Original Research

Reproducibility of Single-Subject fMRI Language Mapping With AMPLE Normalization

James T. Voyvodic, PhD*

Purpose: To evaluate the reproducibility of presurgical functional MRI (fMRI) language mapping based on test–retest scans, comparing traditional activation t-maps to relative activation maps normalized by activation mapping as percentage of local excitation (AMPLE).

Materials and Methods: Language fMRI scans were performed by 12 healthy volunteer subjects undergoing a standard clinical presurgical mapping protocol in multiple independent scan sessions. Objective relative AMPLE activation maps were generated automatically by normalizing statistical t-value maps to the local peak activation amplitude within each functional brain region. The spatial distribution of activation was quantified and compared across mapping algorithms, subjects, scanners, and pulse sequences.

Results: The spatial distribution of traditional blood oxygen level-dependent (BOLD) t-value statistical activation maps was highly variable in test–retest scans of single subjects, whereas AMPLE normalized maps were highly reproducible in terms of the location, hemispheric laterality, and spatial extent of relative activation. AMPLE map reproducibility was good regardless of scanner, field strength, or pulse sequence used, but reproducibility was best for scans acquired on the same scanner using the same pulse sequence.

Conclusion: Reproducibility of the spatial pattern of BOLD activation makes relative amplitude fMRI mapping a useful normalization tool for clinical imaging of language function, where reproducibility and quantitative measurements are critical concerns.

Key Words: fMRI; reproducibility; clinical language mapping; AMPLE
J. Magn. Reson. Imaging 2012;000:000–000.

LOCALIZATION OF BRAIN regions essential for receptive and expressive language function is critically important in neurosurgical treatment planning for brain tumors, epilepsy, and other diseases. Blood oxygen level-dependent (BOLD) functional MRI (fMRI) has recently become a routine clinical procedure for localization of essential language and motor brain regions, largely replacing more invasive presurgical diagnostic methods (1,2). However, an obstacle to broader clinical application is the fact that standard fMRI methodologies tend to produce results that are neither quantitative nor reproducible. Even for simple hand movement tasks, previous studies have reported that multiple scans of a single individual performing the same motor behavioral task typically produce generally similar brain activation maps using standard statistical mapping methods, but with significant variability in the details of active regions identified in different scans (3–5). Quantitative reproducibility of single subject motor cortex mapping has been shown to improve when a local normalization approach, AMPLE, is used for defining activation thresholds, rather than an arbitrary fixed statistical threshold criterion (6,7). For language mapping tasks, previous studies with fMRI have reported relatively good intrasubject reproducibility for determining hemispheric laterality of language dominance (8–11) but relatively poor reproducibility for localization of expressive and receptive language centers (putative Broca's and Wernicke's areas, respectively) within the dominant hemisphere (12,13).

Assessing the quantitative reproducibility of language localization is considerably more difficult than for motor mapping because language function typically involves multiple different brain areas, none of which are anatomically well-defined (14,15). Many different behavioral paradigms have been developed that attempt to localize major language centers (e.g., Ramsey et al, and Pillai and Zaca) (16,17). Each language task activates language-specific areas, but in doing so also engages a complex network of sensory, motor, attention, and decision-making brain functions. Although all paradigms attempt to control for nonlanguage task components, even well controlled tasks typically result in multiple brain regions with statistically significant task-dependent BOLD signals, which can vary significantly across individuals.

Brain Imaging and Analysis Center, Radiology Department, Duke University Medical Center, Durham, North Carolina USA.

Contract grant sponsor: NIH; Contract grant number: P01NS041328; Contract grant sponsor: the Quantitative Imaging Biomarker Alliance of the Radiological Society of North America.

*Address reprint requests to: J.T.V., Brain Imaging and Analysis Center, Duke University Medical Center, 2424 Erwin Road, Suite 501, Durham, NC 27705. E-mail: jim.voyvodic@duke.edu

Received August 17, 2011; Accepted March 20, 2012.
DOI 10.1002/jmri.23686
View this article online at wileyonlinelibrary.com.

Ideally, a single individual performing the same task on different MRI scanners should yield similar brain activation maps, but differences in scanner model, magnetic field strength, pulse sequence, and image reconstruction algorithms all influence fMRI activation results (18–20). Activation patterns within single subjects are also highly dependent on such variables as task performance, head motion, attention, anxiety, and other physiological factors (e.g., Abbott et al) (21). The activation map itself is the result of the specific image analysis methods used, and is very sensitive to statistical activation threshold settings (6). Making fMRI reproducible, therefore, depends on identifying and controlling as many sources of variability as possible.

Assessing reproducibility requires objective ways of quantifying fMRI results to compare activation across repeated scans. Language mapping, for example, can be quantified in at least three clinically useful ways. The first is a laterality index for language dominance, calculated separately for frontal (expressive) and temporoparietal (receptive) areas as a ratio of active voxels in the left and right hemispheres (e.g., 22–26). Knowing whether a lesion is in the functionally dominant hemisphere is critical for treatment risk assessment. The second way is to localize the brain location (in three-dimensional [3D] coordinates) for the center or peak of individual language areas. For treatment within the dominant hemisphere, identifying language centers allows clinicians to assess risk and plan their treatment approach, and it facilitates intraoperative localization of eloquent cortex during surgery (27). The third is to quantify the spatial extent of individual active areas, which provides an additional treatment risk assessment for planning how much of a lesion can be safely resected. The ability to obtain objective and reproducible values for any of these spatial metrics would significantly enhance confidence in the reliability and usefulness of clinical fMRI language mapping.

The current study measured reproducibility of all three of these quantitative metrics of language function in repeated fMRI scans of healthy volunteer subjects. Each subject underwent a standard clinical fMRI language mapping protocol in two or more separate scan sessions. The goal of the study was to quantify multiple aspects of task-dependent brain activation using a variety of different sampling methods to see which yielded the most consistent results. Clinical fMRI mapping was tested in healthy subjects to establish baseline reproducibility measures independent of disease variables. To obtain objective quantification all image analysis procedures were automated and the same analysis was applied to the images for each scan. This was done to avoid a subjective bias due to manual adjustment of activation thresholds or selection of brain regions of interest (ROIs). The same behavioral task design was used in each scan session so that the pattern of brain activity for each subject should be similar across scans. Across the twelve subjects, however, different MRI scanners and pulse sequences were used in different scan sessions to test the robustness of the fMRI mapping metrics to brain function itself, rather than to the particular imaging device being used.

MATERIALS AND METHODS

Subjects

Twelve healthy volunteer subjects (female:male = 4:8; right:left handed = 11:1; age = 18–56 years; mean = 28.7 years) gave informed consent and underwent 2 or more fMRI scan sessions. All subjects were fluent speakers of English. Intervals between scan sessions ranged from 1 h to 6 years. Four MRI scanners were used: a GE 1.5 Tesla (T) Signa, GE 4T Signa, GE 3T Excite, and GE 3T 750. BOLD T2*-weighted fMRI scans used either linear echoplanar trajectories, or spiral trajectories (outward or inward spirals). All fMRI scans were 64 × 64 acquisitions. Details of the scanning parameters that varied across scans are shown for each subject in Table 1. A set of 256 × 256 T2-weighted anatomical images coplanar to the functional images was also collected in each scan session. For each subject at least one high-resolution (1 × 1 × 1 mm voxel) whole brain T1-weighted anatomical image data set was obtained.

Language Task

All subjects performed a standard clinical sentence-completion behavioral task for fMRI language mapping. For this task, subjects viewed short incomplete sentences or nonsense black text projected on a white screen viewed by means of a mirror. Example stimuli included sentences such as "The current month is ____", and nonsense text such as "Fvp swvflmmjr smw fvp ____." Stimuli were presented in a block design, with alternating 24-s blocks of sentences and nonsense; each block consisted of 4 stimuli seen for 6 s each. This silent reading task forces subjects to engage both receptive and expressive language centers. It is routinely used for clinical fMRI language mapping (27).

