Variability of Clinical Functional MR Imaging Results: A Multicenter Study[1]

Moritz C. Wurnig, MD, MSc
Jakob Rath, MD
Nicolaus Klinger, MSc
Ilse Höllinger, MSc
Alexander Geissler, MSc, PhD
Florian P. Fischmeister, DSc
Markus Aichhorn, MSc
Thomas Foki, MD
Martin Kronbichler, PhD
Janpeter Nickel, MD
Christian Siedentopf, MD
Wolfgang Staffen, MD, PhD
Michael Verius, MD
Stefan Golaszewski, MD, PhD
Florian Koppelstätter, MD
Engelbert Knosp, MD, PhD
Eduard Auff, MD, PhD
Stephan Felber, MD, PhD
Rüdiger J. Seitz, MD
Roland Beisteiner, MD, MA

Purpose: To investigate intersite variability of clinical functional magnetic resonance (MR) imaging, including influence of task standardization on variability and use of various parameters to inform the clinician whether the reliability of a given functional localization is high or low.

Materials and Methods: Local ethics committees approved the study; all participants gave written informed consent. Eight women and seven men (mean age, 40 years) were prospectively investigated at three experienced functional MR sites with 1.5- (two sites) or 3-T (one site) MR. Nonstandardized motor and highly standardized somatosensory versions of a frequently requested clinical task (localization of the primary sensorimotor cortex) were used. Perirolandic functional MR variability was assessed (peak activation variability, center of mass [COM] variability, intraclass correlation values, overlap ratio [OR], activation size ratio). Data quality measures for functional MR images included percentage signal change (PSC), contrast-to-noise ratio (CNR), and head motion parameters. Data were analyzed with analysis of variance and a correlation analysis.

Results: Localization of perirolandic functional MR activity differed by 8 mm (peak activity) and 6 mm (COM activity) among sites. Peak activation varied up to 16.5 mm (COM range, 0.4–16.5 mm) and 45.5 mm (peak activity range, 1.8–45.5 mm). Signal strength (PSC, CNR) was significantly lower for the somatosensory task (mean PSC, 1.0% ± 0.5 [standard deviation]; mean CNR, 1.2 ± 0.4) than for the motor task (mean PSC, 2.4% ± 0.8; mean CNR, 2.9 ± 0.9) ($P < .001$, both). Intersite variability was larger with low signal strength (negative correlations between signal strength and peak activation variability) even if the task was highly standardized (mean OR, 22.0% ± 18.9 [somatosensory task] and 50.1% ± 18.8 [motor task]).

Conclusion: Clinical practice and clinical functional MR biomarker studies should consider that the center of task-specific brain activation may vary up to 16.5 mm, with the investigating site, and should maximize functional MR signal strength and evaluate reliability of local results with PSC and CNR.

[1] From the Department of Neurology (M.C.W., J.R., N.K., I.H., A.G., F.P.F., T.F., E.A., R.B.), MR Center of Excellence (M.C.W., J.R., N.K., I.H., A.G., F.P.F., T.F., R.B.), and Department of Neurosurgery (E.K.), Medical University of Vienna, Währinger Gürtel 18-20, 1090 Vienna, Austria; Department of Psychology and Centre for Neurocognitive Research, University of Salzburg, Salzburg, Austria (M.A., M.K.); Neuroscience Institute and Centre for Neurocognitive Research (M.K.) and Department of Neurology (W.S., S.G.), Christian-Doppler-Clinic, Paracelsus Private Medical University, Salzburg, Austria; Department of Neurology, Heinrich-Heine-University Dusseldorf, Dusseldorf, Germany (J.N., R.J.S.); Department of Radiology, Medical University of Innsbruck, Innsbruck, Austria (C.S., M.V., F.K.); and Institute for Diagnostic Radiology, Stiftungsklinikum Mittelrhein, Koblenz, Germany (S.F.). Received July 6, 2012; revision requested August 22; final revision received September 20; accepted October 3; final version accepted January 17, 2013. Supported by the Austrian Science Fund (grant P18057). Supported in part by a research cluster grant from the Medical University and University of Vienna (no. FA103FC003). **Address correspondence to** R.B. (e-mail: *roland.beisteiner@meduniwien.ac.at*).

Because of the capability of noninvasive monitoring of brain activity with high spatial resolution, clinical applications of blood oxygen level–dependent functional magnetic resonance (MR) imaging are steadily increasing. Since its introduction 20 years ago, functional MR imaging has become a key technology to avoid invasive presurgical procedures, such as the Wada test, to localize functional brain tissue (1). Functional MR imaging also is increasingly being used as a biomarker for intervention effects in a large variety of brain diseases (2–5). Investigators in previous studies have assessed intrasubject, intrasite, and between-site variability with healthy subjects (6–10); however, to our knowledge, researchers in no previous study have compared the same patients at different functional MR imaging sites.

Owing to the pathophysiologic changes in brains with diseases, as well as patient-specific methodological limitations, functional MR imaging results from healthy subjects are not representative for patient populations. Moreover, the main goal of investigators in previous studies on between-site variability was to prove the feasibility of multicentric functional MR imaging group studies with pooled data (6–15). Therefore, the researchers in none of these previous studies investigated measures of variability relevant for the individual patient, such as the spatial variability of functional MR imaging activations.

We investigated intersite variability of clinical functional MR imaging, including the influence of task standardization on variability and the use of various parameters to inform the clinician whether the reliability of a given functional localization is high or low.

Advances in Knowledge

- Clinical functional MR imaging variability among patients and centers can be substantial, even when the task is highly standardized; in this study, peak activation varied up to 45.5 mm and center of mass (COM) varied up to 16.5 mm between sites.
- Clinical tasks with high activation levels are more important than extensive task standardization.
- High percentage signal change (PSC) and contrast-to-noise ratio (CNR) indicate better reliability of functional MR imaging results compared with the reliability achieved from low PSC and CNR data sets.

Materials and Methods

Patients and Tasks

This prospective study was approved by the local ethics committees of each participating institution. Within a study period of 26 months, 15 patients (for the group, mean age and range were 40 years and 18–55 years; for women, mean age and range were 42 years and 31–55 years; for men, mean age and range were 37 years and 18–52 years) from a series of consecutively referred patients for presurgical evaluation of the primary sensorimotor cortex were prospectively included, with written informed consent (Tables 1, 2). Inclusion criterion was the request for a clinical functional MR imaging report, allowing varying pathologic findings and clinical symptoms. Exclusion criteria were general MR imaging contraindications (n = 0), unjustifiable risks associated with the study efforts (n = 5), or missing agreement (n = 11). All measurements were acquired before planned surgery, but seven patients had already undergone surgery before study inclusion and were scheduled to undergo brain surgery for a second time. Absence of relevant morphologic changes between visits was evaluated through analysis of the anatomic images of all sites (Fig 1; Tables 1, 2) and could be confirmed for every patient by two of us (R.B., a senior neurologist with 20 years of experience; M.C.W., a junior radiologist with 3 years of experience). Absence of relevant clinical changes was also confirmed at the investigational sites by three of us (R.B., a senior neurologist with approximately 20 years of clinical functional MR imaging experience; S.G., a senior neurologist with approximately 15 years of clinical functional MR imaging experience; and F.K., a senior radiologist with approximately 10 years of clinical functional MR imaging experience).

