A Prospective Multi-institutional Study of the Reproducibility of fMRI: A Preliminary Report from the Biomedical Informatics Research Network

Kelly H. Zou[1,2,6], Douglas N. Greve[3,6], Meng Wang[1,6], Steven D. Pieper[1,6], Simon K. Warfield[1,4,5,6], Nathan S. White[3,6], Mark G. Vangel[3,6], Ron Kikinis[1,6], William M. Wells[1,4,6], and First Birn[6]

[1] Surgical Planning Laboratory, Brigham and Women's Hospital,
[2] Department of Health Care Policy, Harvard Medical School,
[3] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital,
[4] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology,
[5] Computational Radiology Laboratory, Brigham and Women's Hospital,
[6] Functional Imaging Research of Schizophrenia Testbed (FIRST), Biomedical Informatics Research Network (BIRN)

{zou,mwang,pieper,warfield,kikinis,sw}@bwh.harvard.edu,
{greve,nwhite,vangel}@nmr.mgh.harvard.edu

Abstract. Functional magnetic resonance imaging (fMRI) has significantly contributed to understanding both normal and diseased human brains. Variability often exists in the magnitude, spatial distribution, and statistical significance of the resulting fMRI maps due to differences in equipment and other site-specific differences. In addition, because of costly imaging, demanding tasks, and analytical burden, understanding the effect of these differences may help develop an efficient pooling and comparison mechanism.

Prospective multi-institutional repeated fMRI data were acquired recently in the first phase of the extensive Functional Imaging Research of Schizophrenia Testbed study, sponsored by the Biomedical Informatics Research Network (BIRN) in the US. Five "human phantoms," who were right-handed healthy males, were included in the study. These subjects repeatedly performed the same sensory-motor task over 10 of the 11 study sites on 2 separate visits per site.

The effects of factors such as subject, study site, field strength, vendor, K-space, visit, and repeated run on the fMRI reproducibility were evaluated. Over 4 repeated runs per visit at each site, at a given binarizing activation threshold, we first calculated a three-dimensional (3D) brain activation map via an intial expectation and maximization (EM) algorithm. Site-to-site differences were then assessed based on a second-level hiearchical EM. Against the estimated gold standard of the 3D activation map, activation percentage, sensitivity, specificity, and receiver operating characteristic curves were then estimated using voxel counts. A statistical regression model was used to assess the significance of accuracy predictors with p-values generated in order to explain those factors contributing towards the variability in repeated brain activation maps.

C. Barillot, D.R. Haynor, and P. Hellier (Eds.): MICCAI 2004, LNCS 3217, pp. 769–776, 2004.

1 Introduction

Functional MRI (fMRI) has significantly contributed to studies of both the normal and diseased human brain. Unfortunately, variability may exist in the magnitude, spatial distribution, and statistical significance of resultant fMRI maps. The reasons for such variability are multi-factorial and are important to study [1-7].

Recently, in the US, the functional subsection of the Biomedical Informatics Research Network (BIRN; http://nbirn.net) aimed at comparing and calibrating the fMRI signals in order to determine whether the inter-relation of fMRI maps from different sites was meaningful. This is an initial extensive effort prior to collecting prospective fMRI data of the Schizophrenic versus control subjects in the next phase of this large multi-institutional prospective study.

In this prospective study with 5 healthy “human phantoms" performing the same tasks during two visits at each of the 11 sites, we investigated the effects of factors such as study site, field strength, vendor, visit, repeated run on the reproducibility of the performance of a sensory-motor (SM) task by these healthy human phantoms in a prospective multi-institutional study. The main goal of our analysis was to characterize the variability seen in a sensory-motor task across runs and sites.

2 Methods

2.1 Study Subjects

A total of 11 sites formed the functional BIRN component of the study. Data were collected from 10 of these 11 sites (five 1.5T, four 3T scanners, and one 4T scanner). Five healthy right-handed male subjects were scanned at each site in two visits on separate days, with 10 task runs per visit. In addition, 3 of those had extra scans in a total of 4 visits only at one of the 10 sites.

2.2 Sensory-Motor (SM) Task

The SM task was performed for 4 out of these 10 fMRI runs during each visit. A block design was used with 15-second epochs of alternating baseline (fixation) and task for a total of 85 (plus the first 2 initally used to reach equalibrum and thus discarded) acquisitions per run. Subjects were instructed to perform bilateral finger tapping on button boxes (1 dummy button box and 1 actual) in time with a 3Hz audio cue and a reversing checkerboard. The subjects pressed buttons 1 through 4 in consecutive order and then back again using both hands, simultaneously and in sync.