Image Processing

Except where specified otherwise, all image processing was performed using the fScan program for fMRI (6,7,28). Before statistical mapping, head motion was measured and overall head motion index metrics were calculated as mean and maximum in-plane translational head motion across all images, relative to the mean head position for each brain slice. To compare scans acquired in different sessions, spatial registration affine transformation matrices were generated to align each functional scan to one reference whole brain T1-weighted anatomical data set for that subject. Another affine transformation matrix was calculated for each subject by aligning the T1-weighted reference scan to the standard MNI-152 average T1-weighted data set. To avoid introducing interpolation errors these alignment steps only generated the transformation matrices without actually transforming image data; all statistical maps were generated using untransformed images. The affine transforms were only used when comparing quantitative mapping results across scans and for visualizing maps

Table 1
Subject and Scanning Details*

Subj	Sex	HP	Sess	Scan	Age	Day	Scanner	PSeq	NZ	NT	DX	DZ	TR	Motion
1	M	R	1	1	50	0	4T	OSprl	22	256	3.75	5	1.5	.13/1.6
1	″	″	1	2	50	0	4T	ISprl	22	256	3.75	5	1.5	.14/0.8
1	″	″	2	3	50	0	1.5T	ISprl	22	256	3.75	5	1.5	.21/1.3
1	″	″	2	4	50	0	1.5T	EPI	22	192	3.75	5	2	.17/3.0
1	″	″	3	5	56	2120	3T2	EPI	24	256	3.75	5	1.5	.22/3.6
1	″	″	4	6	56	2349	3T1	EPI	22	256	3.75	5	1.5	.18/5.2
2	F	R	1	1	26	0	1.5T	EPI	22	192	3.75	5	2	.71/2.5
2	″	″	1	2	26	0	1.5T	ISprl	22	192	3.75	5	2	.14/3.0
2	″	″	2	3	26	0	4T	ISprl	40	192	3.75	3	2	.23/1.9
3	M	R	1	1	18	0	1.5T	EPI	22	192	3.75	5	2	.11/1.2
3	″	″	1	2	18	0	1.5T	ISprl	22	192	3.75	5	2	.15/1.5
3	″	″	2	3	18	0	4T	OSprl	30	192	3.75	3	2	.17/1.1
4	F	R	1	1	19	0	3T1	EPI	22	256	3.75	5	1.5	.19/1.8
4	″	″	2	2	21	528	3T2	EPI	24	256	3.75	5	1.5	.14/0.8
4	″	″	2	3	21	528	3T2	EPI	24	256	3.75	5	1.5	.16/3.0
5	M	R	1	1	22	0	1.5T	EPI	22	192	3.75	5	2	.15/1.2
5	″	″	2	2	22	0	4T	ISprl	20	256	3.75	5	1.5	.13/1.0
6	M	R	1	1	23	0	3T2	EPI	24	256	3.125	4	1.5	.09/0.5
6	″	″	2	2	23	1	3T1	EPI	22	256	3.75	5	1.5	.13/0.8
7	M	L	1	1	21	0	3T2	EPI	22	256	3.125	4	1.5	.13/0.7
7	″	″	2	2	21	2	3T1	EPI	22	256	3.75	5	1.5	.13/1.3
8	F	R	1	1	21	0	3T2	EPI	22	256	3.125	4	1.5	.22/2.7
8	″	″	2	2	21	2	3T2	EPI	24	256	3.75	5	1.5	.25/5.2
9	M	R	1	1	28	0	3T2	EPI	24	256	3.75	5	1.5	.18/2.8
9	″	″	2	2	28	59	3T2	EPI	24	256	3.75	5	1.5	.25/1.3
10	M	R	1	1	31	0	3T1	EPI	22	256	3.75	5	1.5	.17/5.0
10	″	″	2	2	32	298	3T2	EPI	24	256	3.75	5	1.5	.12/4.0
11	M	R	1	1	19	0	3T1	EPI	22	208	3.75	5	1.5	.15/1.0
11	″	″	2	2	20	83	3T1	EPI	22	256	3.75	5	1.5	.14/1.2
12	F	R	1	1	27	0	3T1	EPI	22	256	3.75	5	1.5	.29/1.7
12	″	″	2	2	27	66	3T1	EPI	22	256	3.75	5	1.5	.19/1.1

*Each row is a different language mapping scan. The columns are as follows: Subj = Subject ID number (1–12); Sex = Male/Female; HP = self-reported handedness preference; Sess = Session number; Scan = Language scan number within subject; Age = Subject age (yrs); Day = Number of days since the subject's first scan session; Scanner = Scanner used (3T1 = Excite, 3T2 = MR750); PSeq = pulse sequence (EPI = linear EPI; OSprl = outward spiral trajectory; ISprl = inward spiral trajectory); NZ = Number of slices; NT = Number of time points; DX = Voxel X/Y size (mm); DZ = Slice thickness (mm); TR = TR interval (seconds); Motion = Translational head motion in mm (2 values are mean/maximum motion across all brain slices).

superimposed on the standard 3D brain surface (described below).

Generation of Statistical Activation Maps

To generate t-maps and percent-signal maps using fScan, images were first spatially smoothed using a 7-mm radius kernel; no other image preprocessing was performed. No explicit motion correction was performed to avoid introducing interpolation errors. Statistical t-maps were generated by binning MR images by stimulus condition offset by 4.5 s to allow for hemodynamic delay. Noise in the t-maps was reduced by spatially smoothing the computed map using a 7-mm radius kernel. Percent-signal change maps were generated as the mean change in intensity value between task and rest conditions expressed as percentage of the mean signal for each voxel.

Activation t-maps generated by fScan were validated by comparison to t-maps generated independently using the popular FEAT processing pathway in the FSL analysis package (29). FEAT analysis included image preprocessing where each language scan was motion corrected, smoothed with a 7-mm gaussian kernel, slice time adjusted to compensate for interleaved acquisition, spatially masked to remove non-brain regions, intensity normalized, and temporally filtered to remove low and high frequency signals. FEAT statistical activation maps were generated using a general linear model approach based on the timing of the language task blocks convolved with a standard FSL double-gamma hemodynamic response function; the derivative of the language waveform was added to the model to cope with transition effects. The first 6sec of each language scan was omitted to compensate for T1 effects before reaching steady-state levels. Only 189 image volumes were included in the FEAT analysis to avoid variability due to different sample sizes across scans.

Activation Mapping as Percent of Local Excitation (AMPLE) Normalization of Statistical Activation Maps

The AMPLE normalization algorithm (6,7) was used to normalize statistical activation values within local brain regions of interest (ROIs). All AMPLE

