NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 35 (2007) 105 – 120

Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses

Bertrand Thirion,[a,*] Philippe Pinel,[c] Sébastien Mériaux,[b] Alexis Roche,[b]
Stanislas Dehaene,[c] and Jean-Baptiste Poline[b]

[a]*INRIA Futurs, Service Hospitalier Frédéric Joliot, 4, Place du Général Leclerc, 91401 Orsay cedex, France*
[b]*Département de Recherche Médicale-CEA-DSV Service Hospitalier Frédéric Joliot, 4, Place du Général Leclerc, 91401 Orsay cedex, France*
[c]*INSERM, U562, Cognitive Neuroimaging Unit, Service Hospitalier Frédéric Joliot, CEA/DRM/DSV, 4 Place du Gnral Leclerc, 91401 Orsay cedex, France*

Received 9 June 2006; revised 30 October 2006; accepted 16 November 2006
Available online 18 January 2007

The aim of group fMRI studies is to relate contrasts of tasks or stimuli to regional brain activity increases. These studies typically involve 10 to 16 subjects. The average regional activity statistical significance is assessed using the subject to subject variability of the effect (random effects analyses). Because of the relatively small number of subjects included, the sensitivity and reliability of these analyses is questionable and hard to investigate. In this work, we use a very large number of subject (more than 80) to investigate this issue. We take advantage of this large cohort to study the statistical properties of the inter-subject activity and focus on the notion of reproducibility by bootstrapping. We asked simple but important methodological questions: Is there, from the point of view of reliability, an optimal statistical threshold for activity maps? How many subjects should be included in group studies? What method should be preferred for inference? Our results suggest that *i*) optimal thresholds can indeed be found, and are rather lower than usual corrected for multiple comparison thresholds, *ii*) 20 subjects or more should be included in functional neuroimaging studies in order to have sufficient reliability, *iii*) non-parametric significance assessment should be preferred to parametric methods, *iv*) cluster-level thresholding is more reliable than voxel-based thresholding, and *v*) mixed effects tests are much more reliable than random effects tests. Moreover, our study shows that inter-subject variability plays a prominent role in the relatively low sensitivity and reliability of group studies.

Introduction

Inter-subject variability in neuroimaging and its impact on group analyses

One of the key characteristics of fMRI data is their large inter-subject variability compared to the generally lower intra-subject variability (see, amongst others, Wei et al., 2004). This high degree of variability impacts dramatically the sensitivity of random effects studies often performed with 10–16 subjects, and much effort is therefore spent to obtain statistically significant results by improving the spatial normalization procedures or the statistical tests while controlling for false positives.

The between subject variability is caused by a mixture of random and deterministic or structured factors that are not easily studied. We briefly summarize those.

- Spatial mismatch between subjects cortical structures. It is known that perfect correspondences between two anatomical images cannot be achieved, and that correspondences should generally be considered as approximate, even after rigid or non-rigid spatial normalization. The magnitude order of such local shifts is probably as large as 1 cm in many brain regions (this can be observed for functional regions like the motor cortex or the visual areas (Thirion et al., 2006a; Stiers et al., 2006) or the position of anatomical landmarks (Collins et al., 1998; Hellier et al., 2003). Across subjects, this effect typically yields a structured but variable pattern.
- The activation magnitude recorded at the same location for several tasks is variable across subjects, and sometimes across sessions (Smith et al., 2005), and the precise nature of this variability is not clear. Part of this variability may be related to physiological fluctuations, motion, resting-state activity (Fox et al., 2006), and more generally, what is usually called structured noise (Lund et al., 2006). It should be recalled that fMRI is not a quantitative neuroimaging modality, and that the standard use of reporting percent of signal increase is also problematic, due to the ambiguous definition (voxel-based or global average) of the baseline reference.
- Finally, there could be global differences in the brain networks elicited by a given task or experimental condition, related to genetic or epigenetic differences between subjects, or to different cognitive strategies (for non-trivial tasks). This is of course an interesting phenomenon to be studied, but clearly

* Corresponding author.
E-mail address: thirion@shfj.cea.fr (B. Thirion).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2006.11.054

difficult to demonstrate and to account for in standard studies given the subject sample size.

While this is only a superficial account of the possible sources of variability across subjects, it is important to note that *all* these effects are equally treated as *confounds* and globally modelled as the second-level variability (Friston et al., 2002) terms in current random effects analyses. Note also that the intra-subject variability across scanning sessions is not generally measured, but is generally less than inter-subject variability (Wei et al., 2004).

Because of this (generally) large inter-subject variance compared with the relatively small increase of Blood Oxygen Level Dependent (BOLD) activity, voxel-based random effects analyses that assess the significance of an effect by comparing its mean value to its variability across subjects are typically not sensitive (Friston et al., 1999; McNamee and Lazar, 2004). Several factors have a direct influence on the sensitivity for a given effect size. First, the quality of the model, including preprocessing, choice of noise and signal model, amount of smoothing performed etc. Second, the power of the statistical test chosen for detection (local maxima, cluster size, combination of the two, parametric or non-parametric testing etc). Third, the number of subjects included in the study. Vast differences in sensitivity can be observed depending on those parameters. In particular, Desmond and Glover (2002) report that about 25 subjects are necessary to achieve 80% power for a 0.5% increase of activity (based on the variability measured on a group of 12). Note that groups are often half this size. We illustrate this issue in Fig. 1 with an example taken from a dataset presented below composed of 78 subjects. We observed that the analysis of 6 different groups of 13 subjects would lead to different reports of the set of activated regions for the same experimental conditions at a standard threshold, and also observe the striking increase in sensitivity with the pooled analysis (all 78 subjects).

