Reproducibility of Functional MR Imaging Results Using Two Different MR Systems

Erik-Jan Vlieger, Cristina Lavini, Charles B. Majoie, and Gerard J. den Heeten

BACKGROUND AND PURPOSE: In the application of functional MR imaging for presurgical planning, high reproducibility is required. We investigated whether the reproducibility of functional MR imaging results in healthy volunteers depended on the MR system used.

METHODS: Visual functional MR imaging reproducibility experiments were performed with 12 subjects, by using two comparable 1.5-T MR systems from different manufacturers. Each session consisted of two runs, and each subject underwent three sessions, two on one system and one on the other. Reproducibility measures D (distance in millimeters) and Rsize and Roverlap (ratios) were calculated under three conditions: same session, which compared runs from one session; intersession, which compared runs from different sessions but from the same system; and intermachine, which compared runs from the two different systems. The data were averaged per condition and per system, and were compared.

RESULTS: The average same-session values of the reproducibility measures did not differ significantly between the two systems. The average intersession values did not differ significantly as to the volume of activation (Rsize), but did differ significantly as to the location of this volume (D and Roverlap). The average intermachine reproducibility did not differ significantly from the average intersession reproducibility of the MR system with the worst reproducibility.

CONCLUSION: The location of activated voxels from visual functional MR imaging experiments varied more between sessions on one MR system than on other MR system. The amount of the activated voxels is independent of the MR system used. We suggest that sites performing functional MR imaging for presurgical planning measure the intersession reproducibility to determine an accurate surgical safety margin.
Experimental Protocol

Each subject underwent three sessions on three different days, twice on one system and once on the other. A session consisted of two runs, without head repositioning, so that each subject underwent six runs in total. In every session, a high-resolution 3D T1-weighted data set was acquired between the two runs. Five subjects had two sessions on system A and one on system B, and seven subjects had two sessions on system B and one on system A.

Experimental Setup

Visual stimuli were presented to the subjects by using the Integrated Functional Imaging System (MR imaging Devices Corporation, Waukesha, WI). This system consisted of a liquid crystal display (LCD) mounted above the head coil, connected by optical fibers to a computer placed outside the magnet room. The stimuli were generated by Eprime (version 1.065; Psychology Software Tools, Pittsburgh, PA). The size of the LCD screen was 17 X 13 cm, and the screen was positioned at an effective distance from the eyes of 40 cm. This yielded a viewing angle of 22°. Mean luminance was 7.8 cd/m² and contrast was 82%. If necessary, ocular refraction was corrected with MR-compatible glasses.

The block-design paradigm consisted of blocks of 45 seconds. The experimental condition consisted of an 8-Hz radial flickering black-and-white checkerboard, which was alternated with a black screen (the rest condition). In both conditions, a small white cross was presented at the center of the screen, and subjects were instructed to focus on this cross. Both the experimental and the rest conditions were repeated twice.

Section Positioning

Full-brain coverage was achieved by using a stack of 22 axial sections, without angulation. In functional MR imaging studies, more advanced section-positioning procedures have been described (9, 19), but these procedures may last up to 15 minutes per subject. We considered this time to be unacceptably long in a clinical situation and decided not to use these methods. Because of our section-positioning protocol, we expected an additional reproducibility error due to the partial-volume effect.

MR Imaging Acquisition

Two 1.5-T MR imaging systems were used: Magnetom Vision (software Numaris VB33; Siemens, Erlangen, Germany) and Signa Horizon (software LX 8.3; GE Medical Systems, Milwaukee, WI). Both systems had echo-planar imaging capabilities. On both systems, the standard quadrature head coils were used. Single-shot echo-planar imaging was used for functional imaging, with the following parameters for both systems: 4500/66 (TR/TE), 90° flip angle, 230-mm field of view, number of sections = 22, 128 X 128 matrix (full k-space), frequency direction left to right, 5-mm section thickness, no intersection gap, 189-second imaging time, bandwidth ~ 1 kHz/pixel. For anatomic reference, a 3D T1-weighted image was acquired. On the Siemens system, the magnetization-prepared rapid acquisition gradient-echo (MP-RAGE) sequence was used (9,7/4/300 [TR/TE/TI], 12° flip angle), and on the GE system the 3D fast spoiled gradient-recalled acquisition in the steady state (FSPGR) sequence was used (30/6, 45° flip angle). In the remainder of the article, the brands of the systems are unnamed.
Results

For each of the 12 subjects, six activation maps were created. An example for one subject is shown in Fig 1.