During each visit, every patient performed two separate tasks with the contralesional hand. These were a highly standardized somatosensory

Implications for Patient Care

- For presurgical planning (including estimation of resection margins), clinicians should consider functional MR imaging variability (eg, the COM of task-specific brain activation may vary up to 16.5 mm, with the investigating site).
- Clinical functional MR imaging setups should maximize PSC and CNR; such parameters can be used as indicators for the reliability of the local clinical report.

Published online before print
10.1148/radiol.13121357 **Content codes:** MR NR

Radiology 2013; 268:521–531

Abbreviations:
ANOVA = analysis of variance
CNR = contrast-to-noise ratio
COM = center of mass
ED = euclidian distance
FWE = family-wise error
ICC = intraclass correlation coefficient
IQR = interquartile range
OR = overlap ratio
PSC = percentage signal change
PV = peak activation voxel
ROI = region of interest
SPM = statistical parametric mapping
SR = size ratio

Author contributions:
Guarantors of integrity of entire study, M.C.W., R.B.; study concepts/study design or data 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, M.C.W., J.R., N.K., A.G., F.P.F., S.G., E.A., S.F., R.J.S., R.B.; clinical studies, J.R., J.N., M.V., S.G., F.K., E.A., S.F., R.J.S., R.B.; experimental studies, N.K., I.H., A.G., M.A., T.F., M.K., C.S., W.S., M.V., S.G., F.K., S.F., R.J.S., R.B.; statistical analysis, M.C.W., F.P.F., J.N.; and manuscript editing, M.C.W., F.P.F., S.G., E.A., R.J.S., R.B.

Conflicts of interest are listed at the end of this article.

Table 1

Patient Characteristics

Patient No./Sex/Age (y)	Admission Diagnosis*
1/M/18	Infarction of the right middle cerebral artery; intractable epilepsy, surgery was planned
2/F/55	Glioblastoma mutliforme, right insula
3/F/42	Gliosis of the right hemisphere, surgery was planned
4/F/40	Astrocytoma WHO grade II, left operculum
5/M/27	Astrocytoma WHO grade II, left central region
6/M/53	Astrocytoma WHO grade II, right frontal lobe
7/F/43	Oligodendroglioma WHO grade II, left insula
8/F/46	Oligodendroglioma WHO grade II, right temporal lobe
9/M/36	Pleomorphic xanthoastrocytoma, left temporal lobe
10/F/45	Mixed glioma WHO grade II, left frontotemporal lobe
11/M/43	Astrocytoma WHO grade II, left operculum; after intracerebral hemorrhage
12/F/31	Astrocytoma WHO grade II, left frontotemporal lobes
13/F/36	Oligoastrocytoma WHO grade II, left frontal lobe
14/M/31	Oligodendroglioma WHO grade II, left insula
15/M/52	Glioblastoma multiforme, left perirolandic area

* WHO = World Health Organization.

Figure 1

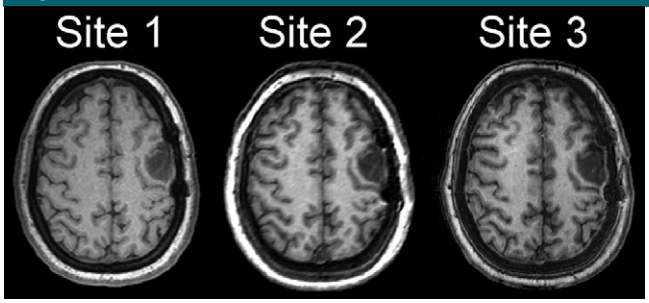

Figure 1: Representative anatomic images from all three sites for patient 13, who had the longest interval between visits. No relevant morphologic changes could be detected.

task and a nonstandardized motor task, and these tasks differed in various aspects of standardization. With the somatosensory task, identical vibrotactile stimulators (16) and stimulation protocols were used at each functional MR imaging site. The following instruction was given: "Provide a localization of the primary sensory hand area according to the predefined procedures." The "predefined procedures" consisted of finger cuffs placed on the second and third finger of the contralesional hand with identical cuff sizes and positions; identical pseudorandom stimulation (16), and 10 runs per session with a 150-second duration for each (blocked design, 10-second dummy MR images, three activation periods, and four rest periods of 20 seconds each). With the motor task, the following instruction was given: "Provide a localization of the primary motor hand area according to your best local procedures." At each site, fist clenching was used (site-specific blocked design, 1–10 runs per session, three to eight activation periods, four to nine rest periods).

Data Acquisition

The following MR units were used at the three sites: site 1, 1.5-T Magnetom Vision (Siemens, Erlangen, Germany); site 2, 1.5-T Gyroscan ACS-NT Powertrak 6000 (Philips Healthcare, Best, the Netherlands), with 3-T Achieva

Table 2

Chronology of Visits

Sequence	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7	Patient 8	Patient 9	Patient 10	Patient 11	Patient 12	Patient 13	Patient 14	Patient 15
1st Visit	S2/0	S2/0	S3/0	S3/0	S1/0	S3/0	S3/0	S1/0	S3/0	S3/0	S3/0	S2/0	S2/0	S2/0	S3/0
2nd Visit	S1/1	S1/2	S2/21	S1/7	S2/21	S1/5	S1/13	S2/8	S1/6	S1/4	S1/13	S3/21	S1/1	S1/1	S2/4
3rd Visit	S3/13	S3/32	S1/22	S2/8	S3/22	S2/6	S2/14	S3/9	S2/8	S2/5	S2/14	S1/110	S3/131	S3/107	S1/5

Note.—Data are sites visited/number of days of interval to first functional MR imaging investigation. S1 = site 1, S2 = site 2, and S3 = site 3.

(Philips Healthcare) used in patients 11, 14, and 15; site 3, 3-T Medspec (Bruker Medical, Ettlingen, Germany), with 3-T Magnetom Trio (Siemens) used in patients 14 and 15.

The spatial distances between sites were 186 km (between sites 1 and 2), 476 km (between sites 1 and 3), and 296 km (between sites 2 and 3). The sites interact regularly with scientific information exchange; however, each site uses different hardware and functional MR imaging methods. Functional images were acquired parallel to the anterior commissure–posterior commissure plane with identical parameters, as follows: an echo-planar imaging sequence, repetition time of 4000 msec, a flip angle of 90°, a field of view of 230 × 230 mm, 30 sections, 3-mm section thickness, and no gap. Echo time (30–55 msec) and voxel size (1.6 × 1.6 × 3.0 mm^3 to 3.6 × 3.6 × 3.0 mm^3) varied for locally optimized blood oxygen level–dependent signal. No local smoothing during reconstruction was performed.

