driven [Le et al., 1997; Mattay et al., 1996; Moser et al., 1996; Yetkin et al., 1996] and data-adaptive [Moser et al., 1996; Noll et al., 1997; Wexler et al., 1997] thresholds have been used with the t-test [Le and Hu, 1997; Mattay et al., 1996; Wexler et al., 1997] and cross-correlation [Noll et al., 1997; Moser et al., 1996; Yetkin et al., 1996] analysis methods. In this study we investigated the effect of different data analysis models on intra- and intersubject reproducibility. Two statistical approaches were taken: one was a standard univariate t-test with different preprocessing steps; the other was a multivariate Fisher's linear discriminate analysis (FLDA) applied to the principal components resulting from data preprocessing according to the scaled subprofile model (SSM) [Moeller and Strother, 1991]. To compare both models, a fixed percentage of all brain voxels with the highest statistical values was classified as active. We found that reproducibility is dependent on both the model and the brain region being investigated.

MATERIALS AND METHODS

Subjects

Six healthy subjects (3 male and 3 female) were studied according to the guidelines approved by the institutional review board of the University of Minnesota; informed consent was obtained from all subjects. Average age was 36 $\pm$ 6 years. All subjects were right-handed according to the Edinburgh Inventory [Oldfield, 1971]. They were recruited from the academic environment of the University of Minnesota Medical School.

MRI

Functional MRI was studied on a 4 T whole-body imaging system with a 1.25-m-diameter horizontal bore (SISCo., Palo Alto, CA / Siemens, Erlangen, Germany) and a head gradient insert operating at a gradient strength of 30 mT / m and a slew rate of 150 T / m / sec in all three axes. For radio frequency transmission and detection, a homogeneous quadrature birdcage coil was used. To reduce head motion, foam padding was used. Manual shimming was performed to improve homogeneity before the image data collection.

In all imaging studies, conventional T_1-weighted anatomic images of the whole brain were collected using the turbo fast low angle shot (FLASH; inversion time TI = 1.2 sec) technique with in-plane resolution of 1.875×1.875 mm^2, a field of view of 24×24 cm^2, and 32–35 slices with a thickness of 5 mm. The blood oxygenation level dependent (BOLD) functional images of the whole brain were acquired with the single-shot echo planar imaging (EPI) technique with a repetition time (TR) of 5 sec, an echo-time (TE) of 30 msec, in-plane resolution of 3.75×3.75 mm^2, and slice thickness of 5 mm. The slices were selected coronally because the gradient noise is less in this direction in our system. To reduce EPI artifacts, we used the image reference method [Le and Hu, 1997].

To investigate within-session reproducibility of functional images, each subject repeated the fMRI study (referred to as a ''run'') three times in one session; i.e., Run1, Run2, and Run3. During each run, T_2*-weighted EPI images were collected continuously during four ''nonstimulated'' resting, or control, and three stimulated task periods. Typically, 12 images were acquired during each 60-sec period except during the first control period, which had 15 images. The first three images acquired during the first control period served as a reference to correct phase differences between odd and even echoes [Hu and Le, 1996] and were discarded

before further data analysis. The waiting time between runs was between 3–10 min. The task was left-handed finger opposition between the thumb and the remaining four digits (digit order: 2, 3, 4, 5, 4, 3, 2, and so on); an auditory cue paced the finger opposition at 1 Hz.

For 2 subjects, fMRI was accompanied by simultaneous recording of the respiration and cardiac signal for retrospective reduction of physiological noise. The respiratory signal was monitored with a flexible pressure belt placed around the abdomen of the subject. The cardiac signal was monitored with a pulse oximeter placed on the finger of the subject.

Data processing

Data preparation

All data sets were visually inspected for head movements by using a CINE movie of the functional images. The multislice two-dimensional images were converted to a three-dimensional (3D) image with an isotropic voxel size of $(3.75\ mm)^3$ and a field of view of $(24\ cm)^3$. The resulting 3D-images were resliced axially for better identification of motor cortical areas. To minimize head motion artifacts, each 3D-image was realigned to the first image of the first run (i.e., the fourth image acquired) using the automated image registration (AIR) program [Woods et al., 1998]. The maximal mean movement between and within runs of all voxels of all subjects was less than 2.5 mm (i..e., less than one voxel). Finally the images were smoothed by averaging a voxel with its adjacent voxels using a 3D $3 \times 3 \times 3$ boxcar function. To generate a mask for voxels within the brain, a semiautomated algorithm was used: the mean volume of all aligned functional volumes from each subject was thresholded at 15% of the maximum value in each slice, and the resulting mask was visually inspected and manually improved if necessary.

Plots of the standard deviation vs. the means of each volume for all runs were generated to examine the interdependency of both parameters and to assess the suitability of a multiplicative global effects model such as the scaled subprofile model (SSM) [Moeller and Strother, 1991]. Furthermore, the reproducibility of the control periods, or baselines, was investigated for each run with an ANOVA: the means of each volume were examined with the four baselines as the treatment and the 12 data points for each baseline as repeated measures. If a significant difference between baseline periods was observed, the maximal percentage difference was calculated.

Statistical analysis

To generate activation maps for each run, two statistical methods were used: a standard univariate unpaired t-test and a multivariate Fisher's linear discriminate analysis (FLDA) applied to the principal components from scaled subprofile model (SSM) preprocessing.

Univariate analysis: t-test. The t-test was performed between control and task periods on a voxel-by-voxel basis with a variance pooled across rest and activation periods. The pooled variance σ^2 is

$$\sigma^2 = \frac{(N_{act} - 1)\sigma^2_{act} + (N_{rest} - 1)\sigma^2_{rest}}{N_{act} + N_{rest} - 1} \quad (1)$$

where σ^2_{act} and σ^2_{rest} are variances during activation and resting periods and N_{act} and N_{rest} the number of volumes during the activation and resting periods.

Two different thresholds were used: one was a fixed t-value for all runs, and the other was determined by a fixed number of voxels with the highest statistical values. All voxels with values higher than the threshold were regarded as nominally ''activated,'' with this designation to be tested using reproducibility across independent runs to provide additional control for false-positive activations where needed (see Appendix). As a fixed threshold we chose $t = 5$ and compared this with a fixed number of activated voxels, i.e., the top 2% of all voxel t-values inside the brain.

To determine the effect of different preprocessing techniques on reproducibility, data were analyzed by the t-test 1) with and without detrending and 2) with and without logarithmic transformation. For the trend correction, first- and third-order polynomial functions were fitted to the time course of each voxel, and the time-dependent component was subtracted.

In another preprocessing step, the data sets of 2 subjects were corrected for 1) respiratory and 2) respiratory and cardiac fluctuations according to a retrospective technique [Hu et al., 1995]. The corrected data sets were then analyzed with both models, the t-test and SSM/FLDA.

Multivariate analysis: SSM/FLDA. For the SSM/FLDA, the data were preprocessed according to the SSM framework [Moeller and Strother, 1991; Strother et al., 1995a,b], which after a logarithmic transformation removes a spatially independent global scaling factor and a temporally independent spatial pattern, followed by a principal component analysis (PCA) of

the remaining space-time (voxel-time) interaction term plus noise. The signal of voxel j in scan q (1 to Q), v_{jq}, is described by

$$v_{jq} = g_q(r_j + i_{jq}) \quad (2)$$

where g_q is a global scaling factor for scan q, r_j the group mean pattern for voxel j, and i_{jq} the voxel-time interaction term with error. The interaction term is described by

$$i_{jq} = \sum_{k=1}^{Q} h_{jk} ssf_{kq} \quad (3)$$

where h_{jk} is the value of the k^{th} orthonormal eigenimage for voxel j and ssf_{kq} is the q^{th} scan scaling factor for the k^{th} eigenvector. Note that this type of model involving a variance decomposition of an interaction term is appropriate when dealing with signals in which there is a high degree of unknown nonstationary structure, as occurs in fMRI time series.

In order to exclude some ''noise'' from the FLDA, only the first 30 eigenvectors from the SSM preprocessing of each run were analyzed (Q = 30). The FLDA was applied to these eigenvectors from each run to find the linear combination of eigenvectors that defined the direction in the 30 dimensional subspace that separated baseline and activation scans best; this ''discriminant eigenvector'' contains a time series of 84 weights, one for each scan in the run. There is an equivalent linear combination of eigenimages to the eigenvectors, which defines the ''discriminant eigenimage'' for that run [Ardekani et al., 1998; Rottenberg et al., 1996]. Note that the FLDA of the SSM eigenvectors represents the simplest two-group example of the more general multigroup analysis of eigenvectors using canonical variates analysis [Friston et al., 1995; Strother et al., 1996].

The PCA in the SSM preprocessing orders the resulting eigenvectors according to the total variance they contribute to the voxel-time interaction term (i_{jq}). This order may have little to do with the order of importance of the first 30 eigenvectors in forming the discriminant between baseline and activation scans [Flury, 1995; Rottenberg et al., 1996]. To obtain this ''FLDA ordering'' we calculated the variance (R^2) contributed by each of the 30 eigenvectors to the discriminant eigenvector and reordered the eigenvectors according to the largest to smallest R^2 values. To further eliminate eigenvectors that contain more noise than predictable baseline-activation signal (i.e., that do not enhance prediction across repeated runs), for each subject we assessed the reproducibility across repeated runs as a function of the number of reordered eigenvectors, accounting for a cumulative R^2 of 80%, 90%, 99%, and 100% of the discriminant eigenvector variance. The cumulative R^2 percentage that on average maximized activated voxel reproducibility was found, and these SSM/FLDA results were analyzed further. Only the top-2% threshold was applied to the discriminant eigenimages of the SSM/FLDA, because there is no standard means of converting eigenimages to statistical parametric maps based on parametric distributions such as t-tests.