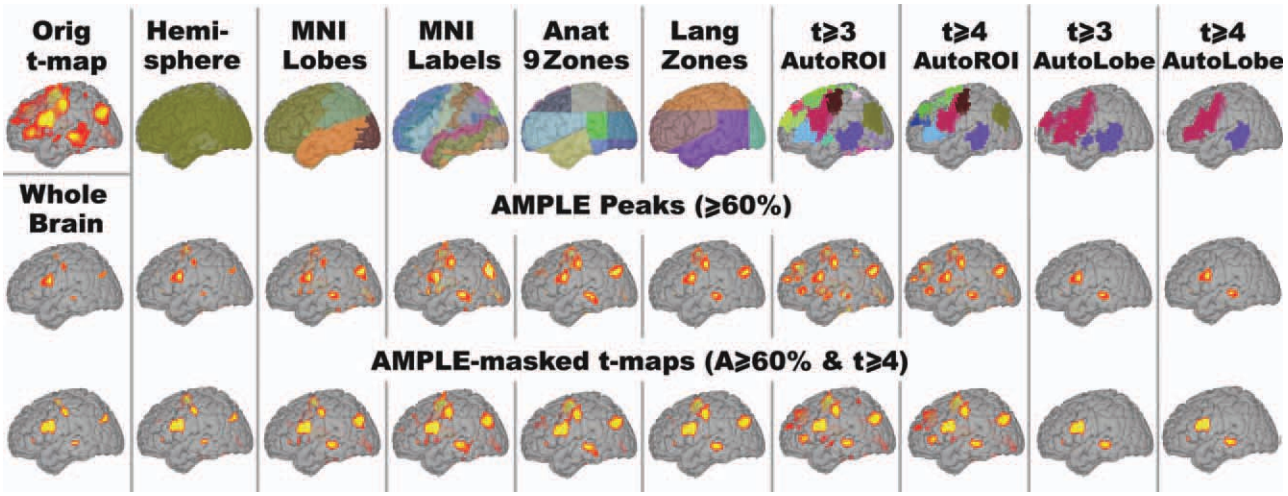

Figure 1. Region of interest masks and AMPLE maps. In the top row, a representative t-value thresholded ($t \geq 4$) activation map (Orig t-map) and nine atlas and auto-segmented cluster ROI masks are shown superimposed on a FreeSurfer reconstruction of the MNI brain surface. The second row shows the AMPLE normalized maps (AMPLE threshold 60%) created by using a whole brain ROI mask or the nine atlas and auto-segmented masks directly above. The third row shows the AMPLE-masked results obtained by selecting subsets of the Orig t-map voxels using the middle row of AMPLE maps (with AMPLE value $\geq$ 60%) as masks.

normalization was performed using fScan-generated t-maps. The AMPLE algorithm converted all voxel activation t-values to a percentage of the peak activation t-value within each ROI. Because of the heterogeneity of fMRI language maps across different subjects, a variety of different types of ROI masks were tested to determine which provide the most robust and reproducible results (Fig. 1). For all ROI masks used except the Left/Right ROIs in the Hemisphere atlas, AMPLE normalization was performed by treating homologous brain regions in the left and right hemisphere as the same ROI to maintain relative differences in hemispheric laterality; when those atlases were used for quantifying activation, left and right hemisphere portions of each ROI were assessed separately. The simplest type of ROI mask used was a whole-brain ROI, in which the activation of every voxel was normalized to the most active voxels in the brain. We also used standardized atlas-based ROIs extracted from the Hemisphere, Lobe, and anatomic Label data sets of the WFU PickAtlas (30) (http://fmri.wfubmc.edu/software/PickAtlas).

Two new customized anatomical atlas masks were created specifically designed to subdivide the brain into ROIs appropriate for language mapping. One, referred to as the 9Zones mask (Fig. 1), segmented the cerebral cortex of the standard MNI brain into nine zones (anterior/central/posterior X superior/middle/inferior regions), using the Sylvian fissure as one boundary, and placing other boundaries in axial and coronal planes chosen to separate middle frontal and temporoparietal regions from superior motor regions and posterior visual areas. The second custom atlas, LangZones (Fig. 1), was a refinement in which the nine zones were reduced to only four regions in each hemisphere: large frontal and temporoparietal zones for language areas, and superior and posterior zones to isolate motor and visual areas.

In addition to these static ROI masks, language activation t-maps were also segmented into scan-

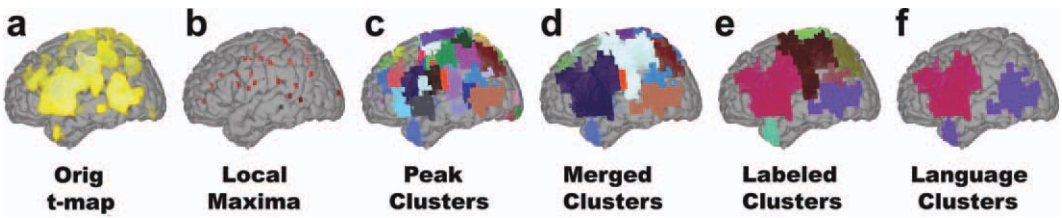

Figure 2. Auto-segmentation of fMRI cluster ROIs. The segmentation algorithm involved: **a**: setting a minimum statistical activation threshold (e.g., $t \geq 3$). **b**: identifying local intensity maxima based on a nearest neighbor analysis of the super-threshold voxels. **c**: Color-coding and then dilating local maxima to create ROIs by adding adjacent voxels with equal or lower t-value until all super-threshold voxels are included. **d**: Merging adjacent clusters if the depth of the intensity valley between the cluster maxima is less than 60% of the intensity of the lower peak, repeated until no ROIs met the merging criterion. **e,f**: Merged ROIs were (E) manually assigned labels based on anatomical lobes, and then (F) language-related clusters were created by merging frontal and temporoparietal clusters into two large ROIs in each hemisphere.

specific cluster-based ROIs using a locally developed automated algorithm (Fig. 2). This algorithm involved first setting a minimum statistical activation threshold and then identifying all local activation maxima based on a nearest neighbor analysis of the super-threshold voxels. Each maximum was assigned a unique ROI number, and then each of those ROIs was dilated by adding adjacent voxels with equal or lower t-values; this dilation process was repeated until all super-threshold voxels were assigned to a cluster ROI. Neighboring ROIs were then merged if the depth of the activation valley between the cluster maxima was less than 60% of the activation value of the lower peak; this process was repeated until no neighboring ROIs satisfied the merging criterion. Each language map was auto-segmented into two different cluster ROI maps: one based on an initial activation threshold of t-value > 3.0 and one with t-value > 4.0.

To compare auto-segmented clusters across different language scans, frontal and temporoparietal cluster ROIs in each hemisphere were manually relabeled so that clusters in similar relative positions would have similar labels (Fig. 2E). This manual step did not modify the cluster ROIs; it simply assigned consistent labels and color coding to facilitate comparing homologous ROIs across scans.

The result of statistical thresholding, auto-segmentation, and cluster labeling was thus an ROI mask with up to 12 putative language areas (1–3 frontal or temporoparietal ROIs in each hemisphere) customized for each fMRI scan; these are referred to as the AutoROI atlases. Each of these AutoROI masks was then used to create another atlas with only 4 putative language areas (referred to as an AutoLobe atlas) simply by merging the labeled clusters into a single ROI for each frontal or temporoparietal lobe and ignoring unlabeled ROIs elsewhere in the brain (Fig. 2F).

Calculation of Quantitative Activation Metrics

The activation signal for each fMRI scan was quantified by combining the original fScan-generated t-maps with 10 separate ROI masks, with and without AMPLE normalization. The 10 ROI masks were: whole brain, PickAtlas Hemisphere, Lobe, and Label masks, locally defined 9Zone and LangZone masks, and auto-segmented AutoROI and AutoLobe masks using segmentation t-value threshold of 3.0 and 4.0 (Fig. 1). For each AMPLE map, a set of AMPLE-masked t-maps was generated by selecting those voxels in the original t-map that were above an AMPLE threshold (AMPLE $\geq$ 40%, 60%, 80% of the peak). These AMPLE-masked t-maps (Fig. 1, bottom row) combined the local-relative spatial properties of the AMPLE maps with the global-relative activation amplitude information contained in the original t-maps.

