Technical Note

Multisite fMRI Reproducibility of a Motor Task Using Identical MR Systems

Sergi G. Costafreda, MD,[1]* Michael J. Brammer, PhD,[1] Ricardo Z.N. Vêncio, MD,[2] Maria L. Mourão, MD,[3] Luiz A.P. Portela, MD,[4] Claudio Campi de Castro, MD,[5] Vincent P. Giampietro, PhD,[1] and Edson Amaro, Jr, MD, PhD[6]

Purpose: To estimate between-scanner functional MRI (fMRI) reproducibility in a multisite study.

Materials and Methods: A total of five identical 1.5T MR systems were used to repeatedly scan five subjects while performing a finger tapping task. A two-way (scanners, subjects) random effects analysis of variance (ANOVA) was used to estimate between-scanner and between-subject variability on two outcome variables: task-related mean blood oxygenation level dependent (BOLD) signal change and volume of activation within a motor region-of-interest (ROI).

Results: Between-scanner variability of fMRI data accounted for a small proportion of the total variation in the BOLD signal change (8.34%, $P = 0.114$) and volume of activation (5.46%, $P = 0.203$). Between-subject variation accounted for more than half of the total variation for both measurements (57.17% and 54.46%, respectively, $P < 0.01$).

Conclusion: These results support the feasibility of multisite studies using identical scanner systems.

Key Words: fMRI; multisite studies; reproducibility of results; variance components analysis; motor task
J. Magn. Reson. Imaging 2007;26:1122–1126.

MULTISITE STUDIES may remediate some of the limitations in current functional magnetic resonance imaging (fMRI) research, in particular the difficulties in assembling large sample sizes in clinical populations. Sufficient reproducibility of fMRI measurements across scanners is necessary for multisite investigations. In an extreme situation, if most of the variance in the measurements were attributable to differences between scanners rather than the factors of interest (i.e., subjects, tasks), fMRI results would be largely site-dependent and the validity of pooling results over sites would be in doubt. Estimating variance components is thus important to assess the validity of multisite measurements.

Previous quantitative estimates of the effect of between-scanner variability on functional measurements are, to our knowledge, limited to the results of an experiment performed by the Biomedical Informatics Research Network (1,2). This study identified two scanner-related variables (field strength and type of k-space) in addition to the particular subject under study as significant predictors of result accuracy. However, no estimate of the variance attributable to each of these factors was reported. Other multiscanner investigations have assessed reproducibility using qualitative comparison of the activation maps (3,4).

In the present work, we report the results of a pilot investigation of between-scanner reproducibility. Five subjects were scanned in five magnetic resonance systems on the same day, while performing a blocked design finger-tapping paradigm. Our analysis focuses on a region of interest (ROI) in the motor strip. Reproducibility was estimated using random effects variance components such that the total variance was partitioned into between-scanner, between-subject, and residual components (5).

MATERIALS AND METHODS

Participants and Task

Participants were five healthy, right-handed volunteers (four female), aged 25 to 30 years. The experimental condition was self-paced finger tapping with the right index finger, and the control condition was silent rest. Before scanning, subjects were trained to perform the finger tapping at a uniform rate. Subjects performed the active task for 30 seconds and the control task for 30

[1]Brain Image Analysis Unit, Institute of Psychiatry, King's College London, UK.
[2]Instituto do Cérebro, Hospital Israelita Albert Einstein, and Instituto de Matemática e Estadística, Universidade de São Paulo, Brazil.
[3]Medimagem, Hospital Beneficiência Portuguesa, São Paulo, Brazil.
[4]Hospital do Coração, São Paulo, Brazil.
[5]Instituto do Coração (InCor), Faculdade de Medicina, Universidade de São Paulo, Brazil.
[6]Instituto de Radiologia (InRad), Faculdade de Medicina, Universidade de São Paulo, Brazil.
*Address reprint requests to: S.G.C., PO Box 89, Institute of Psychiatry, De Crespigny Park, SE5 8AF London, UK.
E-mail: s.costafreda@iop.kcl.ac.uk
Received September 26, 2006; Accepted July 10, 2007.
DOI 10.1002/jmri.21118
Published online in Wiley InterScience (www.interscience.wiley.com).

seconds in a blocked design with three repetitions of each condition.

fMRI Acquisition and Analysis

Participants were tested on five different MRI scanners with fMRI capabilities on the same day, from 10 AM to 10 PM. The scanners were all 1.5 Tesla General Electric systems (GE Healthcare Technologies, WI, USA), located in separate hospitals in Sao Paulo, Brazil. Subjects visited all scanner sites together as a group, and the order in which the subjects were scanned was counter-balanced between scanners.