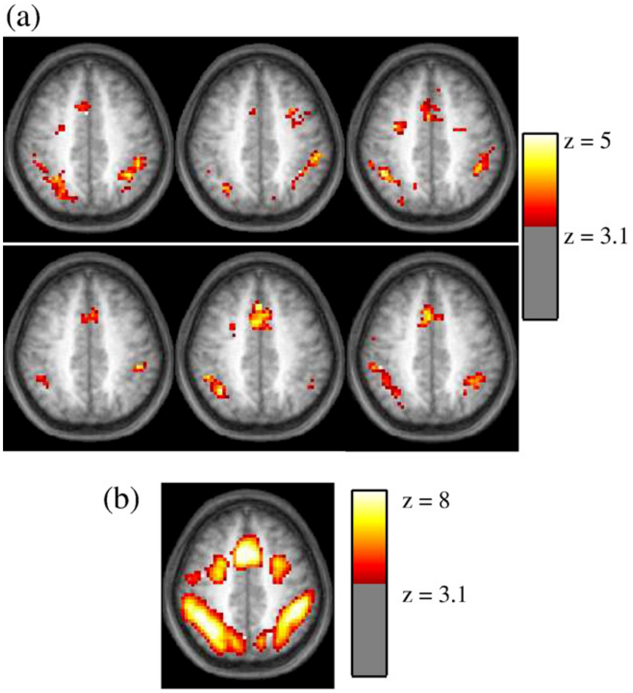

Fig. 1. Illustration of the low sensitivity and weak reliability of supra-threshold patterns in standard group studies. (a) For a functional contrast that shows regions involved in a computation task, we show activity maps thresholded at a $p<0.001$ level, uncorrected for multiple comparisons, after a random effect analysis on 6 disjoint groups of 13 subjects; the position of the view is $z=37$ mm in the MNI normalized space. (b) In the same plane, we present the same map computed from all the subjects combined. Note the low sensitivity and weak reliability of the maps in (a): different regions would be reported.

Reproducibility measures

Beyond the poor sensitivity of group analyses, there is an apparent lack of reliability in brain mapping studies (Jernigan et al., 2003), that may be seen as one of the key problems of this domain. This notion has been used in very few brain imaging papers (Murphy and Garavan, 2004; Liou et al., 2003; Genovese et al., 1997). A more systematic approach that combines the prediction of the activation states–or inverse inference–and the measure of the reproducibility of brain maps obtained from univariate or multivariate models has been presented in Strother et al. (2002), and applied in the optimization of pre-processing choices (LaConte et al., 2003; Shaw et al., 2003; Strother et al., 2004). To our knowledge, these methods have not provided any conclusion on the best way to perform statistical tests in group studies.

There are two main reasons why reproducibility is not systematically studied. First, studying reproducibility requires a large sample of subjects and second, it is less widely used in the medical or biological literature than the standard hypothesis testing framework. Nevertheless, the notion does seem to be at the heart of what would be needed by researchers or clinicians, because it can give a direct and interpretable answers to questions such as "how likely is this result to hold on a new dataset?", "What is the chance to observe this effect on a new subject?"

Reproducibility (or reliability[1]) analysis is based on binary i.e. thresholded maps obtained from distinct subgroups of subjects. In this work, we will use two reproducibility measures. The first is based on the modelling of the "activated" or "non-activated" label of the voxels across groups through a mixture of binomial distributions to assess the reproducibility of this labelling[2] The second is based on the distance between the position of large clusters in the thresholded binary maps. A large distance means that no correspondence can be found between supra-threshold clusters across groups of subjects, and hence that the maps are not very reliable (Murphy and Garavan, 2004). Importantly, notions of reproducibility and sensitivity are different and cannot be confounded. In particular, false positive occurring in (too) sensitive analyses will not be reproducible.

Reproducibility depends on the analysis performed

Those measures will clearly depend on the choice of the analysis that precedes thresholding and on the threshold. In this respect, a large number of methods are available in the literature and show increasing sophistication in defining an appropriate

[1] In this paper we will use both words to describe the same notion considering that a result with high reproducibility is reliable and a reliable result has high reproducibility measure.

[2] We exclude trivial or non-informative cases, e.g. cases in which the entire brain is "activated", or "non-activated".

threshold (Worsley, 2005). The classical random effects model is a particular instance of mixed effects models, which recently gained popularity (Worsley et al., 2002; Friston et al., 2002; Beckmann et al., 2003; Neumann and Lohmann, 2003; Woolrich et al., 2004). In this work, we will use tests based either on the voxel intensity or cluster size, with mixed effect models or standard random effects, and investigate the use of non-parametric versus parametric statistics:

- Choice of the threshold. Usually, statistical maps (SPMs) are thresholded to control for the rate of false positives.[3] Hereafter, we study the impact of the threshold on reproducibility measures.
- Voxel versus cluster based tests. While sensitivity and specificity have been studied for thresholding procedures based on the voxel or at the cluster level, reproducibility of those procedures is unknown.[4] We also investigate the use of parcel-based random effects maps (Thirion et al., 2006a), with a possibly double advantage: if parcels adapt to individual anatomy, they can cope with some parts of the inter-subject variability; second, this procedure considerably alleviates the multiple comparison problem.
- Parametric versus non-parametric tests. While parametric tests are particularly efficient and computationally cheap, they are based on possibly unrealistic hypotheses that may reduce their sensitivity (e.g. normal distribution). These hypotheses cannot be checked in the usual, small datasets. Non-parametric tests may avoid these issues (Holmes et al., 1996; Brammer et al., 1997; Bullmore et al., 1999; Nichols and Holmes, 2002; Hayasaka and Nichols, 2003; Mériaux et al., 2006a), but at a higher computational cost.
- Spatial filtering. Amongst the standard pre-processing steps, the smoothing kernel size (often chosen between 8 and 12 mm FWHM) is known to have large impact on sensitivity. It is already has been shown (Shaw et al., 2003; LaConte et al., 2003) that cross-validation schemes could help optimizing this choice; we simply use two different filter sizes and report the effect of this choice on reliability.

Note that the tests considered in this paper are signed, so that supra-threshold areas have a positive sign.

Taking advantage of a very large number of subjects

Another fundamental parameter of a study is the number of subjects that should be included in a study. While this question has been addressed with sensitivity measures and power analyses, it has not been studied with reproducibility measures. This number typically represents a trade-off between (a) the cost of conducting neuroimaging experiments on large cohorts of subjects and (b) the necessity of having enough subjects for the significance of statistical tests (Desmond and Glover, 2002; Murphy and Garavan, 2004). Reproducibility measures could answer the fundamental question: “how many subjects are enough to make the analysis reliable, in terms of avoiding false negatives while still controlling the false positives?”. This would yield the confidence level that can be given to a result as a function of the number of subject included in the study. Following (Liou et al., 2003) we define it practically as the agreement between independent measurements of a given effect size. While such a study would still be difficult with groups of 30 or 40 subjects, because of the limited number of subgroups–two or three–that can be drawn from such populations, it becomes feasible with 80 subjects.

It is clear that very strong activity will show higher reproducibility compared to weaker signals. We therefore also study the effect of the activation size on the number of subjects necessary to achieve high reproducibility using different cognitive contrasts which have different characteristics in terms of contrast-to-noise ratio or spatial variability.