The activation signal for each fMRI scan was quantified by sampling the original t-map and AMPLE-masked t-maps, as well as the percent-signal change map, for all 10 ROI masks using five different t-value thresholds (t $\geq$ 3, 4, 6, 8, 10) and three AMPLE thresholds (A $\geq$ 40, 60, 80%). For each ROI mask, this involved first resampling both the original statistical map and the ROI mask to standard 1 mm^3 voxels in MNI152 brain coordinates so that all results would be expressed in the same spatial units. Then the thresholded statistical map was combined with the ROI mask to calculate the number of super-threshold voxels, mean voxel intensity, hemispheric laterality index, and 3D spatial location of the peak center of activation in each ROI. Laterality index (LI) was calculated as the number of super-threshold voxels in homologous ROIs in the left and right hemisphere as:

$$\mathrm{LI} = (\mathrm{NVoxLeft} - \mathrm{NVoxRight})/(\mathrm{NVoxLeft} + \mathrm{NVoxRight})$$

To remove the need for thresholding, activity-weighted LI values (WLI) were calculated by comparing summed voxel intensity values in corresponding ROIs in each hemisphere as:

$$\mathrm{WLI} = (\mathrm{Sum}(\mathrm{VoxLeft}) - \mathrm{Sum}(\mathrm{VoxRight}))/(\mathrm{Sum}(\mathrm{VoxLeft}) + \mathrm{Sum}(\mathrm{VoxRight}))$$

where VoxLeft/Right represents voxel intensities in the left/right ROI.

The 3D location of the ROI peak of activation was expressed as the spatial coordinates in MNI-space of the voxel with the highest (smoothed) activation value, and an intensity-weighted center of activation was calculated for all super-threshold voxels along each axis (XYZ) as:

$$\mathrm{WCtr_X} = \mathrm{Sum}(\mathrm{X} * \mathrm{I})/\mathrm{Sum}(\mathrm{X})$$

where X is voxel location in standard MNI coordinates along a principal axis and I is activation intensity at that voxel. Separate sets of activation statistics were generated for every labeled ROI in each of 200 distinct thresholded and masked t-map combinations for each language scan.

Calculation of Reproducibility Metrics

These quantitative activation metrics were compared across all scans for each subject to assess reproducibility as a function of acquisition methods and as a function of traditional versus normalized analysis methods. Because subjects performed different numbers of language scans and many scans were acquired under different scanning conditions, reproducibility metrics were calculated separately for every possible pair combination of two scans for the same subject. Thus, data for subject 1 involved 15 pair comparison combinations of six scans, subjects 2–4 had 3 pair comparisons of three scans, and subjects 5–12 each had only 1 pair of scans to compare. For each ROI we calculated the percent change in number of active voxels ($\mathrm{NVoxel_{pct}}$) and the 3D distance between the intensity-weighted centers of activation (dWCtr, in mm). The percent agreement in laterality index was calculated as:

$$\mathrm{LI_{pct}} = \mathrm{Min}(\mathrm{Abs}(\mathrm{LI_1}, \mathrm{LI_2}))/\mathrm{Max}(\mathrm{Abs}(\mathrm{LI_1}, \mathrm{LI_2}))\ \mathrm{X}\ 100$$

where $\mathrm{LI_{pct}}$ was set to zero if $\mathrm{LI_1}$ and $\mathrm{LI_2}$ had different signs. We calculated the percent voxel overlap as the

ratio of the number of voxels that were active in both maps to the total number active in either map. All reproducibility metrics were calculated by resampling each map to standard 1 mm^3 voxels in MNI space and comparing activation statistics in corresponding ROIs in each map pair.

Visualizing Activations

To facilitate comparison of results across multiple scans of multiple subjects, all color-coded activation and ROI maps were rendered superimposed on a standard FreeSurfer (31) "pial surface" reconstruction of the single-subject MNI brain. For these renderings activation maps were resampled in standard MNI coordinate space and then overlaid into the semi-transparent 3D surface to show color-coded voxels below the pial surface.

RESULTS

Details of the 12 subjects, scan timing, scan parameters, and head motion estimates for all 31 language fMRI scans obtained are shown in Table 1. Subjects performed each task successfully with little head motion (mean in-plane translational motion < 1 mm) resulting in statistically significant BOLD activation signal in every scan (Fig. 3). All 11 right-handed subjects showed clearly left-hemisphere language dominance, whereas the left-handed subject was clearly right-hemisphere dominant for language function. Within the dominant hemisphere, averaging the sentence completion task t-maps for all scans for all subjects resulted in significant BOLD activation in inferior-frontal and superior temporal language regions as well as more superior frontal motor and supplementary motor areas (Fig. 3A). Spatial overlap of active areas averaged across all scans was somewhat higher in the frontal lobe than for temporal areas. Because the spatial overlap of the temporal language activation spread across the temporal-parietal boundary, in our analysis we refer to this as the temporoparietal language region. The magnitude of the task-dependent BOLD signal (expressed as percent signal change) averaged across all scans showed peaks in frontal and temporoparietal regions similar to the t-value maps, with an additional signal peak in the anterior Sylvian fissure. The fact that this large BOLD signal peak is not seen in the t-maps presumably reflects large signal variability in this highly vascular brain area, which reduces the statistical significance of the signal in that region.

Qualitative Reproducibility of Activation Maps

Visual comparison of the individual t-maps (thresholded at $t \geq 4$) obtained for each subject showed that the overall pattern of activation was similar across multiple scans of the same subject, but there was considerable variation in the number and distribution of statistically significant active voxels in each scan (columns 1 and 2 in Fig. 3B–M). This variability was partly due to global differences in activation strength across different scans, but also involved differences in the spatial distribution of active areas even when overall activation levels were comparable. Similarly, the percent-signal change maps (BOLD signal $\geq 1\%$) varied in both magnitude and spatial distribution across different scans of the same subject (column 3 in Fig. 3).

To verify that the variability across repeat language scans was not due to the minimal preprocessing approach used to generate fScan t-maps, another complete set of activation t-maps was independently generated for each language scan using the full FEAT preprocessing and general linear model approach (Fig. 4). Only the first 189 time points from each scan were included to control for different acquisition series lengths. The activation maps produced by the FEAT analysis demonstrated variability across multiple scans for each subject similar to that observed in the fScan maps. The FEAT results thus confirm that variability across repeat scans was not due to differing TRs, scan lengths, or some aspect of fScan's processing software, and that variability is not eliminated by motion correction or other standard preprocessing steps.

Normalizing t-value activation maps to local peak activation by applying the AMPLE algorithm, however, did greatly reduce intra-subject variability. Figure 1 shows the results for a single representative fScan-generated t-map normalized using all ten ROI mask data sets. Atlas masks with large anatomical ROIs (whole brain, Hemisphere, MNI Lobe, LangZone, and AutoLobe) tended to result in only a few peaks in AMPLE-masked maps because only the strongest signals reached threshold. Because the superior eye-movement areas, and sometimes visual areas, produce strong and variable BOLD signals in the sentence-reading task, those areas were typically the most prominent peaks in any ROI that included them. Normalization using large ROI mask atlases did not result in consistent maps across repeated scans, confirming that the differences between scans of the same subject were not simply due to differences in global activation levels. Normalization based on atlas data sets with small ROIs (MNI Labels, 9Zones, and AutoROIs) resulted in larger numbers of active areas in AMPLE maps, with the pattern of active areas varying somewhat depending on where ROI boundaries were with respect to local peaks. In general, the 9Zone, LangZone, and AutoLobe atlases with relatively large ROIs, which explicitly separate superior and posterior brain areas from frontal and temporoparietal regions, resulted in the most consistent normalization of language areas. The AutoLobe atlas generated for each scan at $t \geq 3.0$ was the most consistent of these, primarily because its automated algorithm avoided splitting clusters across arbitrary anatomical boundaries. Its large ROIs were dominated by the most active brain areas in the frontal and temporoparietal language regions. AutoLobe cluster ROIs created using $t \geq 4.0$ were somewhat less consistent than those created using $t \geq 3.0$ because the higher threshold sometimes resulted in no ROI clusters associated with small or weak activation peaks.

