Radiology

Kelly H. Zou, PhD
Douglas N. Greve, PhD
Meng Wang, MSE
Steven D. Pieper, PhD
Simon K. Warfield, PhD
Nathan S. White, BS
Sanjay Manandhar, MS
Gregory G. Brown, PhD
Mark G. Vangel, PhD
Ron Kikinis, MD
William M. Wells III, PhD
For the FIRST BIRN Research Group

Published online
10.1148/radiol.2373041630
Radiology 2005; 237:781–789

Abbreviations:
A_z = area under ROC curve
BIRN = Biomedical Informatics Research Network
ERS = estimated reference standard
FIRST = Functional Imaging Research of Schizophrenia Testbed
ICC = intraclass correlation coefficient
ROC = receiver operating characteristic
SM = sensory motor

[1] From the Surgical Planning Laboratory, Dept of Radiology, Brigham and Women's Hosp (K.H.Z., M.W., S.D.P., S.K.W., S.M., R.K., W.M.W.); Dept of Health Care Policy (K.H.Z.); and Athinoula A. Martinos Center for Biomedical Imaging, Dept of Radiology, Massachusetts Gen Hosp (D.N.G., N.S.W., M.G.V.), Harvard Medical School, 75 Francis St, L-2, Boston, MA 02115; Isomics, Cambridge, Mass (S.D.P.); Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Mass (S.K.W., W.M.W.); Computational Radiology Laboratory, Dept of Radiology, Brigham and Women's Hosp, Boston, Mass (S.K.W.); Dept of Radiology, Children's Hosp, Boston, Mass (S.K.W.); Laboratory of Cognitive Imaging, Dept of Psychiatry, Univ of California, San Diego, La Jolla, Calif (G.G.B.); and Veterans Affairs San Diego Health Care System, San Diego, Calif (G.G.B.). Received Sep 21, 2004; revision requested Nov 29; revision received Jan 24, 2005; accepted Feb 24. The BIRN study at two sites supported by NIH grants NCRR P41RR13218 and P41RR14075. Supported in part by NIH grants R01LM007861-01A1, R03HS013234-01, R21MH67054, and R21CA89449-01. **Address correspondence to** K.H.Z. (e-mail: *zou@bwh.harvard.edu*).

Authors stated no financial relationship to disclose.

Author contributions:
Guarantor of integrity of entire study, K.H.Z. The complete list of author contributions is cited at the end of this article.

Reproducibility of Functional MR Imaging: Preliminary Results of Prospective Multi-institutional Study Performed by Biomedical Informatics Research Network[1]

PURPOSE: To prospectively investigate the factors—including subject, brain hemisphere, study site, field strength, imaging unit vendor, imaging run, and examination visit—affecting the reproducibility of functional magnetic resonance (MR) imaging activations based on a repeated sensory-motor (SM) task.

MATERIALS AND METHODS: The institutional review boards of all participating sites approved this HIPAA-compliant study. All subjects gave informed consent. Functional MR imaging data were repeatedly acquired from five healthy men aged 20–29 years who performed the same SM task at 10 sites. Five 1.5-T MR imaging units, four 3.0-T units, and one 4.0-T unit were used. The subjects performed bilateral finger tapping on button boxes with a 3-Hz audio cue and a reversing checkerboard. In a block design, 15-second epochs of alternating baseline and tasks yielded 85 acquisitions per run. Functional MR images were acquired with block-design echo-planar or spiral gradient-echo sequences. Brain activation maps standardized in a unit-sphere for the left and right hemispheres of each subject were constructed. Areas under the receiver operating characteristic curve, intraclass correlation coefficients, multiple regression analysis, and paired Student *t* tests were used for statistical analyses.

RESULTS: Significant factors were subject ($P < .005$), k-space ($P < .005$), and field strength ($P = .02$) for sensitivity and subject ($P = .03$) and k-space ($P = .05$) for specificity. At 1.5-T MR imaging, mean sensitivities ranged from 7% to 32% and mean specificities were higher than 99%. At 3.0 T, mean sensitivities and specificities ranged from 42% to 85% and from 96% to 99%, respectively. At 4.0 T, mean sensitivities and specificities ranged from 41% to 73% and from 95% to 99%, respectively. Mean areas under the receiver operating characteristic curve ($\pm$ their standard errors) were 0.77 ± 0.05 at 1.5 T, 0.90 ± 0.09 at 3.0 T, and 0.95 ± 0.02 at 4.0 T, with significant differences between the 1.5- and 3.0-T examinations and between the 1.5- and 4.0-T examinations ($P < .01$ for both comparisons). Intraclass correlation coefficients ranged from 0.49 to 0.71.

CONCLUSION: MR imaging at 3.0- and 4.0-T yielded higher reproducibility across sites and significantly better results than 1.5-T imaging. The effects of subject, k-space, and field strength on examination reproducibility were significant.

Functional magnetic resonance (MR) imaging has substantially contributed to our understanding of normal and diseased brains in humans (1). A large degree of variability in the

magnitude, spatial distribution, and statistical significance of the corresponding functional MR imaging maps often exists owing to differences in the equipment used and to subject- and site-specific differences (2–12). Therefore, understanding the effects of these factors is desirable, particularly given the costs of imaging and the complexity of performance tasks.

The first phase of the Functional Imaging Research of Schizophrenia Testbed (FIRST) Biomedical Informatics Research Network (BIRN) study *(http://www.nbirn.net)* was aimed at comparing and calibrating functional MR imaging signal intensities to determine whether the interrelation of functional MR imaging maps across different study sites was meaningful. This preliminary effort was made prior to collecting prospective functional MR imaging data from subjects with schizophrenia and from control subjects during the next phase of our planned multi-institutional prospective study.

The purpose of our current study was to prospectively investigate the factors—including subject, brain hemisphere, study site, field strength, MR imaging unit vendor, imaging run, and examination visit—that affect the reproducibility of functional MR imaging activations based on a repeated sensory-motor (SM) task.

MATERIALS AND METHODS

Study Subjects

The institutional review boards of all of the participating sites approved this Health Insurance Portability and Accountability Act–compliant study. Informed consent was obtained from all subjects at all sites. A total of 11 sites formed the FIRST BIRN component of the study. Data were collected from 10 of these 11 sites. Five 1.5-T MR imaging units, four 3.0-T units, and one 4.0-T unit were used. The MR imaging unit used at the 11th site was not functional when the subjects were traveling around the country. Only a few data sets were collected at this site, which was excluded from the analysis.

Five healthy right-handed men aged 20–29 years underwent MR imaging at each site during two visits on separate days; they engaged in 10 task runs per visit. These subjects were volunteers who responded to advertisements for participation in the study. In addition, three of these subjects were randomly chosen to undergo additional MR imaging examinations, so there were a total of four visits at two of the 10 sites. Because this was a preliminary study of repeated functional MR imaging performed at all of these sites, the limited sample size of five subjects was derived on the basis of the projected study cost rather than by means of a formal statistical power calculation.