Finally, because the number of subjects included in group analyses is usually small it is practically impossible to study the population distribution of activity in response to an experimental condition. Several techniques have been employed to quantify inter-subject differences globally (across brain regions) (Kherif et al., 2004), and this can be used to show that only one subject’s pattern of activity can significantly impact the group results (outlier detection). As a first analysis, our large number of subjects allows us to statistically test for normality of the activation level across subjects, and in the future could be use to detect subpopulations with univariate or multivariate procedures. Indeed, it is possible that inhomogeneous populations may be encountered in neuroimaging studies, and this can only be studied with a large number of subjects.

To summarize, in this paper we present reproducibility measures with various statistical procedures and thresholds and study the statistical properties of activation across subjects using an unusually large cohort.

Materials and methods

Dataset

We used an event-related experimental paradigm consisting of 10 conditions. Subjects were presented with a series of stimuli and were engaged in tasks such as passive viewing of horizontal or vertical checkerboards, left or right button press after audio or video instruction, computation (subtraction) after video or audio instruction and sentence processing, from the audio or visual modality. Events were randomly occurring in time (mean inter-stimulus interval: 3 s), with 10 occurrences per event type (except button presses for which there are only five trials per session). Note that contrasts of experimental conditions rely in fact on the sum of number of trials of each condition. For instance, the *left–right button press* contrast combines four experimental conditions (left/right button press after audio/video instruction) and relies on 20 trials. Similarly, the *audio–video* and *computation–sentences* contrasts rely on 60 and 40 trials respectively.

[3] In general, the chosen threshold does not reflect a trade-off between the necessity of controlling both the number of false positives and the number of false negatives. One straightforward reason is that it is relatively simple to model the statistical distribution under the null hypothesis, but not under the alternative hypothesis.

[4] For cluster size tests, the map is first thresholded at a (relatively lenient) significance level, and a second, the size of the resulting connected components is assessed against its distribution under the null hypothesis. This is usually considered as a safe procedure, but fully neglects the possibility of small yet significant activation foci. One of the reasons is that intensive smoothing of the data simply removes the possibility of finding such peaks.

Eighty-one right-handed subjects participated in the study. The subjects gave informed consent and the protocol was approved by the local ethics committee. Functional images were acquired on a 3T Bruker scanner using an EPI sequence (TR=2400 ms, TE=60 ms, matrix size=64×64, FOV=24 cm×24 cm). Each volume consisted of n_a 3-mm- or 4-mm-thick axial slices without gap, where n_a varied from 26 to 40 according to the session. A session comprised 130 scans. The first four functional scans were discarded to allow the MR signal to reach steady state. Anatomical T1 images were acquired on the same scanner, with a spatial resolution of $1\times1\times1.2$ mm^3.

fMRI data processing consisted in 1) temporal Fourier interpolation to correct for between-slice timing, 2) motion estimation. For all subjects, motion estimates were smaller than 1 mm and 1°, 3) spatial normalization of the functional images, re-interpolation to $3\times3\times3$ mm^3, and 4) smoothing (5 mm FWHM). This pre-processing was performed with the SPM2 software (see e.g. Ashburner et al., 2004). Datasets were also analyzed using the SPM2 software, using standard high-pass filtering and AR(1) whitening. For further analysis, the voxel-based estimated effects for several contrasts of interest were retained.

We determined a global brain mask for the group by considering all the voxels that belong to at least half of the individual brain masks defined with SPM2. It comprises approximately 60,000 voxels (this is the average size of individual brain masks). Note that considering the strict intersection of the individual masks yields about 34,000 voxels only and a large part of the brain–mostly cortical voxels!–is not included in the intersection mask. In what follows, the estimation procedures take this into account by considering that data is not available in some subjects. In such cases, mean signal and standard deviations are computed on the subsample of subjects that have data in this part of the mask. When necessary, appropriate corrections for the degrees of freedom are performed.

Elementary statistical description of the dataset

In this section, we select a few contrasts of interest, and study the statistical distribution of the corresponding parameters in each voxel. Using a first level, subject-specific, General Linear Model (GLM), one can obtain parametric estimates of the BOLD activity at each voxel in each subject: For each subject $s\in\{1, \ldots, S\}$ and each voxel $v\in\{1, \ldots, V\}$, we have a parameter estimate $\hat\beta(s, v)$, and a variance estimate $\hat\sigma^2(s, v)$.

The first question that may arise is whether the effects $\beta(s, v)$ are normally distributed or not, since this is a key assumption in standard (random effects) group analysis. We have used the D'Agostino–Pearson test (Zar, 1999), based on the computation of the skewness and the kurtosis (third and forth order cumulants) of the values $\{\hat\beta(s, v)\}$, $s=1\ldots S$ in each voxel v. This provides the p-value of the D'Agostino–Pearson statistic under the null (normal) hypothesis. For the sake of visualization, we convert the p-value into a z-value. We have then repeated the procedure based on the normalized effects $\left\{\tau(s,v)=\dfrac{\hat\beta(s,v)}{\hat\sigma(s,v)}\right\}$, $s\in\{1, \ldots, S\}$, $v\in\{1, \ldots, V\}$ which removes a potential variability in signal scaling across the population. At the group level, the normalization through the residual magnitude has a much greater impact than the deviation from normality on the resulting tests due to the fact that $\hat\sigma(s, v)$ is estimated with a finite ($\nu=100$) number of degrees of freedom.

Then, assuming a two-level normal model of the data

$$\hat\beta(s,v) = \beta(s,v) + \varepsilon(s,v), \text{ with } \varepsilon(s,v) \sim N(0, \sigma^2(s,v)) \quad (1)$$

$$\beta(s,v) = \bar\beta(v) + \zeta(s,v), \text{ with } \zeta(s,v) \sim N(0, v_g(v)) \quad (2)$$

where $\beta(s, v)$ is the true effect for subject s, $\hat\beta(s, v)$ is the estimated effect for subject s, and $\bar\beta(v)$ is the average effect in the population at voxel v; $\varepsilon(s, v)$ and $\zeta(s, v)$ are first-level (estimation) and second-level (inter-subject) normal residual terms. The first equation represents thus the subject-specific estimation of the signal and the second, the group-level model. We have estimated the second level variance v_g in each voxel, since it plays a central role in many group-level statistics. In particular, an interesting question is whether $\bar\beta(v)$ and $v_g(v)$ are independent or not. Note that v_g is estimated by maximizing the likelihood of the data given $\beta(s, v)$ and $\sigma(s, v)$. Newton or EM estimation schemes can be used (Worsley et al., 2002; Mériaux et al., 2006b). In this work, we use a Newton estimation scheme (see Appendix A).

Group-level analysis methods: voxel-based statistics

We review here different techniques used for voxel-based inter-subject activation detection. We consider a given contrast of interest. For each subject $s\in\{1, \ldots, S\}$ and each voxel $v\in\{1, \ldots, V\}$, we have a parameter estimate $\hat\beta(s, v)$, and a variance estimate $\hat\sigma(s, v)^2$.

