Impact of cognitive performance on the reproducibility of fMRI activation in schizophrenia

Olivier Maïza, MD; Bernard Mazoyer, MD, PhD; Pierre-Yves Hervé, PhD; Annick Razafimandimbí, PhD; Sonia Dollfus, MD, PhD; Nathalie Tzourio-Mazoyer, MD

Maïza, Mazoyer, Hervé, Razafimandimbí, Dollfus, Tzourio-Mazoyer — Centre d’imagerie–neurosciences et applications aux pathologies, Universités de Caen et Paris Descartes; Maïza, Mazoyer, Dollfus — Centre Hospitalier Universitaire de Caen, Caen; Mazoyer — Institut universitaire de France, Paris, France

Previously published at www.jpn.ca

Background: Longitudinal functional magnetic resonance imaging (fMRI) studies in patients with schizophrenia allow exploration of the course of the illness and brain activity after therapy. A crucial question, however, is whether fMRI findings are reliable, because they can be affected by performance deficits in patients with schizophrenia. Our aim was to evaluate the reproducibility of fMRI activations in highly integrated language areas in patients with schizophrenia, taking into account task performance.

Methods: Ten patients with schizophrenia and 10 matched healthy controls were scanned twice, 21 months apart, while performing a story comprehension task. The reproducibility of the activations in each participant was evaluated globally by the percentage of spatial overlap between the 2 sessions and locally by a voxel-wise computation of the between-session relative standard deviation. We performed between-group comparisons both with and without the inclusion of comprehension scores (measuring task performance) as a covariate.

Results: On average, patients with schizophrenia had significantly lower comprehension scores than controls (4.5/12 v. 7.8/12, \( p = 0.002 \)). The mean spatial overlap between fMRI sessions was 30.6% in the patient group and 47.0% in the control group (\( p = 0.017 \)). Locally, the lower reproducibility in patients was most prominent in the left posterior middle temporal gyrus, inferior frontal gyrus and medial prefrontal cortex (\( p < 0.001 \) uncorrected for multiple comparisons). Comprehension scores were positively correlated with both reproducibility measures in patients (overlap: \( r = 0.82, p = 0.004 \); relative standard deviation: several significant clusters at \( p < 0.001 \)). When we included the comprehension scores as a covariate, most of the local between-group differences in reproducibility were removed, and the difference in overlap was not significant.

Limitations: Owing to the small sample size, we could not investigate the impact of clinical subtypes and different types of medications on reproducibility.

Conclusion: Our findings suggest that the greater variability in activation in patients with schizophrenia compared with controls concerns high-level areas and is mainly attributable to deficient task performance. Consequently, cognitive performance must be carefully controlled when longitudinal fMRI studies are undertaken.

Introduction

Functional magnetic resonance imaging (fMRI) is the method of choice for evaluating neurofunctional correlates of symptoms and cognitive impairments in patients with schizophrenia. Most studies have had a cross-sectional design, but a few fMRI studies have adopted a longitudinal design, with 2 types of objectives. Some authors have attempted to identify functional biomarkers that are stable through the course of the illness and that may represent trait markers. Others have assessed the effects of cognitive therapies or medications on brain activity during cognitive tasks. However, these studies have had conflicting findings because of several methodologic limitations. In particular, they did not address the question of whether fMRI activation was reproducible in patients, which is crucial for establishing the validity of a longitudinal experimental paradigm.

In healthy participants, the test–retest reliability of fMRI activation has been investigated in the context of sensory and motor tasks as well as for various cognitive tasks addressing...
memory, executive functions, and language. Classically, the reproducibility of fMRI activations has been assessed by considering indices such as the percentage of spatial overlap between individual statistically thresholded activation maps at different time points. Reproducibility, as assessed by the percentage of spatial overlap, appears to be higher for motor tasks (54%–84%) than for sensory tasks (49%–64%). Memory tasks yielded less reproducible activations (with spatial overlap ranging from 36% to 41%). For language tasks, similar to the greater reproducibility of motor compared with sensory tasks, word generation elicited more reproducible activations (spatial overlap 75%) than story comprehension (spatial overlap 50%).

However, this type of reproducibility index can be affected by a dependence on the arbitrary statistical threshold that is applied to define activated areas. Despite the good reliability of estimated blood oxygen level–dependent (BOLD) signals, the difference between the task and reference may be slightly above the statistical threshold in one session and slightly below threshold in another session. Several authors have circumvented this problem by computing the percentage of overlap at different thresholds. Another way of alleviating this bias is to threshold the activation maps so as to keep the number of activated voxels constant across sessions. Otherwise, reproducibility indices that are independent of a statistical threshold can be used. In particular, the intraclass correlation coefficient (ICC) and coefficient of variation (CV) are both computed on a region-of-interest or voxel-wise basis and thus allow local assessment of reproducibility. Because of methodologic differences and marked functional differences between the various cerebral regions investigated, it is difficult to compare the results obtained with these indices across studies.