2.3 Data Acquisition

Anatomical T2W images were acquired at all sites with FSE/TSE or equivalent RARE: oblique axial, FOV 22 cm, 35 slices, 4 mm, TR/TE 4000 ms/68 ms, train length 12, 256×192 matrix, scan time 2:24 min. In addition, 3DSPGR was acquired at 1 site: axial, FOV 22×16.5 cm, 124-128 slices, 1.2 mm, TR/TE/FA 9.8 ms/min/$15\,\mathrm{deg}$, T1 300 ms, 256×192 matrix, BW $\pm$ 15.625 khz, NEX 2, scan time 9:02 min.

Table 1. A list of variables examined in the funcational BIRN study

No	Variable	Value	No	Variable	Value
1	Subject	1, ..., 5	5	Strength	1.5T, 3T, 4T
2	Site	1, ..., 10	6	Vendor	Siemens, GE, Picker
3	Visit	1, 2 (all); 1, ...,4 (3 subjects at 1 site)	7	K-space	Raster, Spiral, Dual-Echo Raster
4	Run	1, ..., 4/Visit			

Functional images were acquired with block-design EPI or spiral GRE: oblique axial, FOV 22cm, 35 slices, 4 mm, TR 3000 ms, TE 30 ms (3T; 4T) 40 ms (1.5T), FA $90\,\mathrm{deg}$, BW$> \pm 100$ khz, 64×64 matrix, 1 shot, 2 dummy frames. The pulse sequences were allowed to vary the K-space trajectory by site. A bite bar was used to minimize head movement.

2.4 Per-voxel fMRI Analysis

Motion correction at each run was applied to middle time point using AFNI (http://afni.nimh.nih.gov/afni). Smoothing was based on FWHM 5mm. Fourier model was used to conduct an F-test to compute the statistical significance at each voxel.

Subject-specific registration was performed over the repeated runs and across the sites in FreeSurfer (http://surfer.nmr.mgh.harvard.edu). Image registration of the anatomical volume with the functional volume was conducted to convert the subject's anatomical volume to the corresponding functional space.

2.5 Statistical Methods

We examined the factors impacting the activation patterns. These included subject ($n = 5$), site ($n = 10$), visit ($n = 2$ or 4), run ($n = 4$), field strength ($n = 3$), vendor ($n = 3$), and K-space ($n = 3$) (Table 1).

Task-related significance (Y) at each voxel was computed using an F-test on the Fourier componant of the task fundamental frequency. At each fixed voxel significance threshold (γ), an estimation-maximization algorithm, developed previously, called the Simultaneous Truth and Performance Level Estimation (STAPLE) [8,9], was applied across the 4 runs to optimally derive a composite 3D gold standard activation map, under a Level 1 STAPLE EM. This algorithm combined all of the factors and enabled visualization of the gold standard in the software, 3D Slicer (http://slicer.org) [10].

Furthermore, a Level 2 STAPLE EM was applied to compare site-to-site differences (see the hiearchical EM-algorithm illustrated in Fig. 1).

Following the Level 1 EM, voxel fractions in the whole brain were used to compute the sensitivity and specificity, for fixed γ, defined respectively as follows:

$$\begin{aligned}\text{Sensitivity} &= \text{True Activation Fraction}\\ &= \Pr(Y > \gamma \mid \text{Gold Standard=Activated Voxel}),\\ \text{Specificity} &= \text{True Non-Activation Fraction}\\ &= \Pr(Y \leq \gamma \mid \text{Gold Standard=Non-Activated Voxel}).\end{aligned}$$

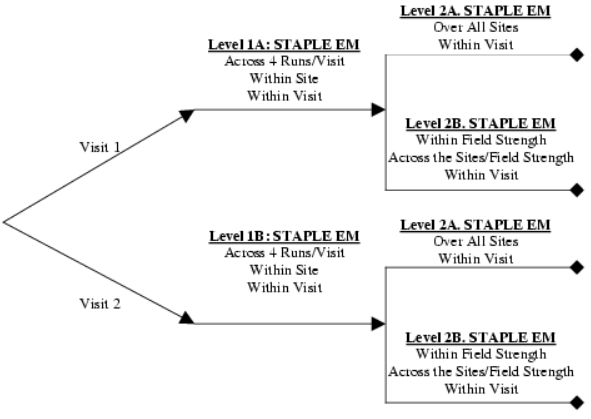

Fig. 1. Subject-specific flowchart of a hierarchical scheme to apply the STAPLE EM algorithm, stratified by visit; *Level 1*: within-site EM was performed, and *Level 2*: between-site EM was performed to generate 3D gold standard activation maps

Following the Level 2 EM, site-specific bi-normal parametric receiver operating characteristic (ROC) curves, a plot of sensitivity vs. (1−specificity), were generated from the activation data on a continuous scale. The area under each ROC curve (AUC) represented the overall classification accuracy, where $\text{AUC} = \Phi\left(\alpha/\sqrt{1+\beta^2}\right)$, (α, β) are the bi-normal ROC parameters based on their maximum likelihood estimates [11-16], and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal.