The same team of field engineers maintained all MR systems, and the manufacturer specifications were matched. Although in two of the systems a daily routine measurement of echo-planar imaging (EPI) ghosting and signal-to-noise ratio (SNR) is part of the institutional quality assurance program (6), there was no special procedure taken regarding the results during the weeks preceding the experiment. The same quadrature head coil was used for all measurements.

A total of 60 scans consisting of eight noncontiguous T2*-weighted echo planar axial slices (TR = 3 seconds, TE = 40 msec, flip angle = 90°, and voxel size = 3.3 × 3.3 × 7.7 mm) were collected during the three active/control task cycles. A single T1-weighted spoiled gradient-recalled (acquisition in the steady state SPGR) structural image of the whole brain with 1.5-mm axial slices was acquired for subsequent registration to the standard stereotactic space of Talairach and Tournoux (7).

Software developed at the Institute of Psychiatry, King's College London, UK was used throughout the fMRI preprocessing and analysis (8–12). Subject motion was corrected by realignment to the average scan. Piecewise baseline linear correction was used to remove linear trends within each active and rest blocks. The data were spatially smoothed using an in-plane Gaussian filter with standard deviation (SD) = 1 voxel (full width at half maximum [FWHM] approximately 7.8 mm). To investigate whether reproducibility results were affected by filter size, we repeated the analysis with a smaller filter (SD = 0.5 voxels, FWHM approximately 3.9 mm). Finally, functional data were registered to each subject's own anatomical scan.

Responses to the experimental paradigm were detected by time-series analysis using a convolution of a box-car function representing the experimental design with two gamma-variate functions with peak responses at four and eight seconds. The weighted sum of the two convolutions giving the best fit (minimizing least squares) to the time series for each voxel was computed. The statistical significance of the resulting ratio of model over residual sum of squares (SSQ ratio) was evaluated against a distribution derived under the null-hypothesis of no experimental effect and generated by 20 cyclical permutations of the time series after orthogonal transformation into the wavelet domain (11). Repeated application of this method at each voxel followed by recalculation of the goodness of fit statistic from the permuted data allows the data-driven calculation of the null distribution of the statistic. Using this distribution it is possible to compute the critical value of the goodness of fit statistic to threshold the maps for any desired type I error. The detection of activated voxels is then extended from voxel to cluster level (12). We used the conventional $P = 0.05$ threshold level, corrected for multiple comparisons, throughout individual and group analysis, except for cluster-level analyses, in which the threshold was less than one expected type I error cluster per brain. In addition to the SSQ ratio, we computed for each voxel an estimator of the change in blood oxygenation level dependent (BOLD) response attributable to the experimental effect normalized to the mean amplitude at that voxel. The maps of the goodness of fit statistic and BOLD signal change were then transformed into the standard space of Talairach and Tournoux (7).

Group activation maps based on the 25 observations were computed by determining the median of the goodness of fit statistic at each voxel (over all individuals) in the observed and permuted data maps. The distribution of the median goodness of fit statistic over all intracerebral voxels from the permuted data was then used to derive the null distribution. In this manner, a group and 25 individual unthresholded BOLD signal change maps and thresholded SSQ ratio maps (defined by combinations of scanner site and subject) were obtained. As expected, a single cluster of activation in the primary motor area was derived from the group map. This cluster was then used as a mask to functionally define the ROI where our analysis would be focused. Within the mask, two outcome variables were retrieved for each individual observation: the mean experimental BOLD percent signal change over the whole ROI and the number of voxels declared activated in the ROI, or volume of activation.