Data Analysis

Data analysis was performed by two of us (R.B., a senior neurologist with 20 years of experience, and M.C.W., a junior radiologist with 3 years of experience) at the principal investigation center (Department of Neurology, Medical UIniversity of Vienna, Vienna, Austria) by using statistical parametric mapping (SPM) (SPM8; Wellcome Department of Cognitive Neurology, London, England) and local scripts to compute peak activation, center of mass (COM) of activation, activation size ratio (SR), overlap ratio (OR), percentage signal change (PSC), and contrast to noise ratio (CNR). Statistical support was provided by a statistician who was not an author. Because the primary study focus was on providing an estimate of functional MR imaging variability for clinical practice (which typically includes different hardware and different procedures at each site), all data were equally weighted and handled, and site-specific factors such as local MR unit hardware, field strengths, and local patient handling were not considered.

On the basis of our previous methodological investigations about registration issues (17,18) and after evaluating various alternative image registration procedures, we minimized postprocessing steps and performed no normalization to the Montreal Neurological Institute space to keep the voxel signals of the functional images true to the original. For image coregistration, the following procedure was applied: *(a)* For every patient, all functional volumes were first realigned to the first volume of the individual patient at the respective site (head motion control). *(b)* Then, all site-specific data sets were coregistered to the first volume of the first experimental run of this patient recorded at site 2. This step resulted in identical but patient-specific brain positions over all three sites. Successful registration was checked with every patient.

Each coregistered functional data set was smoothed with a Gaussian kernel (full width at half maximum doubled echo-planar imaging voxel size). Individual SPM *t* maps were then calculated by using a general linear model.

To consider only clinically relevant activity, perirolandic regions of interest (ROIs) containing the primary sensorimotor cortex were drawn by a neurologist (R.B., with about 20 years of functional MR imaging experience) for each patient (19) Fig 2). Because the criterion for ROI borders was based on neuroanatomy, the validity of our neuroanatomical analysis was double-checked by evaluating whether the primary sensorimotor activation really comprised the neuroanatomical ROIs. This was conducted at a threshold widely used in functional MR imaging literature ($P <$ .001, uncorrected). The resulting ROIs comprised between 488 and 2977 voxels (mean, 2010 voxels). For all subsequent analyses, only voxels within these ROIs were considered.

Quantification of Intersite Variability of Functional MR Imaging Localizations

For each task, the intersite variability of functional MR imaging localizations was quantified through commonly used variability metrics (8–15,20–22), including the following: *(a)* peak activation variability, *(b)* COM variability, *(c)* intraclass correlation coefficients (ICCs), *(d)* OR, and *(e)* activation SR.

For calculation of peak activation variability, peak activation was defined as the ROI voxel with the highest *t* value per patient, task, and site. Peak activation should represent the most likely position where task-specific functional tissue can be found. Therefore, this metric is commonly used in functional MR imaging somatotopic studies to quantify localization differences (20,21). Intersite variability was determined by obtaining the euclidian distance (ED) between the peak activation voxels (PVs), comparing all pairs of sites per patient and task. From these 45 EDs (15 patients multiplied by three site pairs), the median and the interquartile range were calculated per task. In addition, results of a morphologic analysis were used to determine the distance between the border of tissue that appeared to be pathologically affected on echo-planar images (the types of neurologic diseases are detailed in Tables 1 and 2) and the PV.

As peak activation variability may be influenced by noise, we also computed the COM variability, which is a more stable feature (20,21). The COM of active voxels was calculated for each experiment. First, a threshold at a family-wise error (FWE)-corrected *P* value of less than .05 was set to separate active voxels from nonactive voxels on SPM *t* maps. Only ROI voxels above this threshold then entered the COM calculation. Then, intersite variability of the COM was determined, corresponding to the peak activation analysis (ie, COM analysis was performed in the same way as the peak activation analysis), except that measurements with no suprathreshold voxels were not considered for this evaluation.

We also calculated the ICC (23). The ICC can be used to measure variability between functional MR imaging sites (10,12–15). The between-site ICC measures the fraction of the total variance of a dependent variable that is due to the between-subject variance. A higher between-site ICC indicates a lower contribution of the sites to the