Inclusion criteria were as follows: *(a)* male subject older than 18 years but younger than 45 years; *(b)* ability to speak English fluently; *(c)* eyesight either 20/20 uncorrected or corrected with contact lens wear (the visual displays at most sites could not be seen otherwise); *(d)* normal hearing in both ears, as tested with an audiometer; *(e)* right handed; and *(f)* ability to travel for the period required to complete the study. Exclusion criteria were as follows: *(a)* claustrophobia; *(b)* presence of metal implants or other contraindications to MR imaging; *(c)* tattoos on the upper half of the body; *(d)* current daily use of cigarettes or narcotic drugs, as self-reported; *(e)* history or first-degree family history of mental illness, as diagnosed by using a structured clinical interview; *(f)* epilepsy, multiple sclerosis, diabetes, and/or other medical illness; *(g)* history of cancer or chemotherapy; and/or *(h)* current use of prescribed medication.

SM Task

The subjects performed the SM task in four of 10 functional MR imaging runs performed during each visit (13). For this study, only the SM task data were analyzed. The six remaining runs, which were not included in the current study analysis owing to limited repetitions at each site, were two cognitive and two breath-hold tasks, as well as two rest periods. Not all subjects had the same cognitive paradigm. The breath-hold task involved a large degree of activation, and, thus, it was difficult to assess the reproducibility. The rest task involved little activation and thus was not appropriate for study inclusion.

The study subjects were imaged during two visits on two separate days at each site. They had a normal night's sleep, had no more than one alcoholic drink the night before the MR imaging examination, and abstained from drinking coffee within 2 hours before the examination. There were two versions of the tasks used to collect functional MR imaging data: One version of the task prompted the start of the MR imaging unit, whereas the start of the other version was prompted by the unit. The task version used at each site depended on how the given MR imaging unit was operated.

All button box data were saved and viewed. Button box responses were observable at run time and were monitored to ensure that the subject was alert and responding. During the first visit, the subjects provided MR imaging compatibility screening, quick mood scale, and consent forms. During this visit, the SM and cognitive tasks they would perform were explained to them and they practiced these tasks in front of the monitor. The same task protocol was followed during the second visit.

The following audio setup phases were used: headphone and volume setting adjustments for tone testing, tone balance setup, and tone volume setup. The introductory screen for the audio task contained a menu from which the examiner could select the phase to run. Each subject was placed in a sitting-up position in the imaging unit with headphones on. No earplugs were used with the headphones. A tone was then played in the subject's left ear, and the subject was visually instructed to press a button on the button box. Pressing the button triggered the tone to terminate in the left ear and start in the right. Likewise, the subject was prompted to hit a button to confirm that he could hear the tone in the right ear. This second button press triggered the tone to be played in both ears at equal volumes. The subject then hit a button again if he could hear the tone in both ears. Subjects were allowed to signal the examiner by using hand cues and verbal responses. The examiner advanced the experiment by pressing the respective buttons on the computer keyboard.

The subject was then placed in the supine position on the imaging table with headphones on and a bite bar in place in the head coil. The button response box was placed in the subject's dominant hand, and the imaging unit squeeze ball or another nonverbal patient alert system was placed in the other hand, because only nonverbal communication was used during the imaging examination owing to the presence of the bite bar.

During functional MR imaging, a dummy button box was used in the nondominant hand, and the squeeze ball could be set aside for emergency communication with the examiner—for example, to stop imaging. The dummy button box was designed so that its size and shape mimicked those of the actual button box. The subject was advanced inside the magnet bore to confirm that he was comfortable—that is, not claustropho-

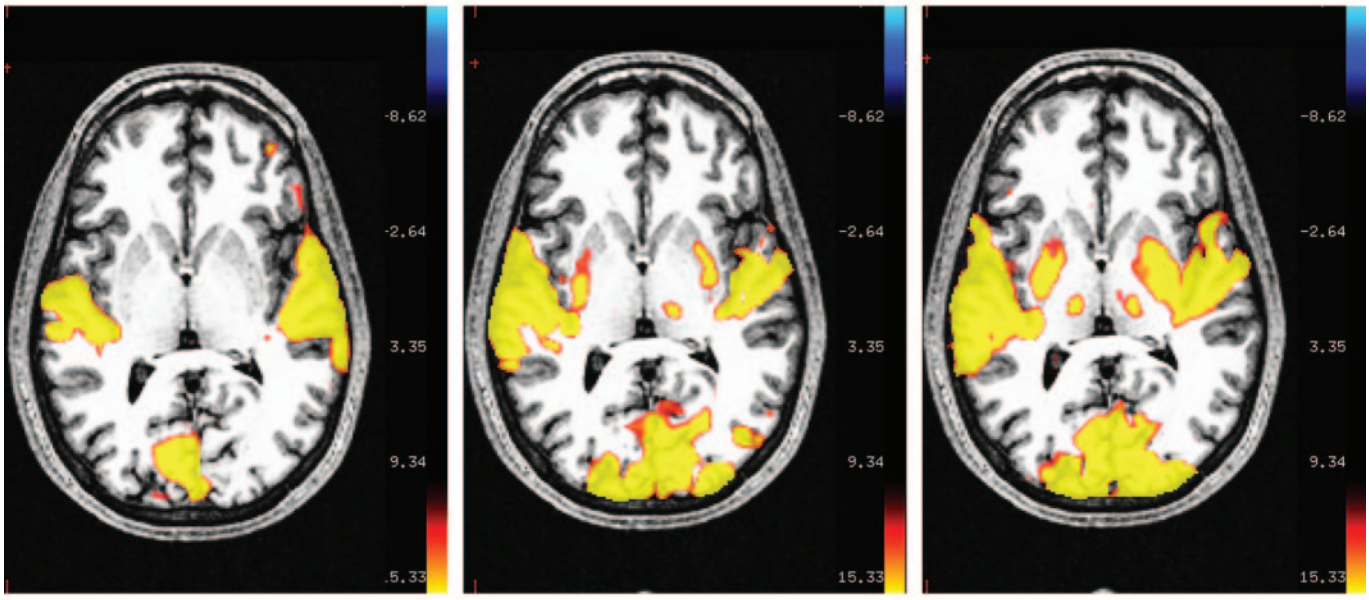


Figure 1. Subject 3. Registered MR imaging activation maps of anatomic data obtained during visit 1 at the study sites with **(a)** 1.5-T, **(b)** 3.0-T, and **(c)** 4.0-T MR imaging units.

bic—and could communicate with individuals in the control room.