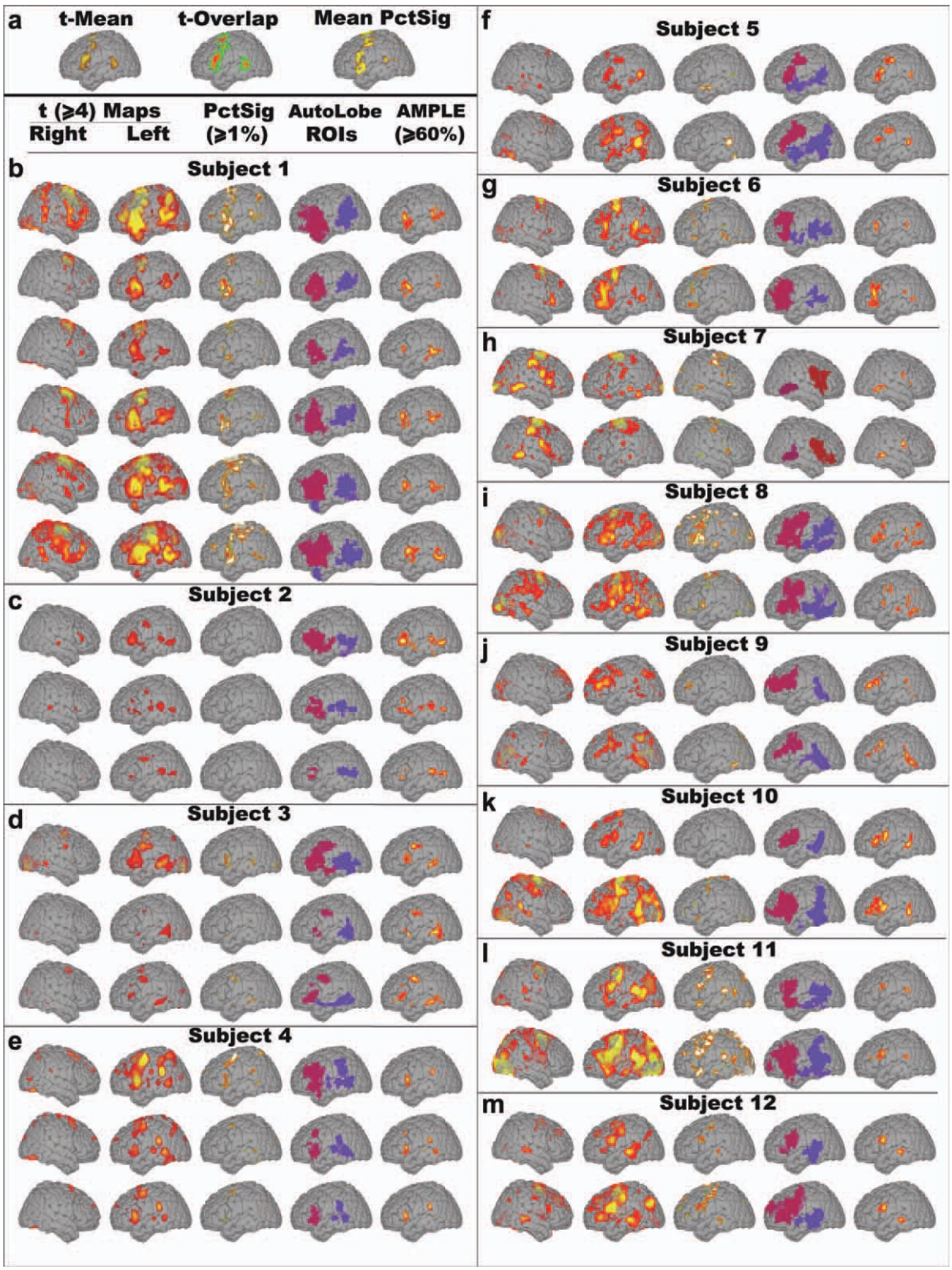

Figure 3. fScan language activation maps. Panel **a** shows the averaged sentence-reading fMRI t-maps for the left hemisphere obtained from all 31 scans of the 12 subjects (t-Mean), as well as a percent-overlap map (t-overlap) of superthreshold ($t \geq 4$) voxels, and an averaged raw BOLD percent signal-change map (Mean PctSig). Panels **b–m** show the individual results for all language scans for subjects 1–12; each row shows the results for a single scan. The first two maps in each row are the language t-maps ($t \geq 4$), the third map is the raw BOLD percent signal-change map (PctSig $\geq 1\%$), the fourth is the AutoLobe mask ($t \geq 3$) generated for that scan, and the last map in the row is the AMPLE-masked map obtained by normalizing the original t-map by the AutoLobe ROI mask and displaying all voxels with AMPLE values $\geq 60\%$. Only the dominant language hemisphere is shown in columns 3–5.

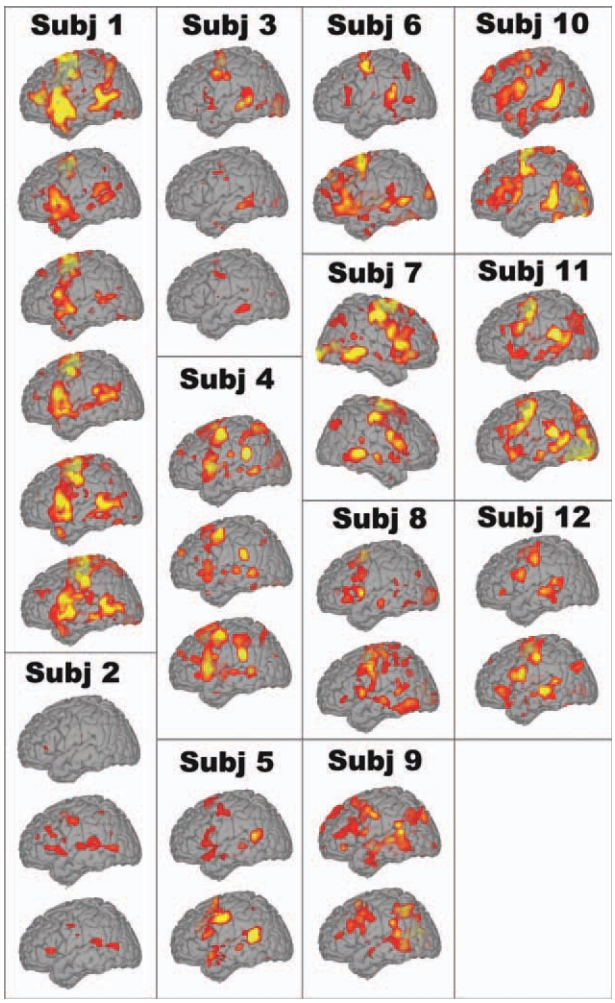

Figure 4. Language activation maps generated using FEAT processing pipeline. Panels show the results in the dominant hemisphere for all language scans for all 12 subjects. FEAT tstat maps are shown thresholded at $t \geq 4$ and overlaid on a standard FreeSurfer reconstructed MNI brain as in Figure 3.

Figure 3 (columns 4 and 5) show the AutoLobe atlas ROI masks and AMPLE normalized activation maps created for every language scan for all 12 subjects. Within each subject, the pattern of activation in the AutoLobe AMPLE maps was highly reproducible across repeated scans. Although there remained some variability across scans in the relative strength of multiple active areas within the same ROI, in general the location and approximate spatial extent of active areas was quite consistent from scan to scan. This consistency was most striking in the six scans performed by subject 1 over a period of almost 6.5 years. Those six scans included both linear EPI and spiral pulse sequence acquisitions performed on four different MRI scanners (1.5T, 4T, and 2 different 3T scanners).