A random effects (RFX) statistic is based on model (1) and (2), in which the first level variance is neglected. It is defined as

$$t(v) = \sqrt{S}\,\frac{\text{mean}_{s\in\{1;\cdots,S\}}\hat\beta(s,v)}{\sqrt{\text{var}_{s\in\{1;\cdots,S\}}\hat\beta(s,v)}} \quad (3)$$

Under the null hypothesis, assuming a normal distribution for $\{\hat\beta\ (s,\ v)\}$, s=1…S, $t(v)$ is t-distributed with $(S-1)$ degrees of freedom, and the p-value under the null hypothesis can be assessed with or without correction for multiple comparisons.[5] Alternatively, a non-parametric scheme can be used to estimate the distribution of $t(v)$ under the null hypotheses, based on milder assumptions (Hayasaka and Nichols, 2003; Mériaux et al., 2006a). In this work, we use the analytical threshold.

A mixed effects (MFX) statistic takes into account the first-level variance: assuming a group (or second-level) variance $v_g(v)$ at each voxel v, the MFX is the quotient of the group mean $\bar\beta = \sum_{s=1}^{S} \frac{\bar\beta(s)}{\sigma^2(s)+v_g}\left(\sum_{s=1}^{S}\frac{1}{\sigma^2(s)+v_g}\right)^{-1}$ [see Eq. (11) in Appendix A] by its standard deviation $\sqrt{\sum_{s=1}^{S}\frac{1}{\sigma^2(s)+v_g}}$ In a Bayesian setting (Beckmann et al., 2003), these quantities can be termed the posterior mean and variance. Thus the MFX statistic is written as:

$$\mu(v) = \sum_{s=1}^{S}\frac{\hat\beta(s,v)}{\hat\sigma(s,v)^2+v_g(v)}\left(\sum_{s=1}^{S}\frac{1}{\hat\sigma(s,v)^2+v_g(v)}\right)^{-\frac{1}{2}} \quad (4)$$

Intuitively, MFX may perform better than RFX since it down-weights the observations with high first-level variance. The

[5] Given our definition of the group mask, it may occur that functional data is available in a sub-sample of the population of size n, with $\frac{S}{2}\le n\le S$. In such a case, S is replaced by n in the formulas. In the present work, we systematically apply such corrections.

distribution of the quantity $\mu(v)$ under the null hypothesis is difficult to assess (Woolrich et al., 2004). We rely on an non-parametric scheme as in Mériaux et al. (2006a,b): we tabulate the values of $\mu(v)$ for different sign swaps of each subject's dataset in order to generate a distribution under the null hypothesis, and compare the actual values with their estimated null distribution. A quicker but very conservative approximation ($\mu \sim t_{S-1}$, t_{S-1} being the Student law with $S-1$ degrees of freedom) is also possible.

One can also construct another statistic by neglecting the group variance v_g in Eq. (4). This yields a pseudo-MFX statistic, which is just a weighted average of the effects of the subjects. We denote it henceforth as ΨFX:

$$\Psi(v) = \sum_{s=1}^{S} \frac{\hat{\beta}(s,v)}{\hat{\sigma}(s,v)^2} \left(\sum_{s=1}^{S} \frac{1}{\hat{\sigma}(s,v)^2} \right)^{-\frac{1}{2}} \tag{5}$$

Note that this is the statistic proposed in Neumann and Lohmann (2003). The difference is that we assess the value of ΨFX through a frequentist approach by estimating the distribution of $\Psi(v)$ under the null hypothesis by random sign swaps of the individual data (which we refer to as a non-parametric approach), exactly as we do for the MFX statistic. In that case it is necessary to use a voxel-based assessment of the statistic value (i.e. voxels may not be exchangeable under the null hypothesis).

We also have used Wilcoxon's signed rank statistic (WKX) (Hollander and Wolfe, 1999), which sorts the absolute effects in ascending order, then sums up the ranks modulated by the corresponding effect's sign:

$$W(v) = \sum_{s=1}^{S} \operatorname{sign}(\hat{\beta}(s,v)) \operatorname{rank}(\hat{\beta}(s,v)) \tag{6}$$

The behaviour of this statistic under the null hypothesis is data-independent, thus its significance is assessed very easily. Unlike the previous statistics, it does not assess the positivity of the average effect, but the asymmetry of the estimated effects $\hat{\beta}(s, v)$ with respect to 0, the null hypothesis being that $\hat{\beta}(s, v)$ are distributed symmetrically about 0. The main interest of this statistic is that it is not based on the hypothesis that the $(\hat{\beta}(s, v))$, $s=1\ldots S$ are normally distributed.

Group-level analysis methods: higher-level statistics

Higher-level or non-voxel-based analyses statistical inference methods include cluster-based inference and parcel-based inference.

Cluster-based inference (Hayasaka and Nichols, 2003; Mériaux et al., 2006a) is simply an extension of the voxel-based procedures, based on a double thresholding of a statistic map: first, a threshold is performed at the voxel level, then supra-threshold clusters are kept whenever their size is statistically significant. In our implementation, we measure connectivity using the 18-nearest neighbours of each voxel in 3D, and estimate the p-values at the cluster-level using the non-parametric framework.

Parcel-based inference is a different scheme in which parcels are defined across subjects using anatomical and/or functional information. Two possible schemes have been presented in Flandin et al. (2002), Flandin (2004) and Thirion et al. (2006a), based on a Gaussian Mixture Model (GMM) and a hierarchical approach respectively. A key issue of both techniques is to obtain functionally and spatially connected parcels that adapt to the subject's anatomical or functional variability. Statistics, such as the t statistic, can then be computed at the parcel level by working on parcel-based signal average instead of voxel-based signal (PRFX statistic). The advantage is that some spatial relaxation is possible in the definition of the parcels, allowing for a better spatial registration of functional information. Care must be taken when the same functional data is used to build the parcels and perform the test to control appropriately for the false positives rate (Thirion et al., 2006a). Here we assess the reliability of PRFX maps and compare it to other techniques.

Assessing the reliability of activation maps

In this work, we propose two measures to assess the reliability of the activation maps derived from group analysis. The first, based on a mixture of binomial distributions, characterizes the stability of the status (active/inactive) of each voxel of the dataset. The second measures how frequently clusters of voxels are found at similar locations in the normalized MNI/Talairach space across subjects. We use these measures in a bootstrap framework that enable us to characterize the reproducibility of activation maps obtained at the group level.