Despite the interest in using longitudinal fMRI studies in clinical populations, few studies have assessed the reproducibility of activations in patients or populations at high risk of illness. Compared with healthy controls, patients at high risk of Alzheimer disease or schizophrenia had similar reproducibility of activations, whereas patients who had experienced a stroke had less reproducible activations.

Only 1 study has investigated the test–retest reliability of measuring fMRI activation within a working memory paradigm in patients with schizophrenia. Relying on ICC to assess reproducibility, Manauch and colleagues showed that patients with schizophrenia exhibited less consistent activations across sessions in areas involved in high-order cognitive processes compared with healthy participants. This raises the question of whether cognitive performance has an impact on the reproducibility of activation. Indeed, some patients performed poorly, but the authors suggested that poor performance was not the cause of low reproducibility, and they concluded that the variability in activation might be intrinsic to schizophrenia. Nevertheless, given the potential of longitudinal fMRI studies in schizophrenia, the effect of performance on reproducibility deserves precise evaluation. Furthermore, ICC, in contrast to CV, assumes that the groups being compared have similar between-subject variability. However, between-subject variability is higher among patients with schizophrenia than among healthy participants, particularly for working memory tasks. Thus, CV appears to be a more relevant measure than ICC for comparing the reproducibility of fMRI activations between patients with schizophrenia and healthy participants.

In the present study, we assessed the effects of performance on the reproducibility of fMRI activations in patients with schizophrenia and healthy participants. We used 2 different reproducibility indices, allowing both global and local assessment of the reproducibility of fMRI activations. We assessed the effect of task performance on both of these indices.

Methods

Participants

We selected participants from the 21 pairs of patients with schizophrenia and matched healthy controls who participated in a previous cross-sectional fMRI study. Ten pairs underwent a second fMRI session and were included in the present study. Eleven of the patients were not included: 3 did not give their consent, 3 were admitted to hospital between the 2 sessions, 4 did not stay in contact with their practitioner and in one case, the data from the second session were not usable owing to an excessive intersession difference of positioning in the scanner, which prevented accurate intersession coregistration.

All patients were free of auditory deficits, neurologic disorders and cerebral abnormalities, as assessed by a brain T2-weighted MRI. All participants reported that French was their primary language and used their right hand for writing. They all had a verbal intelligence quotient (IQ) above 80 (Wechsler Adult Intelligence Scale). Patients and healthy participants (controls) were matched one-to-one for sex, age, at the first fMRI session, level of education and handedness (Edinburgh Inventory score). Schizophrenia was diagnosed by clinicians according to the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV) criteria. The patients were stabilized out-patients with no hospital admissions or psychotic exacerbations during the follow-up period. At each session, their clinical state was evaluated by use of the Positive and Negative Syndrome Scale (PANSS). Control participants were free of psychotic disorders and substance dependence (including alcohol) as assessed by the Structured Clinical Interview of the DSM III-revised (SCID). The local ethics committee (Comité de Protection des Personnes Nord-Ouest, France) approved this study, and the participants provided informed, written consent. Consent was obtained from the schizophrenia patients after an interview with a practitioner. All were outpatients who were capable of understanding the information that was given to them.

Experimental paradigm

All participants underwent 2 fMRI sessions, separated by an average of 21 months. The intersession interval was significantly longer in patients than in controls (patients: mean 17 mo, standard deviation [SD] 5 mo; controls mean 26 [SD 5] mo; t = 4.79, p < 0.001, 2-sample t-test).

At each session, each participant was invited to listen to a
A factual story in French that described a sports competition involving intricate social interactions between the characters (Appendix 1, available online at www.jpn.ca). The French story condition (hereinafter the story condition) was contrasted with listening to the same story in Tamil, a language to which no participant had been exposed to before the experiment. Phonologically, Tamil is relatively similar to French, in the sense that it does not involve tonal patterns, and words and phrases can be segmented by the participants even though the language sounds foreign to them. The Tamil condition was chosen to enable the subtraction of activations elicited by phonologic processing of auditory stimuli. The detection of activity in high-order language integration areas was enhanced upon contrasting the two conditions (story v. Tamil).

For each condition, presentation of the auditory stimuli followed a block design with 9 alternating 30-second duration blocks of either the Tamil condition (5 blocks) or the story condition (4 blocks), starting and ending with the Tamil condition. Participants were in a dark room and were instructed to listen attentively to the stimuli while keeping their eyes closed.