Linear models were used to compute the p-values for assessing the significance of the factors. Analytic software used included Matlab (http://www.mathworks.com) and S-Plus (http://www.insightful.com).

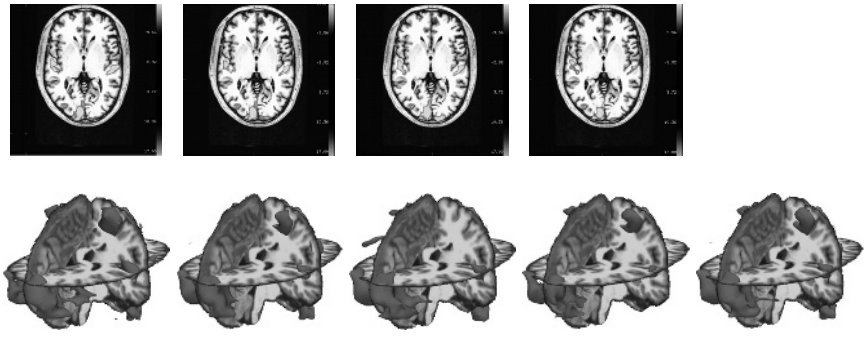

Fig. 2. Registered activation maps on anatomical data of Subject 3 during Visit 1 at one site with a 3T scanner: *top 4 panels*: Runs 1 to 4 in 2D; *bottom left 4 panels*: Runs 1 to 4 in 3D over 35 slices, and *bottom right panel*: the estimated 3D gold standard activation map derived by Level 1 STAPLE EM

Table 2. P-values indicating the significance of the factors on sensitivity and specificity, respectively, based on Level 1 STAPLE EM and regression analysis

Variable	n	p-Value on Sensitivity	p-Value on Specificity
Subject	5	**0.01**	**0.04**
Site	10	0.66	0.47
Strength	3	0.26	0.57
Vendor	3	0.79	0.85
K-space	3	0.93	0.40
Visit	{2; 4}	0.38	0.91
Run	4	0.35	**0.04**

Table 3. Mean activation percentage of all voxels in the brain, sensitivity, and specificity by field strength and subject, based on Level 1 STAPLE EM and regression analysis

Strength	Subject	Activation Percentage	Sensitivity	Specificity
1.5T	1	0.0922	0.5548	0.9997
	2	0.1777	0.6845	0.9996
	3	0.2468	0.5691	0.9992
	4	0.0626	0.5544	0.9998
	5	0.0606	0.4185	0.9498
3T	1	0.7484	0.6558	0.9981
	2	1.2848	0.7543	0.9982
	3	3.0342	0.6199	0.9883
	4	0.8781	0.5792	0.9967
	5	1.1136	0.6834	0.9970
4T	1	0.7550	0.5694	0.9977
	2	1.9181	0.7972	0.9973
	3	1.7149	0.6268	0.9946
	4	0.4274	0.4365	0.9979
	5	0.8185	0.6727	0.9986

3 Results

Of all scanners, 5 were 1.5T; 4 were 3T; 1 was 4T. Significant factors for sensitivity included subject (p=0.01) and for specificity included subject (p=0.04) and run (p=0.04) (Table 2). Registered data for a subject and site were provided in Fig. 2.

At the threshold of $\gamma = 10^{-9}$ to minimize false discovery rates [17], the mean activation percentage of all voxels in the brain, sensitivity, and specificity are presented (Table 3 and Fig. 3). At 3T, the mean sensitivity per subject ranged $0.58 - 0.76$ while the mean specificity ranged $0.99 - 1.00$. At 4T, available at only one study site, the mean sensitivity per subject ranged $0.44 - 0.80$ while the mean specificity ranged $0.99 - 1.00$. At 1.5T, however, the mean sensitivity only ranged $0.42 - 0.69$ while the mean specificity ranged $0.95 - 1.00$.