Next, the subject heard a tone that was 5-dB louder in the left ear than in the right. He was instructed to make the tone sound balanced between the two ears by using the button box: Pushing button 1 moved the tone to the left ear, and pushing button 2 moved the tone to the right ear. Each button press triggered a 1-dB step change in the direction of the tone according to the given button pressed. When the tone sounds were balanced, the subject pressed button 3.

Before continuing, the examiner manually started a functional MR image acquisition to generate background noise without saving any imaging data. After the imaging unit was started, a 440-Hz tone was played at 50 dB. The subject pressed button 1 to increase the volume and button 2 to decrease the volume. The minimum step size was 2 dB. Button 3 was pressed when the subject believed that the tone volume was set to a comfortable level. The decibel level at which the tones in the left and right ears were balanced was recorded.

After all of the above task phases were completed, the examiner chose option 4 from the audio setup task menu to exit the task. While the subject was still in the MR imaging unit, the examiner accessed his audio setup data. The auditory balance and volume level were recorded. If the volume level result was lower than 30 dB, the examiner needed to turn down the overall volume and repeat phase 3 (ie, tone volume setup) of the audio setup task. The new volume level was used to set the volume levels for the SM tasks.

The subjects performed bilateral finger tapping on button boxes when prompted by a 3-Hz audio cue and a reversing checkerboard visual cue. The block design involved the use of 15-second epochs of alternating baseline and task that yielded 85 acquisitions per run. The subjects were instructed how to do the tapping and were allowed to practice extensively. They tapped on a button box, and their responses were recorded.

Data Acquisition

Anatomic imaging.—At one site, transverse three-dimensional magnetization-prepared rapid acquisition gradient-echo images (9.8/minimum [repetition time msec/ineffective echo time msec], 15° flip angle, 22.0 × 16.5-cm field of view, 124–128 sections, 1.2-mm thickness, T1 of 300 msec, 256 × 192 matrix, bandwidth of ±15.625 kHz, two signals acquired) were obtained.

Functional imaging.—Functional MR images were acquired with echo-planar or spiral gradient-echo sequences (transverse oblique plane; repetition time, 3000 msec; echo time, 30 msec [at 3.0 and 4.0 T] or 40 msec [at 1.5 T]; flip angle, 90°; field of view, 22 cm; 35 sections; thickness, 4 mm; bandwidth ≥ ±100 kHz; matrix, 64 × 64; one shot; two dummy frames). The MR images in Figure 1 show how the appearance of the raw data can change from one site to another with different field strengths. These particular image sections were chosen because they show a large susceptibility artifact and distortion of the orbital frontal region. Although one would not expect activation on this section in response to finger tapping, one would expect activation in response to visual and auditory stimuli. The stability data obtained during the course of the human phantom study were incomplete because such data were not collected at all of the sites.

Per-Voxel Functional MR Imaging Analysis

With use of the AFNI software package (National Institute for Mental Health, Bethesda, Md, *http://afni.nimh.nih.gov/afni*), each functional MR imaging task run of 85 frames was motion corrected to the first image obtained during the first run of the visit. The functional data were spatially smoothed with a full width half maximum of 5 mm. Time series analysis was performed by using a Fourier-generalized linear model during the task period (30 seconds), and second-order polynomial regressors were used to model drift. Registered two-dimensional activa-

tion data were obtained. These analyses were performed by one of the authors (D.N.G.), who had 6 years of experience in functional MR imaging analysis.

Anatomic Analysis

T1-weighted anatomic MR image data were processed (by D.N.G.) by using the FreeSurfer software package (CorTech Labs, La Jolla, Calif; Athinoula A. Martinos Center for Biomedical Imaging, Charlestown, Mass, *http://surfer.nmr.mgh.harvard.edu*) to reconstruct the cortical surfaces in each subject (14–16). The surfaces were registered to a unit-spherical atlas, which was then used as a common coordinate space within which subjects were spatially compared. The anatomic and functional images were linearly spatially registered with each other to resample the functional significance maps onto the common-space (ie, spherical) surface.

Statistical Analyses

The level of activation at each voxel was assessed by using an *F* test on the sine and cosine task components. We examined the factors affecting the functional MR imaging brain activation patterns in the left and right hemispheres by using the spherical modeling described earlier. These statistical factors included subject ($n = 5$), study site ($n = 10$), examination visit ($n = 2$ or 4), imaging run ($n = 4$), field strength ($n = 3$), MR imaging unit vendor ($n = 3$), and k-space ($n = 3$) (Table 1). The analyses described in the paragraphs that follow were conducted jointly by two authors with 2 years (M.W.) and 3 years (K.H.Z.) of experience in statistical methods for functional MR imaging research. We used a *P* value of .05 to indicate statistical significance. The analytic software used included Matlab 7.0 (The MathWorks, Natick, Mass, *http://www.mathworks.com*) and S-Plus 6.0 (Insightful, Seattle, Wash, *http://www.insightful.com*).

Activation and estimated reference standard map.—The task-related significance at each voxel was computed by using an *F* test of the Fourier component of the task fundamental frequency. At each fixed voxel significance threshold, a recently developed expectation-maximization algorithm, simultaneous truth and performance level estimation, or STAPLE (17,18), was applied hierarchically. This algorithm combined all of the runs into a single composite activation that was visualized by using the three-dimensional Slicer software package (Massachusetts Institute of Technology, Cambridge, Mass; Brigham and Women's Hospital, Boston, Mass, *http://slicer.org*) (19). This algorithm implicitly relies on the assumptions that the user-defined regions of interest matched the spatial extent of the true activation region well and that a single region of interest (or set of regions of interest) was appropriate for all subjects. A series of thresholded three-dimensional brain activation maps was used to estimate the underlying reference standard by using an expectation-maximization algorithm and to assess the performance of each activation map. For the SM task, we assumed that there was a true underlying activation map, with each voxel either activated or not activated. Moreover, we assumed that each of the included runs was an imperfect version that was characterized directly by its sensitivity (ie, probability of correctly demonstrating an activated voxel) and specificity (ie, probability of correctly demonstrating an inactivated voxel).

STAPLE algorithm–derived estimated reference standard maps.—The hierarchical approach involved the following three steps, separately for the MR imaging examination visits of each study subject: In step 1, within each subject and at each site, all two-dimensional sections were combined to optimally derive a composite three-dimensional estimated reference standard (ERS) map for the four runs per visit. In step 2, within each subject, the ERS maps derived in step 1 were combined across all 10 sites. Finally, in step 3, the ERS maps constructed in step 2 were combined across all five subjects.