Variance Components Analysis

An analysis of variance (ANOVA) model was fitted to the data for each of the outcome variables, to investigate whether scanner and subject considered as categorical factors were introducing significant variability in the measurement of the outcome variables. These factors were modeled as random effects (5) to allow population inferences. For any of the two outcome variables, the random-effects model may be written:

$$y_{ij} = \mu + a_i + b_j + \varepsilon_{ij} \quad (1)$$

where:

- $i = 1,2, \ldots, I$, $j = 1, 2, \ldots, J$ with I = number of scanner sites, J = number of subjects. Therefore, a_i represents the random site effect, b_j the random subject effect, and ε_{ij} represents the residual error.
- $E(a_i) = E(b_j) = E(\varepsilon_{ij}) = 0$, where $E()$ is the expectation operator.
- $\sigma_A^2 = E(a_i^2), \sigma_b^2 = E(b_i^2), \sigma^2 = E(\varepsilon_{ij}^2)$,
- a_i, b_j, and ε_{ij} are independent.

After fitting the model, a plot of the residuals vs. the fitted values did not reveal any substantial departure from these assumptions.

Estimates for each of the variance components σ_A^2, σ_b^2, and σ^2 were derived using the method of moments. This method often generates negative estimates, when the true variance of the population mean is small. Although we have reported these negative estimates for completeness, they are often considered to represent negligible or null variance (5). The mean squares (MS) expectations for the model in Eq. [1] are as follows:
$E(MS_{scanners}) = \sigma^2 + J\sigma_A^2$
$E(MS_{subjects}) = \sigma^2 + I\sigma_B^2\ E(MS_{residual}) = \sigma^2$,
leading to the estimates:
$\hat{\sigma}_A^2 = MS_{scanners} - MS_{residual}/J$
$\hat{\sigma}_B^2 = MS_{subjects} - MS_{residual}/I\ \hat{\sigma}^2 = MS_{residual}$
The hypotheses of $\sigma_A^2 = 0$ or no significant between-scanner variance, and $\sigma_b^2 = 0$ or no between-subject variance were tested using F-like statistics from the ANOVA model, which distribution was determined by permutation testing with 999 permutations. Following Edgington (13) recommendations for two-way ANOVA, the scanner and subject labels were each permuted independently of each other, that is, permuting the scanner labels within each subject and vice versa. The 95% confidence intervals (CIs) for ratios of between-scanner and between-subject over total variance were computed using the percentile method of the nonparametric bootstrap (14) with 999 samples. The limits of the CI are given by the two values that encompass the central $100(1 - \alpha)\%$ of the bootstrapped distribution of the parameter. In our analysis, the 2.5 and 97.5 percentiles of the ordered distribution of the bootstrapped and observed variances then give the 95% CI. Computations were implemented using the R statistical language (15).

RESULTS

fMRI Activation Maps

Significant activations in motor areas were detected in each of the 25 measurements. The group analysis revealed as expected only one statistically significant cluster of activation at the group level in motor areas, of volume 421 voxels. The coordinates of the peak of maximum activation of that cluster in standard Talairach space were: $x = -29$, $y = -30$, $z = 48$, which define a location in Brodman Area 4 in the precentral gyrus, the primary motor cortex. This cluster was used as a mask to measure the outcome variables (mean BOLD signal change and volume of activation). The individual values for each subject and site are presented in Table 1 and plotted in Fig. 1. In the following, sites are referred to by a number, while capital letters reference subjects.

In spite of important variability, both mean BOLD signal and volume of activation were remarkably reproducible within subjects across scanners. For example, subject A showed in each system the largest signal change and volume of activation. As well, mean BOLD signal change for subject B varied within a relatively small range across all five MR systems (0.15–0.24%). No such a strong effect was apparent for scanners, although scanner 2 showed higher clustering of the results across subjects than the other systems (Fig. 1).