Reproducibility of Quantitative Activation Metrics

Brain activation was quantified automatically for all super-threshold voxels in terms of number of active voxels (Nvoxels), location of weighted activation centers (WCtr), mean and peak activation intensity for all ROIs for every language map. Laterality indices (LI) were computed by comparing activations in homologous ROIs across hemispheres. Activation values were computed separately for each combination of ROI and activation threshold levels. As expected, values varied considerably based on threshold levels and ROI size and location. Both within and across scans, NVoxels, WCtrs, and LIs varied more when calculated based on t-value threshold alone than when calculated based on combining t-value threshold and normalized AMPLE level thresholds. For AMPLE-masked maps the most consistent results were obtained for a t-value threshold of 4.0 to include all statistically significant activated voxels, combined with an AMPLE threshold of 60% to restrict quantitation to only the 40% most active voxels within each region. The auto-segmented ROI clusters again resulted in the most consistent quantitative results across repeated scans (Figs. 5 and 6). Results for other ROI atlas masks showed similar trends but were somewhat less consistent and are not shown to due space limitations. Except for laterality measurements, all results shown are from the dominant brain hemisphere only.

Of the three principle quantitative metrics, laterality (LI) and location (WCtr) were the most reproducible across repeated scans of the same subject, whereas the spatial extent of activation (NVoxels) was less consistent. Laterality was very consistent, with all 12 subjects having one side clearly dominant in both expressive (frontal) and receptive (temporoparietal) language areas. LI values measured based on fixed t-maps varied from scan to scan because overall t-value signals varied across scans (Fig. 5). Calculating activation-weighted laterality indices (WLI) from t-maps, with or without a minimum activation threshold, still varied with overall t-value amplitudes and did not significantly improve reproducibility compared with voxel counting. Calculating laterality after performing AMPLE normalization with bilateral ROIs, however, was much more consistent across scans (Fig. 5A1,A2). Calculating percent agreement for all repeat scan-pair combinations, the average agreement for LangZone ROIs was 72.9 ± 3% (SEM) using standard t-maps compared with 88.6 ± 2% after AMPLE normalization. LI values varied with ROI selection and again the most reproducible results were obtained for auto-clustered ROIs that avoided nonlanguage areas. Across all repeat scan-pairs, LI values based on AutoLobe language ROIs agreed by 92.4 ± 2% using AMPLE level 40% (Fig. 5A3), and increased to 95.2 ± 2% when only voxels above the 60% AMPLE level were considered.

Quantifying and comparing the center of activation locations using ROIs depended on having each ROI contain only one major activation peak. The auto-segmented clusters in the AutoROI atlases were each defined by a single major active peak, and thus resulted in the most consistent WCtr values when homologous ROIs were compared across multiple scans. The distance between such WCtr locations compared in every pairing of scans for each subject (Fig. 5B) was 9.0 ± 0.5 mm for maps masked at AMPLE level 40% and 8.2 ± 0.5 mm when masked at AMPLE level

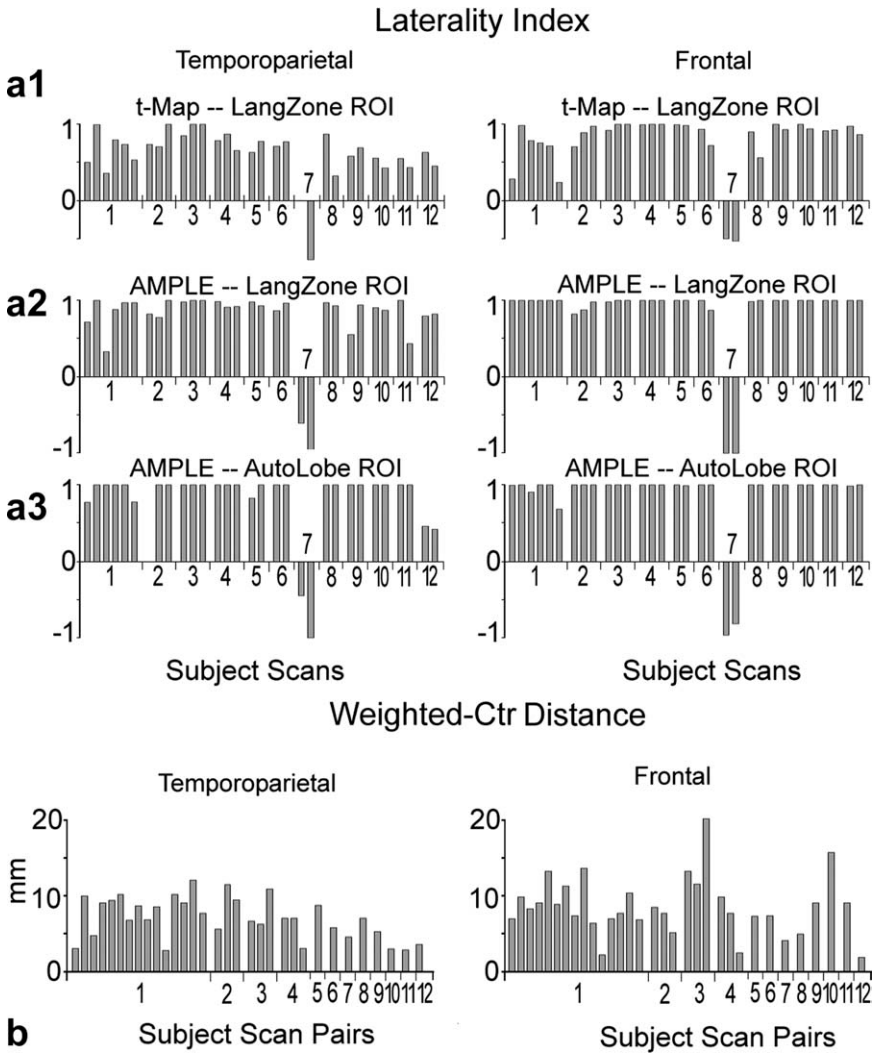

Figure 5. Quantitative metrics of language activation calculated separately for temporoparietal (left graphs) and frontal (right graphs) ROI masks. **a**: Language laterality indices (LI) for the 31 language scans. Each bar represents a separate scan, and scans are grouped by subject. **a1**: LI values based on original t-maps thresholded at $t \geq 4$ and sampled in the anatomical LangZone mask. **a2**: AMPLE-masked t-maps ($t \geq 4$ and AMPLE $\geq 60\%$) for the LangZone ROIs. **a3**: AMPLE-masked t-maps ($t \geq 4$ and AMPLE $\geq 40\%$) for auto-cluster AutoLobe ROIs. **b**: Pair-wise reproducibility metrics for the distance (mm) between the MNI brain locations of the t-value weighted center of activation (dWCtr) obtained by comparing all pairings of 2 language scans for each subject. Each bar represents a different scan pair, grouped by subject. Temporoparietal and frontal AutoROIs are shown for AMPLE-masked ($t \geq 4$ and AMPLE $\geq 60\%$) maps.

60%. Peak center locations were similar in standard t-maps and AMPLE maps using the same AutoROI clusters.

The spatial extent of active regions showed the most variability across repeated scans, quantified either as the percent change in the number of active 1 mm^3 voxels per ROI ($NVoxel_{pct}$) or as a percent-overlap index per ROI for pairs of different scans of the same subject (not shown). Spatial extent was so variable when t-maps were compared directly that it was not possible to make meaningful quantitative comparisons at any fixed t-value threshold level. In AMPLE-masked maps, however, spatial extent showed fairly good reproducibility in language-related ROIs (see Fig. 3). Overall, agreement in the number of active voxels in pair-wise comparisons ($NVoxel_{pct}$) across subjects in anatomical LangZone frontal and temporoparietal ROIs was 53 ± 3% for AMPLE level 40% and 58 ± 3% above AMPLE 60%. In auto-cluster ROIs agreement was similarly 54 ± 3% at AMPLE 40% and 55 ± 3% at level 60%. On average, Nvoxel consistency was approximately 10% higher in temporoparietal ROIs than in frontal cortex ROIs.