Shortly after scanning, task performance was assessed with a 12-item questionnaire, designed to investigate both the overall understanding of the story and the comprehension of selected details. Oral answers to each item were scored 1, 0.5 or 0 for an exact, partial, or inexact or no response, respectively.

Neuroimaging data acquisition and preprocessing

At both sessions, all participants underwent both anatomic MRI and fMRI scanning with a General Electric Signa 1.5 T magnetic resonance imager. Both scans covered the same field of view (240 mm). We acquired anatomic data with a high-resolution, T1-weighted MRI scan using a fast, 3-dimensional, spoiled gradient echo sequence that used a spectral selective inversion recovery pulse (matrix size 256 × 256 × 124, slice thickness 1.5 mm). The fMRI data were acquired with an echo-planar imaging BOLD sequence (repetition time 6 s, echo time 60 ms, flip angle 90°, matrix size 64 × 64 × 32, 50 volumes, slice thickness 3.8 mm).

The data preprocessing was specifically designed for this longitudinal study and has been previously described. We performed the preprocessing using statistical parametric mapping software (SPM99; Wellcome Department of Imaging Neuroscience). To achieve optimal coregistration of fMRI data from both sessions in the same stereotaxic space, we first corrected the fMRI data for differences in time acquisition between slices and then for head motion; the data were then coregistered with the contemporary anatomic MRI data. We coregistered the anatomic MRI data between sessions, and the coregistration parameters were applied to the contemporary fMRI data. We estimated spatial normalization parameters for the Montreal Neurological Institute T1-weighted template using the anatomic MRI data from session 1, and we applied these parameters to the corresponding fMRI data from both sessions. These spatially normalized fMRI data were then smoothed with a Gaussian kernel of 8 mm and a high-pass filter (0.0102 Hz cut-off).

Maïza et al.
Statistical analyses

Task performance
To assess the possible effects of illness and repetition on task performance, we performed a 3-way mixed-model analysis of variance (ANOVA) with the fixed effects of group (patients v. controls) and session (session 1 v. session 2), and subject as random effects. We evaluated the reproducibility of task performance across sessions separately for each group with Pearson correlation analyses.

fMRI data
We entered the preprocessed fMRI data for each participant from both sessions into a SPM99 regression model for computation of individual contrast maps (story v. Tamil) and the corresponding t maps. We subsequently used both of these maps to assess local and global reproducibility.

Global reproducibility
We assessed global reproducibility of the spatial distribution of activation by computing the percentage of spatial overlap at both sessions (V_overlap) to the number of voxels activated at session 1 (V1) or session 2 (V2). It was computed as follows: percentage of spatial overlap = [(2 × V_overlap) / (V1 + V2)] × 100 (Fig. 1).

The percentage of spatial overlap was computed by thresholding the t maps at a constant activation volume (V1 = V2) to alleviate the bias of an arbitrary statistical threshold. The constant volume was set at 5000 voxels because, in both groups, it maximized the average percentage of spatial overlap and represented the best compromise between the inclusion of irrelevant voxels at high activation volumes (false positives) and the exclusion of relevant voxels (false negatives) at low activation volumes (Fig. 2). We computed the average t values of the 5000 most activated voxels of each individual t map. We assessed the stability across sessions of these average t values by computing correlation coefficients in controls and patients with schizophrenia.

For comparability with other studies, we also computed a traditional “fixed threshold” percentage of spatial overlap with the following significance thresholds: p < 0.001, p < 0.01 and p < 0.05 uncorrected for multiple comparisons (corresponding to t value thresholds of 3.15, 2.35, 1.66, respectively).

We compared global reproducibility, which was assessed by the percentage of spatial overlap, between patients and controls with a Welch 2-sample t test. To assess the potential effect of task performance on reproducibility, we entered the percentage of spatial overlap data as a dependent variable in a linear model, with group, comprehension score and the interaction between the 2 as explanatory variables; thus, one regression slope was estimated per group. We centred the comprehension scores on the mean of the controls in our statistical model; this allowed us to make comparisons between the 2 groups at the same performance level, corresponding to that of the controls.

To assess the possible influence of other factors on reproducibility, we also analyzed whether the percentage of spatial overlap was correlated to age, verbal IQ, intersession duration, controls, n = 10

Patients with schizophrenia, n = 10

Fig. 2: Percentage of spatial overlap between the activation maps from 2 sessions as a function of a given activation volume threshold in patients with schizophrenia and controls.

J Psychiatry Neurosci 2010;35(6)
absolute value of the difference between session comprehension scores, duration of illness, doses of antipsychotic medication and positive or negative PANSS scores. We performed correlation analyses separately for patients and controls.