Figure 2

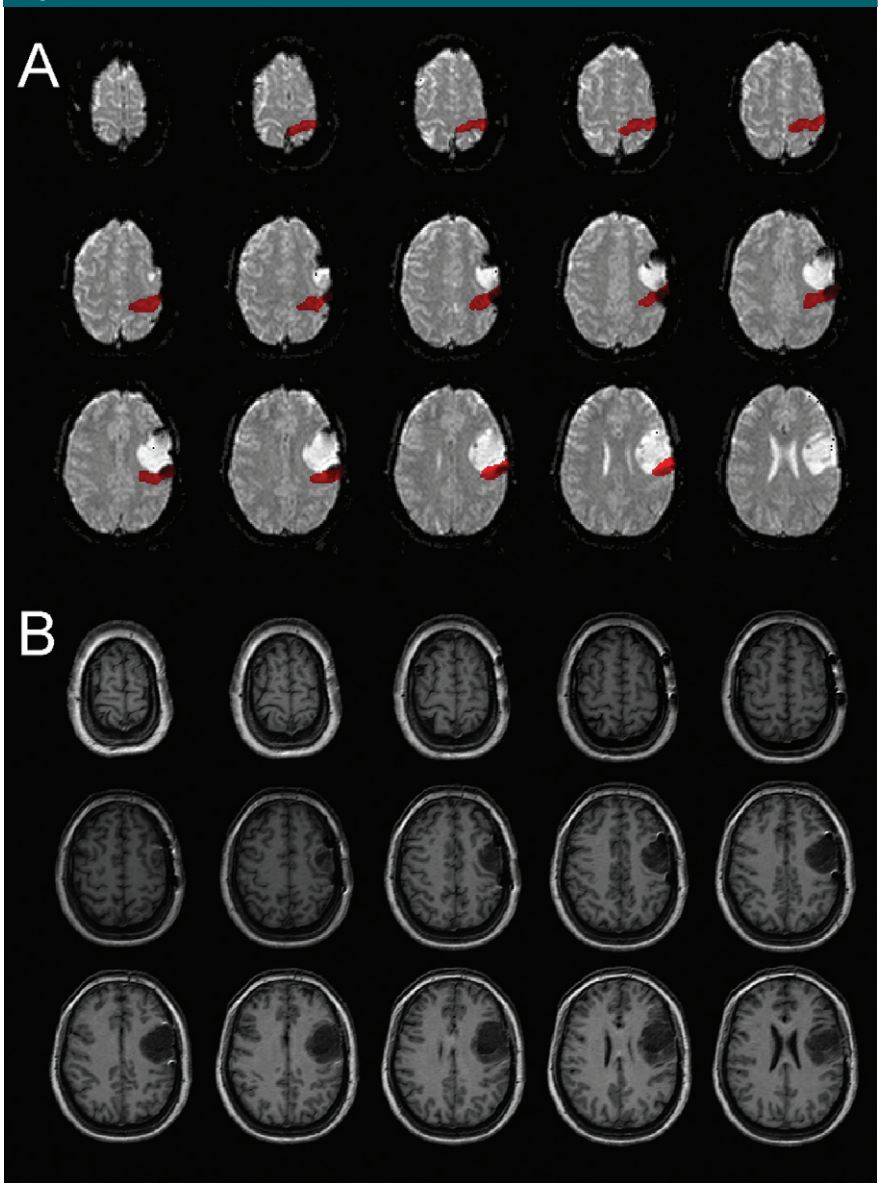

Figure 2: Example of neuroanatomic ROI definitions on axial echo-planar images, which show central distortions, in patient 13. The patient previously underwent brain surgery. *A,* Functional echo-planar images with ROIs in red. *B,* Corresponding anatomic images.

overall variability and therefore a higher between-site reliability. The extreme case (ICC of one) would mean that all sites produced identical results. As dependent variable for an analysis of variance (ANOVA), the mean value of beta regressors of the SPM general linear model (obtained during model estimation) was used. However, only highly reliable ROI voxels (with $P < .05$, FWE corrected) entered analysis for each patient, site, and task.

The following ANOVA model was chosen for each task: $Y_{ij} = M + \text{Sub}_i + \text{Site}_j + U_{ij}$, with Y_{ij} denoting the dependent variable for subject i and site j, M signifying the mean, Sub_i signifying subject i, Site_j signifying site j, and U_{ij} signifying unexplained for subject i and site j. Subject and site were treated as random effects. The ICC was then computed as ICC(2,1) in the notion of Shrout and Fleiss (23).

A further analysis was used to check variability of the activation volume with calculation of OR and SR, derived according to Rombouts et al (8). The OR quantifies reproducibility of localization of activation between different measurements (in terms of cluster overlapping), whereas the SR quantifies the reproducibility of the size of activation (in terms of the number of active voxels) (8). Both ratios range between zero (no reproducibility) and one (complete reproducibility).

For ROIs with suprathreshold voxels ($P < .001$, uncorrected) (8,14,15), SR and OR were computed for all pairs of functional MR imaging sites, patients, and tasks separately. Correspondingly, SRs and ORs were computed over all three sites per patient and task, and the median of these values was determined.

Quantification of Data Quality Effects

We also looked for data quality parameters that might inform the clinician whether the functional MR imaging result is reliable. PSC, CNR, and motion parameters (translation, rotation) were computed as data quality measures. PSC and CNR calculation was based on ROI voxels with $P < .05$ FWE corrected (19).

Statistical Evaluation of Calculated Parameters

To evaluate the influence of tasks, sites (or site comparisons), and subjects on the parameters of intersite variability, as well as the parameters of functional MR imaging data quality, ANOVAs were used. The following ANOVA model was chosen for each parameter: $Y_{ijk} = M + \text{Sub}_i + \text{Site}_j + T_k + (\text{Sub}_i \cdot \text{Site}_j) + (\text{Sub}_i \cdot T_k) + (\text{Site}_j \cdot T_k) + U_{ijk}$, where Y_{ijk} is the dependent variable for subject i, site j, and task k; T_k is task k; and U_{ijk} is unexplained for subject i, site j, and task k.

For parameters that were based on two-site comparisons (eg, peak activation variability, COM variability, two-site OR, two-site SR), Site_j was replaced by Site-comparison j. For parameters that were based on three-site comparisons (eg, three-site OR, three-site SR), the

Table 3

Activity within ROI and Correlation Analysis Results

A: Activity in ROI at Various Thresholds			
Task	Activity at $P < .01$, Uncorrected (%)	Activity at $P < .001$, Uncorrected (%)	Activity at $P < .05$, with FWE Correction (%)
Somatosensory	88.9	84.4	64.4
Motor	97.8	95.6	88.9

B: Correlation Coefficients for Tasks				
Parameter	Mean CNR	Mean PSC	Mean Translation*	Mean Rotation*
Sensorimotor task				
Mean ED PV	−0.16	−0.20	0.39	0.35
Mean ED COM	−0.40	−0.02	0.28	−0.79
Three-site OR	0.54†	0.48	−0.21	−0.15
Three-site SR	0.33	0.27	−0.02	−0.23
Motor task				
Mean ED PV	−0.52†	−0.56†	−0.23	−0.05
Mean ED COM	−0.29	0.49	0.02	−0.21
Three-site OR	0.71‡	0.62†	−0.19	−0.53†
Three-site SR	0.75‡	0.77‡	−0.01	−0.54†

Note.— ED COM = ED between COM values, ED PV = ED between PVs.

* Head motion during functional MR imaging acquisition.

† Correlation was significant at $P = .05$ (two-tailed test).

‡ Correlation was significant at $P = .01$ (two-tailed test).

model was adjusted in the following way: $Y_{ij} = M + \text{Sub}_i + T_j + U_{ij}$, where T_j is task j.

As the significance level, $P < .05$ (two-tailed test) was chosen. For parameters showing significant influence of the factors task and site additional paired t tests were performed to check whether task differences were present in all sites (P value Bonferroni corrected to .016 [two-tailed test]).

To check for correlations between data quality and functional MR imaging variability, Pearson correlation coefficients between the measures of functional MR imaging variability (peak activation variability, COM variability, three-site OR, three-site SR; mean values per patient and task) and functional MR imaging data quality (PSC, CNR, motion parameters; mean values per patient and task) were also obtained.

Results

Individual Brain Activation Maps

We checked all data sets for above-threshold activity within the respective ROIs (Table 3). After detailed instructions and training, all patients fully collaborated (as monitored throughout the measurements), but not all measurements showed active voxels at all tested thresholds. The nonstandardized motor task generated more above-threshold activity: At least 88.9% of the measurements were successful, as opposed to 64.4% with the somatosensory task (Table 3).

Quantification of Intersite Variability of Functional MR Imaging Localizations

For the highly standardized task, the median localization error in terms of peak activation variability (median ED between PVs across all patients; comparing sites 1 vs 2, 1 vs 3, and 2 vs 3 per patient and task) was 7.9 mm (interquartile range [IQR], 5.4–16.5 mm). The single patient with the smallest variation demonstrated the following differences: 1.8 mm (site 1 vs 2), 2.5 mm (site 1 vs 3), and 4.0 mm (site 2 vs 3). The single patient with the greatest variation demonstrated the following differences: 27.6 mm, 45.5 mm, and 20.1 mm. For the motor task, the median peak localization error was 8.3 mm (IQR, 5.8–14.0 mm). The single patient with the smallest variation had differences of 3.6 mm, 4.0 mm, and 6.5 mm. The single patient with the greatest variation had differences of 1.8 mm, 25.4 mm, and 26.4 mm. Distances between PVs and borders with pathologically affected tissue ranged between zero and 53.5 mm. Median COM variability for the highly standardized task (comparison performed as with peak voxels; however, ROIs with absent activity were not considered) was 5.8 mm (IQR, 3.4–8.0 mm), with a maximum of 12.0 mm. Median COM variability for the motor task was 5.7 mm (IQR, 3.2–7.9 mm), with a maximum of 16.5 mm (Fig 3).