To investigate if any SSM eigenvectors, particularly those retained for calculation of SSM/FLDA results, represented residual movement effects, each of the first 30 eigenvectors was correlated with the time course of maximal voxel displacement derived from the movement correction.

Reproducibility across the whole brain

To determine the reproducibility of activation patterns across all voxels in the whole brain for repeated runs, correlation coefficients were calculated for each pair of t-value maps and discriminant eigenimages: Run1 vs. Run2, Run2 vs. Run3, and Run1 vs. Run3. Scatter plots for each pair were generated to visualize pattern similarities [Strother et al., 1997].

Reproducibility of active voxels

Reproducibility of active voxels between three runs was examined in the whole brain and within regions of interest (ROI). Three reproducibility categories were used: 1) no reproducibility: the voxel was activated only in a single run; 2) medium reproducibility: the voxel was active in two runs; and 3) high reproducibility: the voxel was active in all three runs. Note that nonreproducibly activated voxels based on 2% thresholds should be interpreted in light of the Type I error caveats discussed in the Appendix.

The ROIs were 1) the right (contralateral) motor area including the primary motor cortex, sensory motor cortex, and lateral premotor areas, and parts of the parietal areas along the postcentral sulcus (SM); 2) the left (ipsilateral) cerebellum (CER); 3) the bilateral medial motor area including the supplementary motor area (SMA); 4) the area around the sylvian fissure in the region of the planum temporale (SF); and 5) the right thalamus (Th). A sixth ROI (X) demonstrating some moderate and high reproducibility was individually chosen for every subject. All remaining activated voxels formed the miscellaneous group.

The extent of each ROI was determined by the cluster of contiguous active voxels that were classified as active in at least one run. In 2 subjects the clusters for SM and SMA were separated based on brain anatomy. Except for the thalamus, the minimum number of voxels for one cluster was five.

For the three largest ROIs (SM, CER, and SMA) in each subject, the centroid and volume of connected active voxels in each run were determined. The range of volumes was divided by the mean volume across runs, and a two-factor ANOVA performed to investigate if the means of these ratios (volume range/volume mean) for each subject varied between ROIs. Similarly, a two-factor ANOVA of the mean centroid distances for each subject was performed.

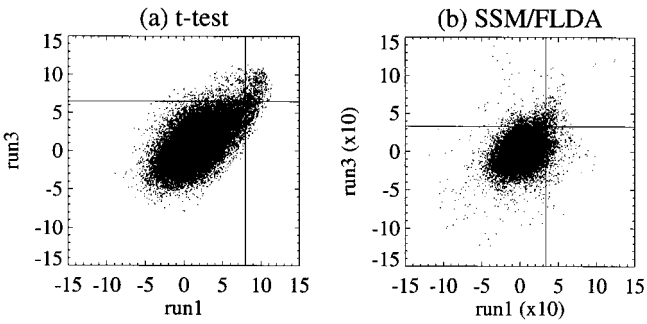

Figure 1.

Scatter plots for pairs of (**a**) t-value maps of t-tests and (**b**) discriminate eigenimages of SSM/FLDA. Each dot represents a voxel, and all brain voxels are plotted. Axes are statistical values of Run1 and Run3. Solid lines represent the thresholds for 2% of all voxels, with the highest statistical values for each run.

RESULTS

Analysis of the baseline

In all except one of the 18 runs, the slopes of the standard deviation vs. the means of each volume were significantly different from zero ($P < 0.05$) and ranged from 0.07–0.42. This indicates that the standard deviation has some linear dependency on the mean.

In 11 of the 18 runs, at least two baseline periods differed significantly in their means ($P < 0.01$, ANOVA); the maximal mean difference ranged from 0.4–2%.

In 13 of the 18 runs, one SSM eigenvector from the movement-corrected data was highly correlated with the time course of the maximal voxel displacement of the movement correction ($r > 0.7$); in 10 runs this was the first eigenvector, which accounts for 24.9–72.7% of the voxel-time-interaction term variance. In five runs (2 subjects), the SSM eigenvector that accounted for the most variance in the discriminant eigenvector of the FLDA correlated with the movement time course ($r > 0.6$); the eigenvectors accounted for 6.6–15.3% of the interaction-term variance.

Scatter plots

Figure 1 shows representative scatter plots of (a) t-values and (b) discriminant eigenimage values in two repeated runs for subject D. Linear detrended data were used for the t-test. Each dot represents one voxel in a whole-brain image. If all voxels were completely reproducible during repeated runs, all points would lie along the diagonal line. Dispersion indicates variations of t-values or discriminant eigenimage values between runs.

From the scatter plots, several observations can be made. 1) The clouds are elongated along the diagonal to the upper right, indicating that voxels with higher statistical values in one run also tend to have higher values in the other run. There is less elongation towards the lower left corner, which means that only a few voxels show potentially significant decreases. 2) Three groups of voxels can be differentiated: voxels that lie above a chosen threshold for Run1 and above the corresponding threshold for Run3 are activated in both runs, while voxels with values higher than the particular threshold only for Run1 or Run3 are only found active in one run. As expected, most voxels belong to the third group and are activated in neither of the two runs. 3) The modes of the whole-brain t-value distributions for each of the 18 runs lie between 1.2–4.2. Because the diagonal extent of the clouds does not vary as much as their modes, these mode shifts result in a high variability of the number of activated voxels with a fixed threshold such as $t = 5$. If the mode is shifted to low values, the fixed threshold classifies fewer voxels active than when the mode is shifted to higher values. Therefore, we chose a data adaptive top-2% threshold to derive activation images when comparing the reproducibility between different preprocessing steps and between both models.

Effects of different preprocessing steps

Figures 2 and 3 show the average of the mean percentages of highly reproducible, moderately reproducible, and nonreproducible voxels for different preprocessing methods. Detrending with a first-order polynomial shows a slight increase in highly reproducible and a slight decrease in nonreproducible voxels (Fig. 2). A logarithmic transformation did not improve the t-test reproducibility compared to no preprocess-

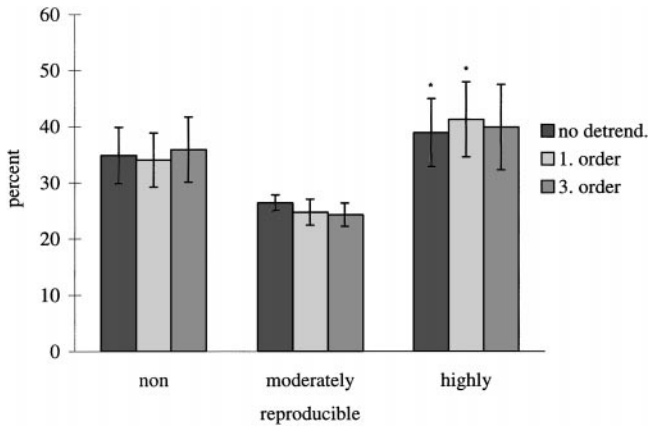

Figure 2.

Averages of the mean percentage of non-, moderately, and highly reproducible voxels (see text) for the t-test without any preprocessing, and detrending with a first- and third-order polynomial. Error bars indicate the standard errors for the averages across 6 subjects. Between the columns marked by an asterisk, a paired t-test indicated a slightly significant difference in the means ($P = 0.047$).

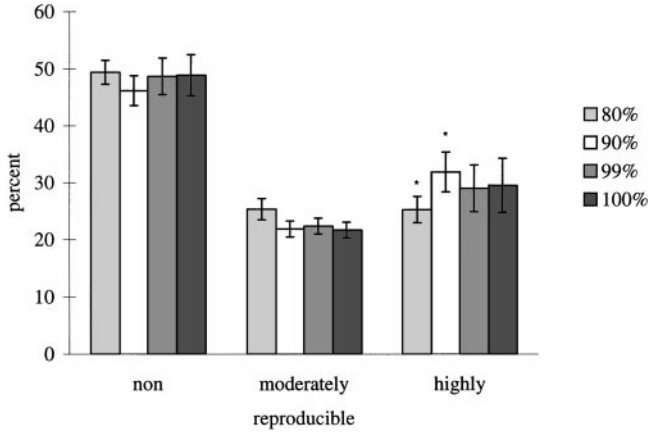

Figure 4.

Averages of the mean percentage of non-, moderately, and highly reproducible voxels (see text) for FLDA with different numbers of SSM eigenvectors accounting for 80%, 90%, 99%, and 100% of the discriminant eigenvector variance. Error bars indicate the standard errors for the averages across 6 subjects. Between the columns marked by an asterisk, a paired t-test indicated a significant difference in the means ($P = 0.011$).

ing (Fig. 3). Therefore, further reproducibility analyses were performed on linear detrended data sets for t-tests.