To display the functional network used to evaluate global reproducibility, individual t maps of the 5000 most activated voxels were binarized, and we computed probability maps of activation for each group at each session.

**Local reproducibility**

To assess local reproducibility on a voxel-wise basis, we computed individual relative standard deviation (RSD) maps across sessions from the 2 contrast maps of each participant (session 1 and session 2). Relative standard deviation is the absolute value of the CV, expressed as a percentage. We defined CV as the ratio of the SD to the mean. For each participant, and at each voxel, we computed the SD between the 2 sessions, divided it by the absolute value of the average of the 2 sessions, and then multiplied this value by 100.

To avoid extremely large RSD values, attributable to a very small average of the 2 sessions, the maps were thresholded at an RSD value of 500%; the maps were then smoothed with a Gaussian kernel of 8 mm.

To identify regions with different patterns of reproducibility between patients and healthy participants, we compared the RSD maps between the groups (2-sample t tests, SPM 99).

To restrict this analysis to the regions belonging to the network involved in this task, we masked the individual RSD images with a binary image representing the network of areas activated during session 1 or 2. This mask was computed by performing a group analysis (1-sample t test on the contrast maps of the 20 participants for session 1 and 2). We applied a liberal threshold of \( p = 0.01 \), uncorrected for multiple comparisons, to the resulting map, which was then binarized. To map the effects of comprehension scores on reproducibility in each group, we computed the correlation between RSD and comprehension scores. Finally, to control for the effect of performance on the difference in RSD between groups, we compared the patient and control groups with a SPM99 multiple regression model, which included the comprehension scores as a covariate, centred over the mean of the healthy participants. The resulting analysis of covariance design was equivalent to the one used for the percentage of spatial overlap (i.e., the difference between groups was estimated at the performance level of the controls, with separate slopes in the 2 groups).

All analyses that dealt with the effect of the comprehension score on reproducibility included the scores from session 1. We repeated these analyses with the scores from session 2.

All voxel-wise statistical analyses were performed with SPM99. The significance threshold was set at \( p < 0.001 \), uncorrected for multiple comparisons, unless otherwise specified. Other analyses were performed with R (http://cran.r-project.org).

**Results**

**Participants**

Tables 1 and 2 show the clinical and demographic data of the participants. At both sessions, 3 patients were classified as paranoid, 2 as undifferentiated and 5 as experiencing residual symptoms. At both sessions, 7 patients were taking an atypical antipsychotic, 2 were taking a typical antipsychotic and 1 was taking both medicines. One patient changed antipsychotic medicines between the 2 sessions (haloperidol followed by flupenthixol).

For the patients with schizophrenia, the PANSS scores and the antipsychotic medication doses in chlorpromazine equivalents were not significantly different (\( p > 0.05 \)) between sessions and were highly correlated across sessions (Table 2). There was...

---

**Table 1: Sociodemographic characteristics of the study groups**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Controls, ( n = 10 )</th>
<th>Schizophrenia patients, ( n = 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of men</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>≥ 12 years of education, no. of participants</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Age at session 1, yr</td>
<td>34.2 (10.1) [25–50]</td>
<td>34.2 (8.1) [24–48]</td>
</tr>
<tr>
<td>Edinburgh Inventory Score(^*)</td>
<td>94 (12) [67–100]</td>
<td>87 (10) [78–100]</td>
</tr>
<tr>
<td>Verbal IQ (^*)</td>
<td>100 (8) [89–115]</td>
<td>98 (12) [82–110]</td>
</tr>
<tr>
<td>Illness duration, yr</td>
<td>NA</td>
<td>12.1 (8.6) [4–28]</td>
</tr>
</tbody>
</table>

\(^*\)IQ = intelligence quotient; NA = not applicable; SD = standard deviation.

---

**Table 2: Clinical characteristics and comprehension scores for the schizophrenia patients and controls at each sessions**