Neither PV variability nor COM variability, respectively, differed significantly between tasks ($P = .669$, $P = .501$), site comparisons ($P = .514$, $P = .403$), or subjects ($P = .641$, $P = .574$). Quantitative values are delineated in Table 4. However, a significant interaction between task and site comparison ($P = .032$), as well as between task and subject comparison ($P = .001$), was observed for peak activation variability. A follow-up of the significant interaction

Figure 3

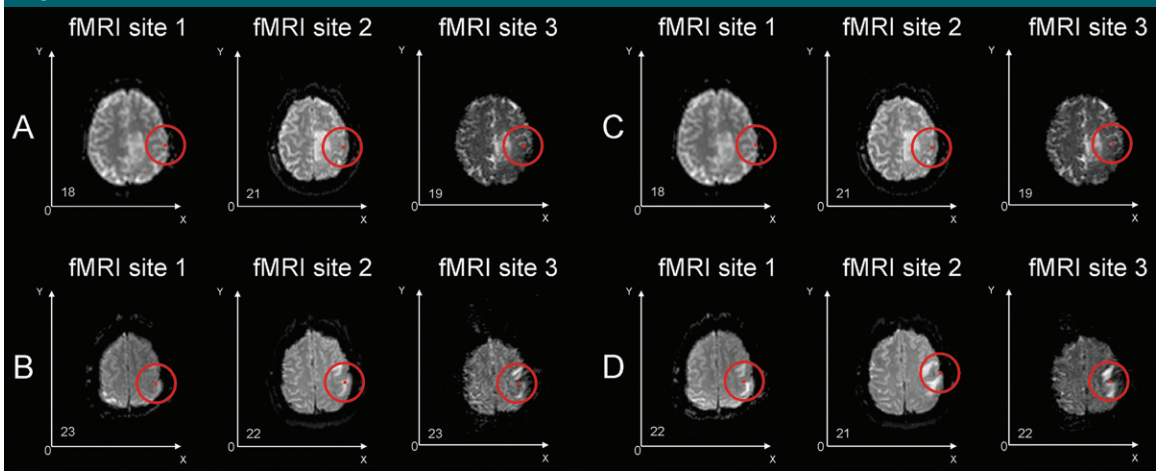

Figure 3: Intersite variability of functional localizations shown for the patient with maximum COM differences. COM voxels (red) are projected on the respective axial echo-planar image. *A,* COM variability for the motor task, with maximum ED between two sites of 16.5 mm. *B,* COM variability for the somatosensory task, with maximum ED between two COM voxels of 12.0 mm. *C,* Peak voxel variability for the motor task, with maximum ED between two peak voxels of 17.4 mm. *D,* Peak voxel variability for the somatosensory task, with maximum ED between two peak voxels of 23.3 mm. *fMRI* = Functional MR imaging.

between task and site comparison did not show significant task differences for any site comparison.

The between-site ICCs were low, with 0.20 (highly standardized task) and 0.23 (motor task).

The median OR between two sites across all patients varied among values of 9.2% (site 2 vs 3), 20.6% (site 1 vs 2), and 29.4% (site 1 vs 3) for the somatosensory task and among values of 45.0% (site 1 vs 3), 52.9% (site 1 vs 2), and 64.3% (site 2 vs 3) for the motor task (Fig 4). The median three-site OR was 4.0% (IQR, 1.1%–10.2%) for the somatosensory task and 33.0% (IQR, 26.6%–43.2%) for the motor task. For the SR, the corresponding values were 34.1% (site 2 vs 3), 35.3% (site 1 vs 2), and 75.3% (site 1 vs 3) for the somatosensory task and 52.4% (site 1 vs 3), 73.4% (site 1 vs 2), and 83.3% (site 2 vs 3) for the motor task. The median three-site SR was 30.8% (IQR, 14.0%–39.3%) for the somatosensory task and 48.7% (IQR, 40.8%–66.1%) for the motor task.

The following significant results were found through ANOVA calculations: The two-site OR, as well as the three-site OR, differed significantly between tasks, indicating a significantly smaller overlap for the highly standardized somatosensory task ($P < .001$ for both) (Table 5). For the two-site OR, a significant interaction between task and site comparison ($P < .001$), as well as between task and subject comparison ($P = .020$), was found. For the two-site SR, a significant interaction between task and site comparison could be found ($P = .003$). Quantitative values are delineated in Table 4.

Quantification of Data Quality Effects

ANOVA calculations showed that PSC, CNR, head translation, and head rotation differed significantly between tasks ($P < .001$, $P < .001$, $P = .002$, and $P = .004$, respectively), indicating a lower PSC and CNR, as well as less motion, for the highly standardized task (Table 5). PSC, CNR, and head rotation also differed among sites ($P = .001$, $P < .001$, $P = .019$, respectively), and head translation differed among subjects ($P = .033$). For PSC and CNR, a significant interaction between task and site ($P = .002$, $P = .005$, respectively) and, for head translation, a significant interaction between subject and site ($P < .001$) could be observed (Table 5). The results of additional t tests showed significant task differences at all sites for PSC and CNR ($P < .005$, for all comparisons) (Table 6).

For both tasks negative correlations of peak voxel variability and COM variability with PSC and CNR were observed. A positive correlation of OR and SR with PSC and CNR for both tasks was observed, as well. In regard to the influence of motion during measurement, a negative correlation of translation, as well as of rotation, with OR and SR was observed. Although not all correlations were significant ($P < .05$, two-tailed test) (Table 3), results indicate that measurements with higher PSC, higher CNR, and less motion are more reliable.