The ROC curves and their AUCs (Table 4 and Fig. 4) demonstrated moderate to high classification accuracy, which was generally higher at 3T (AUC ranged $0.69 - 0.92$) and 4T (AUC ranged $0.77 - 0.96$), than at 1.5T (AUC ranged $0.52 - 0.77$).

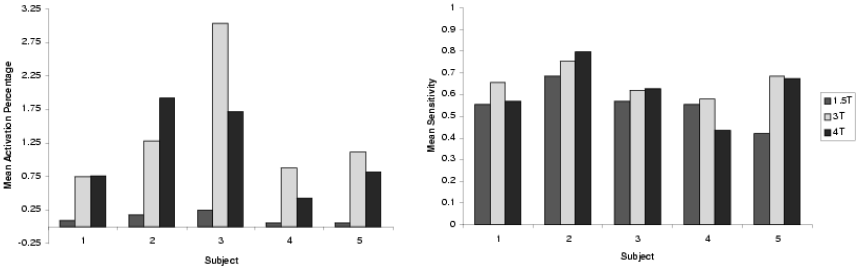

Fig. 3. *Left panel*: mean activation percentage and *right panel*: mean sensitivity, by subject and field strength using Level 1 EM; *left to right bins:* 1.5T, 3T, and 4T for each subject. See Table 3 for actual values, with specificities all close to 1.

Table 4. Estimated ROC parameters (α, β) and the corresponding areas under the ROC curves, AUC, based on Level 2 STAPLE EM

Strength	Site	Vendor	K-space	Visit 1			Visit 2		
				α	β	AUC	α	β	AUC
1.5T	1	Siemens	Raster	9.1029	12.3057	0.7695	4.1368	8.6258	0.6831
	2	Siemens	Raster	0.2703	1.7500	0.5533	3.5568	6.6278	0.7022
	3	GE	Raster	2.3450	4.0026	0.7151	2.5243	5.0277	0.6888
	4	GE	Spiral	3.8472	6.5266	0.7199	5.6221	7.5293	0.7704
	5	Picker	Raster	0.0621	0.7772	0.5195	0.0772	0.9449	0.5224
3T	6	Siemens	Dual-Echo Raster	3.9681	4.0588	0.8288	4.1202	4.1097	0.8350
	7	Siemens	Raster	9.4499	6.7944	0.9156	6.4577	5.7678	0.8650
	8	GE	Spiral	1.6394	2.3350	0.7407	2.3597	2.8091	0.7856
	9	GE	Raster	4.1490	6.0816	0.7496	1.9062	3.6907	0.6909
4T	10	GE	Spiral	5.5703	2.9369	0.9637	2.4119	3.1739	0.7657

4 Conclusions

In this unique multi-institutional prospective fMRI reproducibility study, we discovered the effects of the following factors in terms of the estimated mean activation percentage, sensitivity, specificity, and AUC:

The effect of individual subjects: There was a significant between-subject variability; however calibration may be feasible as part of the pooling mechanism of different cohorts.

The effect of field strengths: Both 3T and 4T were better than 1.5T, yielding more activation and less variability in terms of sensitivity and specificity.

The effect of repeated runs: The activation patterns were variable over the runs after the rest and task periods.

The effect of site vs. subject: The variability across subjects appeared greater than that across sites. This finding may help develop a calibration plan to minimize the variability introduced by the sites themselves, ultimately enabling us to pool independent functional data of normal and diseased subjects across different institutions.

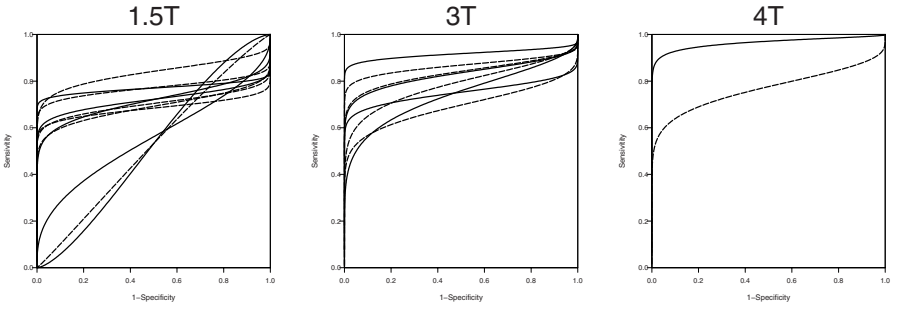

Fig. 4. ROC Curves for Subject 2, by field strength and visit using Level 2 EM; *left panel*: 1.5T, *middle panel*: 3T, and *right panel*: 4T; in each panel, *solid lines*: Visit 1 and *dashed lines*: Visit 2

The effect of visit on different days: Less activation was observed and more robust and systematic activation under different thresholds for the second vs. the first visit. For those three subjects who participated in 4 visits at one site only, less activation was observed for the latter two days. However, there was higher specificity and less variability on these days. A learning effect was not apprarent.