Reliability measure at the voxel level

In order to estimate the reliability of a statistical model, we need a method to compare statistical maps issuing from the same technique, but sampled from different groups of subjects. We use the reliability indexes elaborated in Genovese et al. (1997) and Liou et al. (2003, 2005). Assume that a statistical procedure (e.g. thresholding) yields binary maps $g_1, \ldots, g_R$ for different groups of subjects. At each voxel v, an R-dimensional binary vector $[g_1(v), \ldots, g_R(v)]$ is thus defined. At the image level, the distribution of $G(v) = \sum_{r=1}^{R} g_r(v)$ is modelled by a mixture of two binomial distributions, one for the null hypothesis, one for the converse hypothesis: Let π_A^1 be the probability that a truly active voxel is declared active, $\pi_A^0 = 1 - \pi_A^1$ the probability that a truly active voxel is declared inactive, π_I^1, the probability that a truly inactive voxel is declared active, $\pi_I^0 = 1 - \pi_I^1$ the probability that an truly inactive voxel is declared inactive, and λ the proportion of truly activated voxels. Then, using a spatial independence assumption, the log-likelihood of the data is written as

$$\begin{aligned}\log(P(G)|\lambda, \pi_A^0, \pi_I^0) = \; & cst + \sum_{v=1}^{V} \log(\lambda(\pi_A^0)^{R-G(v)}(\pi_A^1)^{G(v)} \\ & + (1-\lambda)(\pi_I^0)^{R-G(v)}(\pi_I^1)^{G(v)})\end{aligned} \tag{7}$$

Assuming $R \geq 3$ the three free parameters, π_A^0, π_I^0, λ can be estimated using EM or Newton's methods. Note that optimizing the model over its different parameters sequentially, and using an adequate initialization, we could run the model for $R=2$, though with higher variability in the estimation. An example of mixture of binomial distributions is given in Fig. 2.

Given these estimates, the coherence index κ, known as Cohen's kappa is computed to measure the concordance of the different observations with the mixture model. Let $p_0 = \lambda\pi_A^1 + (1-\lambda)\pi_I^0$ be the fraction of voxels that are correctly classified by the mixture model. p_0 should be compared to the fraction of correct classifications that occur by chance $p_C = \lambda\pi^0 + (1-\lambda)(1-\pi^0)$, where $\pi^0 = \lambda\pi_A^0 + (1-\lambda)\pi_I^0$ is the proportion of voxels declared

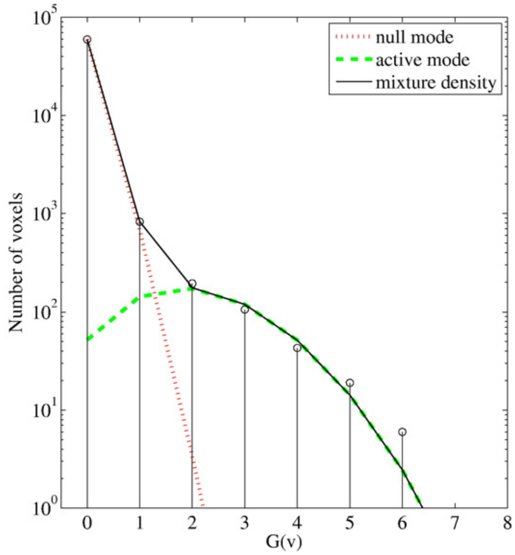

Fig. 2. Example of mixture of binomial distributions. The empirical histogram of $G(v)$ is modelled by the model in Eq. (7), with $R=8$. The Yaxis is in log-coordinates for the sake of clarity.

inactive. The fraction of correct classifications corrected for chance is thus

$$\kappa = \frac{p_0 - p_C}{1 - p_C} \tag{8}$$

In this setting, $0 \leq \kappa \leq 1$ measures the fit of the mixture model to the data, which in turns reflects the concordance of the binary maps given as input to the model (7). If there is very little agreement on which voxels are active, the components of the mixtures have a strong overlap, and κ is close to 0, whereas the separation between the components of the mixture increases and κ is close to 1 if there is a good agreement between binary maps. For instance, $\kappa=0.45$ for the data presented in Fig. 1. λ can also be retained as an index of the test sensitivity.

Note that more complex–and realistic–models have been proposed in the literature (Maitra et al., 2002), in which the parameter λ is allowed to vary spatially. However, our main purpose is not activation detection, but obtaining a global reliability measurement; for this reason, we keep the basic setting.

Reliability measure at the cluster level

Another way to assess the reliability of the results is to compare the positions of the clusters of supra-threshold voxels that arise through any group analysis. Assuming that the binary maps $g_1, \ldots, g_R$ are obtained from different groups of subjects through a thresholding procedure, one can post-process them in order to yield connected components. The connected components with a size greater than a given threshold η are then retained, and their centre of mass (*cm*) is computed: let x_i^r, $i=1 \ldots I\ (r)$ be the spatial coordinates of the *cms* derived from map g_r, we propose the following average distance between any two maps:

$$\Phi = \frac{1}{R(R-1)} \sum_{r=1}^{R} \sum_{s \in \{1; \cdots, R\} - \{r\}} \frac{1}{I(r)} \sum_{i=1}^{\mathrm{I}(r)} \min_{j \in \{1; \cdots, I(s)\}} \varphi(\|x_i^r - x_j^s\|), \tag{9}$$

where $\varphi(x) = 1 - \exp\left(-\frac{x^2}{2\delta^2}\right)$ is a penalty function that is close to zero when the cluster centroids are properly matched and close to 1 otherwise. Φ represents the average mismatch between the *cm* of a supra-threshold component in a given map and the closest cm of any supra-threshold cluster obtained from another map. Appropriate penalty terms are used to handle the case $\mathrm{I}(r)=0$. We have performed some experiments using $\eta=10$ voxels or $\eta=30$ voxels, and use $\delta=6$ mm.

Procedure for the assessment of reliability

The procedure consists in dividing the population of 81 subjects in $R=2$, 3, 4, 5, 6 or 8 disjoint groups of $S=40$, 27, 20, 16, 13 and 10 subjects respectively. The computation of different statistics, the derivation of an adequate threshold and the thresholding are performed in the different subgroups, and global reliability measures are derived from the ensuing binary maps. This procedure is repeated 100 times for each instance, yielding a distribution of the indexes κ, λ and Φ for each possible technique/parameter.

First, we choose the traditional RFX analysis procedure [see Eq. (3)], thresholded at $p<0.001$, uncorrected using an analytical threshold and evaluate the distribution of the different indexes for three contrasts of interest. This is important to understand how well the indexes are characteristic of the amount, the spread and the variability of supra-threshold activity. In particular, it is important that the estimated reliability indexes are less variable for a given contrast than across contrasts.