For the SSM we found that across the 18 runs, the first 30 eigenvectors accounted for 89–99% of the total variance of the voxel-time-interaction term, i_{jq}, of Eq. 2. The number of SSM eigenvectors needed to account for 80%, 90%, and 99% of the discriminant eigenvector

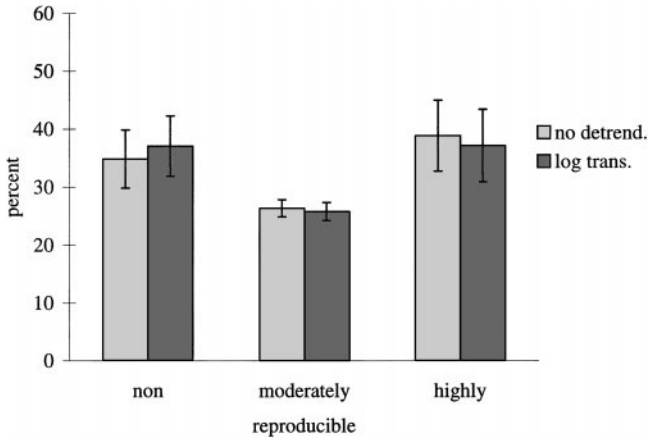

Figure 3.

Averages of the mean percentage of non-, moderately, and highly reproducible voxels (see text) for the t-test without any preprocessing, and logarithmic transformation as preprocessing. Error bars indicate the standard errors for the averages across 6 subjects.

built from 30 components is (3–10), (7–14), and (19–25), respectively. The reproducibility of the four FLDAs (80%, 90%, 99%, and 100% included variance) is shown in Figure 4. Clearly the maximum of high reproducibility and the minimum of nonreproducibility lies at 90% included variance, and therefore this SSM/FLDA was used for further analysis.

Figure 5 allows a comparison of the reproducibility of the fMRI data with and without retrospective correction for physiological noise for both statistical methods. Each bar represents the average for the 2 subjects, and range bars demonstrate the spread of the exact values for both subjects. The reproducibility for the t-test remains almost the same, while the effect of physiological correction on the SSM/FLDA is much stronger. Respiratory correction increases the fraction of highly reproducible voxels for SSM/FLDA by approximately 50% while decreasing the fraction of nonreproducible voxels. Further correction for cardiac fluctuations improves the reproducibility only slightly. Because only two data sets were corrected for physiological noise, these data were not further used.

Reproducibility across the whole brain

The average correlation coefficient of all voxels between runs is 0.56 ± 0.14 (range, 0.31–0.76) for the t-tests and 0.38 ± 0.09 (range, 0.20–0.52) for the SSM/FLDA. The correlation values are significantly

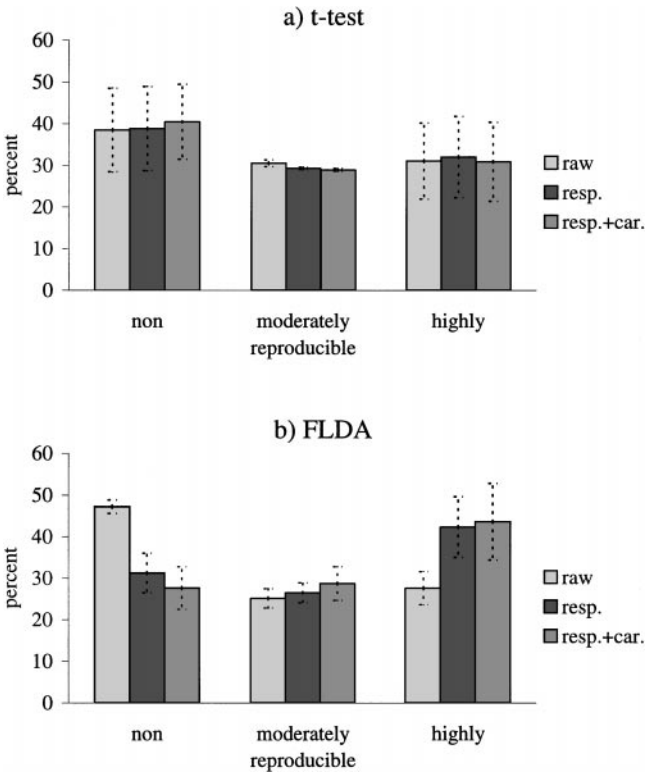

Figure 5.
Averages of the mean percentage of non-, moderately, and highly reproducible voxels (see text) for (**a**) the t-test and (**b**) SSM/FLDA for raw data and data corrected for respiratory noise and respiratory and cardiac noise. Dotted bars indicate range for the 2 subjects.

higher for the t-test than for the SSM / FLDA ($P < 0.005$, Wilcoxon matched-pairs signed-rank test), as illustrated in Figure 1.

However, for the 2 subjects after physiological correction, the correlation coefficients for SSM / FLDA exceed those for the t-test (SSM / FLDA, 0.70 ± 0.10; t-test, 0.66 ± 0.11 for respiratory and cardiac correction).

Reproducibility of active voxels

The number of active voxels for the t-test with a fixed threshold of t = 5 is listed in Table I. The t-values for the top-2% thresholds and the corresponding *P* values are tabulated in Table II. Clearly, large variations in the number of active voxels and *P* and t-values are seen even in repeated runs of one subject, suggesting difficulties for inter- and intrasubject comparisons with a fixed threshold.

When the top-2% threshold was applied, the average fraction of highly reproducible voxels was 0.41 ± 0.16 for the t-test of linear detrended data (see Fig. 2) and 0.32 ± 0.09 for the SSM / FLDA (see Fig. 4). No significant differences were found ($P > 0.05$, paired t-test) in the fraction of highly reproducible voxels between the t-test and SSM / FLDA.

TABLE I. Number of voxels per run classified as active by t-test with a fixed threshold of t = 5

Subject	Run1	Run2	Run3
A	5,317	3,910	6,543
B	1,154	499	958
C	2,071	241	145
D	3,895	2,132	2,185
E	3,243	379	710
F	6,828	12,816	4,858

Figures 6 and 7 show representative reproducibility maps of subject D calculated by the t-test and the SSM / FLDA using the top-2% threshold. Voxels activated in all three runs (highly reproducible) are colored in yellow, voxels activated in two runs (moderately reproducible) are red, and voxels activated only in one run (nonreproducible) are blue. The general activation pattern of the motor circuit is similar for both maps, but they differ substantially in their details. All subjects showed highly reproducible activation in the contralateral SM and bilateral SMA areas, and 5 subjects had highly reproducible activation in the ipsilateral cerebellum. Moderately reproducible activated voxels were found mainly on the border of the activation foci. The map generated by the t-test reveals activation in at least one run in the basal ganglia (A34) and parts of the ipsilateral premotor cortex (A48–A50), and highly reproducible activation in the thalamus (A34). In contrast to the t-test, the SSM / FLDA shows highly reproducible activation in the region of the right occipital gyrus (A28), moderately reproducible activa-

TABLE II. t- and *P*-values of three runs and 6 subjects for the thresholds classifying 2% of all voxels as active

	Run1		Run2		Run3	
Subject	t-value	*P*-value ($\times 10^{-7}$)	t-value	*P*-value ($\times 10^{-7}$)	t-value	*P*-value ($\times 10^{-7}$)
A	7.44	<1	7.08	<1	7.15	<1
B	5.79	<1	4.25	274	5.45	2
C	6.35	<1	3.47	4,083	3.08	13,930
D	7.97	<1	6.76	<1	6.48	<1
E	6.60	<1	4.25	303	4.79	40
F	8.59	<1	8.87	<1	8.36	<1