ROI reproducibility metrics of AMPLE-masked activation t-maps in auto-cluster ROIs were compared as a function of magnetic field strength, pulse sequence, mean t-value, raw BOLD signal amplitude, head motion, days between scan sessions, sex, and age to see which parameters were correlated with reproducibility. Figure 6 shows those parameters that were clearly correlated in AutoLobe maps (results for frontal and temporoparietal ROIs were similar and were combined in the graphs for simplicity). Both the distance between activation-weighted peak locations (dWCtr) and the agreement in spatial extent ($Nvoxels_{pct}$) across all pairs of scans of the same subject were best (lowest dWCtr and highest $Nvoxels_{pct}$) when the two scans were acquired on the same scanner using the same pulse sequence. Consistency of peak location was better for scans acquired on the same scanner regardless of pulse sequence, compared with the same pulse sequence regardless of scanner. The opposite was seen for spatial extent; it was more reproducible when comparing scans acquired with the same pulse sequence regardless of scanner (or magnetic field), than for the same scanner regardless of pulse sequence. Reproducibility was also strongly correlated with overall BOLD signal amplitude, whether expressed as raw percent signal change or as mean t-value. Reproducibility of peak locations (dWCtr) was

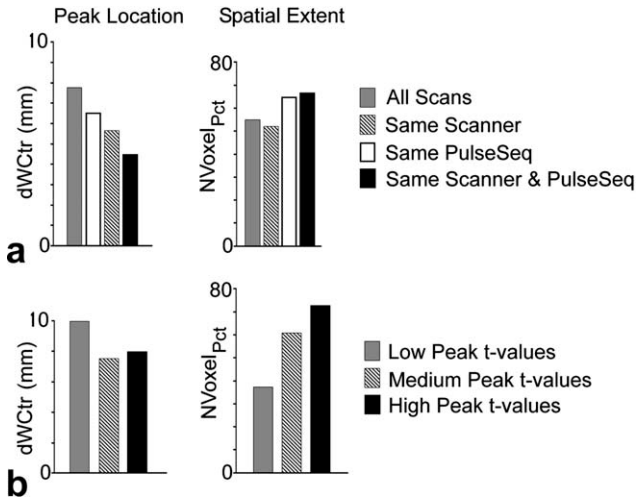

Figure 6. Reproducibility metrics as a function of acquisition method (scanner and pulse-sequence) and mean activation amplitude. Each graph shows the averaged values for reproducibility metrics calculated from all test–retest pairing combinations of AMPLE-masked t-maps ($t \geq 4$ and AMPLE $\geq$ 60%) for all subjects. **a**: The average distance (mm) between activation-weighted peak locations (dWCtr) and spatial extent ($NVoxel_{pct}$) of all test–retest pairs as a function of the similarity in scanner and pulse-sequence used in the 2 scans in the pair; "All Scans" = any combination of scanner (1.5T, $3T_A$, $3T_B$, or 4T) and pulse sequence (linear EPI or spiral), "Same Scanner" = both scans performed on the same scanner using any pulse sequence, "Same PulseSeq" = scans performed on any scanner but both performed using the same pulse sequence, "Same Scanner and PulseSeq" = both scans performed on the same scanner using the same pulse sequence. **b**: Average peak distance and spatial extent shown as in (a) as a function of overall t-value amplitude of the 2 scans in each test–retest pair; t-value amplitude was calculated as the average of the peak t-value for the 2 scans, averaged in 3 groups: "Low" (peak t-values < 9.0), "Medium" ($9.0 <$ t-value < 14.0), and "High" (t-value > 14.0).

poorer when scans had relatively low overall t-values compared with those with either medium or high values. These quantitative relations in peak location and spatial extent were virtually identical for AMPLE-masked maps at both 40% and 60% AMPLE levels.

DISCUSSION

This report demonstrates that BOLD fMRI mapping of language function as performed clinically for neurosurgical planning can produce reproducible brain maps when analyzed using the AMPLE normalization method. A single subject performing the same language paradigm (sentence-completion in this case) in different scan sessions activated very similar frontal and temporal language areas. By normalizing differences in BOLD signal intensity within local brain regions, brain maps from different scans were much more similar than when using standard fixed statistical activation thresholds. Even fMRI scans performed on scanners with different magnetic fields strengths (ranging from 1.5T to 4T) and using different pulse sequences (linear or spiral echoplanar imaging) were highly reproducible in AMPLE maps. Of the ROI sampling methods tested, the most reproducible results were obtained using auto-segmented cluster ROIs and comparing only active voxels ($t \geq 4$) within 40% of the most active values in that ROI (AMPLE $\geq$ 60%). These language mapping results confirm and generalize similar results on fMRI reproducibility previously reported for clinical fMRI motor mapping using the AMPLE approach (6,7).

Obtaining reproducible results is essential for improving fMRI as a clinical tool for single-subject studies. Although reproducibility will not in itself mean that fMRI maps are accurate, the ability to make consistent objective maps is a prerequisite for fully understanding the clinical significance of brain activation maps. In this respect it is noteworthy that the 60% AMPLE level observed to be optimal for reproducibility in the current study for language mapping, agrees well with the results previously reported for clinical fMRI motor mapping, which found that 60% was the optimal AMPLE level that consistently mapped hand movement activity to the central sulcus region of the sensorymotor cortex (7). This quantitative agreement, coupled with appropriate anatomical specificity for the motor maps, suggests that the lower 60% of the statistical fMRI signal in each active area may represent a relatively nonspecific BOLD response. In that case using AMPLE normalization to exclude the lower 60% of the signal provides a mechanism for improving not only the spatial consistency but also the functional specificity consistency of fMRI maps.

An important aspect of the current study is that reproducibility was achieved using an objective and completely automated analysis protocol. Previous reports that fMRI results could be improved by considering the most active voxels used ad-hoc approaches for how to identify which voxels were the most active (7,32). Unlike motor mapping, language mapping involves multiple brain areas that cannot be predicted by anatomical landmarks, especially in the context of disease. This study addressed this issue by comparing multiple different types of ROI masks. The fairly straightforward automated clustering algorithm used here for empirically identifying functional activation peaks appears to be capable of adapting to intersubject variability in the location of language centers, without sacrificing objectivity.

In contrast to the success of AMPLE normalization, other processing approaches tested here did not dramatically improve reproducibility across repeated scans. For example, raw BOLD percent signal level reproducibility was not noticeably better than statistical t-map reproducibility. Standard fMRI preprocessing steps such as motion correction, temporal filtering, and intensity normalization also did not make activation maps reproducible across scans, nor did standardization of the number of time points analyzed for each scan.

Normalization of the BOLD activation signal did succeed in improving and quantifying fMRI reproducibility, but even that did not eliminate variability across repeated language scans of the same subject.

In Figure 3, for example, the AMPLE maps are not identical across all scans, especially for subjects 4, 6, and 9. Reproducibility of the center location and spatial extent of activation was best when comparing repeated scans performed on the same scanner using the same acquisition methods. This is not surprising given the differences in *k*-space trajectories used in linear and spiral echoplanar imaging and in the sensitivity of these ultra-fast sequences to local field nonuniformity. This finding is consistent with published results of the FBIRN multi-center imaging project, which recommends strict standardization across MRI acquisition methods to minimize scanner-dependent variability in multi-site fMRI studies (18–20). In the current study, however, scanner and pulse sequence only accounted for a portion of the residual variability in AMPLE maps, as demonstrated by the results in Figure 3 for subjects 8, 9, 11, and 12 who each had both scan sessions on the same scanner using the same scan parameters (Table 1). The data revealed no significant correlations between any reproducibility metric and scanning parameters such as subject age, days between scans, duration of scans, or task stimulus set.