Second, we evaluate the choice of the threshold on the different indexes, in the case of the voxel-based *t*-test. While the sensitivity index certainly decreases while the threshold increases, the behaviour of the reliability may be more complex, due to the trade-off between false positive and false negative rates (non-standard behaviours due to extremely low or high thresholds are not considered here).

Third, we study the behaviour of the different measurements when the number of subjects in the group varies; while it is obvious that reliability increases with the group size, it is not clear whether there exists a plateau and at which level. Previous studies (Desmond and Glover, 2002; Murphy and Garavan, 2004) suggest a steady increase of sensitivity with the group size.

Finally, we choose the following statistics: RFX, RFX on smoothed (12 mm FWHM instead of 5 mm) effect maps (SRFX), MFX, Wilcoxon(WKX), Cluster-level RFX (CRFX), Parcel-based RFX (PRFX) and ΨFX. RFX, SRFX, MFX, ΨFX and PRFX maps are thresholded at the $p<0.001$ level, uncorrected for multiple comparisons. CRFX is thresholded at $p<0.01$, uncorrected level at the voxel level, then at $p<0.01$, at the cluster level. Note that these choices are made in order to roughly balance the specificity of the methods, while using them in a standard way.

PRFX maps are computed for $Q=500$ parcels. Since the parcel centres are defined at the group level in Talairach space, the voxels in the group result map are assigned to the parcel with the closest centre in Talairach space. This results in a piecewise constant map, the pieces resulting from a Voronoi parcellation of the group mask into parcels. Note that in our bootstrap procedure, such boundaries are defined independently in each subgroup of subject. For parcellation, we use the hierarchical procedure presented in Thirion et al. (2006a) and a number of parcels optimized according to cross validation (Thyreau et al., 2006).

Results

Statistical model of the inter-subject data

We performed the D'Agostino–Pearson test on the effects $\hat{\beta}(v)$ of all the voxels, as well as the normalized effects $\frac{\hat{\beta}}{\hat{\sigma}}(v)$, which yields two maps for each contrast. We present them for *left–right button press*, *audio instructions–video instructions* and *computation–reading*, thresholded at the $p<0.001$ uncorrected level. We also present the inter-subject variance maps $v_g(v)$ computed in a mixed-effect model (see Appendix A). We present these maps together with the RFX map (converted to a z-variate) based on 81 subjects in Figs. 3–5. Note that other contrasts, e.g. *horizontal–vertical checkerboards, sentence reading–low-level vision, cognitive trials–motor trials*, and the opposite ones, not presented here due to space limitations, yield qualitatively similar results.

In each case, the regions with highest group variance are found in the regions with highest random effects statistics in absolute values; some of them are absent in the maps 3–5, where signed statistics are presented.

Inspection of these maps suggests that

- Areas of high variance tend to co-localize with the activated areas. This implies that the parameters $v_g(v)$ and $\overline{\beta}(v)$ are certainly not independent, and that statistics that are penalized by the group variance may not be very efficient in general.
- Non-normality is very significant in wide regions of the brain: deviation from normality of $\hat{\beta}$ across subjects concerns 22% of the brain voxels at ($p<0.001$, uncorrected) for the *computation–reading* contrast, 27% for the *left–right button press* contrast and 30% for the *audio instructions–video instructions* contrast.
- Deviation from the normality hypothesis is much lower for the normalized effects $\tau=\frac{\hat{\beta}}{\hat{\sigma}}$ than for the raw effects $\hat{\beta}$. For instance, the rate of voxels with normality rejected at ($p<0.001$, uncorrected) drops from 22% to 9.2% for the *computation–reading* contrast, from 27% to 2.9% for the *left–right button press* contrast and from 30% to 10% for the *audio instructions–video instructions* contrast. This means that dimensionless first-level statistics yield more homogeneous quantities across subjects than effects expressed in percents of baseline signal increase.
- Deviation from normality of the effects does not specifically co-localize with activated areas, but, in several cases it coincides with the boundaries of activated areas.

Reliability measurements for different cognitive contrasts

We computed the random effects z-variate for different cognitive contrasts, using $R=5$ groups of $S=16$ and a threshold