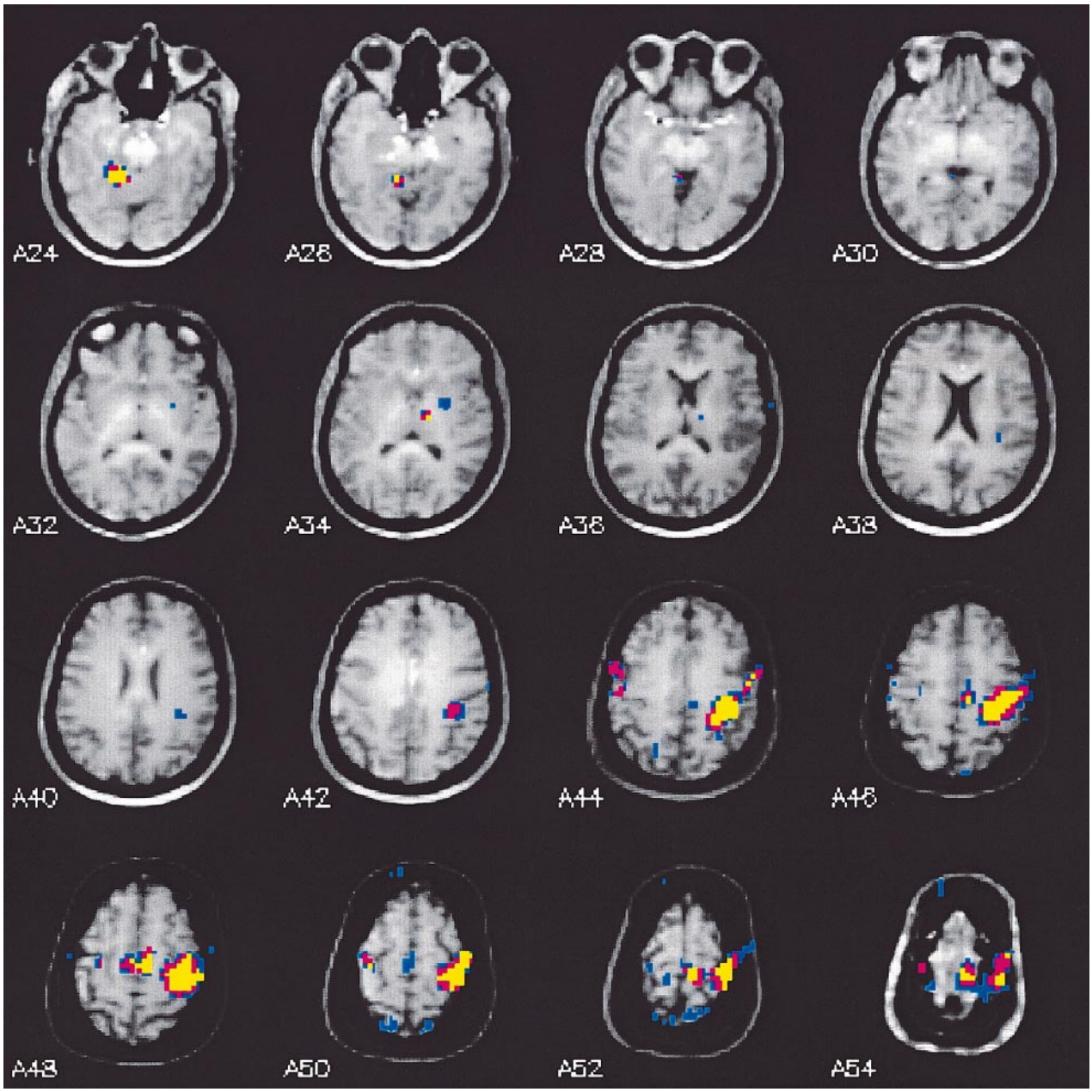

Figure 6.
Activation map for subject D generated with t-tests using the top-2% threshold for each individual run. Yellow voxels are highly reproducible, red voxels moderately reproducible, and blue voxels nonreproducible. At the bottom of each image the slice number is given. Slices are 2 × 3.75 mm apart. (Image left = subject left)

tion in the area around the sylvian fissure (SF), the SMA, the region of the planum temporale, and the parietal area (A42–A48), and no activation in the ipsilateral premotor cortex. Also, the highly reproducible activated area of the contralateral SM extends more anteriorly in the t-test map than in the SSM/FLDA map. Furthermore, the SSM/FLDA map has many more nonreproducible activation sites scattered over the entire brain, including the thalamus (A32–A36) and parietal areas (A44).

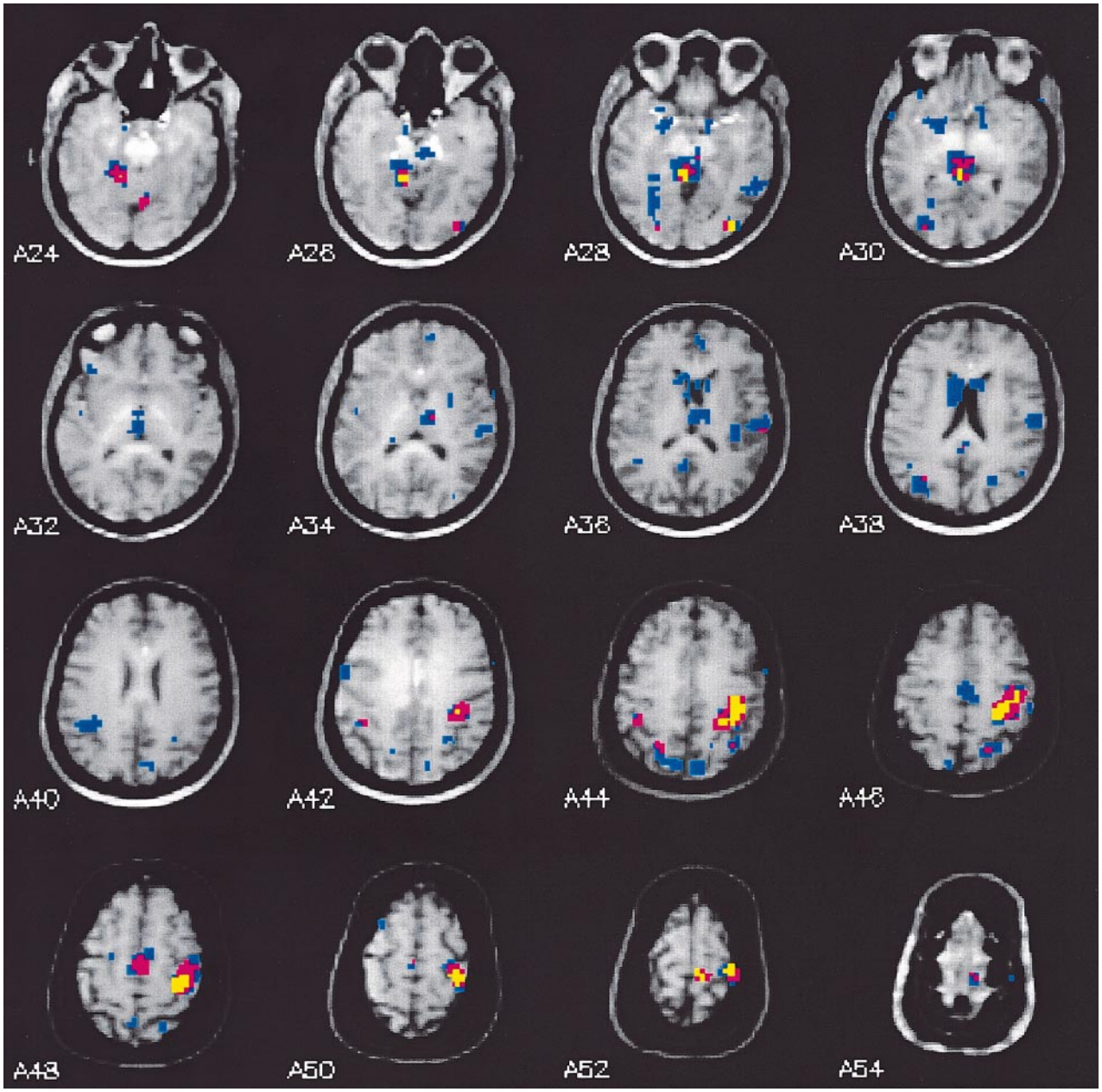

Figure 7.
Activation map for subject D generated with SSM/FLDA, using the top-2% threshold for each individual run. Yellow voxels are highly reproducible, red voxels moderately reproducible, and blue voxels nonreproducible. At the bottom of each image the slice number is given. Slices are 2 × 3.75 mm apart. (Image left = subject left)

Figure 8 displays the reproducibility of activated voxels within ROIs. Each bar represents the number of active voxels in the ROI as a fraction of the total number of active voxels within the whole brain; the sum of the fractions in all ROIs for one subject is 1.0. For each bar three categories exist: one is activation in all three runs (white), the second is activation in two runs (gray), and the third is activation only in one run (black). Because the thalamus was classified as no more than nonreproducibly activated in only four

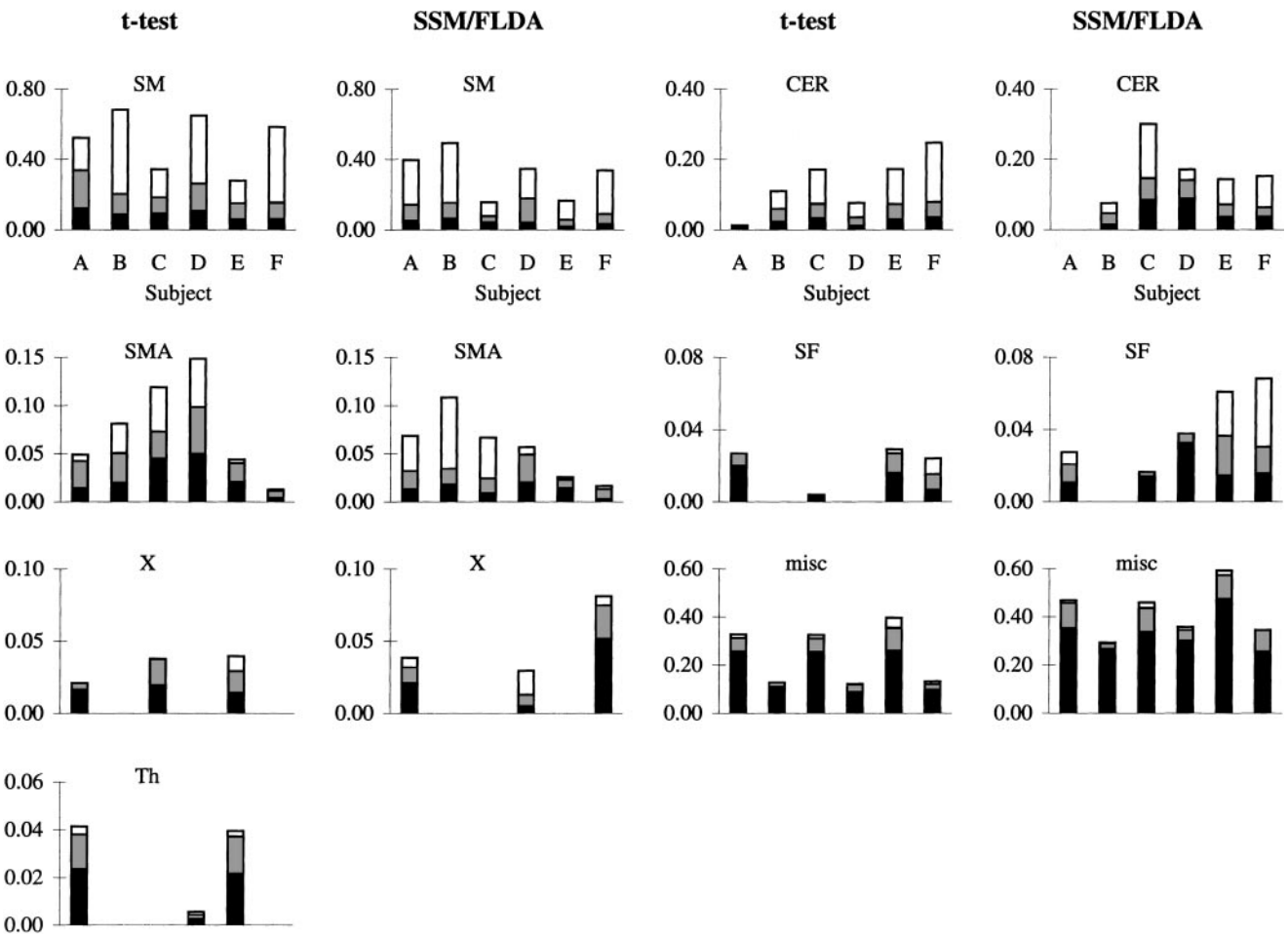

Figure 8.

