◆ Human Brain Mapping 29:461–477 (2008) ◆

Group Analysis and the Subject Factor in Functional Magnetic Resonance Imaging: Analysis of Fifty Right-Handed Healthy Subjects in a Semantic Language Task

Mohamed L. Seghier,[1,2] François Lazeyras,[1] Alan J. Pegna,[3,4,5] Jean-Marie Annoni,[4] and Asaid Khateb[3,4,5]*

[1]*Department of Radiology, Geneva University Hospitals, Geneva, Switzerland*
[2]*Wellcome Trust Centre for Neuroimaging, Institute of Neurology, UCL, London, United Kingdom*
[3]*Laboratory of Experimental Neuropsychology, Department of Neurology, Geneva University Hospitals, Geneva, Switzerland*
[4]*Department of Neurology, Neuropsychology Unit, Geneva University Hospitals, Geneva, Switzerland*
[5]*Geneva Neuroscience Center, University of Geneva, Geneva, Switzerland*

Abstract: Before considering a given fMRI paradigm as a valid clinical tool, one should first assess the reliability of functional responses across subjects by establishing a normative database and defining a reference activation map that identifies major brain regions involved in the task at hand. However, the definition of such a reference map can be hindered by inter-individual functional variability. In this study, we analysed functional data obtained from 50 healthy subjects during a semantic language task to assess the influence of the number of subjects on the reference map and to characterise inter-individual functional variability. We first compared different group analysis approaches and showed that the extent of the activated network depends not only on the choice of the analysis approach but also on the statistical threshold used and the number of subjects included. This analysis suggested that, while the RFX analysis is suitable to detect confidently true positive activations, the other group approaches are useful for exploratory investigations in small samples. The application of quantitative measures at the voxel and regional levels suggested that while ~15–20 subjects were sufficient to reveal reliable and robust left hemisphere activations, >30 subjects were necessary for revealing more variable and weak right hemisphere ones. Finally, to visualise inter-individual variability, we combined two similarity indices that assess the percentages of true positive and false negative voxels in individual activation patterns relative to the group map. We suggest that these measures can be used for the estimation of the degree of 'normality' of functional responses in brain-damaged patients, where this question is often raised, and recommend the use of different quantifications to appreciate accurately the inter-individual functional variability that can be incorporated in group maps. *Hum Brain Mapp 29:461–477, 2008.*

Contract grant sponsor: Swiss National Science Foundation; Contract grant numbers: 3151A0-102271/1 and 3200BO-100717/1; Contract grant sponsor: Center for Biomedical Imaging (CIBM) of Geneva and Lausanne.

*Correspondence to: Dr. Asaid Khateb, Laboratory of Experimental Neuropsychology, Department of Neurology, Geneva University Hospitals, 24 Rue Micheli-du-Crest, 1211 Geneva 14, Switzerland. E-mail: asaid.khateb @hcuge.ch

Received for publication 10 November 2006; Revision 9 February 2007; Accepted 23 March 2007

DOI: 10.1002/hbm.20410

Published online 30 May 2007 in Wiley InterScience (www.interscience.wiley.com).

Key words: language areas; semantic categorisation; left hemisphere lateralisation; functional magnetic resonance imaging; inter-individual variability; group analysis; fixed and random effects; gold standard map; sample size; clinical fMRI

INTRODUCTION

The use of functional magnetic resonance imaging (fMRI) to assess brain activation during different sensory, cognitive, and affective tasks is in constant growth. In addition to its increasing popularity within the cognitive and basic neuroscience communities, it is also becoming increasingly popular for clinical purposes since the first applications in epileptic [Jackson et al., 1994] and schizophrenic patients [Wenz et al., 1994], which have led to the introduction of the concept of 'clinical fMRI' [Levin et al., 1995; Thulborn et al., 1996]. Clinical fMRI has thus been applied to patients with various pathologies to assess the functional reorganisation and plasticity after brain damage [for review, see Detre and Floyd, 2001; Matthews et al., 1999; Powell et al., 2004; Price and Crinion, 2005]. More specifically, it has become routinely employed in the assessment of critical language areas prior to brain surgery [Balsamo and Gaillard, 2002; Khateb et al., 2004; Seghier et al., 2001; Woermann et al., 2003].

To consider any given fMRI paradigm as a valid clinical tool, one should first assess the reliability, reproducibility, and robustness of functional responses obtained across subjects [Herholz et al., 1996]. Indeed, different methodological and ecological factors are known to contribute to the variability of fMRI responses within identical tasks [for more details, see Desmond and Chen, 2002; Hennig et al., 2003; Raz et al., 2005; Thulborn and Davis, 2001]. One approach for assessing the reliability of functional responses is achieved through the creation of a functional normative database for a given population and paradigm. The normative database consists of a representative activation map that defines the main brain regions that are engaged when healthy control subjects perform a given task. When considering patients with brain lesions, such reference maps basically allow typical or atypical activations in brain-damaged subjects to be identified, so that the brain regions involved in reorganisation and recovery processes can be characterised [e.g., Cao et al., 1999; Fernandez et al., 2004; Hugdahl et al., 2004; Pizzamiglio et al., 2001; Thulborn et al., 1999; Xu et al., 2004].

For a specific focus of activation to be characterised as a manifestation of brain reorganisation because of a given lesion, a reliable reference map for neurologically healthy subjects is necessary. However, the assessment of a reliable good reference (gold standard) map is a complex issue that can be hindered by inter-individual functional variability [e.g., Miller and Van Horn, 2007]. For instance, the gold standard map, at a given statistical threshold, can depend on the functional sessions and subjects included. Assuming that the inter-session variability is of limited magnitude [see Smith et al., 2005], the 'subject factor' can thus be considered as the major cause of the commonly observed functional variability [Kherif et al., 2003; Rimol et al., 2006]. The influence of this factor, which may among other things reflect the different cognitive strategies used by the subjects to perform the same task [Edwards et al., 2005; Nadeau et al., 1998; Miller et al., 2002; Miller and Van Horn, 2007; Noppeney et al., 2006; Tsukiura et al., 2005], has recently been investigated using different methods of analysis [Liou et al., 2003; Maldjian et al., 2002; Otzenberger et al., 2005; Seghier et al., 2004; Tzourio-Mazoyer et al., 2004]. In particular, different studies have attempted to characterise the effect of the number of subjects (*N*-sub) necessary in a functional study to make inferences about the typical activation in a given task and population [Desmond and Glover, 2002; Friston et al., 1999b; Grabowski et al., 1996; Kiehl et al., 2005; Murphy and Garavan, 2004; Strother et al., 1997].

In this study, our aim was to determine the effect of *N*-sub on the definition of a representative group activation map (or normative database) and to quantify inter-individual functional variability that can be incorporated in the group analysis. For this purpose, we used a semantic language task that has previously been shown to provide reliable activations in classical language areas and which has a high left hemispheric (LH) lateralizing power [Seghier et al., 2004]. Functional data from 50 neurologically healthy right-handed subjects were first analysed using different group analysis approaches. As group analysis generally neglects the subject factor and extracts mainly the collective effects of neural activations that are spatially coincident across subjects, we then employed additional

Abbreviations

%FNg	percentage of false negatives relative to the gold standard map
FFX	fixed-effect analysis
LH	left hemisphere
PO	percentage of overlap
RFX	random-effect analysis
RH	right hemisphere
ROC	receiver operating characteristic
SPM	statistical parametric mapping
%TPi	percentage of true positives at the individual level

approaches to investigate the inter-individual variability of the functional responses. Specifically, the influence of the *N*-sub included and the choice of the statistical threshold on the definition of a representative group map were assessed using various computations with an increasing number of subjects, including the power of activation map and receiver operating characteristic (ROC) curves. These investigations allowed true positive and false negative rates to be estimated for a given sample size. We then proposed a graphical tool for appreciating the inter-individual variations that are incorporated in the group analysis. We propose that such analyses might be of particular interest during the process of database formation, in particular in the clinical contexts for the estimation of the degree of 'normality' of functional responses in neurological patients in whom this question is often raised.

MATERIALS AND METHODS

Subjects

Fifty-six healthy, right-handed subjects (35 men, 21 women, 27 ± 4 years) from Geneva hospital and medical school gave their informed consent to participate in this study. Subjects were native French speakers and had normal or corrected-to-normal vision. Some of the subjects analysed here have been included in our previous reports [Seghier et al., 2001, 2004]. Because of acquisition problems in four subjects and to the fact that brain anomalies were detected in two other subjects, 50 of 56 subjects were included in this analysis.

Paradigm and Stimuli

A block paradigm that alternated between 'control' and 'activation' sequences was applied. The stimuli, concrete imaginable high frequency nouns, were selected from a French table for word frequency [Content et al., 1990] and presented to the subjects via a video projector, a front-projection screen and a system of mirrors fastened to a head coil. The activation condition, referred to as semantic categorisation task, was composed of a set of 60 pairs of words that were either categorically related (i.e., two words were exemplars of the same semantic category, $n = 40$ pairs) or unrelated (the two words belonged to two different semantic categories, $n = 20$ pairs). The subjects performed a go/no-go task and responded whenever the two words were exemplars of the same semantic category. The control condition that alternated with the activation condition consisted of a perceptual categorisation task in which pairs of either visually similar ($n = 40$ pairs) or different ($n = 20$ pairs) meaningless Greek letter-strings were presented to the subjects. To maintain the rate of motor responses constant as in activation condition, the subjects had to respond when two simultaneously presented strings were visually identical. In both activation and control conditions, stimulus pairs were presented on the screen for 600 ms at 0.5 Hz and in blocks of 24 s, repeated five times per condition. Accordingly, alternating blocks of activation-control conditions yielded a total task duration of 4 min. In about one half of the subjects, the responses were given by pushing an air-mediated button and the experimenter recorded the performance on the computer and in the other half, subjects' responses were directly linked to the computer using a mouse. To ensure that the task was correctly understood, all subjects were provided with detailed instructions before entering the scanner and underwent a short training session.

MRI Acquisition

Experiments were performed on a 1.5-T system (Philips Medical Systems, Best, The Netherlands). Acquired multislice volume was positioned on sagittal scout images. Before the functional MR scans, an anatomical scan (a GRE T1-weighted sequence, TR/TE/Flip = 162 ms/4.47 ms/80°, FOV = 250 mm, matrix = 256 × 256, slice-thickness = 5 mm) was performed to acquire the same volume as in the functional session. Anatomical reference images, acquired after the functional scans, consisted of a 3-D GRE T1-weighted sequence (TR/TE = 15 ms/5 ms, FOV = 250 mm, matrix = 256 × 256, slice-thickness = 1.25 mm). Functional imaging consisted of an EPI GRE sequence (TR/TE/Flip = 2 s/40 ms/80°, FOV = 250 mm, matrix = 128 × 128, 17–19 contiguous 5 mm axial slices). The explored volume was measured 12 times during each condition of the paradigm. Functional scanning was always preceded by 8 s of dummy scans to insure tissue steady-state magnetisation.