Discussion

The major finding of our study was that the diagnostic uncertainty clinicians face with functional MR imaging localizations is in an average range of 6–8 mm for the sensorimotor cortex. However, in individual cases, clinical functional MR imaging variability can be substantial and vary up to 20–40 mm. Because we investigated a rather stable sensorimotor paradigm, it seems likely that our figures will even increase with less

stable cognitive paradigms (eg, language tasks). The largest intersite variability was found with the peak voxel analysis. Because the peak voxel might be affected by random processes, especially when there is a large activation area with similar *t* values, we added a measure representing the complete cluster of overall suprathreshold activity (COM). Here, intersite variability was reduced to a maximum cluster shift of 16.5 mm. This factor is probably because of a reduced influence of random signal fluctuations or draining-vein effects (24). Knowledge of intersite variability affects two important clinical aspects: *(a)* For clinical localization of essential cortex (eg, primary motor, language, memory areas), it allows the clinician to consider the uncertainty of functional localizations, given the fact that shifting resection margins in the range of millimeters may already have drastic consequences resulting in either a reversible or permanent clinical deficit (25). *(b)* For a valid setup of multicentric functional MR imaging biomarker studies, the researchers need to know and consider possible intersite variability, because this factor has an effect on statistical power and sample size (11). The source of our functional MR imaging variability probably was multifactorial, including technical signal fluctuations, local artifact situation, variable patient performance, physiologic fluctuations, and experimental handling. Clinically, it is promising that we found markers that allow inferences about expectable functional MR imaging variability in individual patients. Functional MR imaging variability was reduced with larger PSCs and larger CNRs. These values may serve as indicators that tell the clinician whether the reliability of a given functional localization is high or low.

Previously, functional MR imaging intersite variability has been investigated with healthy subjects; however, in none of these previous studies was the focus on localization variability, and, therefore, no data exist for direct comparison with data in our own work. A comparison is possible concerning the parameters of ICC, OR, and SR. Here, our patient data are at the lower end of

Table 4

Quantitative Values Entered to ANOVA Testing

Task, Comparison, and Location	ED PV (mm)*	ED COM (mm)†	Two-Site OR (%)‡	Two-Site SR (%)§	Three-Site OR (%)‖	Three-Site SR (%)#	PSC	CNR	Translation**	Rotation**
Task										
Somatosensory	11.2 (8.6)	6.0 (3.4)	22.0 (18.9)	50.3 (29.8)	7.2 (7.9)	32.7 (28.1)	1.0 (0.5)	1.2 (0.4)	0.05 (0.03)	0.0005 (0.0001)
Motor	10.2 (6.2)	6.2 (3.6)	50.1 (18.8)	64.7 (25.0)	31.9 (15.4)	49.2 (20.9)	2.4 (0.8)	2.9 (0.9)	0.06 (0.04)	0.0006 (0.0002)
Site comparison										
1 vs 2	10.6 (7.7)	7.4 (3.4)	33.9 (22.2)	54.7 (30.1)	...	...	...	...	...	...
1 vs 3	10.0 (5.9)	6.5 (3.4)	39.7 (28.4)	60.7 (31.7)	...	...	...	...	...	...
2 vs 3	11.6 (8.6)	4.7 (3.3)	38.4 (19.2)	59.0 (22.4)	...	...	...	...	...	...
Location										
Site 1	...	...	...	...	...	...	2.0 (1.3)	2.6 (1.5)	0.06 (0.03)	0.0005 (0.0002)
Site 2	...	...	...	...	...	...	1.4 (0.7)	2.1 (0.9)	0.07 (0.05)	0.0006 (0.0002)
Site 3	...	...	...	...	...	...	2.0 (0.8)	1.9 (0.7)	0.04 (0.02)	0.0004 (0.0001)

Note.—Data are the means, and numbers in parentheses are standard deviations. ED COM = ED between COM values, ED PV = ED between PVs, Three-Site OR = OR among all three sites, Three-Site SR = activation SR among all three sites, Two-Site OR = OR between two sites, Two-Site SR = activation SR between two sites.

* For every task, the mean ED was calculated over 45 EDs corresponding to 15 patients and three site pairs. For every site comparison, EDs were calculated comparing the PV locations at the two given sites. The mean was then calculated over 30 EDs corresponding to 15 patients and two tasks.

† Means were calculated in the same way as for EDs between PVs (using COM instead of PV).

‡ Calculated over all pairs of sites per patient and task. Means were calculated in the same way as for ED between PVs (using two-site OR instead of ED between PVs).

§ Calculated over all pairs of sites per patient and task. Means were calculated in the same way as for ED between PVs (using two-site SR instead of ED between PVs).

‖ Calculated over all three sites per patient and task. Means were calculated in the same way as for ED between PVs (using three-site OR instead of ED between PVs).

Calculated over all three sites per patient and task. Means were calculated in the same way as for ED between PVs (using three-site SR instead of ED between PVs).

** Translation is translational and rotation is rotational head motion during MR acquisition. Means were calculated in the same way as for ED between PVs (using translation or rotation instead of ED between PVs).

Table 5

***P* Values for ANOVA Results**

Factor or Interaction	ED PV	ED COM	Two-Site OR	Two-Site SR	Three-Site OR	Three-Site SR	PSC	CNR	Translation	Rotation
Task*	.669	.501	<.001	.08	.001	.213	<.001	<.001	.002	.004
Site pair†	.514	.403	.379	.702	. . .	. . .	. . .	. . .	. . .	. . .
Site‡	. . .	. . .	. . .	. . .	. . .	. . .	.001	<.001	.101	.019
Subject§	.641	.574	.333	.884	.597	.986	.769	.988	.033	.081
Interaction between										
Site and subject‖	.66	.596	.527	.587	. . .	. . .	.295	.133	<.001	.068
Task and site#	.032	.189	<.001	.003	. . .	. . .	.002	.005	.888	.67
Task and subject**	.001	.454	.02	.068	. . .	. . .	.13	.318	.465	.321

Note.—ED COM = ED between COM voxels, ED PV = ED between PVs, rotation = rotational head motion during MR, 3-site OR = OR among all three sites, Three-Site SR = activation SR among all three sites, translation = translational head motion during MR, Two-Site OR = OR between two sites, Two-Site SR = activation SR between two sites.

* Task refers to the comparison of the motor and somatosensory tasks.

† Site pair refers to the analysis of the three site-comparison results (results of comparison of site 1 vs 2, site 1 vs 3, and site 2 vs 3).

‡ Site refers to the comparison of the results at the three sites.

§ Subject refers to the evaluation of between-subject differences.

‖ Dependence of between-subject differences on the investigating site.

Dependence of between-site differences on the task.