Properties of the Statistical Maps

Motion (as reported by the motion correction algorithm from BrainVoyager) was below 0.2-mm translation and 0.8° rotation for all 72 ETCs (12 subjects, six ETCs per person). The average activated area (with a z-score threshold of 4 and a cluster-size threshold of 250 mm³) was 38 ± 25 cm² (n = 34) for system A and 34 ± 16 cm² (n = 38) for system B (P > .05). The average signal intensity change was 3.28 ± 0.74% for system A and 2.19 ± 0.44% for system B; this difference was significant (P < 10⁻⁷).

Reproducibility Values

The averages of D, Rsize and Roverlap are plotted in Fig 2. Rsize did not differ significantly between the systems or between the compare conditions (same session, intersession, and intermachine) and was 0.72 ± 0.23 on average.

Results for Roverlap and D were as follows: the highest reproducibility values were found in same-session comparisons with Roverlap = 0.54 ± 0.20 (mean ± SD) and D = 6.7 ± 6.0 mm for system A (n = 17); for system B they were Roverlap = 0.58 ± 0.17 and D = 5.5 ± 3.6 mm (n = 19). These values did not differ significantly between the systems.

Intersession reproducibility values of system A were Roverlap = 0.36 ± 0.19 and D = 13.6 ± 8.5 mm (n = 20), and for system B they were Roverlap = 0.51 ± 0.20 and D = 6.7 ± 5.8 mm (n = 28). Both of these differences were significant, with P = .008 and P < .001, respectively.

For system A, intersession reproducibility values were significantly lower than same-session reproducibility values (P = .005 and P < .001, respectively), but for system B, same-session reproducibility values did not differ significantly from intersession reproducibility values.

Intermachine reproducibility values were Roverlap = 0.36 ± 0.18 and D = 10.5 ± 6.9 mm (n = 96). These did not differ significantly from the intersession results for system A, but were significantly larger than the intersession results for system B (P < .001 and P < .001, respectively).

For the above-mentioned results, the z-score threshold was 4.0 and the cluster-size threshold was 250 mm³. The reproducibility values were also calculated with z-scores ranging from 3 to 6 and cluster-size thresholds from 100 to 1000 mm³, but whether or not statistical significance was obtained in the above-mentioned comparisons was not influenced by changing these thresholds.
Influence of Section Positioning and Frequency- and Phase-Encoding Directions

The influence of section positioning and of the frequency- and phase-encoding directions was determined by calculating the (absolute) distance $D$ for the frequency ($D_f$), phase ($D_{ph}$), and section ($D_s$) directions separately. The averages are plotted in Fig 3.

For same-session reproducibility, $D_f$, $D_{ph}$, and $D_s$ did not differ significantly, nor did they differ significantly between the two systems, and the average value was $3.1 \pm 3.4$ mm.

For system A, the intersession $D_f$ was $4.7 \pm 6.2$ mm, which was significantly smaller than $D_{ph}$ and $D_s$, which were $8.8 \pm 7.2$ mm and $6.7 \pm 5.0$ mm, with the $P$ values being .015 and .033, respectively. For system B, the intersession $D_{ph}$, $D_s$, and $D_f$ did not differ significantly, the average was $3.3 \pm 3.9$ mm. The intersession $D_f$ did not differ significantly between the two systems, but $D_{ph}$ and $D_s$ did, the $P$ values being .015 and .001, respectively.

For intermachine reproducibility, the values were $D_f = 4.6 \pm 4.0$ mm, $D_{ph} = 6.6 \pm 6.3$ mm, and $D_s = 4.7 \pm 4.2$ mm. None of these differed significantly from the intersession values of system A. Compared with the intersession results of system B, the intermachine $D_s$ was significantly larger ($P = .05$).

Discussion

If functional MR imaging is to be used for presurgical mapping, it is essential that the results be reproducible. To see whether the reproducibility of functional MR imaging might be MR-system dependent, we performed reproducibility studies on two different MR systems from different manufacturers. This could also be important when functional MR imaging results are exchanged between different centers.

One limitation of this study might be that the two systems produced significantly different signal intensity changes between the rest state and the activated state. However, we found that this did not induce significant differences in activated volume, and therefore any effect on the reproducibility measures presented herein must have been negligible.

Another limitation could be that a small partial-volume effect might have occurred in intersession and intermachine reproducibility; it was only apparent for system A ($D_s$ was significantly larger than $D_f$). A more subtle section-positioning protocol, such as that proposed by Noll and Eddy (9, 19), might have reduced the partial-volume effect, although we believe it is impossible to avoid this effect completely. The more so because, in the field of surgical mapping, partial-volume effects must be taken into account when determining margins.