Table 1
Mean BOLD Signal Change and Volume of Activation in the Motor Activation Cluster for Each Subject by Scanner Site

	Subjects				
	A	B	C	D	E
BOLD signal change (%)					
Site 1	0.35	0.15	0.13	0.17	0.09
Site 2	0.22	0.16	0.13	0.14	0.20
Site 3	0.33	0.24	0.12	0.29	0.25
Site 4	0.40	0.22	0.07	0.19	0.17
Site 5	0.38	0.17	0.26	0.22	0.19
Volume of activation					
Site 1	145	48	57	75	34
Site 2	110	81	52	75	83
Site 3	161	80	36	126	111
Site 4	157	109	30	63	44
Site 5	150	82	127	98	76

Variance Components Analysis

Between-scanner variability accounted for 8.34% of the total variance observed in the experimentally induced BOLD signal change. For the volume of activation, the difference between scanners accounted for a smaller share of the total variance in the measurements (estimate of 5.46%). From the random-effects model and the bootstrap analysis there was no evidence that scanners were introducing statistically significant variability in the results, as both *P*-values were above the conventional threshold of 0.05 and all 95%CI included the zero value (Table 2).

Between-subject variability was the largest contributing factor for both outcomes, accounting for 57.17% of the total variability for BOLD signal change and 54.46% for volume of activation. Between-subject variability was thus, a highly significant factor in the variability in the BOLD signal change and volume of activation.

The previous analysis was repeated with a smaller filter size (Gaussian filter with SD = 0.5 voxels, approximately 3.9 mm FWHM). The results of the variance components analysis were similar (Table 2), although there was a decrease in the amount of variance explained by the factors under study.

DISCUSSION

During a finger-tapping paradigm, between-scanner variance represented a relatively small part of the total variance for two important aspects of the activation within a primary motor ROI: experimentally induced BOLD signal change and volume of activation. The highest point estimate of the between-scanner contribution to variability for either outcome was below 10%, and the upper bound of the CIs was below 15% of the total variance for both measurements. Between-subject differences accounted for approximately half of the total variation.

Variance components analysis produces interpretable estimates of variance proportions attributable to the different sources of variability (5). These estimates can then be used to appraise the reliability of a given

