NeuroImage 53 (2010) 119–131

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Multi-site characterization of an fMRI working memory paradigm: Reliability of activation indices

Anastasia Yendiki [a,*], Douglas N. Greve [a], Stuart Wallace [a,b], Mark Vangel [a,c], Jeremy Bockholt [d,e], Bryon A. Mueller [f], Vince Magnotta [g], Nancy Andreasen [g], Dara S. Manoach [a,b], Randy L. Gollub [a,b]

[a] Athinoula A. Martinos Center for Biomedical Imaging, Dept. of Radiology, MGH, Dept. of Radiology, Harvard Medical School, USA
[b] Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA
[c] Mallinkcrodt GCRC Biomedical Imaging Core, Massachusetts General Hospital, Charlestown, MA, USA
[d] University of New Mexico, Albuquerque, NM, USA
[e] The Mind Research Network, Albuquerque, NM, USA
[f] Department of Psychiatry, University of Minnesota, Minneapolis, MN, USA
[g] University of Iowa, Iowa City, IA, USA

ARTICLE INFO

Article history:
Received 11 September 2009
Revised 25 February 2010
Accepted 28 February 2010
Available online 5 May 2010

Keywords:
fMRI
Multi-center studies
Reliability
SIRP

ABSTRACT

Neuroimaging studies are facilitated significantly when it is possible to recruit subjects and acquire data at multiple sites. However, the use of different scanners and acquisition protocols is a potential source of variability in multi-site data. In this work we present a multi-site study of the reliability of fMRI activation indices, where 10 healthy volunteers were scanned at 4 different sites while performing a working memory paradigm. Our results indicate that, even with different scanner manufacturers and field strengths, activation variability due to site differences is small compared to variability due to subject differences in this cognitive task, provided we choose an appropriate activation measure.

Introduction

Multi-site studies provide an efficient means for collecting neuroimaging data from a large number of subjects. Thus they augment our ability to study conditions that are relatively rare in the general population, allow larger samples for studies of genetic polymorphisms, and increase the generalizability of the findings. However, differences in scanner hardware and acquisition protocols may be a source of variability in the data. It is important to quantify this effect and compare it to the variability introduced by other factors, such as individual subject differences and imaging noise, before embarking on studies where data are pooled across multiple sites.

Several recent studies have shown fMRI activation measures to be highly reproducible across sites with identical scanners in tasks ranging from facial affect processing (Suckling et al., 2008) to motor (Costafreda et al., 2007; Sutton et al., 2008) and visual (Sutton et al., 2008). In particular, these studies have found the proportion of the variance in activation measures that can be attributed to across-site variability to be an order of magnitude smaller than the proportion that can be attributed to across-subject variability.

Pooling data acquired at sites with different scanners poses additional challenges. Initial results from a multi-site study performed by the Biomedical Informatics Research Network (BIRN, http://www.nbirn.net) indicated that scanner differences could result in significant variability in fMRI-derived measures of brain activation (Zou et al., 2005). These results were obtained for a sensorimotor paradigm, performed by 5 subjects at 10 different scanners. The experience from this study led to a series of recommendations on how to mitigate across-site variability. These include a quality assurance protocol to ensure stable scanner performance (Friedman and Glover, 2006) and guidelines for data analysis methods that lead to improved reliability of activation measures (Friedman et al., 2008).

In this work we present results from a study of neuroimaging data reliability conducted by the Mind Research Network (MRN) sponsored Mind Clinical Imaging Consortium (MCIC). For this study 10 healthy volunteers traveled to 4 sites and were scanned twice. Structural, functional, and diffusion-weighted MRI data were acquired at each site. Here we focus on the reliability of the functional data.

At the time of the study the sites had scanners from different manufacturers (GE, Waukesha, WI, USA or Siemens, Erlangen,

* Corresponding author. Athinoula A. Martinos Center for Biomedical Imaging, 149 13th St., Charlestown, MA 02129, USA. Fax: +1 617 726 7422.
E-mail address: ayendiki@nmr.mgh.harvard.edu (A. Yendiki).

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.02.084

Germany) and with different field strengths (1.5 T or 3 T). However, all of the sites are also members of BIRN and thus the present study benefited from the lessons learned by phase I of the BIRN study in addressing some of the factors that may result in site differences. This effort included following the specifications of the quality assurance protocol proposed by the BIRN (Friedman and Glover, 2006), as well as standardizing certain acquisition parameters across sites, as described in more detail later.

Although the study presented here involved healthy subjects, it was performed with the ultimate goal of informing a large-scale, multi-site fMRI study of schizophrenia conducted at the same four sites by the MCIC. To this end, the paradigm studied here is one of particular interest to schizophrenia research. It consisted of a variation of the Sternberg item recognition paradigm (SIRP) (Sternberg, 1966), tailored for use in neuroimaging experiments (Manoach et al., 1997). Performance of the SIRP is relatively stable in healthy participants, even after extensive daily practice (Kristofferson, 1972). In fMRI studies, the SIRP gives rise to activation in a network of brain areas associated with working memory and has been used to characterize working memory deficits in schizophrenia patients (Manoach et al., 1999; Manoach et al., 2000; Ragland et al., 2007). The within-subject reliability of SIRP activations has been found to be high for healthy subjects but low for schizophrenia patients (Manoach et al., 2001). Here we study the across-site reliability of these activations in healthy individuals.