<table>
<thead>
<tr>
<th>Group; measure</th>
<th>Session 1; mean (SD) [range]</th>
<th>Session 2; mean (SD) [range]</th>
<th>t value</th>
<th>p value</th>
<th>r value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schizophrenia patients, ( n = 10 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total PANSS</td>
<td>52.2 (8.3) [41–66]</td>
<td>48.9 (8.1) [39–63]</td>
<td>1.3</td>
<td>0.22</td>
<td>0.52</td>
<td>0.12</td>
</tr>
<tr>
<td>Positive PANSS</td>
<td>11.8 (3.5) [7–18]</td>
<td>10.1 (2.9) [7–14]</td>
<td>2.15</td>
<td>0.06</td>
<td>0.71</td>
<td>0.021</td>
</tr>
<tr>
<td>Negative PANSS</td>
<td>13.8 (5.4) [8–27]</td>
<td>14.4 (4.8) [8–22]</td>
<td>0.38</td>
<td>0.71</td>
<td>0.89</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Medication, CPZ</td>
<td>371.5 (228.7) [40–775]</td>
<td>354.7 (153.4) [125–550]</td>
<td>0.74</td>
<td>0.48</td>
<td>0.92</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Comprehension scores</td>
<td>4.6 (3.4) [0–10]</td>
<td>5.4 (2.6) [0–8.5]</td>
<td>—</td>
<td>—</td>
<td>0.87</td>
<td>0.001</td>
</tr>
<tr>
<td>Controls, ( n = 10 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comprehension scores</td>
<td>7.8 (2.3) [4–11]</td>
<td>8.8 (0.9) [7–10]</td>
<td>—</td>
<td>—</td>
<td>—0.015</td>
<td>0.97</td>
</tr>
</tbody>
</table>

\(^*\)IQ = intelligence quotient; PANSS = Positive and Negative Syndrome Scale; SD = standard deviation.
a trend toward lower positive PANSS scores at the second session ($p = 0.06$), but the difference between session 1 and 2 was low (mean 1.7, SD 2.5) for a scale ranging from 7 to 49.

**Task performance**

The ANOVA of comprehension scores (Table 2) showed no group $\times$ session interaction ($F_{1,18} = 0.002$, $p = 0.96$). Controls performed significantly better than patients in both sessions ($F_{1,18} = 12.32$, $p = 0.002$). The slight increase in scores between sessions was not significant ($F_{1,18} = 2.82$, $p = 0.11$).

Patient comprehension scores were strongly correlated across sessions ($r = 0.70$, $p = 0.024$), but control scores were not correlated across sessions ($r = -0.015$, $p = 0.97$). However, the range (7–10) of comprehension scores from session 2 was very restricted in this group (Fig. 3).

**Global reproducibility**

Patients had a significantly lower percentage of spatial overlap than controls (30.6% [SD 15.8%] v. 47% [SD 11.7%]; $t_{16.5} = 2.64$, $p = 0.017$; Table 3). This difference was not significant when we controlled for comprehension scores ($t_{16} = 0.63$, $p = 0.53$). The percentage of spatial overlap was highly correlated with the comprehension scores in patients ($r = 0.82$, $p = 0.004$) but not in controls ($r = 0.12$, $p = 0.73$; Fig. 4). No other variable was significantly correlated with the percentage of spatial overlap in either group. The same analyses were performed with the percentage of spatial overlap computed at fixed statistical thresholds; this yielded the same results (Table 3). We found a significant correlation between the $t$ values corresponding to the 5000 most activated voxels at the 2 sessions (controls: $r = 0.65$, $p = 0.005$; schizophrenia patients: $r = 0.42$, $p = 0.025$), showing that the threshold selection was reproducible.

Figure 5 shows the most consistently activated brain areas (i.e., areas activated in most patients or controls for each session). Activation patterns were similar for both sessions. In controls, the most consistently activated voxels were located along the left superior temporal sulcus, right anterior superior temporal sulcus, pars triangularis of the left inferior frontal gyrus and left medial prefrontal cortex. In patients, the most consistently activated areas were restricted to the bilateral anterior and middle parts of the superior temporal sulcus.

**Local reproducibility**

Figure 6A shows the reproducibility patterns in both groups. Controls had high reproducibility (i.e., low RSD) in the left cortical areas, particularly along the left superior temporal sulcus but also in the frontal inferior gyrus and medial prefrontal cortex. In contrast, the right hemispheric areas were

---

**Table 3: Percentage of overlap in schizophrenia patients and controls**

<table>
<thead>
<tr>
<th>Measure; session</th>
<th>Controls, $n = 10$</th>
<th>Schizophrenia patients, $n = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of spatial overlap</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed threshold of $p = 0.001^*$ ($t = 3.15$)</td>
<td>46.99 (11.13) [23.2–60.98]</td>
<td>24.11 (18.28) [0.51–57.07]</td>
</tr>
<tr>
<td>Fixed threshold of $p = 0.01^*$ ($t = 2.35$)</td>
<td>45.25 (10.07) [31.07–59.08]</td>
<td>26.35 (16.05) [2.89–51.82]</td>
</tr>
<tr>
<td>Fixed threshold of $p = 0.05^*$ ($t = 1.66$)</td>
<td>43.76 (10.58) [26.97–58.91]</td>
<td>27.60 (13.08) [5.00–43.96]</td>
</tr>
<tr>
<td>Fixed volume of 5000 voxels, %</td>
<td>47.00 (11.58) [21.74–62.92]</td>
<td>30.65 (15.81) [6.26–57.34]</td>
</tr>
<tr>
<td>Average $t$ value corresponding to 5000 voxels</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session 1</td>
<td>3.34 (0.85) [2.47–4.91]</td>
<td>2.26 (0.74) [0.90–3.31]</td>
</tr>
<tr>
<td>Session 2</td>
<td>3.23 (1.07) [2.05–5.09]</td>
<td>2.69 (0.49) [2.09–3.86]</td>
</tr>
</tbody>
</table>