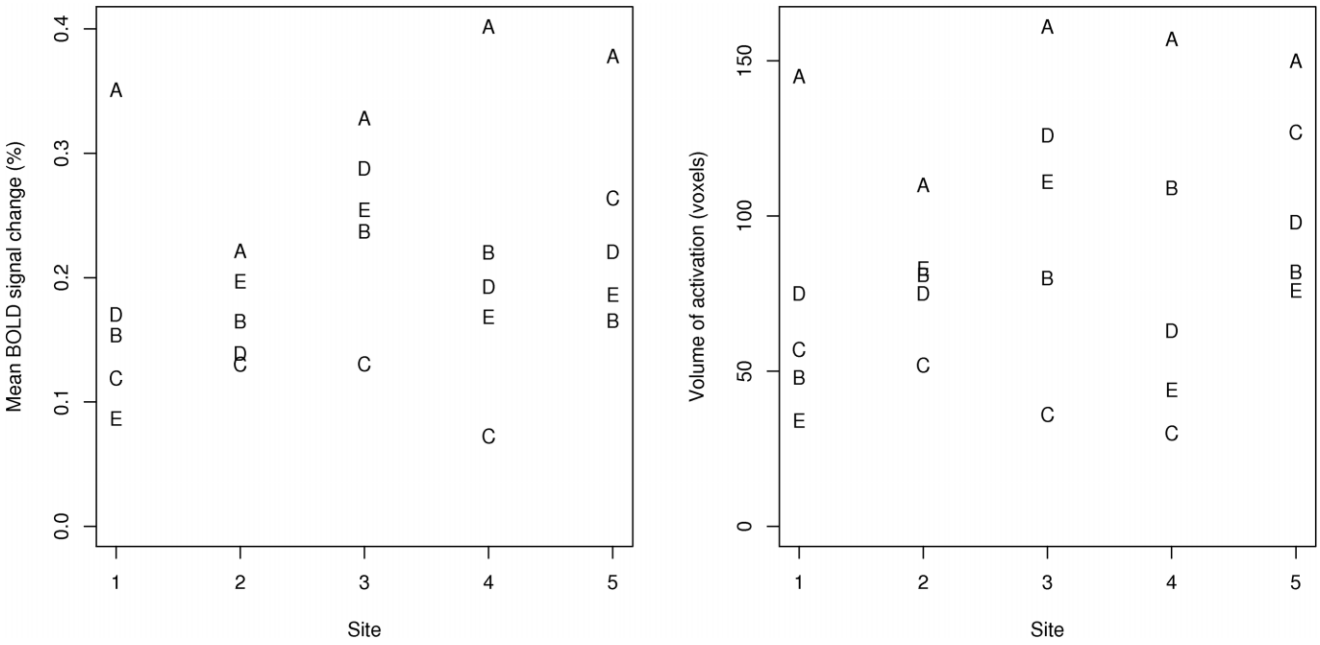

Figure 1. Mean BOLD signal change and volume of activation in the motor activation cluster for each individual (represented by a letter) by scanner site.

measurement system (16) or to design experiments using such a system. Even with this analytic device assessing reproducibility in fMRI activation maps remains problematic. Each factor (subjects, scanners) under study could in principle independently influence at least three parameters in the activation maps: the location of the clusters of activation, the intensity of the signal, and the SNR. We have tried to limit the effect of the first parameter by employing a robust motor paradigm and by restricting our analysis to a motor ROI, which captured most of the relevant activity. Signal intensity variability was estimated in the BOLD signal change estimate. Between-subject or between-scanner variation in statistical noise was addressed indirectly, by including both a statistically thresholded (volume of activation) and an unthresholded (BOLD signal change) measure of activation. We found comparable estimates of variance components for both measurements, suggesting that the variation in statistical noise across subjects or scanners was unlikely to be substantial.

Past work on multisite fMRI reproducibility showed "highly similar results" using a word-stem completion task measured in two laboratories (3) and "reliable patterns of activation" (4) during a working memory task performed across four sites. These assessments were mostly based, however, on qualitative visual judgments of the activation maps and no quantitative estimate of reproducibility was obtained.

To our knowledge, only one previous study has quantitatively examined the issue (1,2). In an experiment performed by the Biomedical Informatics Research Network (BIRN), the same five healthy right-handed males were scanned in 10 study sites with two separate runs at each site while performing a blocked sensory-motor task. The scanners differed in magnet strength and manufacturer. Consistent with our findings, subject was a significant factor, and the authors noted that between-subject variability appeared greater than between-scanner variability. The significant scanner-related factors were noted as field strength, with higher field measurements being more reproducible, and type of k-space. In particular, differences in k-space filtering across scanners, by contributing to the initial spatial smoothness of the images, were found to be an important determinant of reproducibility across scanners (2).

The consequences of k-space filtering are similar to the well-known spatial-domain Gaussian filtering

Table 2
Proportion of Total Variance Due to Between-Scanner and Between-Subject Variability

	Filter FWHM 7.8 mm			Filter FWHM 3.9 mm		
	%	95% CI	*P*	%	95% CI	*P*
Between-scanner						
BOLD signal change	8.34	−9.11 to 13.76	0.114	4.91	−11.21 to 15.02	0.255
Volume of activation	5.46	−10.15 to 14.62	0.203	2.82	−9.96 to 12.71	0.219
Between-subject						
BOLD signal change	57.17	38.81 to 86.87	<0.001	52.14	34.36 to 82.42	0.001
Volume of activation	54.46	34.01 to 85.76	0.001	48.57	30.98 to 80.13	0.004

performed in most fMRI data analysis packages as a preprocessing step (17). This suggests that the amount of Gaussian smoothing employed in our analysis may also have had an impact on reproducibility. We found that a reduction in the size of the smoothing filter from 7.8 to 3.9 mm (FWHM) resulted in a decrease of the variance explained both by differences between subjects and between scanners. That is, the 3.9 mm filter shifted the variability toward the residual, or random, component of the variance. This suggests that the smaller filter may not be optimal for the present analysis, and therefore that the results obtained with the larger (7.8 mm) filter may be more valid.