Data Analysis

Data processing and statistical analyses were carried out with Statistical Parametric Mapping SPM2 software package (Wellcome Trust Center for Neuroimaging, London, UK, http://www.fil.ion.ucl.ac.uk/spm/). All functional volumes were spatially realigned, normalised to the MNI space, and smoothed with an isotropic 6-mm FWHM Gaussian kernel, with resulting voxels size of 3 × 3 × 3 mm^3. Time-series from each voxel were high-pass filtered (1/128-Hz cutoff) to remove low-frequency noise and signal drift. For each subject, the pre-processed functional volumes were then submitted to fixed-effects analyses (FFX, i.e., first level analysis) using the general linear model applied at each voxel across the whole brain. Each condition of interest (activation and control conditions) was modelled by boxcar waveform convolved with a canonical haemodynamic response function (with no dispersion or temporal derivatives) and included in a multiple regression analysis with six covariates of no interest representing the head motion parameters [Friston et al., 1996; Johnstone et al., 2006]. Parameter estimates (i.e., β-images) were assessed with least square regression analysis, and the contrast images (i.e., weighted beta images or the con-images) were computed for the main effect of semantic

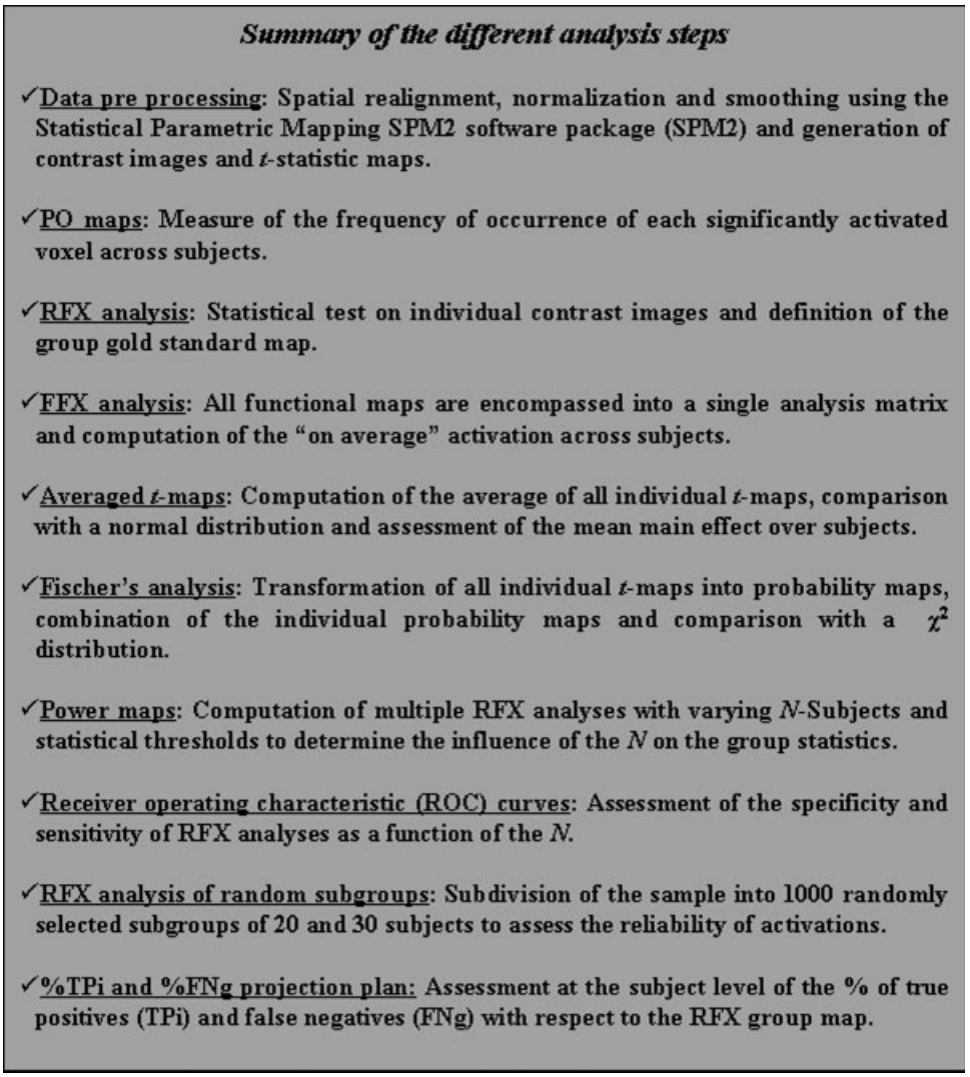

Figure 1.
Highlights of the different analyses carried out at the group and subject levels.

categorisation relative to perceptual matching of Greek letter strings. Accordingly, 50 first level analyses (i.e., 50 individual FFX analyses) were performed, which yielded to the quantification of 50 contrast images. Statistical parametric maps of the t statistics (SPM[t]) were then generated for the contrast 'activation–control'. In the individual functional maps, only clusters of >7 statistically significant contiguous voxels (i.e., 0.2 ml cortical volume) at $P < 0.001$ (uncorrected) were considered [Forman et al., 1995]. These clusters were then projected on a glass-brain or on the T1-weighted SPM2 template brain in the neurological convention. As summarised in Figure 1, different computations were afterwards performed both at the group and the subject levels.

First, to visualise the inter-individual variability at the voxel level, the percentage of overlap (PO) of each activated voxel was assessed across subjects using programs developed locally with Matlab (The MathWorks, Natick, MA). As proposed previously [Machielsen et al., 2000; Specht et al., 2003], the PO measures the frequency of occurrence of each activated voxel over all individual t-maps. To take into account anatomical variability at the voxel scale or the possible contribution of errors (e.g., those due to the data pre-processing normalisation) to the functional variability [Juch et al., 2005], we considered for the PO maps the voxel itself with its 18 nearest neighbours. Practically, this was achieved by applying a dilation operation on each individual binary image (i.e., a t-map thresholded at $P < 0.001$) using a structuring element at 18-connected neighbourhood, and then summing these dilated binary images to form the PO maps. Of note is the fact that dilation operation would decrease false negatives

but will in parallel increase false positives. However, considering the large number of subjects analysed here and the fact that false positives are less reliable across subjects, we assume that the later effect will be of limited impact in our context. The PO maps were shown at a threshold of 10%, thus, each visible voxel in these maps has been 'significantly and concurrently' activated in at least five of the subjects. Basically, the main advantage of assessing PO maps is that they tell us how frequently each visible voxel in these maps was activated in individual subjects. In addition, for illustrating the influence of the statistical threshold on the spatial variability between activated voxels, the PO maps were then generated using different individual statistical thresholds.

Afterwards, to examine the sensitivity of the representative group map to the individual differences, we performed four types of group analyses that use different statistical approaches:

a. A fixed-effect analysis (FFX) was conducted by integrating the pre-processed functional volumes of all 50 subjects into a unique SPM2 design matrix. This FFX design consisted of >400 colons (i.e., regressors) and >6,000 lines (i.e., data volumes). The FFX allows the regions activated 'on average' across the 50 subjects to be quantified by using intra-subject variability [Buchel et al., 1997]. Specifically, FFX analysis assumes that each subject makes a fixed (i.e., the same) contribution to the group main effect of interest, ignoring random variations from subject to subject [Woods, 1996; Penny et al., 2003]. Therefore, this analysis was performed here for exploratory purposes to detect regions that have been implicated in some, but not necessarily in all, subjects. Note that FFX analysis has frequently been employed in studies that have involved a limited number of participants (e.g., <10 subjects).
b. A random-effect analysis (RFX) was performed with SPM2 by using the inference images (i.e., the 50 contrasts images) resulting from the first-level analysis of individual subjects. The RFX second-level analysis (i.e., t test on contrasts images) detects regions that are consistently activated (i.e., less variable across subjects) by incorporating both intra- and inter-subjects variability [see Holmes and Friston, 1998; Penny et al., 2003]. Generally, RFX allows inferences (i.e., implicated regions) to be generalised to the population from which the subjects were drawn. This RFX analysis on all 50 subjects was thenceforth considered as our group's 'gold-standard' map [see Murphy and Garavan, 2004] from which the major activated regions (at $P < 0.001$ uncorrected) were defined. Activated foci in this group map were identified and labelled according to their Talairach coordinates [Talairach and Tournoux, 1988], obtained from the MNI space using the mni2tal tool (http://www.mrc-cbu.cam.ac.uk/Imaging/Common/mnispace.shtml), and to their anatomical landmarks after projection on a normalised T1 volume.
c. An averaged t-map analysis using all the individual statistical t-maps [Bosch, 2000; Lazar et al., 2002] to show the mean main effect over the 50 subjects. As shown previously, the resulting average is then compared to a standard normal distribution [Bosch, 2000], and the threshold is therefore calculated for $P< 0.001$. This method is computationally very simple to implement and therefore worth to comparing with more demanding analyses such as the FFX. However, despite its high sensitivity (such as the FFX approach), this approach ignores random variations from subject to subject.
d. A group analysis using Fisher's method. This method is based on the combination of individual probability maps that can be calculated from the first level analysis of all subjects. Briefly, all individual t-maps are transformed (i.e., expressed) into individual probability maps (p_i, $i = 1 \ldots N$-sub). Then the Fischer's approach combines these probability maps according to the formula: $-2\sum_{i=1}^{Nsub}\log(p_i)$, and this expression is compared to a χ^2 distribution with $2 \times N$-sub degrees of freedom [for more details, see Lazar et al., 2002; McNamee and Lazar, 2004]. The resulting map is finally thresholded at $P < 0.001$ (threshold χ^2 (df = 100) = 149.45). As this method is generally accepted to be less conservative at the group level than RFX analysis, its main advantage is its sensitivity to individual differences. However, this method might be problematic in the presence of outliers.

To characterise the effect of the N-sub on the activated pattern in the representative group map, several RFX analyses are carried out by increasing gradually the N-sub from 6 to 50. We generated 50 different randomisations of our subjects' sample (yielding thus a total of >1000 different RFX analyses). The number of activated voxels and regions at $P < 0.001$ were computed for each RFX analysis with a given N-sub. Then, we quantified the overlap between the obtained map from each RFX analysis with a given N-sub and the gold standard map (generated with all 50 subjects). This measure is equivalent to the power of the activation map, which is defined as the percentage of significantly activated voxels in the map that were also significantly activated in the gold standard map [Desmond and Glover, 2002; Murphy and Garavan, 2004]. This procedure was performed at several statistical thresholds. Although the power of the activation map takes into account true positives only, the measure proposed here also takes into account false positives by quantifying the percentage of voxels in the map that were outside the gold standard map.