Reproducibility within ROIs. One bar is drawn for each subject, ROI, and model. The heights of each bar represent the mean proportion of active voxels in the specific ROI relative to all active voxels in the whole brain. The black, gray, and white parts of each column symbolize the proportion of non-, moderately, or highly reproducible voxels, respectively. SM, sensorimotor cortex; CER, ipsilateral cerebellum; SMA, supplementary motor area; SF, area around the sylvian fissure in the region of the planum temporale; Th, contralateral thalamus; X, left precentral gyrus (subject A), right cerebellum (subject C), right occipital gyrus (subject D), ipsilateral SM (subject E), and right temporal pole (subject F); misc, miscellaneous.

discriminant eigenimages of the SSM/FLDA, this was excluded from Figure 8.

The ROI denoted by "X" is different for every subject: with the t-test, subject A revealed moderately reproducible activation in the left precentral gyrus, subject C in the right cerebellum, and subject E in the ipsilateral SM. SSM/FLDA showed highly reproducible activation in the left precentral gyrus for subject A, in the right occipital gyrus for subject D, and in the right temporal pole for subject F.

The location of the highly reproducible voxels (white part of the bars) lay mainly in the motor cortical areas (SM and SMA) and cerebellum. Their sum ranges from 50–87% for the t-test and from 33–68% for SSM/FLDA. Only subject A shows no reproducible activation in the left cerebellum, in agreement with both methods.

The individual foci of activation (X) were reproducible in most instances only for the particular subject and method. The most striking of these is subject D's highly reproducible activation in the right occipital gyrus, found only with SSM/FLDA (see A28 in Fig. 7). In contrast, subject A reveals moderately to highly reproducible activation in the left precentral gyrus with both analysis methods.

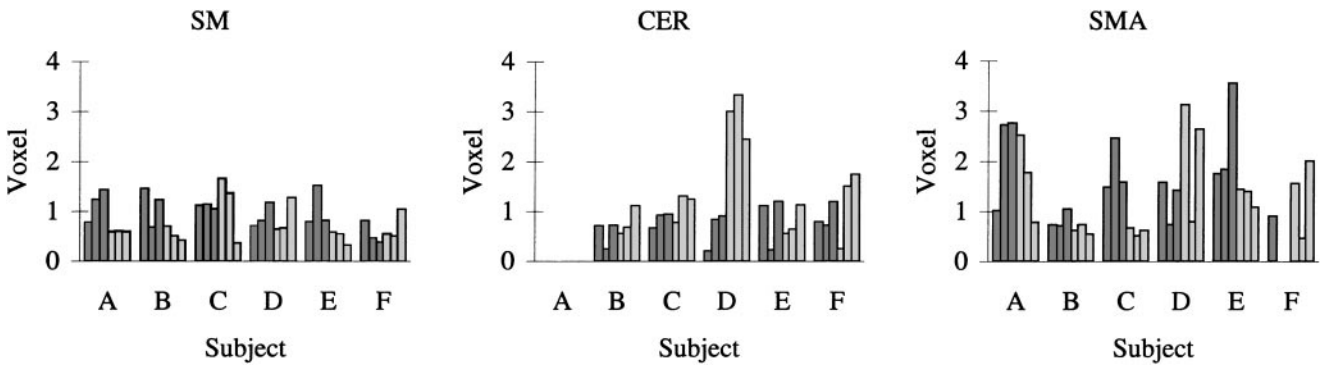

Figure 9.
Distances between centroids for clusters in SM, CER, and SMA of each subject for t-test (dark gray) and FLDA (light gray). Each bar represents the distance between one pair of runs, i.e., Run1-Run2, Run2-Run3, and Run1-Run3.

All other areas were predominantly active in only one run, although some are highly reproducible. The locations of these activations were scattered over the entire brain and not further examined. The total area of this miscellaneous group was smaller for the t-test than for the SSM/FLDA.

For the ROIs of SM, CER, and SMA, the size of the clusters of activated voxels in each run was determined. For each subject the maximum minus minimum volume across runs for each ROI was divided by the volume mean. In the CER, data for subject A were excluded. The volume range/volume mean ratio is significantly different across different ROIs for the t-test ($P < 0.005$, two-factor ANOVA), but not for SSM/FLDA ($P > 0.3$, two-factor ANOVA). The mean values and their standard deviation for the t-test are: SM, 0.3 ± 0.3; CER, 0.51 ± 0.13; and SMA, 1.4 ± 0.4; and for SSM/FLDA: SM, 0.40 ± 0.16; CER, 0.8 ± 0.2; and SMA, 0.7 ± 0.7.

The centroids of the activated cluster within the ROIs were calculated. Distances of centroids for all three pairs of runs (Run1-Run2, Run2-Run3, and Run1-Run3) within one subject in units of voxels are shown in Figure 9. Black columns result from t-tests and light gray columns from SSM/FLDA. The reproducibility of the centroids is generally higher for the SM than for the SMA. The reproducibility of the centroids for CER is similar to that of the SM except for subjects D and F for SSM/FLDA. The two-factor ANOVA revealed significant differences between the ROIs for the t-test ($P < 0.05$) but not for SSM/FLDA ($P > 0.2$). The average distances between centroids in units of a voxel were for the t-test: SM, 1.0 ± 0.2; CER, 0.75 ± 0.15; and SMA, 1.6 ± 0.7; and for SSM/FLDA: SM, 0.7 ± 0.2; CER, 1.3 ± 0.9; and SMA, 1.3 ± 0.6.

DISCUSSION

Thresholds

The scatter plots can help to define appropriate thresholds for individual runs. For both t-test and SSM/FLDA, the scatter plots consist of an elongated cloud of voxels with a higher density of voxels in the middle than at the border and the tails. Reasonable thresholds should be chosen in such a way that voxels near the center of the cloud lie below the threshold, while voxels in the tail of the cloud pass the threshold (see Fig. 1). If this criterion is fulfilled with a certain threshold for one run, the same threshold might not be the best one for another run, because the modes of the scatter plots are nonzero and variable across runs. An individual threshold for each run, such as the top-2% of voxels, takes care of first-order shifts in the t-value distributions of the maps and the coalescing of highly activated areas into single clusters is reduced. In addition, the combination of setting a percent threshold and requiring reproducibility across two or more independent runs provides protection against false-positive activation (i.e., sets Type I error P values), as discussed in the Appendix. Thresholding with $t = 5$ is close to the threshold yielded by the Bonferoni correction for a Type I error of 5% with $t = 5.2$, when 40,000 independent voxels (total number of voxels inside the brain) were compared. Voxels in the gray matter area are about 60% of all voxels. Furthermore, voxels were

averaged with their adjacent voxels, reducing the number of independent voxels. Therefore, the threshold of t = 5 results in $P < 0.05$ with the Bonferroni correction and is a conservative way of declaring voxels active. The t-values for the top-2% threshold all lie clearly above t = 1.99, the cutoff value for a significance of $P = 0.05$ without any correction for multiple comparisons. For 13 runs the t-value for the top-2% cutoff lay above t = 5, and the thresholding was therefore even more conservative than with t = 5.

Effects of preprocessing on reproducibility

We examined the impact of voxel-based polynomial detrending and log transformation on reproducibility. Besides their routine use by some fMRI groups, our testing of voxel-based preprocessing was also motivated by the possibility that the combined effect of all individual voxel transformations could compensate for the significant global effects we observed. The baselines of each run do not always stay constant, and their changes can be as high as the changes between rest and activation for activated voxels. The two runs with the highest baseline shifts showed a sudden global signal increase during the run, which could not be explained by head-movement effects or scanner problems. The cause of these global signal changes needs further investigation. However, the reproducibility increased only slightly with first-order polynomial detrending, and even with detrending, modes of the t-value distributions of the maps were shifted from zero, causing large variations in the number of activated voxels with fixed thresholds. In addition, although a significant linear relationship between scan means and standard deviations suggests the usefulness of a *global* logarithmic transformation as used in the SSM model, a *voxel*-based logarithmic transformation did not improve reproducibility. This indicates that there are sources of relatively large global signal and/or noise variations that are not adequately understood or dealt with, using simple voxel-based detrending and the stationary assumption implicit in t-tests. This observation also applies to the widely used cross-correlation analysis technique, as this is equivalent to t-test analysis for simple two-state experiments [Lange, 1996]. For a recent examination of the issue of intrasubject noise variation and estimation, see Purdon et al. [1998a]. These results provided one of our motivations for examining the more globally dependent processing provided by SSM/FLDA.