Sensitivity and specificity.—After applying the above algorithm to construct ERS maps, voxel fractions in the whole brain were used to compute the sensitivity (SEN) and specificity (SPEC) at a fixed voxel significance threshold (γ) of 10^{-9} adjusted for multiple comparisons: SEN = TAF = $P(Y > \gamma \mid V_{ERS} = ACT)$, and SPEC = TNAF = $P(Y \leq \gamma \mid V_{ERS} = NACT)$, where the activation threshold took into account the issue of multiple comparisons within active regions (20). In the above formulas, TAF is the true activation fraction, TNAF is the true nonactivation fraction, *P* is the probability, *Y* is the task-related significance, V_{ERS} is the voxel of the ERS, ACT means activated, and NACT means nonactivated.

Receiver operating characteristic curve.—Following the second step of the applied STAPLE algorithm, site-specific binormal parametric receiver operating characteristic (ROC) curves—plots of sensitivity versus (1 minus specificity) at all possible levels of activation thresholds—were generated from the activation data on a continuous scale. The area under each ROC curve (A_z) represented the overall classification accuracy, where $A_z = \phi[\alpha/(1 + \beta^2)^{1/2}]$ and $\phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. The binormal ROC parameters (α and β) were computed on the basis of their maximum likelihood estimates (20–23).

TABLE 1
List of Factors Examined

Factor	Value or Description
No. of subjects	5
Hemispheres	Left, right
No. of sites	10
No. of visits	2 (4)*
No. of runs	4 per visit
Field strengths used	1.5 T, 3.0 T, 4.0 T
MR unit vendors	Siemens, GE Healthcare, Picker†
k-Spaces used	Raster, spiral, dual-echo raster

* Three randomly chosen subjects underwent additional MR imaging examinations at two sites, for a total of four visits at two of the 10 sites.

† Siemens, Munich, Germany; GE Healthcare, Waukesha, Wis; Picker, Cleveland, Ohio.

Intraclass correlation coefficient.—With spherical modeling, within- and between-subject intraclass correlation coefficients (ICCs) were computed by performing a two-way analysis of variance of the fractions of the activated voxels, as compared against the ERS and stratified by hemisphere (left or right). The within-subject ICC was the fraction of the total variance due to the subject effect. A higher within-subject ICC would suggest a lower contribution of repetitions (over different runs, visits, and sites) to the overall variability and thus higher inter-subject variability. Conversely, the between-subject ICC was the fraction of the total variance due to the repetition effect. A higher between-subject ICC would suggest a lower contribution of the subjects to the overall variability and thus higher interrepetition variability (12).

Multiple regression analysis.—Multiple regression analyses were conducted to assess the significance of the factors and determine their associated *P* values (24).

RESULTS

Activation and ESR Maps

At the dichotomizing activation fixed voxel significance threshold of 10^{-9} to

TABLE 2
Mean Activation Percentages of All Voxels in the Brain, Mean Sensitivities, and Mean Specificities according to Field Strength and Subject

	Activation (%)		Sensitivity		Specificity	
Field Strength	Left Hemisphere	Right Hemisphere	Left Hemisphere	Right Hemisphere	Left Hemisphere	Right Hemisphere
1.5 T						
Subject 1	0.4312	0.4077	0.1857	0.2908	0.9958	0.9965
Subject 2	0.7264	0.8152	0.0870	0.0624	0.9974	0.9994
Subject 3	1.0440	0.8669	0.0737	0.1187	0.9992	0.9957
Subject 4	0.3308	0.2920	0.2250	0.3217	0.9967	0.9972
Subject 5	0.3115	0.2981	0.1292	0.2464	0.9970	0.9970
3.0 T						
Subject 1	2.8482	2.6434	0.6830	0.5592	0.9720	0.9747
Subject 2	7.7560	6.8143	0.6917	0.4609	0.9586	0.9865
Subject 3	8.9571	8.5409	0.4242	0.5846	0.9614	0.9343
Subject 4	4.2828	3.9779	0.8548	0.6839	0.9572	0.9604
Subject 5	2.0961	3.7310	0.8421	0.8259	0.9644	0.9627
4.0 T						
Subject 1	3.5646	3.1208	0.6565	0.5463	0.9649	0.9699
Subject 2	8.4144	8.3925	0.7319	0.5981	0.9540	0.9876
Subject 3	8.1396	7.0521	0.4273	0.5843	0.9713	0.9498
Subject 4	3.1219	2.0026	0.6250	0.5361	0.9688	0.9801
Subject 5	4.2311	3.8566	0.5600	0.4107	0.9582	0.9614

adjust for multiple comparisons, the mean activation percentages of all voxels were as follows: 0.3%–1.0% of voxels were activated at 1.5-T MR imaging; 2.1%–9.0% of voxels, at 3.0 T; and 2.0%–8.4% of voxels, at 4.0 T (Table 2, Fig 2).

Sensitivity and Specificity

At 1.5 T, mean sensitivities ranged from only 6% to 32% and mean specificities were higher than 99%. At 3.0 T, an improvement was observed, with mean sensitivities ranging from 42% to 85% and mean specificities ranging from 93% to 99%. With use of the 4.0-T MR imaging unit, which was available at one study site, mean sensitivities ranged from 41% to 73% and mean specificities ranged from 95% to 99% in the brain. Specific mean values are cited in Table 2.

ROC Curves

The ROC curves and the associated areas under the curve are presented in Figure 3 and Table 3, respectively. The ROC curves demonstrated moderate to high classification accuracy. Overall, the mean area under the ROC curve was 0.77 ± 0.05 at 1.5-T MR imaging, 0.90 ± 0.09 at 3.0 T, and 0.95 ± 0.02 at 4.0 T. There were significant differences between the 1.5- and 3.0-T values and between the 1.5- and 4.0-T values ($P < .01$ for both comparisons).

ICC Maps

The ICCs among the study sites ranged from 0.49 to 0.71. Between-subject ICCs calculated according to brain hemisphere and examination visit on separate days were as follows: during visit 1, 0.52 for the left hemisphere and 0.49 for the right hemisphere; and during visit 2, 0.71 for the left hemisphere and 0.66 for the right hemisphere. In contrast, the between-site ICCs were only 0.15 for the left hemisphere and 0.17 for the right hemisphere during visit 1 and 0.10 for the left hemisphere and 0.08 for the right hemisphere during visit 2; these values were generally smaller than the between-subject ICCs.

Multiple Regression Analysis

Finally, multiple regression analysis revealed that the factors significant for sensitivity were subject ($P < .005$), k-space ($P < .005$), and field strength ($P = .02$), whereas the factors significant for specificity were subject ($P = .03$) and k-space ($P = .05$) (Table 4).