In the same way, we generated ROC curves of RFX analyses with different N-sub values. ROC curves represent the dependency of the true positive rate (sensitivity) with the false positive rate (1-specificity) for different thresholds [Skudlarski et al., 1999]. During the computation of sensitivity and specificity at a given N-sub, we assume that true

positives are voxels that are observed in the gold standard map. Compared to the power of activation map, the usefulness of the ROC approach is to generate threshold-free curves about the sensitivity and the specificity of RFX for a given sample size.

Furthermore, to assess the reliability of the activated regions during this language task, several RFX analyses were subsequently performed using different sub-samples of our 50 subjects group. For this purpose, we first randomly generated 1,000 different subgroups of 20 subjects and 1,000 other subgroups of 30 subjects, from the original whole sample. For each subgroup size (i.e., 20 or 30), 1,000 RFX analyses were then performed on the different subgroups (yielding a total of 2,000 RFX analyses). Afterwards, PO maps were generated from these RFX analyses (individually thresholded at $P < 0.001$ uncorrected). These PO maps were thresholded at 90% to identify voxels that were visible in at least 900 RFX analyses (i.e., reliable in >90% of the randomised subgroups). The rational for performing such extensive analysis was to assess the detectability of LH or RH activations; when using RFX analysis at $P < 0.001$, with a limited number of subjects randomly selected form a given population.

Finally, different measures were also computed at the individual level with the aim to characterise the pattern of activation in each subject and to assess the inter-individual functional variability (see summary in Fig. 1) that could be incorporated during normative database formation.

The first index represented the percentage of true positives in the individual subject (%TPi) using the following formula:

$$\%\mathrm{TPi} = \left(N_{\mathrm{subject}} \bigcap N_{\mathrm{gold\ standard}}\right)/N_{\mathrm{subject}}$$

The %TPi index (varying between 0 and 1), similar to the Simpson similarity coefficient [Cheetham and Hazel, 1969], quantifies the similarity between the spatial distributions of the activated pattern in each subject with the group gold-standard map. By assuming that all voxels activated in the gold standard map are true positives, the %TPi index will therefore represent the percentage of true positives detected in a given subject. Likewise, 1-%TPi values will approximately indicate the percentage of false positives (i.e., voxels that are activated in a given subject but not observed in the gold standard map).

The second index assessed the percentage of false negatives when compared with the group map (%FNg) in each subject, following the formula:

$$\%\mathrm{FNg} = 1 - \frac{\left(N_{\mathrm{subject}} \bigcap N_{\mathrm{gold\ standard}}\right)}{N_{\mathrm{gold\ standard}}}$$

The %FNg index (varying between 0 and 1) allows the amount of false negatives (true positives present in the gold standard but not observed in a given subject) to be quantified. This index is similar to the Braun-Blanquet coefficient of difference [Cheetham and Hazel, 1969]. Note that both indices depend on the definition of the gold standard map. It is worth reminding here that these two latter indices were computed using the gold standard map at $P < 0.001$. Finally, both indices were used to define a projection plan that represents the position of each subject with respect to other subjects. The rational for computing this projection plan is to appreciate qualitatively the inter-individual variability by identifying subjects with (i) high %FNg and low %TPi: few activations but are different from the activations observed in the gold standard; (ii) low %FNg and high %TPi: subjects whose activation pattern is similar to that observed in the gold standard; (iii) low %FNg and low %TPi: large activated volume but having low overlap with the gold standard; (iv) high %FNg and high %TPi: small activated volume but with good overlap with the gold standard.

RESULTS AND DISCUSSION

The individual functional maps of the 50 subjects as revealed by FFX analyses are shown in Figure 2. With the same statistical threshold ($P < 0.001$ uncorrected), the regions activated varied across subjects in size, localisation, and level of activation. Such inter-individual variability is commonly observed in fMRI studies with language paradigms. The regions that are most consistent across subjects, as identified by the RFX analysis on all 50 subjects (referred to hereafter as the gold standard map), are concordant with those observed in previous studies using semantic tasks [Billingsley et al., 2001; Pugh et al., 1996; Seghier et al., 2004]. Table I summarises these areas and details their size and coordinates as determined from the gold standard map. In the LH, these included predominantly the inferior frontal gyrus, the posterior prefrontal cortex, the mid-dorsolateral prefrontal cortex, the precentral gyrus, the superior/middle temporal gyrus, the supplementary motor area, and the inferior parietal gyrus. As in our previous report [Seghier et al., 2004], weak but significant activations were also observed in the right hemisphere (RH), particularly in the inferior frontal gyrus, the superior frontal gyrus, and the superior/middle temporal gyrus. In addition, the vast majority of the subjects showed LH dominance with a mean laterality index (LI) of 0.72 ± 0.25. By using a cutoff at 0.2 [Springer et al., 1999] or a cutoff equal to the Mean LI-2 × SD [corresponding here to 0.22; see Seghier et al., 2004], we observed that only 4 of 50 subjects showed a bilateral language representation (see subject nos. 7, 9, 13, and 31 in Fig. 2). This proportion corresponds to that already observed in other studies in healthy right-handed subjects [Springer et al., 1999], showing that ~92% of the subjects have an LH language lateralisation. It also confirms our previous conclusion about the high LH lateralizing power of this semantic task [Seghier et al., 2004].

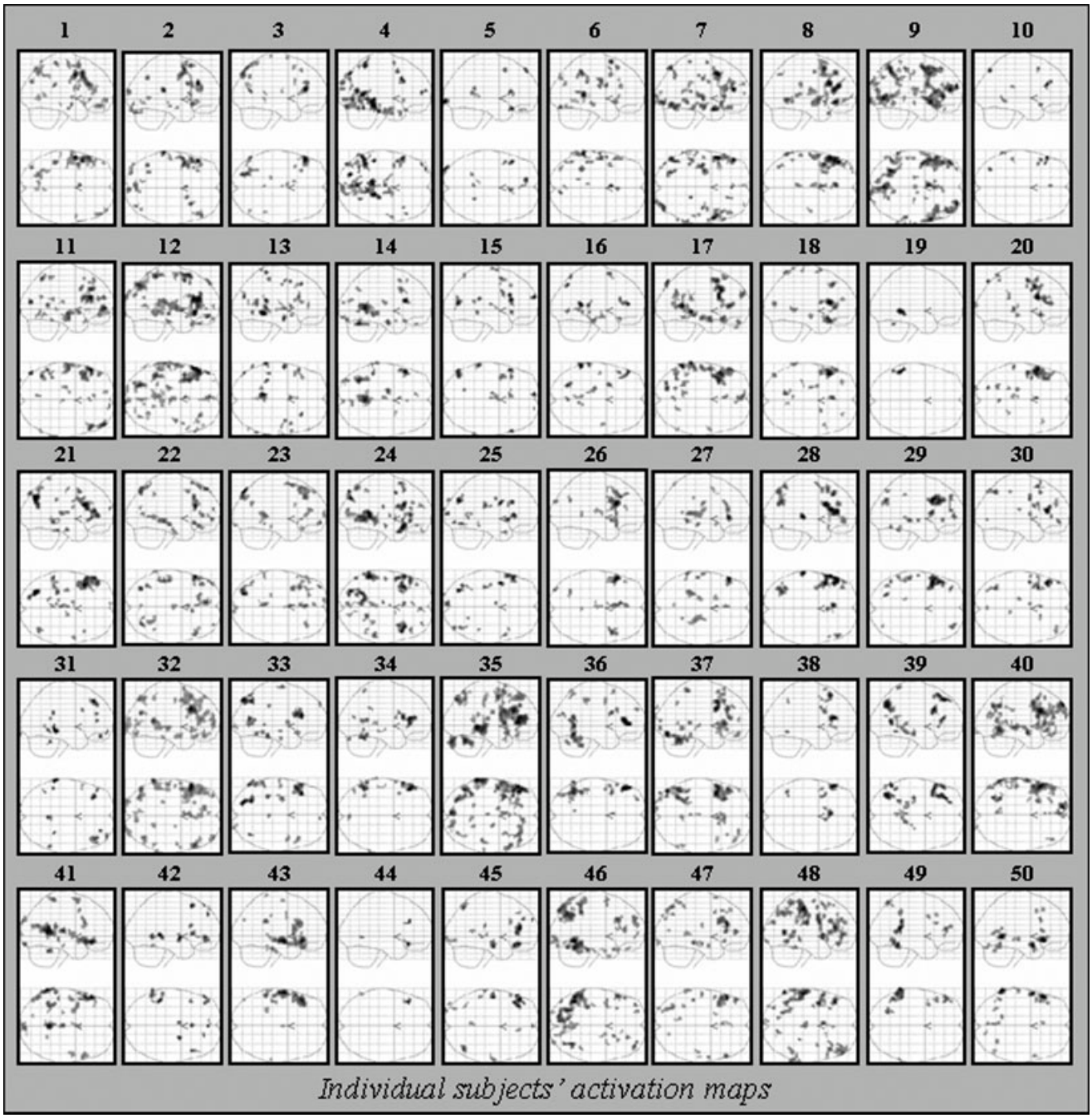

Figure 2.

Functional maps of all individual subjects (1–50) at the same statistical threshold ($P < 0.001$ uncorrected). For each subject, sagittal and axial views of the glass brain are shown in the neurological convention. These maps are generated from the 50 individual FFX analyses.

PO Maps and Statistical Thresholds

The statistical threshold is one of the many different factors that influence the extent and definition of the network involved in a given task [see illustrations in Desmond and Chen, 2002; Loring et al., 2002]. Not surprisingly, as illustrated in Figure 3A, the activated volume varied across subjects at different statistical thresholds. Regarding these different individual activated volumes, it is interesting to find and visualise the voxels that are consistent across sub-

TABLE I. The *x*, y and z Talairach coordinates, the z-score maximum values and the clusters' size of the main activated regions in the gold standard map revealed by RFX group analysis on the 50 subjects (at $P < 0.001$ uncorrected)

Activated regions	*x*, *y*, *z*-Coordinates	*z*-Score max	Cluster size
Left hemisphere			
Inferior frontal gyrus (Broca area)	−50, +14, −13	6.4	320
Mid-dorsolateral prefrontal cortex	−48, +27, +15	6.6	321
Posterior prefrontal cortex	−45, +15, +22	6.5	322
Superior/middle temporal gyrus	−50, −38, +03	6.5	296
Inferior parietal lobule	−30, −65, +42	4.3	59
Precentral gyrus	−48, +05, +47	5.7	151
Supplementary motor area (SMA)	−03, +20, +46	7.3	276
Right hemisphere			
Inferior frontal gyrus	+42, +20, −09	4.8	120
Superior frontal gyrus	+39, +54, +22	4.7	15
Superior/middle temporal gyrus	+50, −32, +02	4.0	48

Major and robust activation are found in the left hemisphere while weak and small clusters are found in the right hemisphere.