Retrospective correction for physiological noise greatly increased the reproducibility of the SSM/FLDA result; the relative number of highly reproducible voxels after physiological correction increased, and the relative number of nonreproducible voxels decreased. For the t-test, little impact of physiological noise correction on reproducibility was found. The effect on SSM/FLDA is strong because it is based on an exploratory variance partition of the voxel-time-interaction term into multiple sources of signal and noise, i.e., the eigenvectors. This process does not cleanly separate physiological and other noise sources from activation signal variation, and as a result the final discriminant eigenvector is contaminated by eigenvectors that partially reflect physiological noise sources. In contrast, the t-test is more conservative because it simply discards any voxel with a large variation over time that is not very strongly coupled to the two states being tested, and there is no attempt to model and/or partition and understand the variation as different sources of potential signal and noise. When images are acquired with shorter TRs, multivariate procedures are capable of providing partitions which allow some physiological variations to be removed [e.g., Mitra et al., 1997].

Model comparison

Activation maps obtained with the t-test and SSM/FLDA show striking differences: some areas might be classified as reproducibly active only for one method but not for the other, and the size of activated areas can vary considerably across methods. For example, the area around the sylvian fissure is more strongly emphasized by the SSM/FLDA model than by the t-test. (see Fig. 8). This general observation also applies to results from the physiological noise corrected data in the two subjects. In Figures 6 and 7 the focus in the right occipital gyrus is reproducibly detected only by SSM/FLDA, and the one in the ipsilateral premotor cortex only by the t-test. Figure 10 shows time courses of single voxels from these two areas. The amplitude of time course A in Figure 10a is clearly much larger than time course B in Figure 10b. In the subspace spanned by the 30 SSM eigenvectors, the direction of the vector representing time course A is almost orthogonal to the direction of the discriminant eigenvector calculated by the SSM/FLDA (Fig. 10c). (The correlation between time course A and the discriminant eigenvector is 0.4, and between time course B and the discriminant eigenvector 0.7.) However, because the amplitude of the vector representing time course A is relatively large, the projection onto the discriminant eigenvector yields an appreciable signal. In contrast, the amplitude

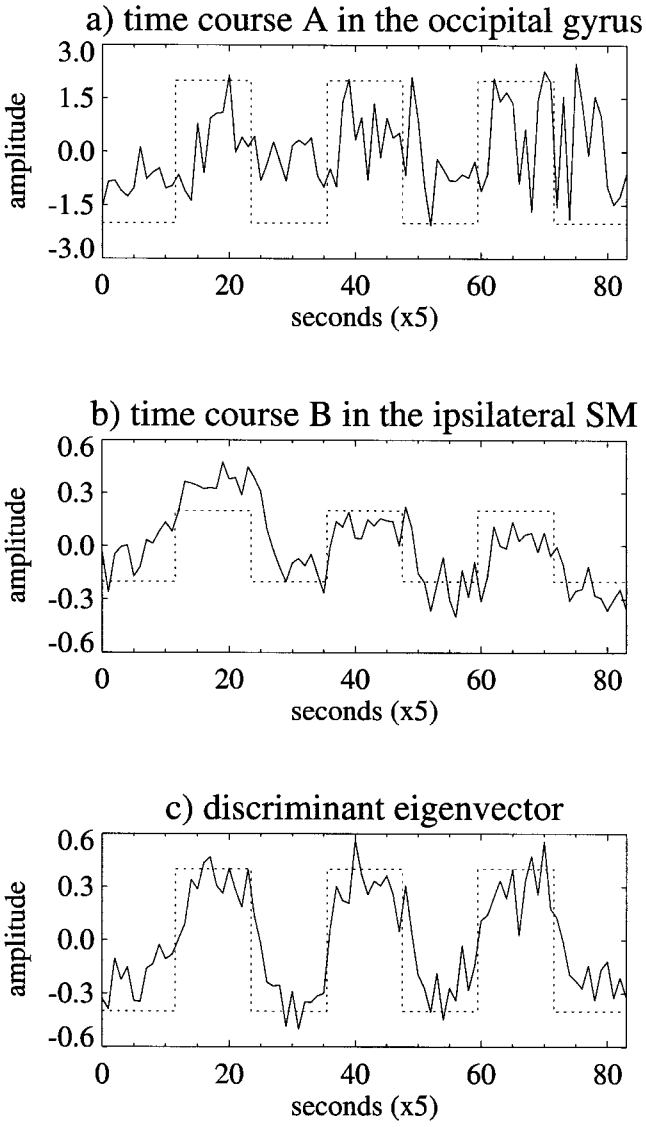

Figure 10.

Time courses from single voxels with highly reproducible signals for (**a**) SSM/FLDA in right occipital gyrus and (**b**) t-test in left premotor area from subject D's first run. **c:** Discriminant eigenvector for the same run. The mean of all time points was subtracted from each time course. The dotted boxcar function symbolizes the on-off-scheme of the paradigm.

of the vector representing time course B is relatively small and the projection onto the discriminant eigenvector gives only a small signal, although the direction of vector B is much closer to the discriminant eigenvector than vector A. Therefore, the values for the particular voxels in the discriminant eigenimage are higher for voxel A than for voxel B (1.2 vs. 0.2). Clearly, time course A exhibits some similarities with the on-off-scheme of the paradigm, but the noise is so large, that the t-value for it is much lower than that for time course B (3.5 vs. 6.7). These results indicate a need to examine different variance weighting schemes in such multivariate analyses (e.g., PCA of correlations vs. covariance matrices; see Nielsen et al. [1998]), in addition to measuring and applying physiological noise corrections.

The correlation of the eigenvectors of the SSM with the time course of maximal voxel displacement from scan to scan reveals that in many runs, residual movement effects that remain after rigid body movement correction are picked up by at least one component. The SSM preprocessing helps to remove variance due to residual movement effects when the FLDA step eliminates the corresponding eigenvector. However, if these components are important in building the discriminant eigenvector, then the movement is related to the on-off-scheme of the paradigm and these runs should be examined with care. Although this is true for five runs in 2 subjects, neither the visual screening nor the analytic estimates of the movement identified these data as outliers, and the reproducibility measures were not different from those of the other subjects. The role of SSM preprocessing and other PCA-based techniques as a means of screening for, and possibly removing, residual movement effects before multivariate and / or univariate tests needs to be studied further. Generally, we found that the t-test analysis yielded better reproducibility than SSM / FLDA for the correlation coefficient between t-value and between discriminant eigenimage statistical maps across runs, and the relative number of active voxels inside the SM, CER, and SMA ROIs. The t-test seems to be a robust, although conservative [e.g., Keenan et al., 1998], analysis model for fMRI data. In addition, the t-test demonstrated significant ROI-specific differences in the variation of the ROI's centroids and volumes, which were not seen in the SSM / FLDA results.

The considerable differences we found across models emphasizes the importance of multiple model comparisons in fMRI studies. Each model explicitly or implicitly incorporates a set of signal and noise assumptions that may respond quite differently to the highly nonstationary nature of fMRI time series. We do not claim that either of the models we used are optimal for fMRI data analysis, and we are well aware that there are a wide range of other approaches, particularly those with more sophisticated parametric descriptions for voxel-based signal and noise structure. [e.g., Lange et al., 1998, 1999; Purdon et al., 1998b]. However, the two rather different models we have compared clearly

indicate the importance of multiple model comparisons for future fMRI studies.

Comparison with previous 1.5 T results

Various test-retest studies have been reported with different approaches for quantifying the extent of reproducibility. All except one [Le and Hu, 1997] were done with 1.5 T magnetic fields [Mattay et al., 1996; Moser et al., 1996; Noll et al., 1997; Wexler et al., 1997; Yetkin et al., 1996]. Most groups used a simple motor task [Mattay et al., 1996; Noll et al., 1997; Wexler et al., 1997; Yetkin et al., 1996]. A few laboratories used visual stimulation [Le and Hu ,1997; Moser et al., 1996], sensory stimulation [Yetkin et al., 1996], and a higher cognitive working memory task [Noll et al., 1997]. A cross-correlation method [Moser et al., 1996; Noll et al., 1997; Yetkin et al., 1996] or a t-test [Le and Hu, 1997; Mattay et al., 1996; Wexler et al., 1997] was commonly used to analyze the data; only the working memory study was examined using an ANOVA. None of these studies compared reproducibility across different models. All studies using a model-driven threshold found a high variability of active voxels within subjects and across runs, and various attempts have been made to determine adaptive thresholds [Arndt et al., 1997; Forman et al., 1995; Genovese et al., 1997]. For example, one study used empirical thresholds [Wexler et al., 1997], and another thresholds based on the percentile of the noise integral of nonactivated pixels assuming a symmetrical Gaussian frequency distribution [Kleinschmidt et al., 1995; Moser et al., 1996].

The most relevant motor study for comparison with our work is from Mattay et al. [1996], who studied reproducibility within the whole brain with an EPI sequence at 1.5 T and an isotropic resolution of 3.75 mm. Eight subjects performed the finger-opposition task in three imaging sessions, which were separated by several weeks. Each session/run was analyzed independently by first correcting for movement and removing any low-frequency time trends from each voxel by fitting a third order polynomial. Afterwards, a t-test analysis was performed and the maps thresholded by a fixed t-value for each subject, classifying about 1% of all voxels within the brain as active.