** Dependence of between-subject differences on the task.

Figure 4

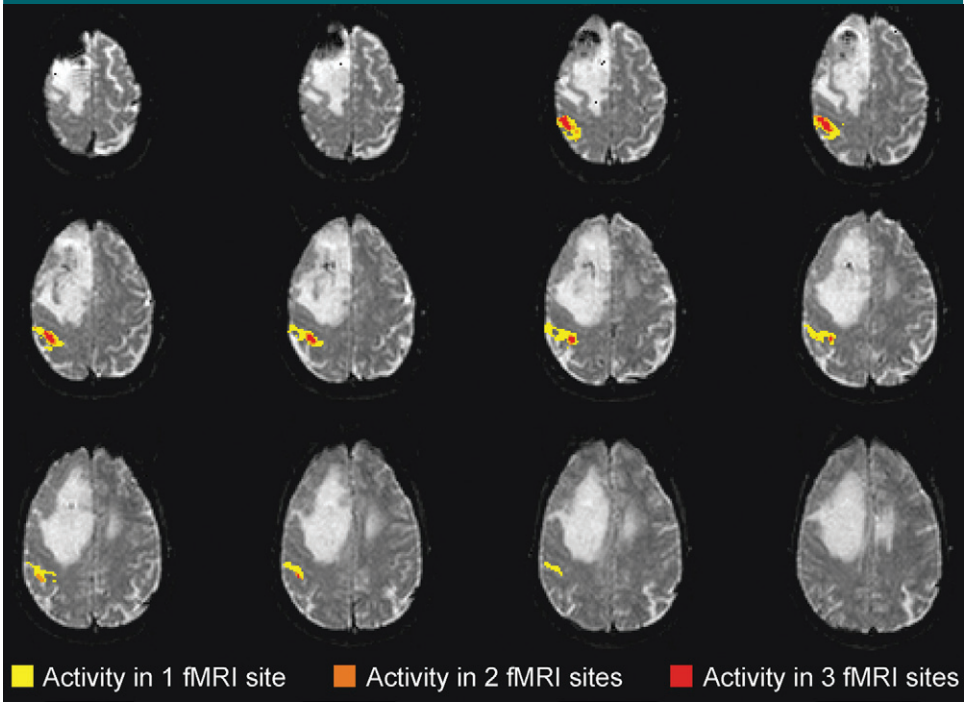

Figure 4: Example for activity overlap when threshold for ROI activation was $P < .001$ (uncorrected) in patient 6. Active voxels are overlaid on the individual echo-planar image reference data (site 2). Yellow voxels indicate that significant activity could be detected at only one site. Orange and red voxels, respectively, indicate replicability of activity at one or two other functional MR imaging sites.

Table 6

Results of Paired *t* Test for Task Differences within Sites

Location	Mean PSC for Somatosensory Task (%)	Mean PSC for Motor Task (%)	*P* Value	Mean CNR for Somatosensory Task	Mean CNR for Motor Task	*P* Value
Site 1	0.86 (0.18)	3.23 (0.62)	<.001	1.28 (0.28)	3.99 (0.88)	<.001
Site 2	0.65 (0.17)	1.35 (0.16)	<.001	1.06 (0.52)	2.53 (0.59)	<.001
Site 3	1.43 (0.56)	2.49 (0.66)	.004	1.29 (0.36)	2.41 (0.57)	.001

Note.—Numbers in parentheses are standard deviations.

previously reported values (9,10,12–15). The investigators in some studies directly compared patients with healthy subjects; however, they did so only at specific single sites and with diverging reliability results (26–33).

There were limitations to the current study. Within one functional MR imaging experiment, it is not possible to control every aspect of a sensorimotor task, and we varied two aspects at the same time: stimulation (motor or somatosensory) and task standardization (low or high). Currently, tasks for localizing the perirolandic area vary much with the local clinical study group. To reflect this situation, we introduced the following task: "Provide a localization of the primary motor hand area according to your best local procedures." We expected this kind of "low-level task standardization" to provide the most realistic representation of current between-site variability for perirolandic localizations. To see how much the factor "level of task standardization" influences intersite variability, a comparative task requires opposite features: maximum stability of stimulation and minimum variability of patient performance. This was the rationale for choosing a fully automated and well-validated vibrotactile stimulation as comparative task design.

Interestingly, within our setup, high task standardization did not reduce intersite variability. Most probably, this was owing to the low signal strength of our highly standardized somatosensory task. In further studies, the researchers might want to separately vary task standardization and task robustness (given that motor signals are more robust). Further sources of variability included in our study concern variable site-specific methods and patients with varying brain pathologic findings. We think that inclusion of different functional MR imaging methods, magnetic field strengths, and brain pathologic findings is important when one wants to provide a realistic estimate of functional MR imaging variability in current clinical practice characterized by different techniques used by different MR groups (18,19,34–38). Therefore, we investigated a series of patients referred for presurgical evaluation of the primary sensorimotor cortex by using equally weighted data, irrespective of local MR unit hardware, field strengths, and experimental procedures. Although significant intersite differences for signal strength and head motion indicate systematic variations, it is not possible to relate our findings to a specific technical factor or disease group. Such data require specific follow-up studies comparing different pathologic types and MR systems (19). A further limitation concerns the sample size of our study: Because of the large efforts involved, the number of participants is typically restricted in round-robin studies (9–15,39).

We conclude that researchers in clinical practice and clinical functional MR imaging biomarker studies should consider that the center of task-specific brain activation may vary up to 16.5 mm with the investigating site and, thus, should maximize functional MR imaging signal strength and evaluate the reliability of local results with PSC and CNR data.

Acknowledgments: Statistical support was provided by Burkhardt Seifert, PhD, Institut für Sozial-und Präventivmedizin, Abteilung Biostatistik, Universität Zürich (Zurich, Switzerland).

Disclosures of Conflicts of Interest: **M.C.W.** No relevant conflicts of interest to disclose. **J.R.** No relevant conflicts of interest to disclose. **N.K.** No relevant conflicts of interest to disclose. **I.H.** No relevant conflicts of interest to disclose. **A.G.** No relevant conflicts of interest to disclose. **F.P.F.** No relevant conflicts of interest to disclose. **M.A.** No relevant conflicts of interest to disclose. **T.F.** Financial activities related to the present article: none to disclose. Financial activities not related to the present article: paid by Novartis for lecture delivered on September 11, 2012; institution received a travel grant from Medtronic for attendance at the Annual Meeting of Movement Disorders Society, Dublin, June 2012. Other relationships: none to disclose. **M.K.** No relevant conflicts of interest to disclose. **J.N.** No relevant conflicts of interest to disclose. **C.S.** No relevant conflicts of interest to disclose. **W.S.** No relevant conflicts of interest to disclose. **M.V.** No relevant conflicts of interest to disclose. **S.G.** No relevant conflicts of interest to disclose. **F.K.** No relevant conflicts of interest to disclose. **E.K.** No relevant conflicts of interest to disclose. **E.A.** No relevant conflicts of interest to disclose. **S.F.** No relevant conflicts of interest to disclose. **R.J.S.** No relevant conflicts of interest to disclose. **R.B.** No relevant conflicts of interest to disclose.

References

1. Wagner K, Hader C, Metternich B, Buschmann F, Schwarzwald R, Schulze-Bonhage A. Who needs a Wada test? present clinical indications for amobarbital procedures. J Neurol Neurosurg Psychiatry 2012;83(5):503–509.

2. Bentley P, Driver J, Dolan RJ. Cholinesterase inhibition modulates visual and attentional brain responses in Alzheimer's disease and health. Brain 2008;131(pt 2):409–424.

3. MacDonald PA, MacDonald AA, Seergobin KN, et al. The effect of dopamine therapy on ventral and dorsal striatum-mediated cognition in Parkinson's disease: support from functional MRI. Brain 2011;134(pt 5): 1447–1463.