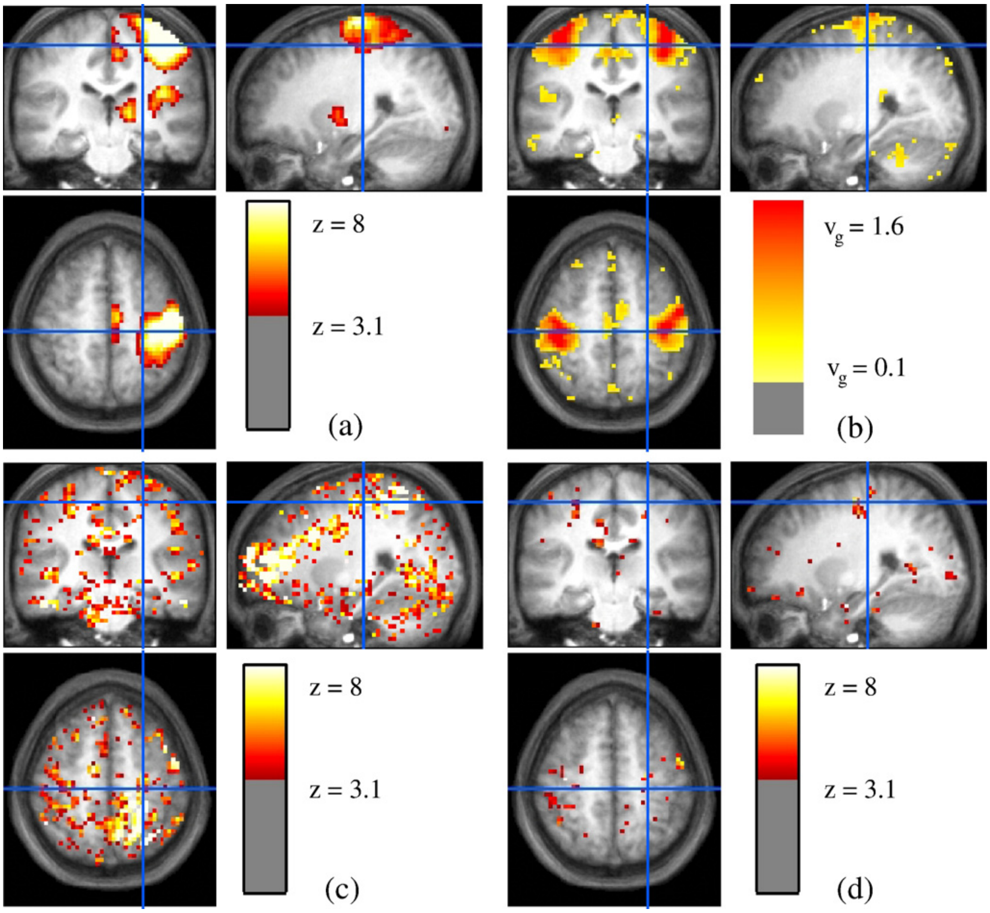

Fig. 3. Statistical model of the effects for the *left–right button press* contrast, on $S=81$ subjects. (a) z-value associated with the RFX test; (b) group variance estimate; (c) z-value of the D'Agostino–Pearson test for normality of the effects $\hat{\beta}$; (d) z-value of the D'Agostino–Pearson test applied to the normalized effects $\tau=\frac{\hat{\beta}}{\hat{\sigma}}$. Note that all the z values are limited to the $[-8, 8]$ range. The color scale of the variance image has been chosen arbitrarily in order to have supra-threshold areas that are comparable with the other maps. The variance is expressed in squared percentage of the BOLD mean signal. Cross position: $(-23, -28, 56)$ mm in the MNI space.

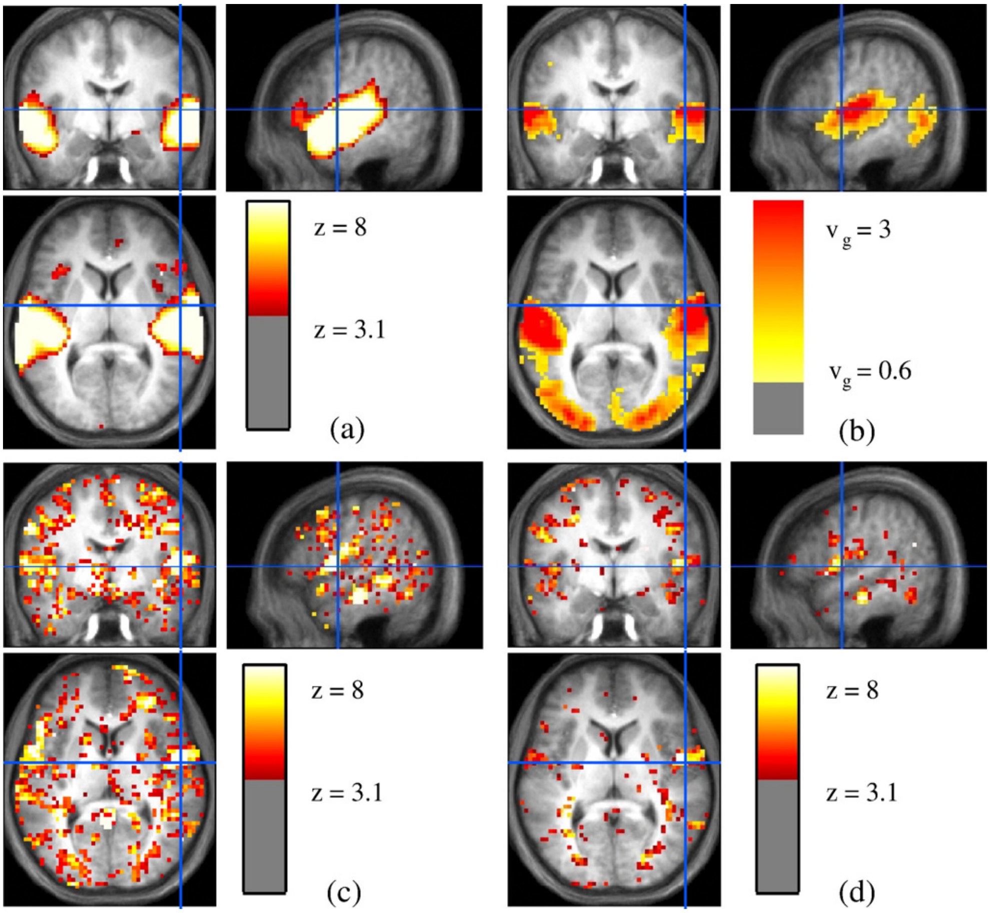

Fig. 4. Statistical model of the effects for the *audio instructions–video instructions* contrast, on S = 81 subjects. (a) z-value associated with the RFX test; (b) group variance estimate; (c) z-value of the D'Agostino–Pearson test for normality of the effects $\hat{\beta}$; (d) z-value of the D'Agostino–Pearson test applied to the normalized effects $\tau = \frac{\hat{\beta}}{\hat{\sigma}}$. Note that all the z values are limited to the [−8, 8] range. The color scale of the variance image has been chosen arbitrarily in order to have supra-threshold areas that are comparable with the other maps. The variance is expressed in squared percentage of the BOLD mean signal. Cross position: (−54, −6, 8) mm in the MNI space.

$\theta = 3.1$ corresponding to $p < 0.001$ uncorrected for the contrasts *left–right button press, audio instructions–video instructions and computation–reading*. The reliability index κ, the proportion of putative true positives λ, and the inter-cluster distance penalty Φ are given in Fig. 6. It shows that κ and λ have different behaviours and are strongly dependent on the cognitive contrast under study. For instance, the left motor contrast activates relatively small regions with a relatively low reliability; the auditory-selective contrast activates larger regions with high reproducibility; the computation-selective contrast activates larger regions, but with low reliability. The inter-cluster distance penalty Φ does not discriminate between the different contrasts as strongly as κ. As could have been expected, it has the opposite behaviour (maximal for the computation contrast, minimal for the auditory contrast).

How the threshold affects the reliability of the analysis

Here we study the behaviour of our reliability measures when applied to a thresholded RFX map, when we let the threshold vary. The reliability measure is computed for 100 different splits of the population of subjects into $R = 5$ groups of $S = 16$ subjects, in the case of the *left–right button press* contrast. The threshold (in z-variate scale) varies from $\theta = 2.2$ ($p < 0.015$, uncorrected) to $\theta = 4$ ($p < 3.2 \times 10^{-5}$, uncorrected) in steps of 0.2.

As expected, the sensitivity parameter λ decreases when θ increases (see Fig. 7(b)). More interestingly, κ reaches a maximum for $\theta^* \sim 2.7$, but the index remains close at least for $\theta < 3.5$ as can be seen in seen in Fig. 7(a). Accordingly, the inter-cluster distance penalty Φ is minimized for a threshold $\theta^* \sim 3$. The correspondence of these results is interesting, given that these two similarity measures are obtained independently, and based on different considerations. Note that we have obtained similar results when studying the other contrasts with slightly higher (auditory contrast) or lower (computation contrast) threshold values. Thereafter, we retain the threshold $\theta = 3.1$ ($p < 0.001$, uncorrected for multiple comparisons) for random effects *z*-statistics.