SD = standard deviation.  
*Uncorrected for multiple comparisons.
Fig. 4: Global reproducibility in patients with schizophrenia and controls, and correlation between global reproducibility and task performance. (Left) The percentage of spatial overlap (%Ovlp) was significantly lower in patients with schizophrenia than in controls ($t_{16} = 2.64, p = 0.017$). (Right) Percentage of overlap was not correlated with comprehension scores in session 1 for controls ($r = 0.12, p = 0.73$) but was positively correlated with comprehension scores for patients with schizophrenia ($r = 0.82, p = 0.004$). Solid lines represent the regression lines (controls: percentage overlap = 0.64 × score + 41.93, $R^2 = 0.02$; schizophrenia patients: percentage overlap = 3.79 × score + 13.60, $R^2 = 0.68$); dotted lines represent the mean comprehension scores in session 1.

Fig. 5: Activation probability maps in patients with schizophrenia and controls for each session. For each voxel, the colour scale indicates the percentage of participants in whom this voxel was activated among the 5000 most activated voxels. The maps were superimposed onto an inflated reconstruction of the Montreal Neurological Institute template.
more variable, with the exception of the anterior middle temporal gyrus. In patients, the areas with greatest reproducibility (i.e., areas exhibiting the lowest RSD) were located bilaterally in the middle part of the superior temporal sulcus.

Patients showed significantly less reproducibility than controls in 3 left hemispheric regions: the posterior part of the middle temporal gyrus up to the angular gyrus, the medial part of the superior frontal gyrus, and the pars triangularis of the inferior frontal gyrus (Table 4, Fig. 6B). In contrast, we did not detect a single area in patients that showed higher reproducibility than in controls.

When we included the comprehension scores as a covariate, all differences between groups were removed, except at the cluster located in the posterior part of the left middle temporal gyrus. That cluster had significantly less reproducibility in patients than in controls; however, after accounting for the comprehension scores, the extent was reduced from 434 to 27 voxels (Table 4).

In patients, the RSD maps were significantly negatively correlated with the comprehension scores in 3 left hemispheric clusters: 2 were located in the posterior part of the middle temporal gyrus and the angular gyrus, and 1 was located in the medial superior frontal gyrus (Table 4, Fig. 6C). These clusters overlapped with 2 of the areas that showed lower reproducibility in patients compared with controls in the between-group comparison analysis.

Discussion

The main finding of our study was that performance during

---

**Fig. 6:** Local reproducibility in patients with schizophrenia and controls, and correlation between local reproducibility and task performance. (A) Average (within group) relative standard deviation (RSD) maps show local reproducibility of activation. Low RSD values correspond to high reproducibility. (B) Cortical areas that exhibited higher activation reproducibility in controls than in patients with schizophrenia: (top) SPM t map and (bottom) selected axial slices of the 2-sample t test that compared RSD maps of controls and patients were superimposed onto the Montreal Neurological Institute template. (C) Cortical areas that showed significant negative correlations in patients between RSD values and task performance score (session 1). The SPM99 t map of the negative correlation between the local RSD and the performance score was converted into an r map (Pearson coefficient). The threshold of p = 0.001 corresponds to an r value of 0.84.
Maiza et al.

A story comprehension task was the only relevant factor that influenced the reproducibility of fMRI activation in patients with schizophrenia compared with healthy participants. Indeed, without taking into account the effect of task performance, patients with schizophrenia exhibited lower global and local reproducibility compared with controls. Three left cortical areas showed lower reproducibility in patients compared with controls: the posterior part of the middle temporal gyrus up to the angular gyrus, the inferior frontal gyrus and the medial prefrontal cortex. Both percentage of spatial overlap and RSD were strongly correlated with task performance in patients. Accordingly, the inclusion of task performance as a covariate removed the intergroup differences in spatial overlap at the global level and in RSD at the local level, except in a small cluster located at the posterior part of middle temporal gyrus.

In this study, we controlled the various factors that might affect the reproducibility of fMRI results. Preprocessing of functional data was specifically designed to reduce signal misregistration by applying the same normalization parameters onto coregistered fMRI data from both sessions. Hence, misregistration between the 2 sessions was not a factor that influenced cross-section reproducibility. Furthermore, we minimized a potential learning effect due to the repetition of the same paradigm by using a long intersession interval of 21 months. Indeed, the differences in intersession task performances were neither significant nor correlated with the percentage of spatial overlap in controls and patients. Finally, the percentage of spatial overlap and RSD computations were not affected by dependence on a statistical threshold.