4. Lahti AC, Weiler MA, Holcomb HH, Tamminga CA, Cropsey KL. Modulation of limbic circuitry predicts treatment response

to antipsychotic medication: a functional imaging study in schizophrenia. Neuropsychopharmacology 2009;34(13):2675–2690.

5. Light SN, Heller AS, Johnstone T, et al. Reduced right ventrolateral prefrontal cortex activity while inhibiting positive affect is associated with improvement in hedonic capacity after 8 weeks of antidepressant treatment in major depressive disorder. Biol Psychiatry 2011;70(10):962–968.
6. McGonigle DJ, Howseman AM, Athwal BS, Friston KJ, Frackowiak RS, Holmes AP. Variability in fMRI: an examination of intersession differences. Neuroimage 2000; 11(6 pt 1):708–734.
7. Yoo SS, Wei X, Dickey CC, Guttmann CR, Panych LP. Long-term reproducibility analysis of fMRI using hand motor task. Int J Neurosci 2005;115(1):55–77.
8. Rombouts SA, Barkhof F, Hoogenraad FG, Sprenger M, Valk J, Scheltens P. Test-retest analysis with functional MR of the activated area in the human visual cortex. AJNR Am J Neuroradiol 1997;18(7):1317–1322.
9. Bosnell R, Wegner C, Kincses ZT, et al. Reproducibility of fMRI in the clinical setting: implications for trial designs. Neuroimage 2008;42(2):603–610.
10. Zou KH, Greve DN, Wang M, et al. Reproducibility of functional MR imaging: preliminary results of prospective multi-institutional study performed by Biomedical Informatics Research Network. Radiology 2005;237(3):781–789.
11. Suckling J, Ohlssen D, Andrew C, et al. Components of variance in a multicentre functional MRI study and implications for calculation of statistical power. Hum Brain Mapp 2008;29(10):1111–1122.
12. Brown GG, Mathalon DH, Stern H, et al. Multisite reliability of cognitive BOLD data. Neuroimage 2011;54(3):2163–2175.
13. Friedman L, Stern H, Brown GG, et al. Test-retest and between-site reliability in a multicenter fMRI study. Hum Brain Mapp 2008;29(8):958–972.
14. Gountouna VE, Job DE, McIntosh AM, et al. Functional magnetic resonance imaging (fMRI) reproducibility and variance components across visits and scanning sites with a finger tapping task. Neuroimage 2010; 49(1):552–560.
15. Gradin V, Gountouna VE, Waiter G, et al. Between- and within-scanner variability in the CaliBrain study n-back cognitive task. Psychiatry Res 2010;184(2):86–95.
16. Gallasch E, Fend M, Rafolt D, et al. Cuff-type pneumatic stimulator for studying somatosensory evoked responses with fMRI. Neuroimage 2010;50(3):1067–1073.
17. Geissler A, Lanzenberger R, Barth M, et al. Influence of fMRI smoothing procedures on replicability of fine scale motor localization. Neuroimage 2005;24(2):323–331.
18. Beisteiner R, Klinger N, Höllinger I, et al. How much are clinical fMRI reports influenced by standard postprocessing methods? an investigation of normalization and region of interest effects in the medial temporal lobe. Hum Brain Mapp 2010;31(12): 1951–1966.
19. Beisteiner R, Robinson S, Wurnig M, et al. Clinical fMRI: evidence for a 7T benefit over 3T. Neuroimage 2011;57(3):1015–1021.
20. Plow EB, Arora P, Pline MA, Binenstock MT, Carey JR. Within-limb somatotopy in primary motor cortex: revealed using fMRI. Cortex 2010;46(3):310–321.
21. Schweisfurth MA, Schweizer R, Frahm J. Functional MRI indicates consistent intradigit topographic maps in the little but not the index finger within the human primary somatosensory cortex. Neuroimage 2011;56(4):2138–2143.
22. Clément F, Belleville S. Test-retest reliability of fMRI verbal episodic memory paradigms in healthy older adults and in persons with mild cognitive impairment. Hum Brain Mapp 2009;30(12):4033–4047.
23. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420–428.
24. Simões-Franklin C, Whitaker TA, Newell FN. Active and passive touch differentially activate somatosensory cortex in texture perception. Hum Brain Mapp 2011;32(7): 1067–1080.
25. Haglund MM, Berger MS, Shamseldin M, Lettich E, Ojemann GA. Cortical localization of temporal lobe language sites in patients with gliomas. Neurosurgery 1994; 34(4):567–576; discussion 576.
26. Ances B, Vaida F, Ellis R, Buxton R. Test-retest stability of calibrated BOLD-fMRI in HIV- and HIV+ subjects. Neuroimage 2011;54(3):2156–2162.
27. Maïza O, Mazoyer B, Hervé PY, et al. Impact of cognitive performance on the reproducibility of fMRI activation in schizophrenia. J Psychiatry Neurosci 2010;35(6):378–389.
28. Manoach DS, Halpern EF, Kramer TS, et al. Test-retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects. Am J Psychiatry 2001;158(6):955–958.
29. Fernández G, Specht K, Weis S, et al. Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. Neurology 2003;60(6):969–975.
30. Atri A, O'Brien JL, Sreenivasan A, et al. Test-retest reliability of memory task functional magnetic resonance imaging in Alzheimer disease clinical trials. Arch Neurol 2011;68(5):599–606.
31. Maïza O, Mazoyer B, Hervé PY, Razafimandimby A, Dollfus S, Tzourio-Mazoyer N. Reproducibility of fMRI activations during a story listening task in patients with schizophrenia. Schizophr Res 2011;128(1–3): 98–101.
32. McKinsey RD, Moritz CH, Meyerand ME, Tomé WA. Assessment of multiple task activation and reproducibility in patients with benign and low-grade neoplasm. Technol Cancer Res Treat 2010;9(4):319–326.
33. Kimberley TJ, Khandekar G, Borich M. fMRI reliability in subjects with stroke. Exp Brain Res 2008;186(1):183–190.
34. Detre JA. Clinical applicability of functional MRI. J Magn Reson Imaging 2006;23(6): 808–815.
35. Jezzard P, Buxton RB. The clinical potential of functional magnetic resonance imaging. J Magn Reson Imaging 2006;23(6): 787–793.
36. Matthews PM, Honey GD, Bullmore ET. Applications of fMRI in translational medicine and clinical practice. Nat Rev Neurosci 2006;7(9):732–744.
37. Bennett CM, Miller MB. How reliable are the results from functional magnetic resonance imaging? Ann N Y Acad Sci 2010;1191: 133–155.
38. Partovi S, Konrad F, Karimi S, et al. Effects of covert and overt paradigms in clinical language fMRI. Acad Radiol 2012;19(5): 518–525.
39. Glover GH, Mueller BA, Turner JA, et al. Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies. J Magn Reson Imaging 2012;36(1):39–54.