DISCUSSION

Reliability studies of functional MR imaging signal intensities have typically involved the use of visual inspection or spatial similarity measures to assess the correspondence of activation maps across imaging runs or sessions. Visual inspection has been criticized for being subjective. Visual inspection measures are also difficult to use with a standardized reliability scale and thus complicate comparisons across studies. Other study investigators have used correspondence metrics to measure reliability. For example, a statistical threshold was applied to group-level activation maps to produce binary activated and nonactivated voxels for each of two replications (2,25). The number of voxels found to be activated in both replications, divided either by the number of activations observed in either condition or by the mean number of activations in each condition yielded the two correlation ratios—that is, the pixels activated in both iterations of the tasks in proportion to the pixels activated by either iteration of the task, and the ratio modified to include first-order neighboring pixels.

Studies of simple motor movement, tactile stimulation, and SM activity have revealed moderately high levels of reliability both within and between sessions. The purpose of collecting the SM data was to evaluate the reliability of multisite functional MR imaging—particularly, so that we could explore how the site-to-site variability compared with the intersubject variability. To do this, we wanted to use a task that the subjects could repeat as accurately as possible and with as few cognitive influences as possible, with the hope that the intersite variability would be the prominent source of variability. Although a language task was not included in this study, other cognitive tasks relevant to schizophrenia were. However, in this study, we were concerned mainly with site-related sources of variability, so we did not evaluate the cognitive data.

The cognitive paradigms studied have also yielded acceptable reliability values (2,26). Contrary to the findings described

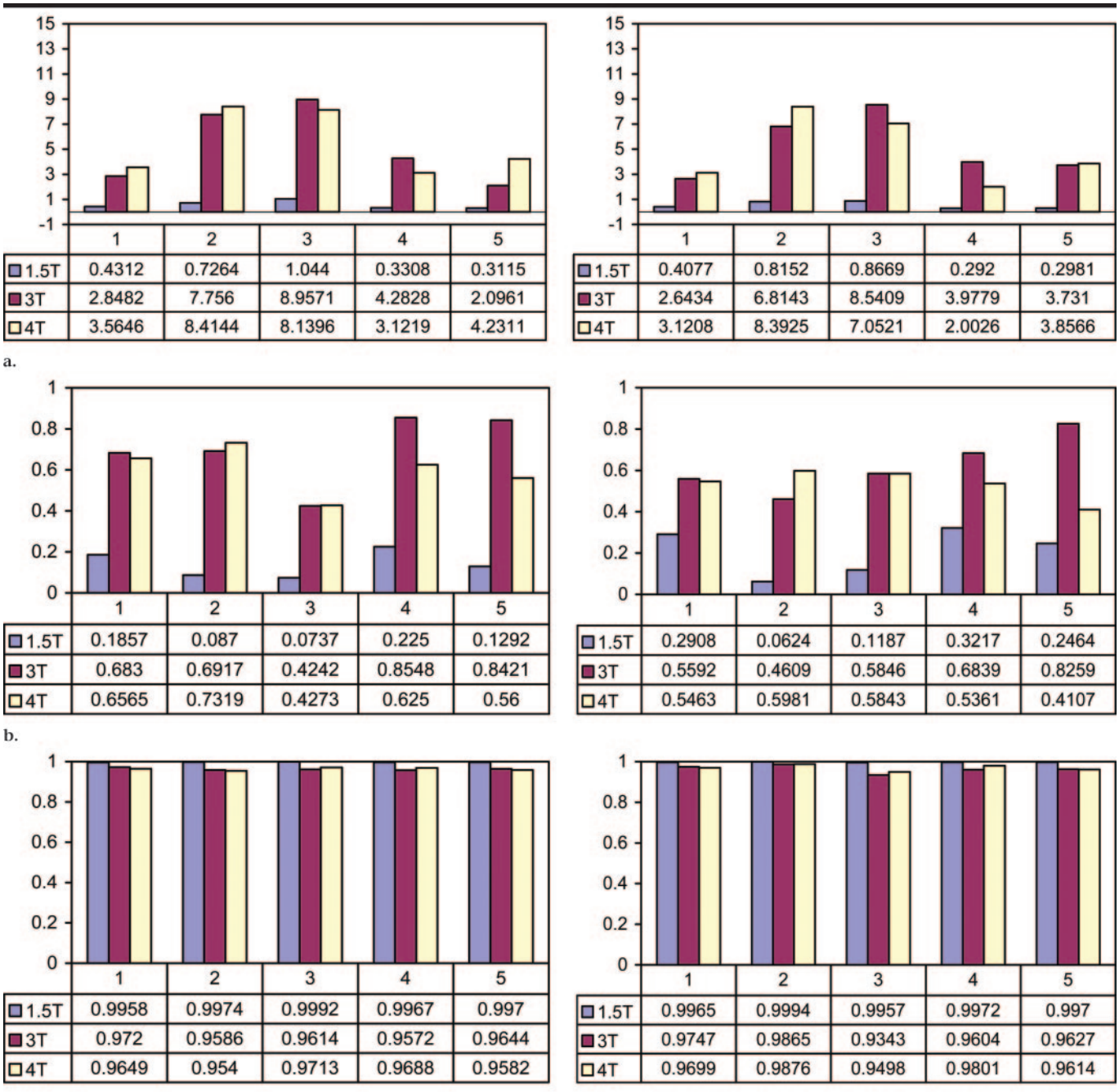


Figure 2. **(a)** Mean activation fractions (x-axis values) in left (left) and right (right) hemispheres at 1.5-, 3.0-, and 4.0-T MR imaging, derived according to subject (y-axis) and field strength by using step 1 in the expectation-maximization algorithm. **(b)** Mean sensitivities (x-axis values) of 1.5-, 3.0-, and 4.0-T MR imaging in left (left) and right (right) hemispheres, derived according to subject (y-axis) and field strength by using step 1 in the expectation-maximization algorithm. **(c)** Mean specificities (x-axis values) of 1.5-, 3.0-, and 4.0-T MR imaging in left (left) and right (right) hemispheres, derived according to subject (y-axis) and field strength by using step 1 in the expectation-maximization algorithm.

herein, emotional stimuli are often associated with habituation effects across runs and sessions and thus lead to poor reproducibility (27,28). A mixture model was used to formalize the relationship between the frequency at which a voxel was observed to be activated in a series of replications, and the underlying model parameters were used to estimate the true probability of voxel-level activations, as well as to determine the error rates conditioned on the true state of activation. With use of this method, the motor and cognitive tasks had similar between-session reliability once the false alarm rate was matched for the two types of tasks (29).