Our results in healthy participants are in line with previous studies that used a similar functional paradigm. With both types of overlap (fixed volume and fixed threshold), we found a mean percentage of spatial overlap of 47%, which was within the range of the percentages (35%–50%) reported in 2 studies contrasting speech comprehension with reversed speech. However, such comparisons should be considered with caution given the differences in methods and tasks. Indeed, to date, there is no consensus about the most appropriate method to study reproducibility or the cut-off points that should be used to make a qualitative judgment about the degree of reproducibility in fMRI.

At the local level, controls exhibited higher reproducibility in the left cortical regions compared with the right regions. This is consistent with the greater involvement of the left hemisphere in language processing in healthy right-handed individuals. Highly activated regions are typically more reliable than less activated regions.

Our main hypothesis was that deficient task performance, reflecting cognitive impairments associated with the illness, could explain the lower reproducibility of activation in patients with schizophrenia compared with controls. That is, the reduced reproducibility of fMRI activation in patients would be correlated to a deficiency in task performance and not a systematic phenomenon due to illness. Our results are consistent with this hypothesis. Indeed, patients had lower global reproducibility (percentage of spatial overlap of 30.6%) than controls (47%), but the local analysis with RSD maps showed that this lower reproducibility did not concern the whole network recruited by the story comprehension task. It was restricted to areas located beyond the language-related perisylvian cortex, in areas that underpinned higher-order processes involved in narrative comprehension.

The left inferior frontal gyrus and posterior middle temporal gyrus are thought to be involved in the resolution of textual ambiguities and the construction of a representation of the global meaning of text. Furthermore, medial prefrontal cortex involvement in the context of a narrative may reflect social cognition processes and especially “theory of mind,” as well as more general inferential processes involved when listening to logically connected sentences. These processes are known to be impaired in schizophrenia. In contrast, the anterior and middle superior temporal sulcus are involved in lexicosemantic processing, and activations in these areas exhibited similar reproducibility in patients and healthy participants. This strengthens the hypothesis that lower reproducibility in schizophrenia patients would preferentially concern integrative areas that sustain high-level cognitive processes. Consistently, in a

<table>
<thead>
<tr>
<th>Table 4: Areas of lower reproducibility in schizophrenia patients (n = 10) compared with controls (n = 10) with and without comprehension scores as a covariate and areas exhibiting a significant* positive correlation between reproducibility and comprehension scores in schizophrenia patients</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comparison; brain region</strong></td>
</tr>
<tr>
<td><strong>Areas of lower reproducibility† in schizophrenia patients v. controls</strong></td>
</tr>
<tr>
<td>Left middle temporal gyrus extending to the angular gyrus</td>
</tr>
<tr>
<td>Left medial superior frontal gyrus</td>
</tr>
<tr>
<td>Left pars triangularis of the inferior frontal gyrus</td>
</tr>
<tr>
<td><strong>Areas of lower reproducibility† in patients v. controls including comprehension scores as covariate</strong></td>
</tr>
<tr>
<td>Left middle temporal gyrus</td>
</tr>
<tr>
<td><strong>Areas of significant positive correlation between reproducibility† and comprehension scores in patients</strong></td>
</tr>
<tr>
<td>Left middle temporal gyrus extending to the angular gyrus</td>
</tr>
<tr>
<td>Left middle temporal gyrus extending to the angular gyrus</td>
</tr>
<tr>
<td>Left medial superior frontal gyrus</td>
</tr>
</tbody>
</table>

MNI = Montreal Neurological Institute.

* The local index of reproducibility was relative standard deviation (RDS); low RSD values correspond to high reproducibility.
less complex task that preferentially targeted regions involved in the early phases of language processing, patients showed re-
producibility similar to that observed in healthy participants.52
This is in line with results observed in patients with schizo-
phrenia compared with controls during a working memory
 task; a similar reproducibility was observed in regions associ-
ated with motor function, whereas a lower reproducibility was
observed in patients in regions associated with higher-level
cognitive functions.52
Comprehension scores give an indirect and global assess-
ment of performance during the story comprehension task.
These scores were highly correlated across sessions, and pa-