Acknowledgement. The funding for the BIRN study was provided by Grant NCRR P41RR13218. The authors are partially supported by NIH R01LM007861-01A1, R03HS013234-01, CA89449-01, R21MH67054, the Harvard Center for Neurodegeneration and Repair, and the Whitaker Foundation.

We acknowledge with thanks constructive comments from investigators and collaborators from all of the 11 participating institutions in the US, particularly the members of the functional BIRN “calibration" group.

References

1. Brannen JH, Badie B, Moritz CH, Quigley M, Meyerand ME, and Haughton VM: Reliability of functional MR imaging with word-generation tasks for mapping Broca’s area. American Journal of Neuroradiology 22 (2001) 1711-1718.
2. Machielsen WCM, Rombouts SARB, Barkhof F, Scheltens P, and Witter MP: fMRI of visual encoding: reproducibility of activation. Human Brain Mapping 9 (2000) 156-164.
3. Le TH and Hu X: Methods for assessing accuracy and reliability in functional MRI. NMR in Biomedicine 10 (1997) 160-164.
4. Genovese CR, Noll, DC and Eddy, WF: Estimating test-retest reliability in fMRI I: statistical methodology. Magnetic Resonance in Medicine 38 (1997) 497-507.
5. Maitra R, Roys SR, and Gullapalli RP: Test-retest reliability estimation of functional MRI Data. Magnetic Resonance in Medicine 48 (2002) 62-70.
6. Casey BJ, Cohen JD, O’Craven K, Davidson RJ, Irwin W, Nelson CA, Noll DC, Hu X, Lowe MJ, Rosen BR, Truwitt CL, Turski PA. Reproducibility of fMRI results across four institutions using a spatial working memory task. NeuroImage 8 (1998) 249-261

7. Wei XC, Yoo S-S, Dickey CC, Zou KH, Guttmann CRG, Panych LP. Functional MRI of auditory verbal working memory: long-term reproducibility analysis. NeuroImage 21 (2004) 1000-1008.
8. Warfield SK, Zou KH, Wells WM III: Validation of image segmentation and expert quality with an expectation-maximization algorithm. Medical Image Computing and Computer-Assisted Intervention-MICCAI, Lecture Notes in Computer Science 2488, Tokyo, Japan (2002) 290-297.
9. Warfield SK, Zou KH, Wells WM III: Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation. IEEE Transactions on Medical Imaging (2004) In Press.
10. Gering DT, Nabavi A, Kikinis R, Hata N, O'Donnell LJ, Grimson WE, Jolesz FA, Black PM, Wells MW III: An integrated visualization system for surgical planning and guidance using image fusion and an open MR. Journal of Magnetic Resonance Imaging 13 (2001) 967-975.
11. Metz CE, Merman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Statistics in Medicine 17 (1998) 1033-1053.
12. Zou KH, Warfield SK, Bharatha A, Tempany CMC, Kaus M, Haker S, Wells WM III, Jolesz FA, Kikinis R: Statistical validation of imagage segmentation quality based on a spatial overlap index. Academic Radiology 11 (2004) 178-189.
13. Zou KH, Warfield SK, Fielding JR, Tempany CM, Wells MW III, Kaus MR, Jolesz FA, Kikinis R: Statistical validation based on parametric receiver operating characteristic analysis of continuous classification data. Academic Radiology 10 (2003) 1359-1368.
14. Zou KH, Wells WM, Kaus MR, Kikinis R, Jolesz FA, Warfield SK. Statistical validation of automated probabilistic segmentation against composite latent expert ground truth in MR imaging of brain tumors. Medical Image Computing and Computer-Assisted Intervention-MICCAI, Lecture Notes in Computer Science 2488, Tokyo, Japan (2002) 315-322.
15. Zou KH, Wells MW III, Kikinis R, Warfield: Three validation metrics for automated probabilistic image segmentation of brain tumors. Statistics in Medicine (2004) In Press.
16. Zou KH, Hall WJ, Shapiro DE: Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. Statistics in Medicine 16 (1997) 2143-2156.
17. Genovese CR, Lazar NA, and Nichols T: Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15 (2002) 870-878.