Following previous studies, our pilot experiment was performed in optimal conditions to minimize between-scanner variability. Subjects performed the same motor task at five sites on a single day. All scanners were identical 1.5 T GE systems, and the same acquisition parameters and sequences were used at each site. The results of the present analysis suggest that multisite studies using homogeneous systems would be valid and interpretable. In "real-life" experimental settings, however, diverse scanner systems are likely to be the norm. It would therefore be important that further experiments were performed to clarify the general conditions for valid multisite studies in heterogeneous environments. Such studies are likely to benefit from a principled combination of measurement standardization and analytical strategies to cope with the extra variability (2).

In our study, the order of the sites was the same for all the subjects, raising the issue of whether a time-related effect could confound the estimates of between-scanner variance. Such a time-related effect could arise from practice effects or boredom. It would be important that future studies on interscanner reproducibility overcome this limitation, for example by randomizing the scanner order across individuals.

In conclusion, our results indicate that the variation in the intensity and volume of the activations attributable to between-scanner variability was small compared to between-subject and residual variation during the performance of a finger-tapping task. Estimates of between-scanner variability are a prerequisite for large-scale multisite studies. Our results support the feasibility of such developments.

ACKNOWLEDGMENT

S.G.C. is supported by a Medical Research Council (UK) fellowship in Neuroinformatics.

REFERENCES

1. 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;237:781–789.
2. 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.
3. Ojemann JG, Buckner RL, Akbudak E, et al. Functional MRI studies of word-stem completion: reliability across laboratories and comparison to blood flow imaging with PET. Hum Brain Mapp 1998;6:203–215.
4. Casey BJ, Cohen JD, O'Craven K, et al. Reproducibility of fMRI results across four institutions using a spatial working memory task. Neuroimage 1998;8:249–261.
5. Brown C, Mosteller F. Components of variance. In: Hoaglin DC, Mosteller F, Tukey JW, editors. Fundamentals of exploratory analysis of variance. New York: Wiley-Interscience; 1991. p 193–251.
6. Simmons A, Moore E, Williams SC. Quality control for functional magnetic resonance imaging using automated data analysis and Shewhart charting. Magn Reson Med 1999;41:1274–1278.
7. Talairach J, Tournoux P. Co-planar stereotaxic atlas of the human brain. New York: Thieme Medical Publishers; 1988. 122 p.
8. Brammer MJ, Bullmore ET, Simmons A, et al. Generic brain activation mapping in functional magnetic resonance imaging: a nonparametric approach. Magn Reson Imaging 1997;15:763–770.
9. Bullmore ET, Brammer MJ, Rabe-Hesketh S, et al. Methods for diagnosis and treatment of stimulus-correlated motion in generic brain activation studies using fMRI. Hum Brain Mapp 1999;7:38–48.
10. Bullmore ET, Long C, Suckling J, et al. Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains. Hum Brain Mapp 2001;12:61–78.
11. Bullmore E, Fadili J, Maxim V, et al. Wavelets and functional magnetic resonance imaging of the human brain. Neuroimage 2004;23(Suppl 1):S234–S249.
12. Bullmore ET, Suckling J, Overmeyer S, Rabe-Hesketh S, Taylor E, Brammer M. Global, voxel and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. IEEE Trans Med Imaging 1999;18:32–42.
13. Edgington ES. Factorial experimental designs. In: Randomization tests, 3rd edition. New York: Marcel Dekker, Inc; 1995. p 121–150.
14. Manly BFJ. The bootstrap. In: Randomization, bootstrap and Monte Carlo methods in biology, 2nd edition. Boca Raton, FL: Chapman and Hall; 1997. p 34–68.
15. R Development Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2005. http://www.R-project.org.
16. Burdick RK, Borror CM, Montgomery DC. A review of methods for measurement systems capability analysis. J Qual Technol 2003; 35:342–354.
17. Lowe MJ, Sorenson JA. Spatially filtering functional magnetic resonance imaging data. Magn Reson Med 1997;37:723–729.