The use of correspondence measures to assess the reliability of functional MR imaging signal intensities has allowed for possible quantitative comparisons of functional MR imaging results across different behavioral paradigms and across studies. The voxel-counting methods based on dichotomizing statistical thresholds had

TABLE 3
Estimated ROC Parameters and Corresponding Areas under ROC Curve

Field Strength and Site No./ Vendor/K-space	Area under ROC Curve			
	Visit 1		Visit 2	
	Left Hemisphere	Right Hemisphere	Left Hemisphere	Right Hemisphere
1.5 T				
1/Siemens/raster	0.769	0.713	0.748	0.698
2/Siemens/raster	0.795	0.817	0.685	0.782
3/GE Healthcare/raster	0.715	0.705	0.776	0.755
4/GE Healthcare/spiral	0.816	0.834	0.770	0.773
5/Picker/raster	0.863	0.859	0.780	0.772
3.0 T				
6/Siemens/dual-echo raster	0.975	0.997	0.926	0.929
7/Siemens/raster	0.957	0.969	0.892	0.898
8/GE Healthcare/spiral	0.958	0.979	0.923	0.936
9/GE Healthcare/raster	0.737	0.740	0.778	0.776
4.0 T				
10/GE Healthcare/spiral	0.959	0.949	0.966	0.927

TABLE 4
***P* Values Derived from Testing the Significance of Factors for Sensitivity and Specificity**

Factor	Number	*P* Value	
		Sensitivity	Specificity
Subject	5	<.005*	.03*
Hemisphere	2	.90	.60
Site	10	.18	.08
Field strength	3	.02*	.11
Vendor	3	.11	.84
k-Space	3	<.005*	.05*
Visit	2(4)†	.17	.93
Run	4	.24	.08

* Given factor significant for sensitivity and/or specificity.
† Three randomly selected subjects underwent additional MR imaging examinations at two sites, for a total of four visits at two of the 10 sites.

less stability compared with the underlying regression slopes (30). In most correspondence studies, the activation maps were derived by aggregating information across the study subjects.

In this FIRST BIRN functional MR imaging project, we discovered the following significant factors in our preliminary analyses: *(a)* The effect of individual subjects had significant between-subject variability. Thus, calibration may be a critical component of the pooling mechanism of different subject cohorts. *(b)* The observed effects of different field strengths suggest that both 3.0-T and 4.0-T MR imaging examinations were significantly better than 1.5-T imaging, yielding more activation and less variability in terms of sensitivity and specificity. *(c)* The effects of k-space might have been due to different degrees of smoothing under varied k-space.

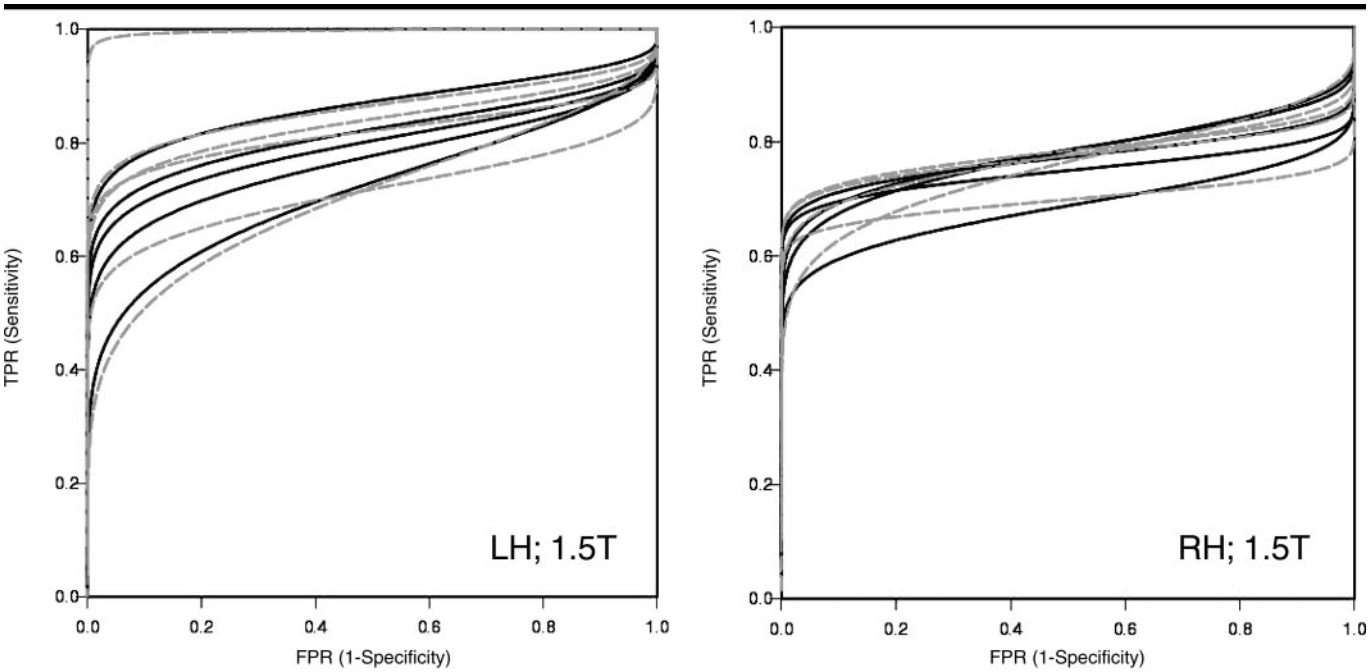
Other factors, although not statistically significant, led to the following observations: *(a)* With regard to the effect of repeated runs, a varied effect was observed across the runs after the rest and task periods; thus, the order of tasks might influence reproducibility. *(b)* With regard to the effect of study site versus subject, the variability across subjects appeared greater than the variability across sites. This finding may help in the development of a calibration plan to minimize the variability introduced by the sites themselves and ultimately enable us to pool the independent functional data of healthy and nonhealthy subjects across different institutions. *(c)* With regard to the effect of examination visit on different days, less activation was observed, and there was more robust and systematic activation at different thresholds during the second visit than during the first visit. In the three subjects who made four visits to one site, less activation was observed during the latter 2 days. However, there was higher examination specificity and less variability on these days.

Our study had several limitations: *(a)* Only five study subjects were included owing to the large volume of image acquisitions and tasks performed. *(b)* We conducted preliminary analyses based on whole-brain voxel counts essentially, without examining the regions of interest anatomically. *(c)* As mentioned earlier, we did not analyze the cognitive data (4). The current ERS required the use of the STAPLE algorithm in which binary activation data obtained with a particular activation threshold were used. When an expectation-maximization algorithm is used, convergence to the globally optimal estimate is not guaranteed. As a result, bias may occur (5). The activation across different test runs varied dramatically, whereas the sensitivity and specificity varied little because we used only hemispheres for stratified analyses rather than confine the analysis to the SM cortical region.