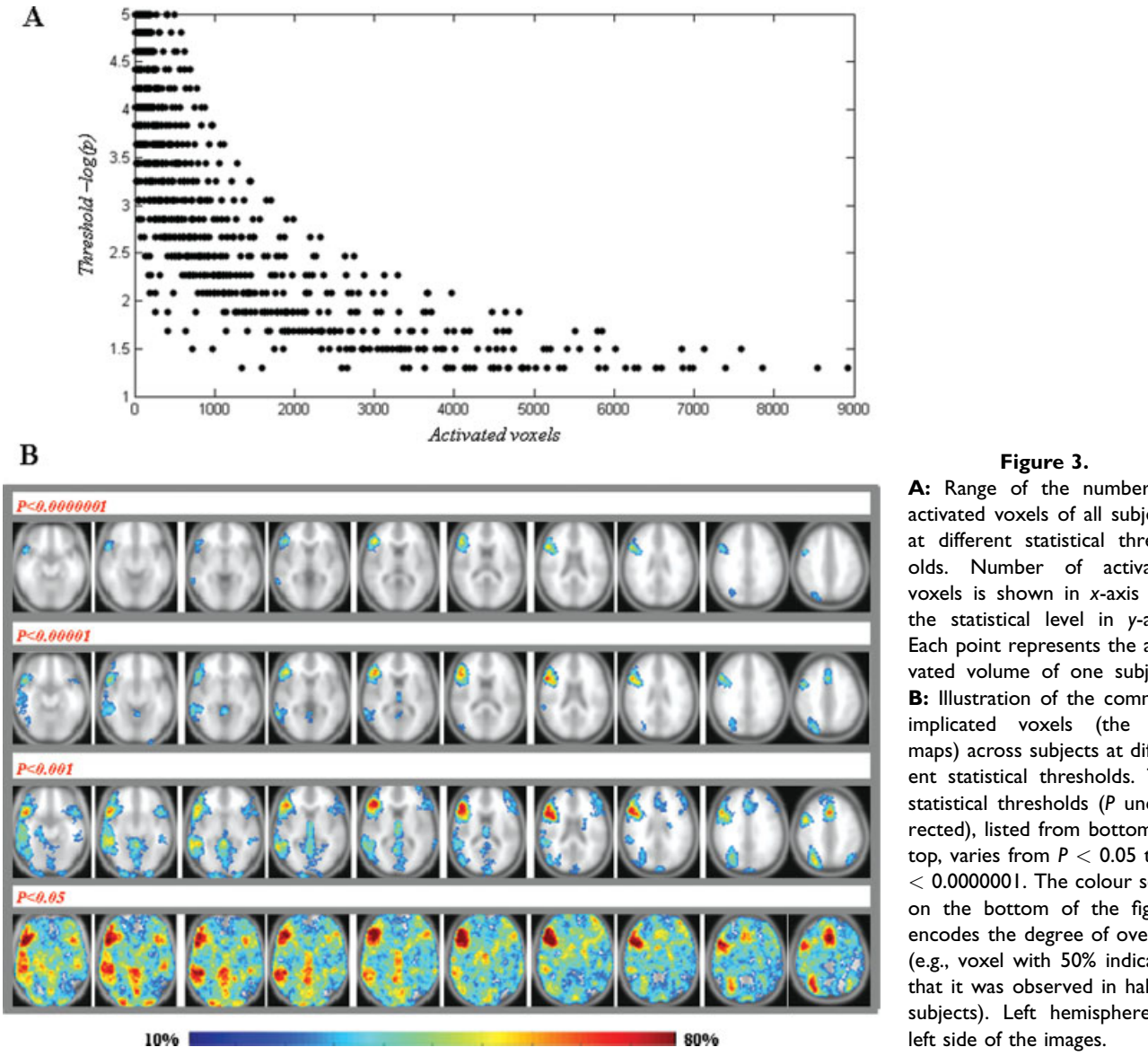

Figure 3.

A: Range of the number of activated voxels of all subjects at different statistical thresholds. Number of activated voxels is shown in *x*-axis and the statistical level in *y*-axis. Each point represents the activated volume of one subject. **B:** Illustration of the common implicated voxels (the PO maps) across subjects at different statistical thresholds. The statistical thresholds (*P* uncorrected), listed from bottom to top, varies from $P < 0.05$ to $P < 0.0000001$. The colour scale on the bottom of the figure encodes the degree of overlap (e.g., voxel with 50% indicates that it was observed in half of subjects). Left hemisphere = left side of the images.

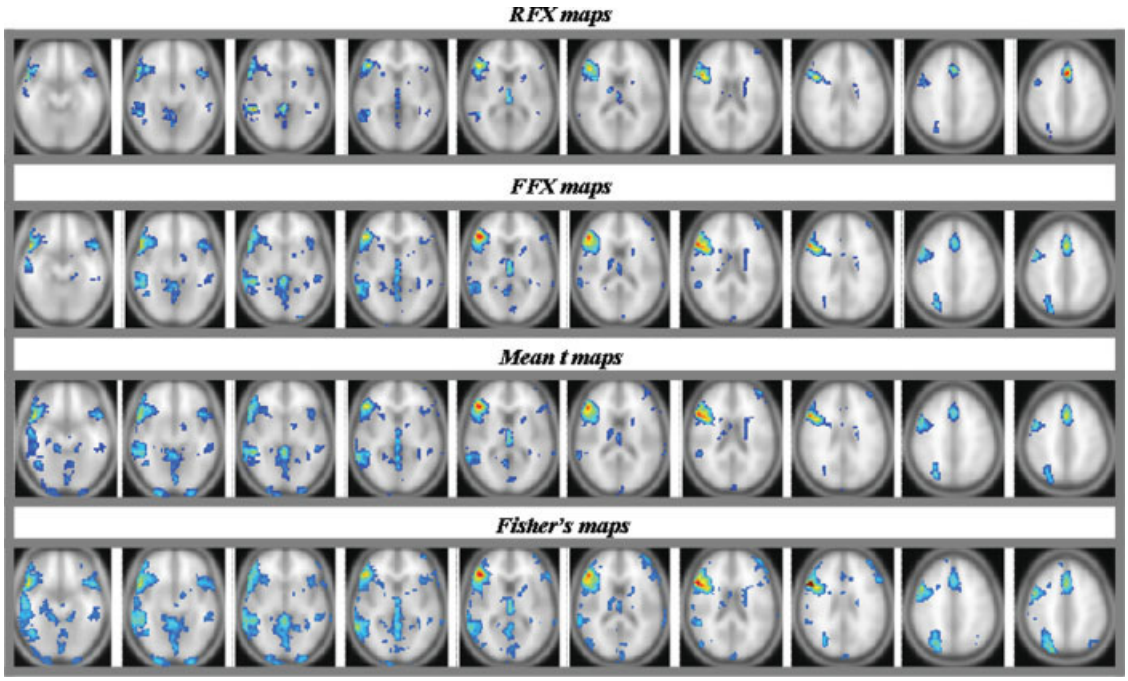

Figure 4.

Group activation maps using different analysis approaches. For all maps a threshold of $P < 0.001$ uncorrected was used. Significance from low to high is coded with blue–green–yellow–red colours' scale.

jects at each statistical threshold. This can be achieved by the use of the PO maps (i.e., the percentage of overlap maps). Figure 3B displays the PO maps generated with four different individual P values. This illustration indicates that the most reliable activations are found in the LH, as previously observed [Seghier et al., 2004]. In accordance with previous observations [Lehéricy et al., 2000; Seghier et al., 2004], we also found that activations are more consistent over subjects in frontal than in temporo-parietal regions. At the individual uncorrected threshold of $P < 0.001$, the frequency of occurrence at the voxel level in the PO maps reaches 80% in the frontal regions, particularly in the inferior frontal gyrus and the prefrontal cortex. The use of PO maps with different statistical thresholds is of considerable interest for exploring the cerebral network, the most dominantly engaged in the task and the robustness of the regional activations, as they indicate how often each voxel has been declared 'activated' across subjects. This provides grounds for proper interpretation of activation patterns observed in brain-damaged patients, for instance, to identify regions that have been activated in all control subjects (i.e., regions with a PO at 100%) but not activated in a patient, or regions that have been activated in the patient but not observed in any control subject performing the same task (i.e., regions with a PO at 0%). Finally, in view of the dependence of certain regions (e.g., temporo-parietal areas) on the threshold used in PO maps (Fig. 3B), it would thus be interesting, when characterising the activation pattern in single case investigations, to lower the statistical threshold (e.g., to $P < 0.05$) before concluding with confidence whether a given brain area has or has not been activated in the patient (i.e., false negatives are usually difficult to appreciate in single case studies).

Comparing the Results of Different Approaches of Group Analysis

Figure 4 illustrates the group activation maps using different analysis approaches. A customary way to define the common networks across subjects is to perform RFX analysis [McNamee and Lazar, 2004; Penny et al., 2003]. This approach is known to incorporate both intra- and inter-individual variability when computing activations across subjects. In line with previous findings [Friston et al., 1999a; Lazar et al., 2002; McNamee and Lazar, 2004], this illustration shows that RFX analysis appears to be the most restrictive compared to other analysis approaches. In fact, the number of observations (i.e., activated voxels at $P < 0.001$ uncorrected) varied as a function of the method used, increasing from 2,346 voxels with RFX analysis to 4,180 voxels with FFX analysis to 5,921 voxels with averaged t-maps analysis, and finally to 8,642 with Fisher's method. Thus, this comparison demonstrates that Fisher's method is the less conservative and consequently is more