A third limitation could be that the method used here to produce statistical maps (Student’s $t$ test) has been shown to suffer from an inherent lower reproducibility than other methods, for example the so-called slope calculation (13). However, even taking into account the noise in the method used here, we found significant differences between the two systems. Therefore, we expect that the differences between the two systems will be even more significant if more advanced methods for the creation of statistical maps are used.

In the literature, two groups of authors describe visual functional MR imaging reproducibility experiments with the reproducibility measures used here. Rombouts et al (7, 12) performed two reproducibility studies. Their first study measured intersession reproducibility by using only two sections and goggles for the visual stimuli; their second study measured same-session and intersession reproducibility with full-brain coverage. Miki et al (16–18) performed one intersession study by using goggles, and two same-session studies, one with goggles and one with checkerboard stimulation. The results of these five articles together with our results are shown in the Table.

<table>
<thead>
<tr>
<th>Study</th>
<th>Stimulus</th>
<th>Same Session</th>
<th>Intersession</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_{size}$</td>
<td>$R_{overlap}$</td>
<td>$R_{size}$</td>
</tr>
<tr>
<td>Current: system A</td>
<td>Checkerboard</td>
<td>0.73</td>
<td>0.53</td>
</tr>
<tr>
<td>Current: system B</td>
<td>Checkerboard</td>
<td>0.76</td>
<td>0.58</td>
</tr>
<tr>
<td>Rombouts et al (7)</td>
<td>Goggles, 2 sections</td>
<td>0.83</td>
<td>0.31</td>
</tr>
<tr>
<td>Rombouts et al (12)</td>
<td>Goggles</td>
<td>0.90</td>
<td>0.74</td>
</tr>
<tr>
<td>Miki et al (16)</td>
<td>Goggles</td>
<td>0.93</td>
<td>0.81</td>
</tr>
<tr>
<td>Miki et al (17)</td>
<td>Checkerboard</td>
<td>0.75</td>
<td>0.61</td>
</tr>
<tr>
<td>Miki et al (18)</td>
<td>Goggles</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Functional MR Imaging Reproducibility Measures from the Current Study and the Literature
from Miki et al seem to be slightly better than our same-session results. This might be caused by their use of goggles and our use of checkerboard stimuli; in Miki et al (18) it is reported that goggles produced better reproducibility results than those of the checkerboard stimuli. The intersession results from the second study from Rombouts et al (12) are similar to our intersession results with system B. The intersession results from Miki et al are not unlike our intersession results with system A.

The most important findings of the present experiments are twofold. First, same-session reproducibility values were independent of the system used. Second, in intersession reproducibility, the amount of activated voxels ($R_{size}$) did not depend on the system, but the positions of these voxels ($R_{overlap}$ and D) did: System B reproduced these positions better than did system A. As was to be expected, intermachine reproducibility was roughly the same as the lowest intersession reproducibility, in our case that of system A.

One of the reasons that system A produced lower reproducibility results is that it suffers more from geometric distortions, as can be noted from the much larger $D_{ph}$. Distortions may have been induced by a less homogeneous magnetic field, perhaps induced by poorer shimming. Another reason for differing results could be that build-in corrections for Nyquist ghost artifacts may differ in effectiveness between the two systems. A robust method to correct both for Nyquist and geometric distortion artifacts is given by Schmitz et al (22), and a method that reduces geometric distortion artifacts is given by Jenkinson (23). Both methods require off-line reconstruction.

The consequences of our study for multicenter functional MR imaging investigations are limited if the focus of the investigation is on volume effects. If, however, the location is important, intermachine reproducibility might be low and hence could decrease statistical power significantly.

The accuracy of functional MR imaging was previously shown (3, 24, 25) to be usually within 1 cm for more invasive methods such as intraoperative mapping. A consequence of our findings for presurgical planning is that the surgical safety margin should be increased by the distance D reported by intersession reproducibility. In our case, this means that the margin to be added is 13.6 mm for system A and 6.7 mm for system B.

For sites performing functional MR imaging as a tool for presurgical planning, we suggest that intersession reproducibility be determined and that the distance D be added to the surgical safety margin. It is most likely that this distance is smaller when either of the above-mentioned methods for correcting Nyquist and/or geometric distortion artifacts is applied.

Other functional MR imaging paradigms than visual ones, including language and motor paradigms, may show other reproducibility data. Intersession reproducibility should therefore be determined for every applied paradigm individually, to obtain a paradigm-dependent surgical safety margin.

Conclusion

Reproducibility of visual functional MR imaging depends on the system used to acquire the functional MR results. The implications for presurgical mapping are that reproducibility measures have to be established per machine. These measures can then be used to determine accurate surgical safety margins.

References