There was, however, a clear relationship between reproducibility of both the center and extent of active areas and the overall magnitude of the BOLD signal (either percent mean or t-value), as seen in Figure 6b. This effect was not driven simply by increased BOLD signal associated with scanning at higher field strengths. It is likely to reflect inconsistency in the task-dependent BOLD response due to cognitive variables such as changing task strategy, attention, or performance levels, or perhaps physiological confounds such as heart rate, respiration, blood sugar, nicotine, or caffeine levels. Further study focusing on careful quality control assessment of fMRI signal components to optimize task-dependent signal amplitude may further improve reproducibility.

For clinical applications, improved reproducibility of fMRI using objective AMPLE normalization should significantly enhance the consistency of brain function imaging and thus facilitate diagnostic interpretation of brain function maps and assessment of treatment risk. The normalization approach attempts to compensate for variations in the functional sensitivity of BOLD signal imaging, but it does not directly address the issue of functional specificity. Because the relationship between functional sensitivity and specificity is not well understood, a limitation of AMPLE normalization as described in this report, is that it risks filtering out small clusters of relatively weakly activated voxels that may nevertheless be clinically important. Systematic validation of clinical fMRI is still needed therefore to better understand the relationship between BOLD activation maps and surgically mapped eloquent cortex, and to optimize image acquisition and analysis methods such as AMPLE to identify the clinically important areas. Improving the consistency of fMRI brain function maps is an important step toward such systematic clinical validation.

In conclusion, this study in healthy volunteer subjects demonstrates that single-subject fMRI language mapping as performed clinically for neurosurgical treatment planning can be reproducible when appropriately analyzed. The quantitative analysis methods presented here are completely objective and highly automated. They enhance reproducibility of fMRI maps and also provide quantitative comparative metrics that enable a systematic assessment of which factors are most directly related to improving reproducibility. Future studies will be able to build on these findings to optimize acquisition and quality assessment procedures with the goal of establishing imaging standards that will allow fMRI to become a more quantitative biomarker of brain function. For patient fMRI exams, other important factors such as tissue pathology, abnormal cerebro-vascular hemodynamic reactivity, and behavioral deficits will also need to be incorporated into any attempt at a quantitative assessment of fMRI images. Establishing methods and standards for improving and measuring reproducibility, however, are critical steps in advancing fMRI as a quantitative biomarker of brain function.

REFERENCES

1. Tharin S, Golby A. Functional brain mapping and its applications to neurosurgery. Neurosurgery 2007;60:185–201.
2. Stippich C. Clinical functional MRI: presurgical functional neuroimaging. In: Baert AL, Knauth M, Sartor K, editors. Medical radiology: diagnostic imaging. Berlin: Springer-Verlag; 2007. p 1–6.
3. McGonigle DJ, Howseman AM, Athwal BS, Friston KJ, Frakowiak RSJ, Holes AP. Variability in fMRI: an examination of intersession differences. Neuroimage 2000;11:708–734.
4. Liu JZ, Zhang L, Brown RW, Yue GH. Reproducibility of fMRI at 1.5T in a strictly controlled motor task. Magn Reson Med 2004; 52:751–760.
5. Costafreda SG, Brammer MJ, Vencio RZN, et al. Multisite fMRI reproducibility of a motor task using identical MR systems. Magn Reson Imaging 2007;26:1122–1126.
6. Voyvodic JT. Activation mapping as percentage of local excitation (AMPLE): fMRI stability within scans, between scans, and across field strengths. Magn Reson Imaging 2006;24:1249–1261.
7. Voyvodic JT, Petrella JR, Friedman AH. fMRI activation mapping as percentage of local excitation: consistent presurgical motor maps without threshold adjustment. J Magn Reson Imaging 2009;29:751–759.
8. Rutten GJ, Ramsey NF, van Rijen PC, Alpherts WC, van Veelen CW. FMRI-determined language lateralization in patients with unilateral or mixed language dominance according to the Wada test. Neuroimage 2002;17:447–460.
9. Woermann FG, Jokeit H, Luerding R, et al. Language lateralization by Wada test and fMRI in 100 patients with epilepsy. Neurology 2003;61:699–701.
10. Adcock JE, Wise RG, Oxbury JM, Oxbury SM, Matthews PM. Quantitative fMRI assessment of the differences in lateralization of language-related brain activation in patients with temporal lobe epilepsy. Neuroimage 2003;18:423–438.
11. Gaillard WD, Balsamo L, Xu B, et al. fMRI language task panel improves determination of language dominance. Neurology 2004; 63:1403–1408.
12. Fernández G, Specht K, Weis S, et al. Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. Neurology 2003;60:969–975.
13. Mayer AR, Xu J, Pare-Blagoev J, Posse S. Reproducibility of activation in Broca's area during covert generation of single words at high field: a single trial FMRI study at 4 T. Neuroimage 2006;32: 129–137.
14. Ojemann G. Individual variability in cortical localization of language. J Neurosurg 1979;50:164–169.
15. Ojemann G. Brain organization for language from the perspective of electrical stimulation mapping. Behav Brain Sci 1983;6:189–206.
16. Ramsey NF, Sommer IE, Rutten GJ, Kahn RS. Combined analysis of language tasks in fMRI improves assessment of hemispheric

dominance for language functions in individual subjects. Neuroimage 2001;13:719–733.
17. Pillai J, Zaca D. Relative utility for hemispheric lateralization of different clinical fMRI activation tasks within a comprehensive language paradigm battery in brain tumor patients as assessed by both threshold-dependent and threshold-independent analysis methods. Neuroimage 2011;54:S136–S145.
18. Friedman L, Glover G, Krenz D, Magnotta V. Reducing inter-scanner variability of activation in a multicenter fMRI study: role of smoothness equalization. Neuroimage 2006;32:1656–1668.
19. Greve DN, Mueller BA, Liu T, et al. A novel method for quantifying scanner instability in fMRI. Magn Reson Med 2011;65:1053–1061.
20. Brown GG, Mathalon DH, Stern H, et al. Multisite reliability of cognitive BOLD data. Neuroimage 2011;54:2163–2175.
21. Abbott DF, Opdam HI, Briellmann RS, Jackson GD. Brief breath holding may confound functional magnetic resonance imaging studies. Hum. Brain Mapp 2005;24:284–290.
22. Liégeois F, Connelly A, Cross JH, et al. Language reorganization in children with early-onset lesions of the left hemisphere: an fMRI study. Brain 2004;127:1229–1236.
23. Swanson SJ, Sabsevitz DS, Hammeke TA, Binder JR. Functional magnetic resonance imaging of language in epilepsy. Neuropsychology 2007;17:491–504.
24. Wilke M, Lidzba K. LI-tool: a new toolbox to assess lateralization in functional MR-data. J Neurosci Methods 2007;163:128–136.
25. Abbott DF, Waites AB, Lillywhite LM, Jackson GD. fMRI assessment of language lateralization: an objective approach. Neuroimage 2010;50:1446–1455.
26. Jones SE, Mahmoud SY, Phillips MD. A practical clinical method to quantify language lateralization in fMRI using whole-brain analysis. Neuroimage 2010;54:2937–2949.
27. Petrella JP, Shah LM, Harris KM, et al. Preoperative localization of language and motor areas with functional MRI: impact upon therapeutic decision making in patients with potentially resectable brain tumors. Radiology 2006;240:793–802.
28. Voyvodic JT. Real-time fMRI paradigm control, physiology, and behavior combined with near real-time statistical analysis. Neuroimage 1999;10:91–106.
29. Smith SM, Jenkinson M, Woolrich MW, et al. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 2004;23(Suppl 1):S208–S219.
30. Maldjian JA, Laurienti PJ, Burdette JB, Kraft RA. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage 2003;19:1233–1239.
31. Dale AM, Fischl B, Sereno MI. Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 1999;9:179–194.
32. Arthurs OJ, Boniface SJ. What aspect of the fMRI BOLD signal best reflects the underlying electrophysiology in human somatosensory cortex? Clin Neurophysiol 2003;114:1203–1209.