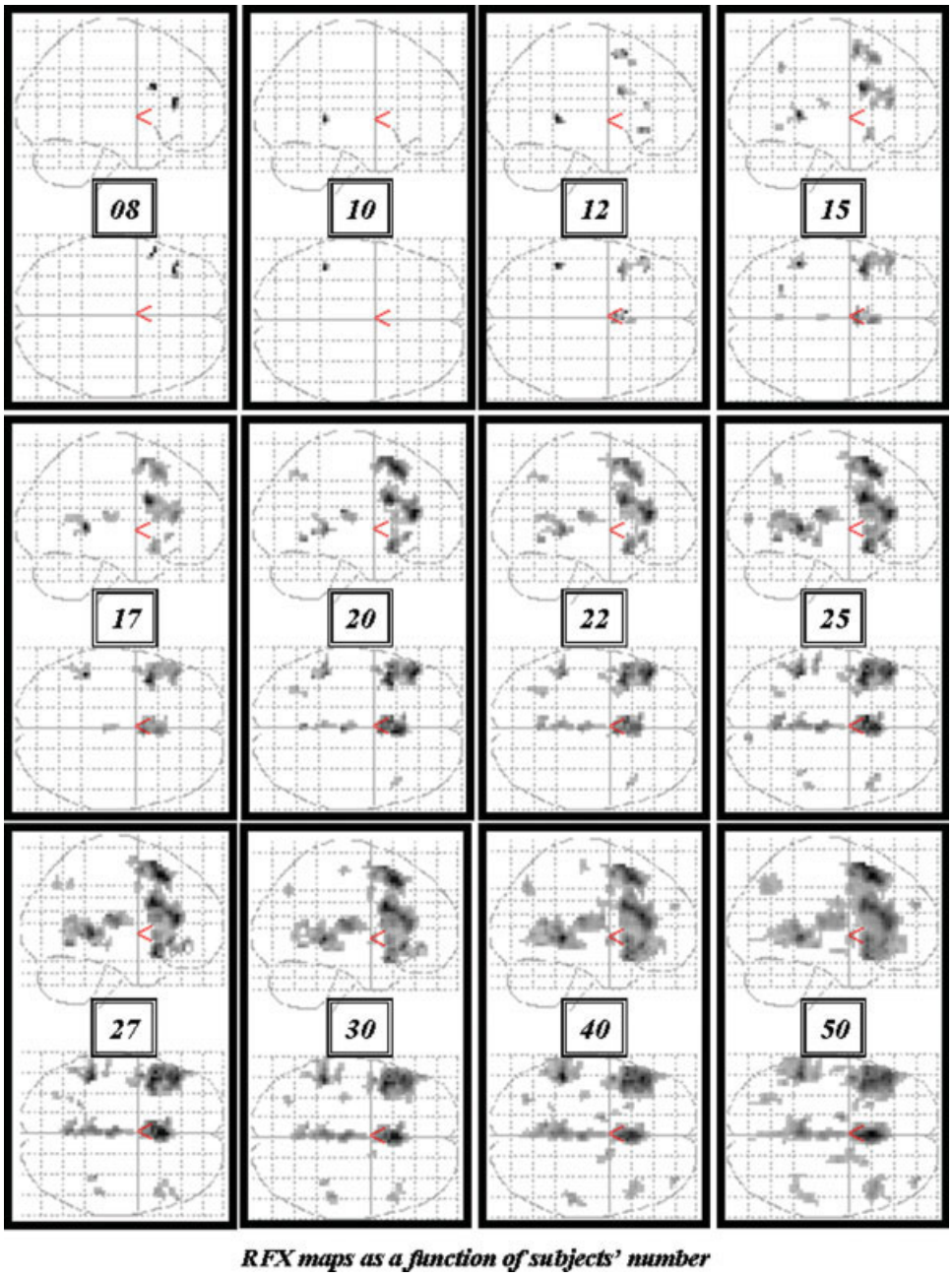

Figure 5.
Activation maps for the different RFX group analysis with *N*-sub varying from 8 to 50. For each RFX analysis, sagittal and axial views are shown in the neurological convention. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

sensitive to individual differences [McNamee and Lazar, 2004]. Specifically, the Fisher's method leads to the detection of more observations compared to the RFX analysis, in particular, in the regions where functional responses are more variable across subjects as shown, for example, in the temporoparietal and RH ones (Fig. 4). The FFX approach and the averaged *t*-maps constitute an intermediate solution between the two previous approaches. Indeed, both maps appear very similar, except in posterior occipital regions, and seem to provide the same information about the average activation across subjects.

In addition, the regions detected with the averaged *t*-maps method are similar to those detected with Fisher's method. In view of that, the averaged *t*-maps method is of particular interest for exploratory studies, and thanks to its computational simplicity, could be recommended for revealing individual differences at the group level during normative database formation. Likewise, PO maps are also suitable for revealing both dominant and less dominant activations across subjects. In the context of the inter-individual functional variability, the use of the averaged *t*-maps and the PO maps is particularly suitable for appreci-

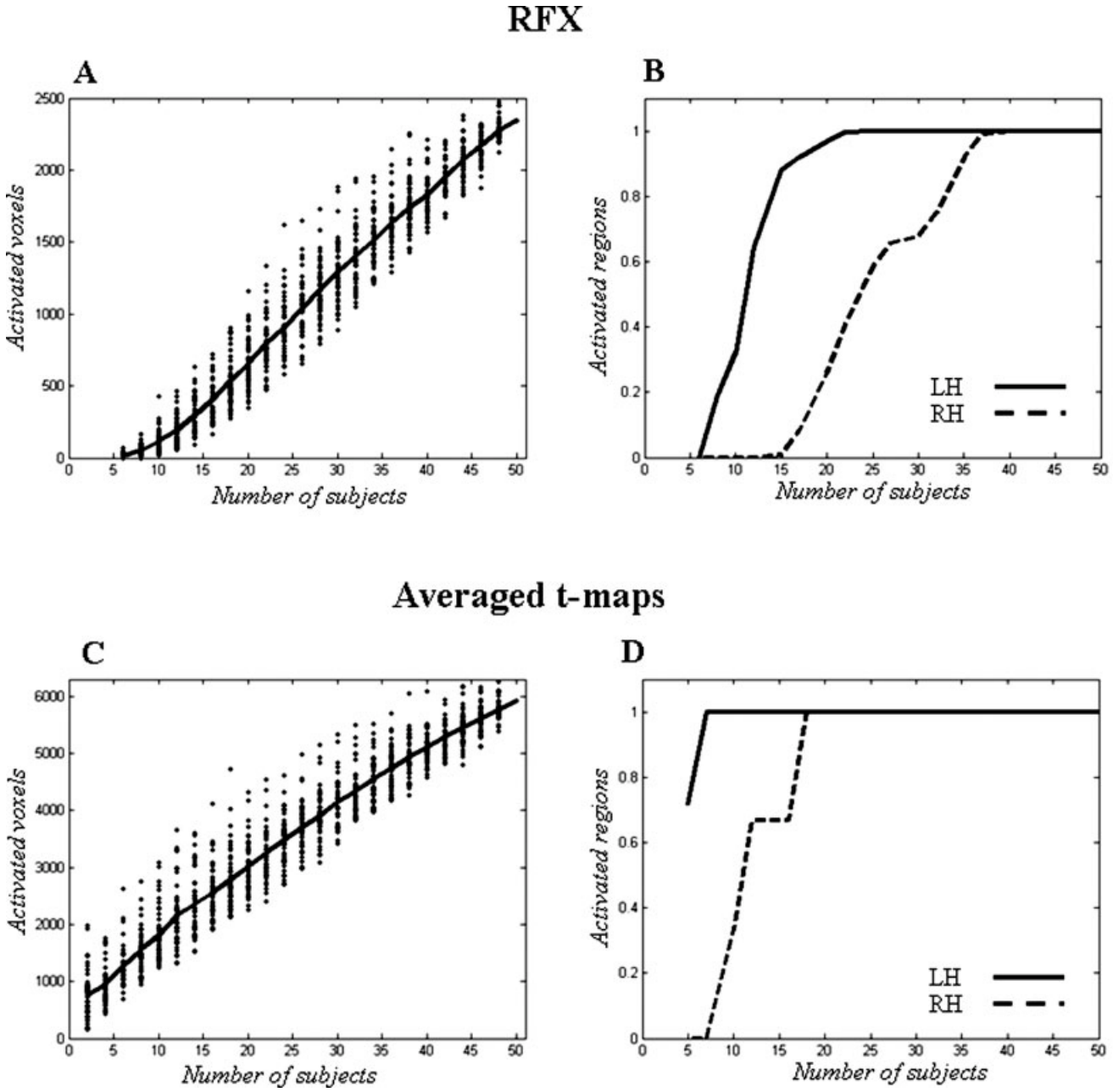

Figure 6.

A: Increase in the number of activated voxels in RFX as a function of increasing *N*-sub. For each *N*-sub value, 50 randomisations of our subjects were analysed with RFX. Each point represents the activated volume of one RFX analysis with a given *N*-sub. Note that we have only one value for number of voxels when *N*-sub reached our group size (50 subjects); we obviously need a larger group size to illustrate also the variability when *N*-sub equal 50. **B:** The number of activated regions in the left (solid line) and the right (dashed line) hemisphere increases with the number of subjects included in the RFX analysis. Note that for both hemispheres, the number of regions was 'normalised' to the total number observed in the RFX analysis with 50 subjects (i.e., seven regions in the left hemisphere; three regions in the right hemisphere, see Table I). **C–D:** Same analysis as in A–B but with averaged *t*-maps approach.

ating all activated foci that have been involved even in a subset of subjects. This issue is crucial when comparing qualitatively the individual activation pattern of patients with the reference group, particularly when the number of control subjects is small [see for instance, Fernandez et al., 2004; Seghier et al., 2001]. These observations are also in line with previous studies that have recommended the use of different group analyses to reveal patterns of activation

that may remain undetected if only one single approach was used [Lange et al., 1999; Lukic et al., 2002].

Number of Subjects and Detectability

Increasing the number of subjects is supposed to improve the significance of the RFX maps. Figure 5 illustrates the extent of activation in RFX maps (at $P < 0.001$ uncorrected) as a function of the *N*-sub included. Figure 6 quantifies this relation and shows that, between 10 and 50 subjects and for all the 50 randomisations of our sample, the extent of activation (expressed in terms of activated voxels) increases almost linearly when the *N*-sub is increased (Fig. 6A). More importantly, the number of regions activated (as defined from the 50 subjects gold-standard map, see Table I) also increases with the number of subjects included (Fig. 6B). In particular, this analysis shows that about 15–20 subjects are needed to reliably detect all LH-activated regions (Fig. 6B, see solid line, see Table I) while about 30–35 subjects are required to reveal the weak and more variable RH regions (Fig. 6B, see dashed line). Consequently, it appears that, in our task, about 20 healthy subjects are necessary for constituting a representative normative database that can be used to assess functional responses in patients with LH language dominance. A greater *N*-sub is required (>30 subjects) when one's objective is also to investigate patients with atypical language representation and where activations are found in areas that are not dominantly involved in healthy controls. On the other hand, it also appears that the estimation of the sample size (i.e., *N*-sub) necessary for the formation of a normative database is evidently linked to the group analysis approach used. For instance, to detect the same language areas (as detailed in Table I), the requested sample size might be smaller when using a less restrictive approach such as the averaged *t*-maps. Figure 6C shows that the extent of activation also increases linearly with *N*-sub. Interestingly, with a sample size of 20 subjects, all LH and RH language areas could be identified with the averaged *t*-maps approach (Fig. 6D). Thus, this comparison confirms that, while the RFX analysis is suitable for detecting true positive activations with high degree of confidence, the averaged *t*-maps approach is particularly useful for appreciating all activated regions of interest during normative database formation.