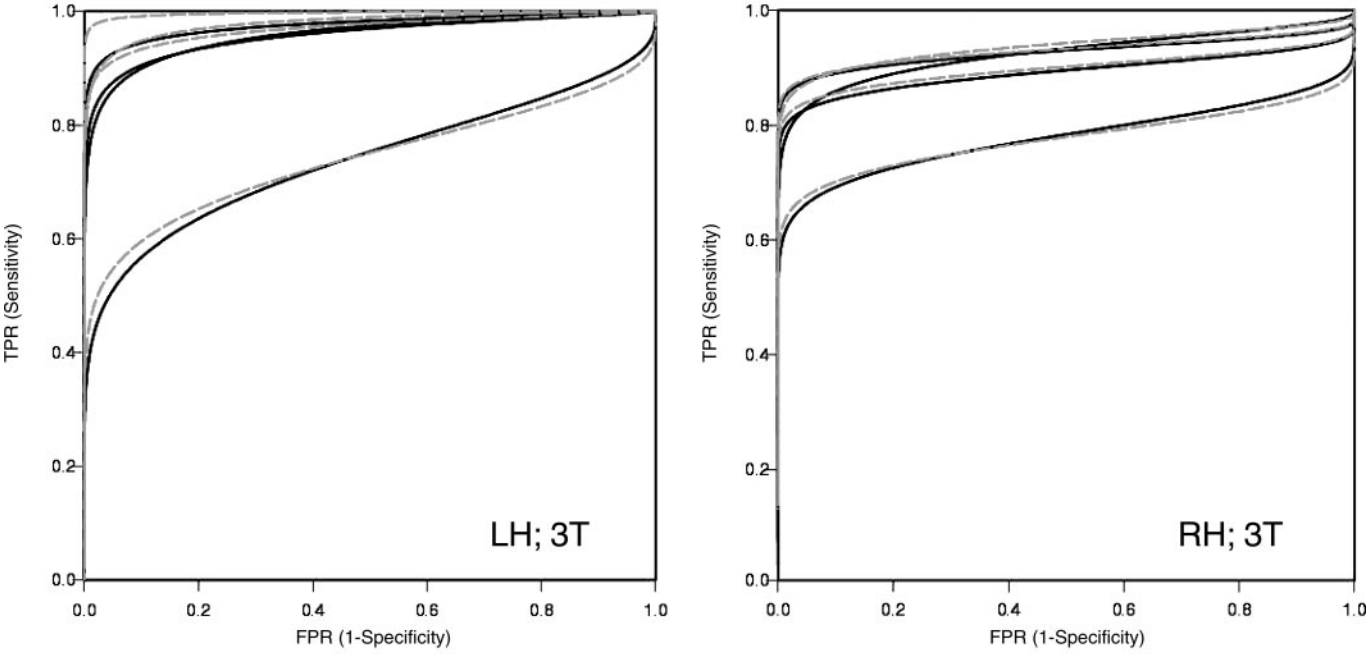
Some of the above limitations might be minimized with the development of a multisite consortium, like that in the BIRN research, with a high-speed broadband network supporting a large imaging database. Use of a common consortium protocol might facilitate calibration and validation across sites. In our research, we also recognized that although many sources of variability were minimized, they might not have been eliminated. Nevertheless, the findings of this prospective research will be useful for studying diseased brains with use of a common test bed and data-mining resource for the application of federated databases to complex clinical problems.

In summary, in a multi-institutional prospective functional MR imaging study, we observed higher reproducibility across study sites with 3.0- and 4.0-T imaging than with 1.5-T imaging, as well as significant effects due to subject and k-space. In general, there was greater activation and higher sensitivity—but lower specificity—at 3.0- and 4.0-T imaging than at 1.5-T imaging. Additional performance issues, such as controlling drift, and greater problems associated with areas of potential interest near the skull base need to be investigated.

Author contributions: Guarantor of integrity of entire study, K.H.Z.; study concepts/study design or data



a.

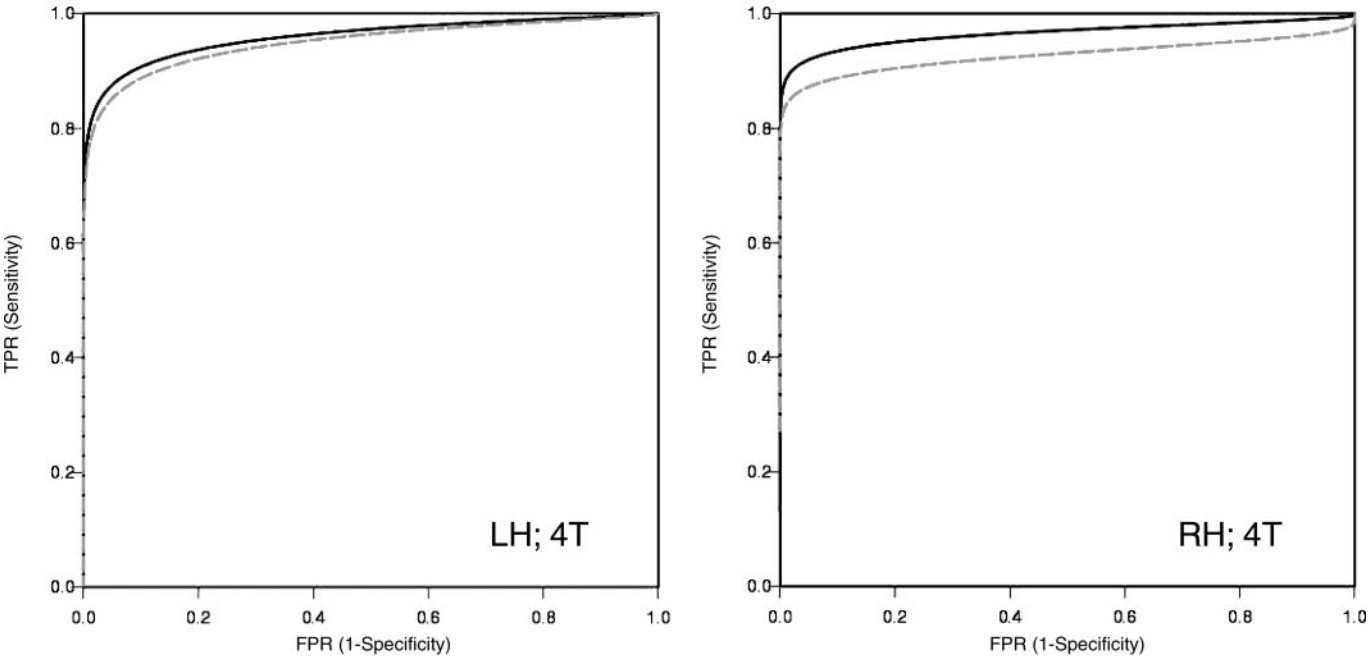


b.