ROIs based on anatomical landmarks were drawn where reproducible activated voxels were expected: the primary sensor motor cortex (PSM), lateral premotor region (PM), parietal region (PAR), SMA, and CER. Although our ROIs were driven by the areas of contiguous activated voxels, they are comparable: our SM ROI is analogous to the combination of PSM, PAR, and PM in Mattay et al. [1996], except that we did not find activation posterior to the posterior sulcus.

In all three runs of both studies, all subjects activated SM and PSM and one subject did not show activation in the CER. In the 1.5 T study, the CER of the particular subject was not covered by the field of view in one run, but in our study no obvious reason could be found for the missing activation in CER. For activation in the SMA, the studies disagree: at 1.5 T all subjects showed activation in all three runs, while at 4 T one subject did not reveal activation in one run. Mattay et al. [1996] found that the relative number of active voxels did not vary significantly across runs (one-way ANOVA). This test demonstrates that there is no specific temporal ordering for variation in sizes, and with the same test we got the same result. However, using the ratio range of sizes / mean size as a measure for reproducibility, we found considerable variations in the sizes which were significantly different across ROIs. We found similar distances between centroids of activation clusters as did Mattay et al. [1996], although we used unweighted centroids and they used weighted centroids. At 1.5 T, despite comparing runs across three sessions separated by several weeks, 75–78% of all activated voxels were found in the five ROIs. In contrast, our study at 4 T revealed a much larger range of this percentage, even though all three runs were performed within one session: from 50–87% of all active voxels in one run were found in SM, SMA, and CER. We found reproducible activation in the thalamus for 3 subjects in contrast to Mattay et al. [1996], which is probably the result of the higher field strength providing a better contrast-to-noise ratio, but may also relate to our use of the top-2% threshold, while they classified only about 1% of all voxels inside the brain as active.

Generally, comparison of our t-test results with the 1.5 T data shows that although the higher field strength of 4 T does provide a better contrast-to-noise ratio, reproducibility is not obviously increased and is highly dependent on the ROI and the subject.

Alternative reproducibility measures

One reason to study the reproducibility of activation sites is to determine ''truly'' activated voxels by separating signal from noise. In this study this was done using an analog of the split t-test [Shaywitz et al., 1995], which classifies voxels as reproducibly active when they pass the individual threshold for all runs, as shown in Figure 1.

Another approach (which includes the technique used in this study as a special case; see Appendix) would be to use a spatially distributed, multivariate

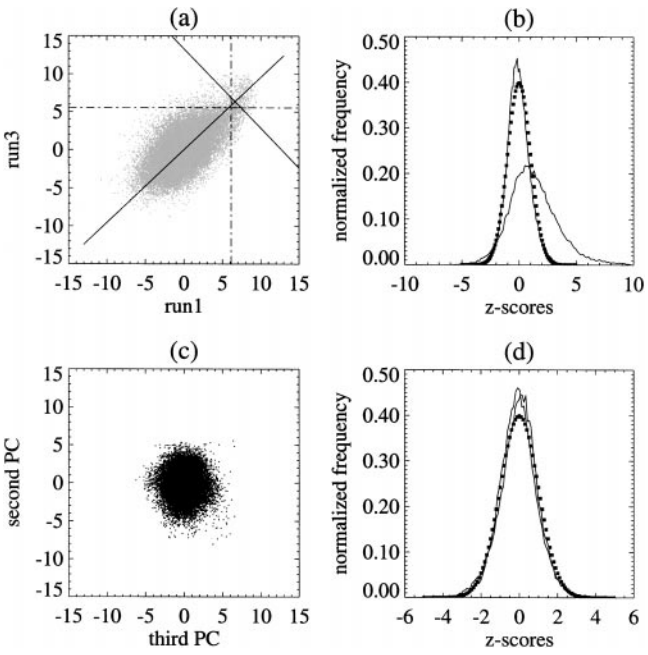

Figure 11.

a: Scatter plot for t-value maps of two runs of subject D. Dashed lines, top-2% threshold for both runs; longer solid line, first PC of the two-dimensional cloud; shorter solid line, threshold perpendicular to the first PC. **b:** Distributions along the first and second PC from a (solid lines) and standard normal curve (dotted line). **c:** Projection of the 3D scatter cloud of the t-test from three runs onto the second and third PC. **d:** Distributions along the second and third PC from c (solid lines) and standard normal curve (dotted line).

analog of the split t-test recently introduced in the PET literature [Strother et al., 1998]. Every voxel is regarded as a point in a multidimensional space with coordinates given by the statistical values of the corresponding analysis method for each of the multiple runs from each subject, one per axis. A multidimensional PCA of this distribution yields the axis along which the maximum common variance of the data is found. Figure 11a demonstrates the procedure for the two-dimensional case for the t-test. While the dashed lines indicate the top-2% threshold for the individual runs, the solid line close to the diagonal is the first principal component (PC) axis, which represents the direction of most similarities. The solid line perpendicular to the first PC indicates a threshold parallel to the second PC. In Figure 11b, the normalized distributions along the directions of the first and second PC scaled by the second PC's noise standard deviation are shown. The distribution along the second PC is very similar to a standard normal curve (dotted line), although it is a little more peaked. Therefore, the distribution along the second PC was normalized to zero mean, but not the distribution along the first PC. This method separates a reproducible signal pattern across all voxels from an orthogonal "noise" component and provides an alternative definition of reproducible signal and noise to that obtained using individual voxel split t-tests. In our case, with three runs we performed a three-dimensional PCA; Figure 11c shows the second vs. third PC. Figure 11d again demonstrates the distributions normalized to unit standard deviation and zero mean along the two PCs (solid lines), together with a normal curve (dotted line). No clear elongation in any direction can be seen in Figure 11c, and the distributions are very similar to the normal curve, indicating that there is a single reproducible pattern of statistical values reflected in all three runs.

These scatter plots are a good tool to visualize the impact of different reproducibility measures and related thresholding procedures [Rehm et al., 1998]. In Figure 11a, the threshold perpendicular to the first PC is chosen so that the same number of voxels are classified as reproducibly activated as with the split t-test. With this measure, voxels are indicated as "truly" active not only when their t- values are high for all individual t-tests but also when a voxel has a very high t-value in only one run and a medium t-value in the other t-value maps. Another measure for reproducibility could be based on the distance of the voxels from the first PC's axis. Voxels that lie close to the first PC would be regarded as highly reproducible, while voxels that lie farther from the axis are less reproducible, whether the voxels have high values for the first PC or not. Other measures based on fitting multidimensional probability density functions are also possible. Which reproducibility measure is the best is the subject of ongoing investigation.

CONCLUSIONS

The t-test is a robust but conservative analysis method compared with SSM/FLDA. Differences in reproducibility across models indicate the need to consider multiple data analysis procedures in future fMRI reproducibility studies. The increased sensitivity of multivariate techniques, such as SSM/FLDA, to physiological noise may be dealt with by retrospective processing to reduce physiological noise effects. The reproducibility measured at 4 T during simple finger opposition is strongly dependent on particular ROIs and subjects, and is not better than at 1.5 T. In the motor circuit, activation of the sensory motor area is most reproducible, the cerebellum is next, and the medial motor area is least reproducible.

ACKNOWLEDGMENTS

We thank Kirt Schaper and Kelly Rehm for technical assistance and David Rottenberg and Kamil Ugurbil for helpful discussions. C.T. thanks the Deutsche Forschungsgemeinschaft for the Forschungsstipendium.

REFERENCES

Ardekani BA, Strother SC, Anderson JR, Law I, Paulson OB, Kanno I, Rottenberg DA. 1998. On the detection of activation patterns using principal components analysis. In: Carson RE, Daube-Witherspoon ME, Herscovitch P, editors. Quantitative functional brain imaging with positron emission tomography. San Diego: Academic Press. p 253–257.

Arndt S, Gold S, Cizadlo T, Zheng J, Ehrhardt JC, Flaum M. 1997. A method to determine activation thresholds in fMRI paradigms. Psychiatry Res 75:15–22.

Bandettini PA, Wong EC. 1995. Effects of biophysical and physiologic parameters on the brain activation-induced R2* and R2 changes: simulations using a deterministic diffusion model. Int J Imaging Syst Technol 6:133–152.

Bandettini PA, Jesmanowicz A, Wong EC, Hyde JS. 1993. Processing strategies for time-course data sets in functional MRI of the human brain. Magn Reson Med 30:161–173.

Flury BD. 1995. Developments in principal component analysis. In: Krzanowski WJ, editor. Recent advances in descriptive multivariate analysis. Oxford: Clarendon Press. p 14–33.

Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintum MA, Noll DC. 1995. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn Reson Med 33:636–647.

Friston KJ, Frith CD, Frackowiak RSJ, Turner R. 1995. Characterizing dynamic brain responses with fMRI: a multivariate approach. Neuroimage 2:166–172.

Gati JS, Menon RS, Ugurbil K, Rutt DK. 1997. Experimental determination of the BOLD field strength dependence in vessels and tissue. Magn Reson Med 38:296–302.

Genovese CR, Noll DC, Eddy WF. 1997. Estimating test-retest reliability in functional MR imaging I: statistical methodology. Magn Reson Med 38:497–507.

Hu X, Le TH. 1996. Artifact reduction in EPI with phase-encoded reference scan. Magn Reson Med 36:166–171.

Hu X, Le TH, Parrish T, Erhard P. 1995. Retrospective estimation and correction of physiological fluctuation in functional MRI. Magn Reson Med 34:201–212.

Keenan JP, Ives JR, Anand K, Cousins J, Pascual-Leone A. 1998. Satterthwaite corrections for homogeneity of variance assumption failures in functional magnetic resonance imaging for a simple motor task. Neuroimage 7:S608.