In relation to this crucial debate on the *N*-sub 'necessary' to constitute a functional neuroimaging study [Desmond and Glover, 2002; Friston et al., 1999b; Murphy and Garavan, 2004], and in order to avoid missing relevant activations when using the RFX approach (type II error), some studies have been carried out with large number of subjects [e.g., 100 subjects or more, see Kiehl et al., 2005; Pujol et al., 1999; Springer et al., 1999; Szaflarski et al., 2006]. However, our analysis points to the fact that the *N*-sub necessary to define with a high degree of confidence the extension of the language network could vary as a function of the cortical regions considered: a small sample is required to reveal LH frontal activations and a large *N*-sub to detect RH ones. Finally, the extent of activation should be used with caution when comparing groups with fMRI. Any statistical parameter that is planned to be estimated should converge rapidly toward its nominal value when the number of subjects increases. As a matter of fact, the linear relation between the extent of activation and *N*-sub indicates that this parameter cannot be easily estimated by fMRI in this type of cognitive tasks, which involve a large and distributed neural network. On the other hand, we observed that the number of regions, reflecting the functional network implicated in the task, is more easily assessed by fMRI, because it reached its nominal value more rapidly when *N*-sub increases (at least in the range of *N*-sub explored here).

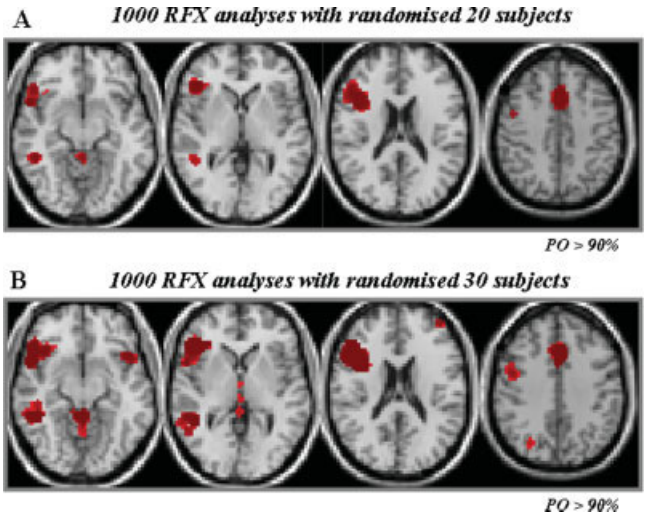

Figure 7.
PO maps on 1,000 RFX analyses on different subgroups. **A:** With a subgroup size of 20 subjects, selected randomly from our sample. **B:** With a subgroup size of 30 subjects selected randomly from our sample. All PO maps were generated at a threshold of 90%, and each RFX analysis was thresholded at $P < 0.001$ (uncorrected). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

To substantiate our conclusion that about 20 subjects are sufficient to reliably detect LH activations with RFX approach, we performed different RFX analyses on 1,000 different subgroups of 20 subjects selected randomly from our original sample of 50 subjects (see Methods section, for more details). Figure 7A highlights the results of these 1,000 RFX analyses, illustrated here in terms of PO maps (threshold at >90%). Mainly, this illustration shows that all LH activations except the left inferior parietal lobule are detected in at least 90% of these RFX analyses with 20 subjects (i.e., voxels involved in 90% of these randomly selected different subgroups). Particularly, the frontal activations are more reliable and significant than parieto-temporal activations, which are in line with our observation on the variability of activation in these areas (see, for instance, Fig. 3B). Similarly, the RFX analyses performed

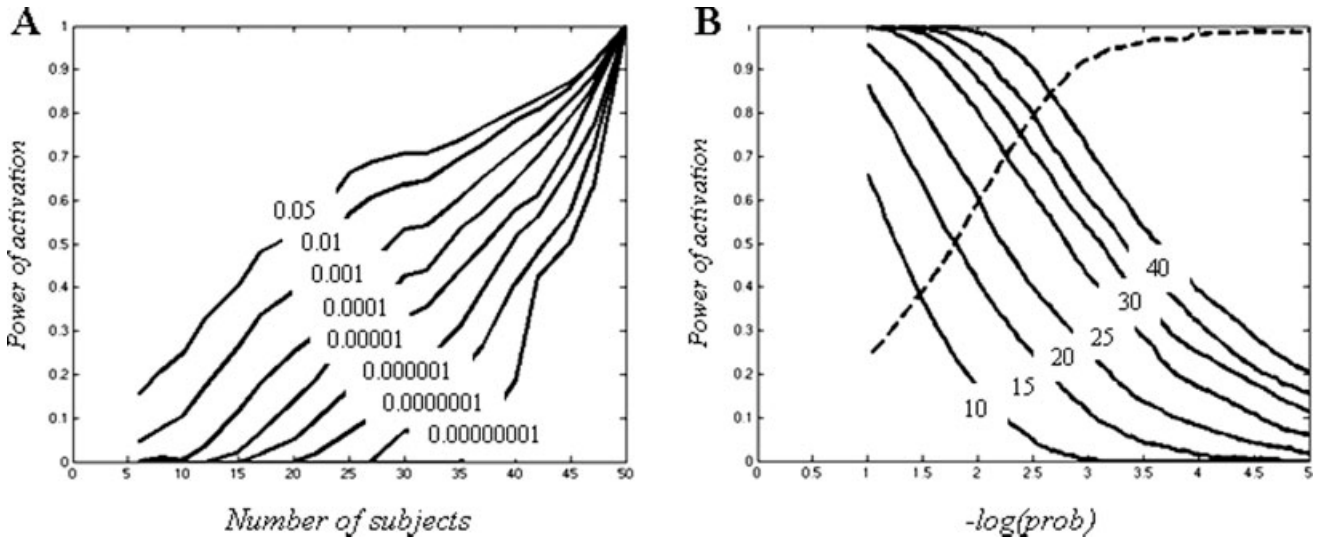

Figure 8.

Results of RFX analysis with different *N*-sub and statistical thresholds. **A:** Power of activation map as a function of the number of subjects at different statistical threshold. **B:** Power of activation map (solid lines) as a function of the regarding to the statistical threshold with different number of subjects. The dashed line illustrates for 30 subjects the rate of false positive activations (i.e., activations not present in the gold standard map).

on additional 1,000 different subgroups of 30 subjects selected randomly from our 50 subjects sample support the reliability of RH activations in larger groups (Fig. 7B), as suggested in our findings described earlier. Indeed, together with the highly reliable LH activations at this size of samples (i.e., 30 subjects), we observe reproducible frontal RH activations in 90% of the subgroups.

Number of Subjects and Power of Activation Map

The advantage of including a large *N*-sub when establishing a normative database can be illustrated by considering the power of activation of the group map with a given *N*-sub at a given statistical threshold (Fig. 8). Concretely, this refers to the overlap between the map obtained from an RFX analysis with *N*-sub and the gold standard map. Figure 8A indicates that with a typical uncorrected statistical threshold of $P < 0.001$, the RFX analysis reveals the need for at least 27 subjects to reach a power of 0.5 (i.e., it implicates 50% of the gold standard activations) and that at least 43 subjects are required to reach a power of 0.8 (Fig. 8A). Figure 8B illustrates another way of characterizing the relation between the power of activation and the statistical threshold (see Methods section, for more details). This relation is particularly relevant when including a relatively limited number of subjects. For example, with 20 subjects (see solid line in Fig. 8B), a power of 0.5 can be reached with a statistical threshold of $P < 0.005$ and a higher power (e.g., 0.8) is attained when the statistical threshold is lowered (e.g., $P < 0.03$). However, as shown in this illustration (see dashed line in Fig. 8B), using low statistical thresholds basically increases the number of activated voxels outside the gold standard map. For instance, with 20 subjects, a power of 0.8 is reached at $P < 0.03$, but only 40% (dashed line in Fig. 8B) of the total activated volume of the 20 subjects' map contributes to this power. This finding suggests that about 60% of the activations detected at $P < 0.03$ in the map with 20 subjects might be considered as false positives (i.e., not present in the gold standard map). In this case, the gain in terms of power of the activation map at low statistical thresholds is unavoidably accompanied by increased false activations when using a small *N*-sub.

It is noteworthy that using the power of the activation map to assess the robustness of activation with different *N*-sub is inherently problematic since the gold standard map is defined from a group analysis that includes the same *N*-sub. For this reason, all the results must converge to a perfect overlap at 50 subjects as illustrated in Figure 8A [see also Murphy and Garavan, 2004]. Ideally, the gold standard map should be assessed with a different set of subjects; however, this will necessitate acquiring and analysing additional 50 subjects from the same population. Moreover, the power map assessment takes into account only the true positives and ignores the number of false positives. For instance, any map that activates the entire brain would have a high power to detect activations that are present in the gold standard map. For this reason, the power of the activation map should be assessed in parallel with other measures that also reflect the percentage of false positives, as performed here (Fig. 8B). In our task, the increase of activation extent (i.e., number of voxels) as a function of *N*-sub (e.g., Fig. 6A) seems to be the major factor responsible of the poor power of activation map. Instead, our results showed that, for the definition of a

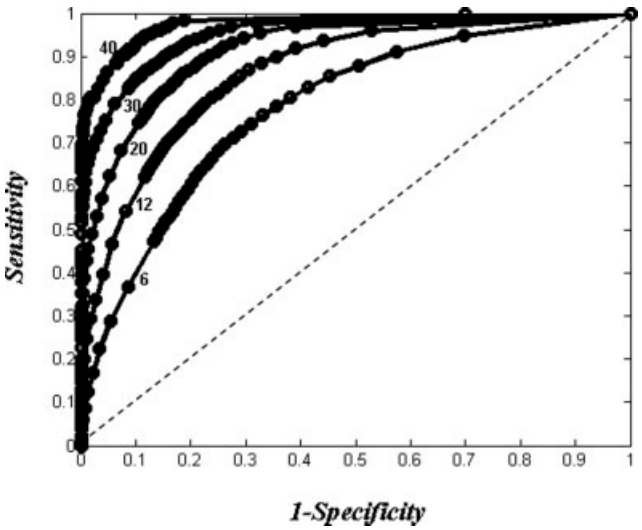

Figure 9.
ROC curves for RFX analysis with 6, 12, 20, 30, and 40 subjects. Diagonal dashed line represents the case of a noninformative random map.

gold standard map, it appears more appropriate to investigate this power at the region scale (as analysed in Fig. 6B) than at the voxel scale.

Number of Subjects and ROC Curves

The results obtained from the power of activation map analysis can alternatively be expressed with ROC curves of RFX analyses with varying *N*-sub. The main advantage of the ROC approach is to generate threshold-free curves about the sensitivity and the specificity of RFX with a given sample size [Skudlarski et al., 1999]. Figure 9, which displays the ROC curves for different *N*-sub, shows that the sensitivity (i.e., true positive rate) and the specificity (i.e., true negative rate) of RFX analysis are improved when the sample size is increased. For instance, a sensitivity of 0.9 with a sample of 12 subjects is accompanied with low specificity (0.65); however, this specificity could be improved to 0.85 without losing in sensitivity if the sample size is increased to 30 subjects. These curves seem thus very useful for assessing the sensitivity and the specificity of group analysis at a given *N*-sub. However, as already mentioned, and since the gold standard used here to compute ROC curves with a given *N*-sub contains these same subjects, the gold standard map should ideally be assessed with a different set of subjects.