Materials and methods

Experimental design and data acquisition

Ten healthy subjects (ages 30–63, 5 males) traveled to four sites and were scanned while performing the SIRP on each of two visits (test–retest). The four sites were: Massachusetts General Hospital (MGH), University of New Mexico (UNM), University of Iowa, and University of Minnesota. Two of the sites used 3 T scanners (Siemens at MGH and Minnesota), while the other two used 1.5 T scanners (Siemens at UNM and GE at Iowa).

The participating sites are also members of the BIRN and in that capacity they had been part of multi-site MRI calibration studies by the Morphometry BIRN (Han et al., 2006; Jovicich et al., 2009) and Function BIRN (Friedman and Glover, 2006; Friedman et al., 2008). The lessons learned from those studies were then applied to reduce disparities in the experimental set-up, data acquisition methods and sequences used for the study presented here. In particular, all sites had matched button press devices, followed common audiovisual set-up calibration methods and paid particular attention to centering each subject's head in the center of the scanner bore to minimize gradient distortion effects. Sequence parameters such as bandwidth and echo spacing were optimized at each site for the best quality images and synchronization of the stimulus onset with the scan start was improved. In addition, the four sites followed the quality assurance procedures recommended by the BIRN to ensure scanner stability (Friedman and Glover, 2006). However, each site followed its own choice of head immobilization strategy (foam packing, soft-strap restraints, or none). The subjects wore Avotec headphones with active noise cancellation (Avotec, Inc., Stuart, FL) during all scans at all sites.

During each visit, a subject performed the SIRP task (EPrime v1.1, Psychology Software Tools, Inc., Pittsburg, PA) during four separate scans. Thus each subject performed the task paradigm a total of 32 times (4 scans × 2 visits × 4 sites). A total of 316 scans were analyzed because data was not available for one of the visits of one of the subjects. Most of the test–retest visits took place on subsequent days. The only exceptions were two cases with 2 days between test and retest and one case each with 6, 7, and 32 days between test and retest.

For each scan we acquired whole-brain, gradient-echo, EPI data along 27 contiguous oblique axial slices, parallel to the AC–PC line (in-plane resolution 3.44 mm, slice thickness 4 mm skip 1 mm, slice order interleaved, TE = 30 ms for 3 T, TE = 40 ms for 1.5 T, TR = 2 sec, FA = 90°, FOV = 22 cm). A total of 177 time frames were collected for a total scan time of 5 min 54 s.

During each scan the subject had to retain in memory a set of 1, 3 or 5 digits during blocks of 46 s, providing a range of task difficulty. First the subject was prompted by the word “Learn” for a time of 1.5 s (*prompt* condition), followed by a blank screen for 0.5 s. Then the targets (digits to be retained in working memory) were presented in red font for a time of 6 s (*encode* condition). The subject was then shown a sequence of probe digits in green font and had to indicate whether each probe digit was a target or a foil, i.e., whether it was a member of the memorized set or not (*probe* condition). The probe condition lasted a total time of 38 s. Each probe digit was presented for up to 1.1 s in a pseudo-randomly jittered fashion within a 2.7 s interval. We presented 14 probe digits in each block, of which 7 were targets and 7 were foils, for a total of 84 probes per scan. Subjects were instructed to respond with a right-thumb button press if the probe digit was a target and a left-thumb button press if it was a foil.

Subjects were instructed to respond as quickly and as accurately as they could. They were told that they would receive a bonus of $0.05 for every correct response. Subjects were trained to perform the task on a computer prior to the first scan session to verify that they achieved a greater than chance performance.

A working memory (WM) block consisting of a single repetition of the *prompt–encode–probe* conditions was then repeated six times per scan. We alternated WM blocks with blocks of fixation. The durations of the fixation blocks were random integer multiples of 2 s, chosen so that the total duration of all fixation blocks within a scan was 78 s. Among the six WM blocks in a scan, there were two blocks of each of the three set sizes (1, 3 and 5) in a pseudorandom order.

We varied the digits that comprised the memory sets for each of the 32 scans to eliminate learning effects. The target digits presented in each block were randomly chosen integers between 0 and 9, with no digit repeated within a single set. To avoid response biases, no digit was used more than 60% of the time as a target digit across the 6 scans in a visit (2 practice scans and 4 experimental scans). Also, in the two sets within a scan that consisted of a single digit, that digit was not the same. The order of targets and foils within a probe epoch was random, but no more than 3 consecutive digits could be targets. Each of the target digits presented during the encode epoch had to be presented at least once during the probe epoch. When the set presented during the encode epoch consisted of 3 target digits, each target digit had to be presented at least twice during the probe epoch.

In addition to the functional data, T1-weighted high-resolution structural scans were collected and we use them here for anatomical localization. Although T1-weighted scans were acquired at all four sites, we used the ones collected at a single site throughout the analyses presented here, as our focus in this work was the variability of the functional data. Specifically, the T1-weighted scans that we used in the present study were acquired at the MGH site on a Siemens 1.5 T scanner with an axial GRE sequence (in-plane resolution 0.625 mm, slice thickness 1.5 mm, FOV = 16 cm, 256 × 256 × 144 matrix, TR = 12 ms, TE = 4.76 ms, FA = 20°, NEX = 3).

Analysis of behavioral data

We recorded the accuracy and latency of the subjects' responses to the probe digits using the same equipment at all sites (EPrime and a NeuroScan response pad, NeuroScan, Charlotte, NC). We performed analysis of variance (ANOVA) on the reaction time (RT) data, modeling subject, site, and run as random effects, and visit, memory load (1, 3 and 5), probe type (target or foil), and site order as fixed effects.