Kleinschmidt A, Requardt M, Merboldt K-D, Frahm J. 1995. On the use of temporal correlation coefficients for magnetic resonance mapping of the functional brain activation: individualized thresholds and spatial response delineation. Int J Imaging Syst Technol 6:238–244.

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

Lange N, Hansen LK, Anderson JR, Nielsen FA, Savoy R, Kim S-G, Strother SC. 1998. An empirical study of statistical model complexity in neuro-fMRI. Neuroimage 7:S764.

Lange N, Strother SC, Anderson JR, Nielsen FÅ, Holmes A, Kolenda T, Savoy R, Hansen LK. 1999. Plurality and resemblance in fMRI data analysis. Neuroimage, in press.

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

Mattay VS, Frank JA, Sanatha AKS, Pekar JJ, Duyn JH, McLaughlin AC, Weinberger DR. 1996. Whole-brain functional mapping with isotropic MR imaging. Radiology 201:399–404.

Mitra PP, Ogawa S, Hu X, Ugurbil K. 1997. The nature of spatiotemporal changes in cerebral hemodynamics as manifested in functional magnetic resonance imaging. Magn Reson Med 37:511–718.

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

Moser E, Teichtmeier C, Diemling M. 1996. Reproducibility and postprocessing of gradient-echo functional MRI to improve localization of brain activity in the human visual cortex. Magn Reson Imaging 14:567–579.

Nielsen FA, Hansen LK, Strother SC. 1998. Canonical ridge analysis with ridge parameter optimization. Neuroimage 7:S758.

Noll DC, Genovese CR, Nystrom LE, Vazquez AL, Forman SD, Eddy WF, Cohen JD. 1997. Estimating test-retest reliability in functional MR imaging II: application to motor and cognitive activation studies. Magn Reson Med 38:508–517.

Ogawa S, Menon RS, Tank D, Kim SG, Merkle H, Ellermann JM, Ugurbil K. 1993. Functional brain mapping by blood oxygenation level-dependent contrast magnetic resonance imaging: a comparison of signal characteristics with a biophysical model. Biophys J 64:800–812.

Oldfield RC. 1971. Assessment and analysis of handedness: the Edinburgh Inventory. Neuropsychologia 9:97–113.

Purdon PL, Solo V, Brown E, Bruckner R, Rotte M, Weisskoff RM. 1998a. fMRI noise variability across subjects and trials: insights for noise estimation methods. Neuroimage 7:S617.

Purdon PL, Solo V, Brown E, Weisskoff RM. 1998b. Signal processing in fMRI: noise estimation with regularization and hemodynamic response modeling. Neuroimage 7:S618.

Rehm K, Lakshminaryan K, Frutiger S, Schaper K, Sumners DW, Strother SC, Anderson JR, Rottenberg DA. 1998. A symbolic environment for visualizing activated foci in functional neuroimaging datasets. Med Image Anal 2:215–226.

Rottenberg DA, Sidtis JJ, Strother SC, Schaper KA, Anderson JR, Nelson MJ, Price RW. 1996. Abnormal cerebral glucose metabolism in HIV-1 seropositive subjects with and without dementia. J Nucl Med 37:1133–1141.

Shaywitz BA, Shaywitz SE, Pugh KR, Constable RT, Skudlarski P, Fulbright RK, Bronen RA, Fletcher JM, Shankweiler DP, Katz L. 1995. Sex differences in the functional organization of the brain for language. Nature 373:607–609.

Strother SC, Anderson JA, Schaper KA, Sidtis JJ, Liow J-S, Woods RP, Rottenberg DA. 1995a. Principal component analysis and the scaled subprofile model compared to intersubject averaging and statistical parameteric mapping: I. "Functional connectivity" of the human motor system studied with [^{15}O] PET. J Cereb Blood Flow Metab 15:738–753.

Strother SC, Anderson JA, Schaper KA, Sidtis JJ, Rottenberg DA. 1995b. Linear models of orthogonal subspaces and networks from functional activation PET studies of the human brain. In: Bizais Y, Barillot C, Di Paola R, editors. Information processing in medical imaging. 14th International Conference. Dordrecht: Kluwer Academic. p 299–310.

Strother SC, Lange N, Savoy RL, Anderson JR, Sidtis JJ, Hansen LK, Bandettini PA, O'Craven K, Rezza M, Rosen BR, Rottenberg DA. 1996. Multidimensional state-spaces for fMRI and PET activation studies. Neuroimage 3:S98.

Strother SC, Lange N, Anderson JR, Schaper KA, Rehm K, Hansen LK, Rottenberg DA. 1997. Activation pattern reproducibility: measuring the effects of group size and data analysis models. Hum Brain Mapp 5:312–316.

Strother SC, Rehm K, Lange N, Anderson JR, Schaper KA, Hansen LK, Rottenberg DA. 1998. Measuring activation pattern reproducibility using resampling techniques. In: Carson RE, Daube-Witherspoon ME, Herscovitch P, editors. Quantitative functional brain imaging with positron emission tomography. San Diego: Academic Press. p 241–246.

Svarer C, Strother SC, Morch N, Law I, Hansen LK, Paulson OB. 1996. Evaluating statistical parametric mapping (SPM) analysis results using leave-one-out resampling in a [^{15}O]water PET functional activation study. Neuroimage 5:S374.

Turner R, Jezzard P, Wen H, Kwong KK, Le Bihan D, Zeffiro T, Balaban R. 1993. Functional mapping of the human visual cortex at 4 and 1.5 T using deoxygenation contrast EPI. Magn Reson Med 29:277–279.

Weiskoff RM, Zuo CS, Boxerman JL, Rosen BR. 1994. Microscopic susceptibility variation and transverse relaxation: theory and experiment. Magn Reson Med 31:601–610.

Wexler BE, Fulbright RK, Lacardie CM, Skularski P, Kelz MB, Todd R, Gore JC. 1997. An fMRI study of the human cortical motor system response to increasing functional demands. Magn Reson Imaging 15:385–396.

Woods RP, Grafton ST, Holmes CJ, Cherry SR Mazziotta JC. 1998. Automated image registration: I. general methods and intrasubject, intramodality validation. J Comput Assist Tomogr 22:139–152.

Yetkin FZ, McAuliffe TL, Cox R, Haughton VM. 1996. Test-retest precision of functional MR in sensory and motor task activation. Am J Neuroradiol 17:95–98.

APPENDIX: STATISTICAL SIGNIFICANCE OF PERCENT-THRESHOLDING WITH REPRODUCIBILITY

The usual practice in functional neuroimaging data analysis is to make binary decisions between activated and nonactivated voxels, using absolute thresholds chosen to reduce Type I errors (i.e., probability of false-positive activation) based on parametric statistical measures such as t-tests. This has been questioned by some authors [e.g., Svarer et al., 1996; Strother et al., 1997, 1998], who noted that the t-values themselves are statistical estimates subject to errors, making any absolute binary decision about activated voxels based on a single thresholded data set an imprecise procedure; this is especially true given the additional arbitrary choice of a *P* value required to fix the threshold value. This imprecision is easily demonstrated using resampling procedures or measures of reproducibility across independent data sets such as the scatter plots in Figure 1a, where many voxels with t-values of 8–10 in Run1 have values < 5 in Run3.

If we abandon the apparent protection of parametric statistical estimates such as voxel-wise t-tests—such parametric estimates may not be readily available (e.g., for eigenimages)—but require reproducibility across independent test sets, do we still have some form of statistical protection from Type I errors?

We may calculate an analytical answer to this question under some reasonable assumptions, using the scatter plot format of Figure 1.

Assume that we have obtained two activation images from independent-but-otherwise-identical noise-only data sets using some modeling procedure, i.e., activation-image noise distributions for a particular experiment and data analysis model. Regardless of the true ensemble distribution of each activation-image voxel, which could be estimated by analyzing many such noise-only data sets, the marginal distributions of the scatter plot along each axis will be close to Gaussian by the central-limit theorem. Assuming that each noise-only activation image is an independent random sample from the noise-only data, the pair of images form a bivariate sample from a two-dimensional Gaussian density defined by the product of the marginal densities as

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-(x-\mu_x)^2/2\sigma_x^2} e^{-(y-\mu_y)^2/2\sigma_y^2} \qquad (4)$$

where, with $\sigma_x = \sigma_y = 1$, x and y should be expressed as Z-values for the percent thresholds of interest, e.g., $Z = 2.054$ for a 2% threshold. The probability of a voxel being above the 2% threshold for two independent data sets is $0.02 \times 0.02 = 0.0004$, which is equal to the volume under the two-dimensional Gaussian density function for $Z_x > 2.054$ and $Z_y > 2.054$. This result generalizes to n-dimensional Gaussians for n independent data sets (see the 3D results in Fig. 11).

Using a 2% threshold in a single noise-only run with activation image volumes of 40k voxels, we expect 800 voxels to be classified as ''active.'' By requiring reproducibility across two and three runs, the expected number of false-positive activations drops to 16 voxels and <1 voxel, respectively. These are actually upper bounds on the number of false positives because the additional requirement that only clusters of at least five contiguous voxels are counted as active will remove some small, isolated voxel clusters above the 2% threshold.

Therefore, the combination of thresholding with reproducibility across independent data sets provides Type I error protection similar to that afforded by choosing *P* values and setting t-value thresholds in individual data sets. The Type I error protection level is set by choosing the percent threshold and the number of independent data sets without having to assume any underlying distribution structure.