Visualizing Inter-Individual Variability

Together with the definition of the group gold standard representative map, the appreciation of the variability between control subjects is an important question that one has to address during the formation of the normative database. This issue is of particular interest in the context of 'clinical fMRI' since it permits determining whether the pattern of activation in a given patient is fundamentally different from the functional pattern dominantly revealed in control subjects, or if it could be considered as part of the normal variation found in healthy control. For assessing the inter-individual variability in terms of activation patterns, we relied here on the use of a graphical tool that presents the position of each subject with respect to the others as a function of two similarity indices computed from the individual *t*-maps and the groups' gold standard map (see the Methods section). Figure 10 illustrates this projection plane and depicts the position of all individual subjects as a function of the %TPi (which assesses the percentage of true positives in an individual map with regard to the gold standard) and the %FNg (which assesses the percentage of false negatives as compared to the gold standard) indices. It can be observed that ~60% of the subjects are concentrated in the intervals [0.21 0.48] and [0.87 0.97] for %TPi and %FNg, respectively, indicating that on average >35% of voxels activated in each of the subjects at $P < 0.001$ can be considered as true positives. This analysis indicates, however, that 92% on average of the activated voxels in the gold standard map are missed in individual subjects. Differences between individual maps and the group map have also been reported in other fMRI studies [Miller and Van Horn, 2007], suggesting that reliance on group analysis alone might be incomplete for our task. Moreover, subjects that are far from these intervals and considerably increase the variance within the group could be easily distinguished. First, subjects with a minimal response, but are in agreement with the gold standard. For instance, at the $P < 0.001$ threshold used here, subject nos. 44 and 19 (see right inset in Fig. 10 for no. 44) had implicated <3% of the total voxels in the gold standard map, but this represented 80% of total volume of these subjects. Conversely, subjects with large activations (good responders) who are not concordant with the gold standard can also be identified. Subject no. 9 (right inset in Fig. 10) is an illustration of such a case, showing a large cortical volume (>20% of total true positives in the gold standard are included in this volume), but with only 35% of this volume overlapping with the gold standard map. Finally, the combination of both, that is, minimal activation and low concordance with the gold standard can also be observed, as for subject nos. 5 and 31 (see right inset in Fig. 10 for no. 5). In such subjects, an activated volume contains <2% of the total true positives in the gold standard map, but where true positives represent <15% of this small volume. Their position in this plane indicates that these subjects, at this specific threshold, behave very differently from the other subjects of the population. These examples show that this plane can provide interesting graphic information about the relative similarity between the functional maps of all subjects, and they can be generated at different statistical thresholds.

Furthermore, our tool could be used with other approaches that have been proposed previously to deal

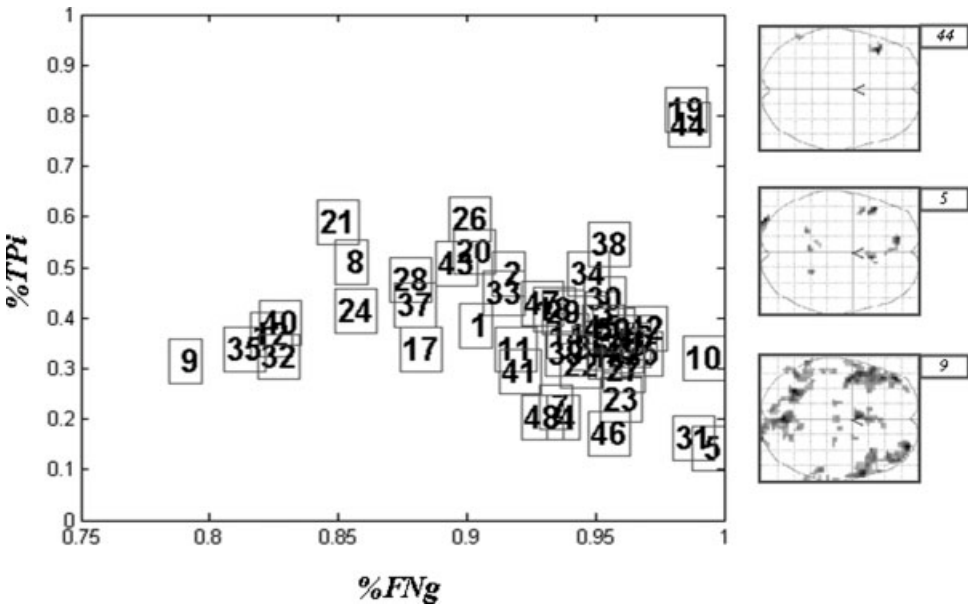

Figure 10.
Projection plan for all subjects according to the %TPi and %FNg indices (see text for definition).

with individual differences. Specifically, these approaches have tried to visualise inter-individual variability [Kherif et al., 2003], evaluate it [Zhang et al., 2006], or down-weight its influence during group effect assessment as done in robust statistics [Wager et al., 2005]. However, in a clinical perspective, this variability is interesting and informative about the potential strategies that can be employed during task execution. The characterisation and the identification of such strategies, reflected in functional responses as individual patterns of activation, will help to understand plasticity and reorganisation mechanisms after brain damage. Actually, our analysis demonstrates that individual differences are significant in this cognitive task that relies on a largely distributed network of brain areas (e.g., 65% on the average of the activated voxels in individual maps are not visible in the group map as shown in Fig. 10). These individual differences, which are probably at the origin of the differences observed in terms of sensitivity between the group analysis (e.g., Fig. 4), emphasise the importance to fully consider the between-subjects response variations when one's objective is to establish a representative activation map for a given task. Finally, it is worth reminding here that the indices computed reflect the similarity between the individual and group maps at the voxel level, and therefore it may be interesting to look to this similarity at the region scale as well.

CONCLUSION

In this study, we analysed brain responses in a large number of subjects who performed a semantic language task. Our aims were to assess the influence of the number of subjects on the definition of a gold standard activation map and to characterise inter-individual functional variability. To achieve this goal, we relied on different group analysis approaches and applied various quantitative indices. We observed that the definition of regions activated (and thus the extent of the activated neural network) by this task and in this population depends not only on the statistical threshold used, but also on the number of subjects included in the analysis and the choice of the group analysis approach. With a typical *N*-sub of 12 or 20 subjects as commonly used in fMRI investigations, the statistical power at the voxel level was relatively poor with RFX analysis. However, at the region scale, our analysis indicated that the most reliable and robust activations, particularly in the LH, could be identified with high degree of confidence from a group of about 16 subjects. By contrast, much more variable and generally weak activations in the RH required at least 30 subjects with the same RFX group analysis. When comparing different group analysis approaches, we found that, relative to RFX analysis, the use of PO maps or averaged *t*-maps allows the network involved in these subjects to be efficiently revealed without missing the individual differences. We thus propose that such simple analyses (i.e., averaged *t*-maps) could be recommended for exploratory investigations when a relatively small *N*-sub (typically 12–20 subjects) is used. Particular care was taken to appreciate the individual variability by qualitatively identifying atypical patterns of activation. The proposed measures permit normal language representation to be assessed empirically, and could thus be used for the estimation of the degree of 'normality' of functional responses in brain-damaged patients, in whom this ques-

tion is often raised. We propose that the different analyses presented here might be of particular interest during the process of database formation and recommend the use of various quantifications to appreciate accurately the inter-individual variability of functional responses.

ACKNOWLEDGMENTS

We thank F. Henry and I. Zimine for their technical assistance and J. Delavelle for her helpful discussions. The authors thank Cathy Price for her comments on the manuscript.

REFERENCES

Balsamo LM, Gaillard WD (2002): The utility of functional magnetic resonance imaging in epilepsy and language. Curr Neurol Neurosci Rep 2:142–149.

Billingsley RL, McAndrews MP, Crawley AP, Mikulis DJ (2001): Functional MRI of phonological and semantic processing in temporal lobe epilepsy. Brain 124:1218–1227.

Bosch V (2000): Statistical analysis of multi-subject fMRI data: Assessment of focal activations. J Magn Reson Imaging 11:61–64.

Buchel C, Turner R, Friston KJ (1997): Lateral geniculate activations can be detected using intersubject averaging and fMRI. Magn Reson Med 38:691–694.

Cao Y, Vikingstad EM, George KP, Johnson AF, Welch KM (1999): Cortical language activation in stroke patients recovering from aphasia with functional MRI. Stroke 30:2331–2340.

Chee MW, Buckner RL, Savoy RL (1998): Right hemisphere language in a neurologically normal dextral: A fMRI study. Neuroreport 9:3499–3502.

Cheetham AH, Hazel JE (1969): Binary (presence–absence) similarity coefficients. J Paleontol 43:1130–1136.

Content A, Mousty P, Radeau M (1990): Brulex: Une base de données lexicales informatisée pour le français écrit et parlé. L'année Psychol 90:551–566.

Desmond JE, Chen ASH (2002): Ethical issues in the clinical application of fMRI: Factors affecting the validity and interpretation of activations. Brain Cogn 50:482–497.

Desmond JE, Glover GH (2002): Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses. J Neurosci Methods 118:115–128.

Detre JA, Floyd TF (2001): Functional MRI and its applications to the clinical neurosciences. Neuroscientist 7:64–79.

Edwards JD, Pexman PM, Goodyear BG, Chambers CG (2005): An fMRI investigation of strategies for word recognition. Brain Res Cogn Brain Res 24:648–662.

Fernandez B, Cardebat D, Demonet JF, Joseph PA, Mazaux JM, Barat M, Allard M (2004): Functional MRI follow-up of language processes in healthy subjects and during recovery in a case of aphasia. Stroke 35:2171–2176.

Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC (1995): Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. Magn Reson Med 33:636–647.

Friston KJ, Williams S, Howard R, Frackowiak RS, Turner R (1996): Movement-related effects in fMRI time-series. Magn Reson Med 35:346–355.

Friston KJ, Holmes AP, Price CJ, Buchel C, Worsley KJ (1999a): Multisubject fMRI studies and conjunction analyses. Neuroimage 10:385–396.

Friston KJ, Holmes AP, Worsley KJ (1999b): How many subjects constitute a study? Neuroimage 10:1–5.

Grabowski TJ, Frank RJ, Brown CK, Damasio H, Ponto LLB, Watkins GL, Hichwa RD (1996): Reliability of PET activation across statistical methods, subject groups, and sample sizes. Hum Brain Mapp 4:23–46.

Hennig J, Speck O, Koch MA, Weiller C (2003): Functional magnetic resonance imaging: A review of methodological aspects and clinical applications. J Magn Reson Imaging 18:1–15.