Figure 3. ROC curves for left *(LH)* and right *(RH)* hemispheres in subject 2 at **(a)** 1.5-T and **(b)** 3.0-T MR imaging, derived according to examination visit. Solid lines represent visit 1 and dashed lines represent visit 2. *FPR* = false-positive rate, *TPR* = true-positive rate *(Fig 3 continues).*

acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; literature research, K.H.Z., M.W., S.K.W., S.M., G.G.B.; clinical studies, M.G.V.; experimental studies, S.D.P., S.K.W., G.G.B., W.M.W.; statistical analysis, K.H.Z., D.N.G., M.W., S.K.W., M.G.V., W.M.W.; and manuscript editing, K.H.Z., D.N.G., M.W., S.D.P., N.S.W., S.M., G.G.B., W.M.W.

Acknowledgments: We acknowledge with thanks the constructive comments from investigators and collaborators from all 11 participating institutions in the United States. We particularly acknowledge the efforts of several members: from Stanford University, Palo Alto, Calif, Gary H. Glover, PhD, and Lara Foland, BA; from the University of California, San Diego, Thomas Liu, PhD, and Anders M. Dale, PhD; from the University of California, Irvine, Steven G. Potkin, MD, Jessica Turner, PhD, Hal Stern, PhD, and David B. Keator, PhD; from the University of Iowa, Iowa City, Vincent A. Magnotta, PhD; from Duke University, Durham, NC, Ershela L. Sims, PhD; from Massachusetts General Hospital, Boston, Randy L. Gollub, MD; from Brigham and Women's Hospital, Boston, Mass, Cynthia G. Wible, PhD; from the University of New Mexico, Albuquerque, Lee



c.

Figure 3 *(continued).* **(c)** ROC curves for subject 2 at 4.0-T MR imaging.

Friedman, PhD; and from the University of Minnesota, Minneapolis, Bryon Mueller, PhD.

References

1. Aguirre GK, Zarahn E, D'esposito M. The variability of human, BOLD hemodynamic responses. Neuroimage 1998;8:360–369.
2. Brannen JH, Badie B, Moritz CH, Quigley M, Meyerand ME, Haughton VM. Reliability of functional MR imaging with word-generation tasks for mapping Broca's area. AJNR Am J Neuroradiol 2001;22:1711–1718.
3. Duann JR, Jung TP, Kuo WJ, et al. Single-trial variability in event-related BOLD signals. Neuroimage 2002;15:823–835.
4. Machielsen WC, Rombouts SA, Barkhof F, Scheltens P, Witter MP. fMRI of visual encoding: reproducibility of activation. Hum Brain Mapp 2000;9:156–164.
5. Miezin FM, Maccotta L, Ollinger JM, Petersen SE, Buckner RL. Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. Neuroimage 2000;11:735–759.
6. Le TH, Hu X. Methods for assessing accuracy and reliability in functional MRI. NMR Biomed 1997;10:160–164.
7. Genovese CR, Noll DC, Eddy WF. Estimating test-retest reliability in fMRI. I. Statistical methodology. Magn Reson Med 1997;38:497–507.
8. McGonigle DJ, Howseman AM, Athwal BS, Friston KJ, Frackowiak RS, Holmes AP. Variability in fMRI: an examination of intersession differences. Neuroimage 2000;11:708–734.
9. Maitra R, Roys SR, Gullapalli RP. Test-retest reliability estimation of functional MRI data. Magn Reson Med 2002;48:62–70.
10. Neumann J, Lohmann G, Zysset S, von Cramon DY. Within-subject variability of BOLD response dynamics. Neuroimage 2003; 19:784–796.
11. Casey BJ, Cohen JD, O'Craven K, et al. Reproducibility of fMRI results across four institutions using a spatial working memory task. Neuroimage 1998;8:249–261.
12. Wei X, Yoo SS, Dickey CC, Zou KH, Guttmann CR, Panych LP. Functional MRI of auditory verbal working memory: long-term reproducibility analysis. Neuroimage 2004;21:1000–1008.
13. Zou KH, Greve DN, Wang M, et al; and FIRST BIRN Research Group. A prospective multiinstitutional study for the reproducibility of fMRI: a preliminary report from the biomedical informatics research network. In: Barillot C, Haynor DR, Hellier P, part 1 eds. Proceedings of 7th International Conference of Medical Image Computing and Computer-Assisted Intervention: MICCAI 2004—lecture notes in computer science. Heidelberg, Germany: Springer, 2004;769–776, 3217.
14. Dale AM, Fischl B, Sereno MI. Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 1999;9:179–194.
15. Fischl B, Sereno MI, Dale AM. Cortical surface-based analysis. II. Inflation, flattening, and a surface-based coordinate system. Neuroimage 1999;9:195–207.
16. Fischl B, Sereno MI, Tootell RB, Dale AM. High-resolution inter-subject averaging and a coordinate system for the cortical surface. Hum Brain Mapp 1999;8:272–284.
17. Zou KH, Wells WM III, Kikinis R, Warfield SK. Three validation metrics for automated probabilistic image segmentation of brain tumors. Stat Med 2004;23:1259–1282.
18. Warfield SK, Zou KH, Wells WM III. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 2004;23:903–921.
19. Gering DT, Nabavi A, Kikinis R, et al. An integrated visualization system for surgical planning and guidance using image fusion and an open MR. J Magn Reson Imaging 2001; 13:967–975.
20. Genovese CR, Lazar NA, Nichols T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 2002;15:870–878.
21. Hanley JA, McNeil BJ. The meaning and use of the area under an ROC curve. Radiology 1982; 143:29–36.
22. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology 2003;229:3–8.
23. Zou KH, Warfield SK, Fielding JR, et al. Statistical validation based on parametric receiver operating characteristic analysis of continuous classification data. Acad Radiol 2003;10: 1359–1368.
24. Zou KH, Tuncali K, Silverman S. Correlation and simple linear regression. Radiology 2003; 227:617–628.
25. Yetkin FZ, McAuliffe TL, Cox R, Haughton VM. Test-retest precision of functional MR in sensory and motor task activation. AJNR Am J Neuroradiol 1996;17:95–98.
26. Kiehl KA, Liddle PF. Reproducibility of the hemodynamic response to auditory oddball stimuli: a six-week test-retest study. Hum Brain Mapp 2003;18:42–52.
27. Phan KL, Liberzon I, Welsh RC, Britton JC, Taylor SF. Habituation of rostral anterior cingulate cortex to repeated emotionally salient pictures. Neuropsychopharmacology 2003;28: 1344–1350.
28. Stark R, Schienle A, Walter B, et al. Hemodynamic effects of negative emotional pictures: a test-retest analysis. Neuropsychobiology 2004;50:108–118.
29. Noll DC, Genovese CR, Nystrom LE, et al. Estimating test-retest reliability in functional MR imaging. II. Application to motor and cognitive activation studies. Magn Reson Med 1997;38:508–517.
30. Cohen MS, DuBois RM. Stability, repeatability, and the expression of signal magnitude in functional magnetic resonance imaging. J Magn Reson Imaging 1999;10:33–40.