tients with schizophrenia had lower scores than controls.
These scores appeared to be the main factor correlated to low
reproducibility in patients. Indeed, comprehension scores in
session 1 accounted for 67% of the spatial overlap variance in
patients.
At the local level, the between-group differences in repro-
ducibility were no longer significant after the scores were used
as covariates. Locally, this performance effect only influenced
areas of greater RSD in patients (left medial prefrontal cortex,
left posterior superior temporal sulcus). When we included the
comprehension scores as a covariate, only a few voxels in the
posterior part of the middle temporal gyrus remained signifi-
cantly more variable in patients than in controls. This residual
difference between patients and controls could be explained
by the fact that comprehension scores are neither a compre-
hensive nor direct assessment of the participants’ behaviour
during the task. There is a recall component in the compre-
hesion scores because there is necessarily a delay between scan-
ning and the comprehension test. It is likely that a more pre-
cise measure would have explained more of the differences.
Indeed, the area that was still significantly less reproducible in
patients was very close to the peak of the effect of comprehen-
sion scores on reproducibility of activations. Nevertheless, we
cannot exclude that some other factors linked to the disease
and its evolution might have contributed to this residual dif-
erence in reproducibility. Our patients, however, were reason-
able stable, as assessed by the symptom scales.
All analyses dealing with the effect of comprehension
scores were also conducted with scores from session 2, and the
same pattern of results was found. Activation probability
maps showed that most patients with schizophrenia did not
recruit all of the cortical regions that healthy participants re-
cruited during the task. One can speculate that the cerebral
activity in these regions would be less correlated with the
functional paradigm that led to poor signal strength and, thus,
poor reproducibility.
In this study, we did not detect any factor, other than per-
formance level, that significantly influenced the reproducibil-
ity of activation. In particular, clinical variables, including
duration of illness, positive and negative PANSS scores and
dose of antipsychotic medication, were not correlated with the
percentage of spatial overlap in patients. All patients were
clinically stable, with no significant changes in symptoms or
medication doses, which were correlated strongly across ses-
sions. The patients included in the present study had been ill
for an average of 12 years; the effects of the illness course on
the brain may be less pronounced at this stage of the disease
than during the first few years.52
Limitations
Several limitations affected the present study. The interses-
sion interval was shorter in patients than in controls. How-
ever, this was not correlated with the percentage of spatial
overlap. The patients were heterogeneous with regard to
clinical subtype and antipsychotic medication. Because of the
small sample sizes, we could not investigate the influence of
these factors on reproducibility. However, medication dose
expressed in chlorpromazine equivalents was not correlated
with spatial overlap. Finally, our results only concerned
language-related tasks, and they should not be directly gen-
eralized to other cognitive domains.
Conclusion
Our findings showed that the lower reproducibility of fMRI
activation elicited by a language paradigm in patients with
schizophrenia compared with healthy participants concerned
only a part of the network of active areas (mostly the
“integrative” areas), and the lower reproducibility was
mainly explained by deficient task performance in patients.
Therefore, when designing a longitudinal fMRI study in pa-

tients with schizophrenia, task performance should be care-
fully controlled.
The fMRI paradigm should yield reliable activations and
be sensitive to illness-related changes. Ideally, there should
not be any ambiguity regarding the cause of the observed dif-
erence, change in task performance or behaviour or evolu-
tion of the illness. If a modification in the terms of task per-
formance occurs between 2 sessions, unless the effects of task
performance and illness can be disentangled statistically,
the results should be interpreted as neural correlates of im-
proved or diminished performance at the task, in a context of
illness. This would be the case when assessing the effects of a
therapeutic intervention on brain function. If no change in
behaviour is observed, and provided the paradigm (in this
context, a “probe”) yields reliable activations in stable pa-

tients, the neuroimaging observations can be interpreted as
reflecting the evolution of the illness. In the latter case, the
performance level of the patients should be high enough to
guarantee reproducibility. Finally, there is nothing intrinsic-
ally associated with the disease that would preclude longitudi-
nal fMRI studies in this population.

Acknowledgements: We thank G. Perchey, N. Delcroix and
F. Lamberton for their assistance. This work was funded by the French
Health Ministry in a Programme Hospitalier de Recherche Clinique and
by the French Research Ministry.
Competing interests: None declared.
Contributors: Drs. Dollfus and Tzourio-Mazoyer designed the study.
Drs. Maïza, Dollfus, Razafimandimbry and Tzourio-Mazoyer ac-
quired the data, which Drs. Maïza, Mazoyer and Hervé analyzed.
Drs. Maïza and Hervé wrote the article, which Drs. Maïza, Mazoyer,
Razafimandimbry, Dollfus and Tzourio-Mazoyer reviewed. All au-
thors approved the article that was submitted for publication.

J Psychiatry Neurosci 2010;35(6) 387
References


69. Corcoran R. Inductive reasoning and the understanding of intention in schizophrenia. Cogn Neuropsychiatry 2003;8:223-35.