Herholz K, Thiel A, Wienhard K, Pietrzyk U, von Stockhausen HM, Karbe H, Kessler J, Bruckbauer T, Halber M, Heiss WD (1996): Individual functional anatomy of verb generation. Neuroimage 3:185–194.

Holmes AP, Friston KJ (1998): Generalisability, random effects and population inference. Neuroimage 7:S754.

Hugdahl K, Gundersen H, Brekke C, Thomsen T, Rimol LM, Ersland L, Niemi J (2004): FMRI brain activation in a Finnish family with specific language impairment compared with a normal control group. J Speech Lang Hear Res 47:162–172.

Jackson GD, Connelly A, Cross JH, Gordon I, Gadian DG (1994): Functional magnetic resonance imaging of focal seizures. Neurology 44:850–856.

Johnstone T, Ores Walsh KS, Greischar LL, Alexander AL, Fox AS, Davidson RJ, Oakes TR (2006): Motion correction and the use of motion covariates in multiple-subject fMRI analysis. Hum Brain Mapp 27:779–788.

Juch H, Zimine I, Seghier ML, Lazeyras F, Fasel JHD (2005): Anatomical variability of the lateral frontal lobe surface: Implication for intersubject variability in language neuroimaging. Neuroimage 24:504–514.

Khateb A, Martory M-D, Annoni J-M, Lazeyras F, de Tribolet N, Pegna AJ, Mayer E, Michel CM, Seghier ML (2004): Transient crossed aphasia evidenced by functional brain imagery. NeuroReport 15:785–790.

Kherif F, Poline JP, Mériaux S, Benali H, Flandin G, Brett M (2003): Group analysis in functional neuroimaging: Selecting subjects using similarity measures. Neuroimage 20:2197–2208.

Kiehl KA, Stevens MC, Laurens KR, Pearlson G, Calhoun VD, Liddle PF (2005): An adaptive reflexive processing model of neurocognitive function: Supporting evidence from a large scale (n = 100) fMRI study of an auditory oddball task. Neuroimage 25:899–915.

Lange N, Strother SC, Anderson JR, Nielsen FA, Holmes AP, Kolenda T, Savoy R, Hansen LK (1999): Plurality and resemblance in fMRI data analysis. Neuroimage 10:282–303.

Lazar NA, Luna B, Sweeney JA, Eddy WF (2002): Combining brains: A survey of methods for statistical pooling of information. Neuroimage 16:538–550.

Lehéricy S, Cohen L, Bazin B, Samson S, Giacomini E, Rougetet R, Hertz-Pannier L, Le Bihan D, Marsault C, Baulac M (2000): Function MR evaluation of temporal and frontal language dominance compared with the Wada test. Neurology 54:1625–1633.

Levin JM, Ross MH, Renshaw PF (1995): Clinical applications of functional MRI in neuropsychiatry. J Neuropsychiatry Clin Neurosci 7:511–522.

Liou M, Su H-R, Lee J-D, Cheng PE, Huang C-C, Tsai C-H (2003): Bridging functional MR images and scientific inference: Reproducibility maps. J Cogn Neurosci 15:935–945.

Loring DW, Meador KJ, Allison JD, Pillai JJ, Lavin T, Lee GP, Balan A, Dave V (2002): Now you see it, now you don't: Statistical and methodological considerations in fMRI. Epilepsy Behav 3:539–547.

Lukic AS, Wernick MN, Strother SC (2002): An evaluation of methods for detecting brain activations from functional neuroimages. Artif Intell Med 25:69–88.

Machielsen WC, Rombouts SA, Barkhof F, Scheltens P, Witter MP (2000): FMRI of visual encoding: Reproducibility of activation. Hum Brain Mapp 9:156–164.

Maldjian JA, Laurienti PJ, Driskill L, Burdette JH (2002): Multiple reproducibility indices for evaluation of cognitive functional MR imaging paradigms. AJNR Am J Neuroradiol 23:1030–1037.

Matthews PM, Clare S, Adcock J (1999): Functional magnetic resonance imaging: Clinical applications and potential. J Inherit Metab Dis 22:337–352.

McNamee RL, Lazar NA (2004): Assessing the sensitivity of fMRI group maps. Neuroimage 22:920–931.

Miller MB, Van Horn JD (2007): Individual variability in brain activations associated with episodic retrieval: A role for large-scale databases. Int J Psychophysiol 63:205–213.

Miller MB, Van Horn JD, Wolford GL, Handy TC, Valsangkar-Smyth M, Inati S, Grafton S, Gazzaniga MS (2002): Extensive individual differences in brain activations associated with episodic retrieval are reliable over time. J Cogn Neurosci 14:1200–1214.

Murphy K, Garavan H (2004): An empirical investigation into the number of subjects required for an event-related fMRI study. Neuroimage 22:879–885.

Nadeau SE, Williamson DJ, Crosson B, Gonzalez Rothi LJ, Heilman KM (1998): Functional imaging: Heterogeneity in task strategy and functional anatomy and the case for individual analysis. Neuropsychiatry Neuropsychol Behav Neurol 11:83–96.

Noppeney U, Penny WD, Price CJ, Flandin G, Friston KJ (2006): Identification of degenerate neuronal systems based on inter-subject variability. Neuroimage 30:885–890.

Otzenberger H, Gounot D, Marrer C, Namer IJ, Metz-Lutz MN (2005): Reliability of individual functional MRI brain mapping of language. Neuropsychology 19:484–493.

Penny WD, Holmes AP, Friston KJ (2003): Random effects analysis. In: Frackowiak RSJ, Friston KJ, Frith C, Dolan RJ, Price CJ, Zeki S, Ashburner J, Penny WD, editors. Human Brain Function. New York: Academic Press. pp 843–850.

Pizzamiglio L, Galati G, Committeri G (2001): The contribution of functional neuroimaging to recovery after brain damage: A review. Cortex 37:11–31.

Powell HW, Koepp MJ, Richardson MP, Symms MR, Thompson PJ, Duncan JS (2004): The application of functional MRI of memory in temporal lobe epilepsy: A clinical review. Epilepsia 45:855–863.

Price CJ, Crinion J (2005): The latest on functional imaging studies of aphasic stroke. Curr Opin Neurol 18:429–434.

Pugh KR, Shaywitz BA, Shaywitz SE, Constable RT, Skudlarski P, Fulbright RK, Bronen RA, Shankweiler DP, Katz L, Fletcher JM, Gore JC (1996): Cerebral organization of component processes in reading. Brain 119:1221–1238.

Pujol J, Deus J, Losilla JM, Capdevila A (1999): Cerebral lateralization of language in normal left-handed people studied by functional MRI. Neurology 52:1038–1043.

Raz A, Lieber B, Soliman F, Buhle J, Posner J, Peterson BS, Posner MI (2005): Ecological nuances in functional magnetic resonance imaging (fMRI): Psychological stressors, posture, and hydrostatics. Neuroimage 25:1–7.

Rimol LM, Specht K, Hugdahl K (2006): Controlling for individual differences in fMRI brain activation to tones, syllables, and words. Neuroimage 30:554–562.

Seghier M, Lazeyras F, Momjian S, Annoni J-M, de Tribolet N, Khateb A (2001): Language representation in a patient with a dominant right hemisphere: fMRI evidence for an intrahemispheric reorganisation. NeuroReport 12:2785–2790.

Seghier ML, Lazeyras F, Pegna AJ, Annoni JM, Zimine I, Mayer E, Michel CM, Khateb A (2004): Variability of fMRI activation during a phonological and semantic language task in healthy subjects. Hum Brain Mapp 23:140–155.

Skudlarski P, Constable RT, Gore JC (1999): ROC analysis of statistical methods used in functional MRI: Individual subjects. Neuroimage 9:311–329.

Smith SM, Beckmann CF, Ramnani N, Woolrich MW, Bannister PR, Jenkinson M, Matthews PM, McGonigle DJ (2005): Variability in fMRI: A re-examination of inter-session differences. Hum Brain Mapp 24:248–257.

Specht K, Willmes K, Shah NJ, Jancke L (2003): Assessment of reliability in functional imaging studies. J Magn Reson Imaging 17:463–471.

Springer JA, Binder JR, Hammeke TA, Swanson SJ, Frost JA, Bellgowan PS, Brewer CC, Perry HM, Morris GL, Mueller WM (1999): Language dominance in neurologically normal and epilepsy subjects: A functional MRI study. Brain 122: 2033–2046.

Strother SC, Lange N, Anderson JR, Schaper KA, Rehm K, Hansen LK, Rottenberg DA (1997): Activation pattern reproducibility: Measuring the effects of group size and data analysis models. Hum Brain Mapp 5:312–316.

Szaflarski JP, Holland SK, Schmithorst VJ, Byars AW (2006): fMRI study of language lateralization in children and adults. Hum Brain Mapp 27:202–212.

Talairach J, Tournoux P (1988): Co-Planar Stereotaxic Atlas of the Human Brain. New York: Thieme.

Thulborn KR, Davis D (2001): Quality assurance for clinical fMRI. Curr Protocols Magn Reson Imaging A 6, 1-4.

Thulborn KR, Davis D, Erb P, Strojwas M, Sweeney JA (1996): Clinical fMRI: Implementation and experience. Neuroimage 4:S101–S107.

Thulborn KR, Carpenter PA, Just MA (1999): Plasticity of language-related brain function during recovery from stroke. Stroke 30:749–754.

Tsukiura T, Mochizuki-Kawai H, Fujii T (2005): The effect of encoding strategies on medial temporal lobe activations during the recognition of words: An event-related fMRI study. Neuroimage 25:452–461.

Tzourio-Mazoyer N, Josse G, Crivello F, Mazoyer B (2004): Interindividual variability in the hemispheric organization for speech. Neuroimage 21:422–435.

Wager TD, Keller MC, Lacey SC, Jonides J (2005): Increased sensitivity in neuroimaging analyses using robust regression. Neuroimage 26:99–113.

Wenz F, Schad LR, Knopp MV, Baudendistel KT, Flomer F, Schroder J, van Kaick G (1994): Functional magnetic resonance imaging at 1.5 T: Activation pattern in schizophrenic patients receiving neuroleptic medication. Magn Reson Imaging 12:975–982.

Woermann FG, Jokeit H, Luerding R, Freitag H, Schulz R, Guertler S, Okujava M, Wolf P, Tuxhorn I, Ebner A (2003): Language lateralization by Wada test and fMRI in 100 patients with epilepsy. Neurology 61:699–701.

Woods RP (1996): Modeling for intergroup comparisons of imaging data. Neuroimage 4:S84–S94.

Xu XJ, Zhang MM, Shang DS, Wang QD, Luo BY, Weng XC (2004): Cortical language activation in aphasia: A functional MRI study. Chin Med J 117:1011–1016.

Zhang H, Luo WL, Nichols TE (2006): Diagnosis of single-subject and group fMRI data with SPMd. Hum Brain Mapp 27:442–451.