How many subjects are necessary to obtain a reliable group map

We study the dependence of κ, λ and Φ when we let the size S of the group vary. We base our investigation on the *left–right button press* contrast, with group maps thresholded at the $\theta = 3.1$ ($p < 0.001$, uncorrected) level. The results are presented in Fig. 8. It shows that the reliability increases with the group size, which was expected. The sensitivity also increases with the group size. Interestingly, the

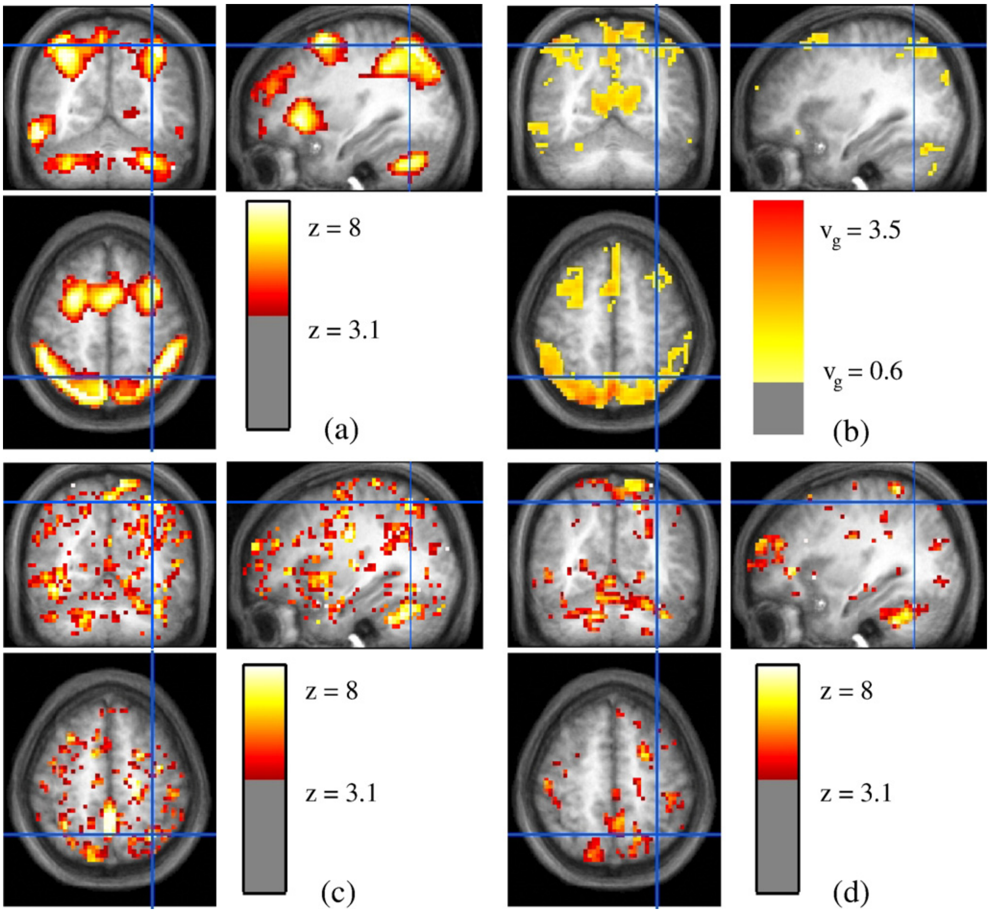

Fig. 5. Statistical model of the effects for the *computation–reading* contrast, on $S=81$ subjects. (a) *z*-value associated with the RFX test; (b) group variance estimate; (c) *z*-value of the D'Agostino–Pearson test for normality of the effects $\hat{\beta}$; (d) *z*-value of the D'Agostino–Pearson test applied to the normalized effects $\tau=\frac{\hat{\beta}}{\hat{\sigma}}$. Note that all the *z* values are limited to the $[-8, 8]$ range. The color scale of the variance image has been chosen arbitrarily in order to have supra-threshold areas that are comparable with the other maps. The variance is expressed in squared percentage of the BOLD mean signal. Cross position: $(-33, -60, 56)$ mm in the MNI space.

reliability reaches a plateau only for $S\approx 25$. The inter-cluster distance penalty Φ has a similar behaviour, with a plateau for $S=27$ subjects when $\eta=10$, while lower values are reached when using $\eta=30$.

Comparison of different group analysis methods

Now we study how the reliability index behaves for different statistical methods: The *t* statistic [RFX, see Eq. (3)], the same test after 12 mm smoothing of the data–instead of 5 mm–(SRFX), the mixed effects statistic, controlled by permutation [MFX, see Eq. (4)], the parcel-based RFX test (PRFX), the *t*-statistic thresholded at the cluster-level (CRFX), the Wilcoxon test (WKX), and the pseudo-MFX test ΨFX. RFX, SRFX, MFX, WKX, PRFX and ΨFX maps are thresholded at the $p<0.001$ level, uncorrected for multiple comparisons. The CRFX map is first thresholded at the $p<0.01$, uncorrected level, then at the $p<0.01$ cluster-level. The results are obtained by bootstrapping in $R=8$ groups of size $S=10$. The results are presented in Fig. 9.

From the point of view of reliability, the WKX and RFX tests have the worst performance overall, while the SRFX performs slightly better. CRFX, PRFX and MFX techniques yield higher reliability, but ΨFX yield the highest values. The results are more variable with PRFX than with other techniques; this reflects the fact that PRFX is based on a smaller number of volume elements, so that statistical tests have a less stable behaviour purely due to fewer number of parcels compared to voxels.

CRFX, MFX, and to a lesser extent, PRFX tests are more sensitive, i.e. have a larger fraction of generally activated voxels, than voxel-based tests. Note however that the specificity control of CRFX matches the other approaches only approximately.

Finally, the average supra-threshold cluster distance Φ is minimal for ΨFX, and relatively low for MFX. It is similar for the other techniques.

Discussion

Normality and second-level variance

From Figs. 3–5, one of the most striking effects is the co-localization of high second-level variance areas with large random effects areas. Numerically, such an effect is not expected since the RFX is defined as the quotient of the estimated mean effect by the standard deviation of this estimate.

The interpretation could be that 1) the contrast-to-noise ratio (CNR) of the BOLD effect is highly variable across subjects, and by definition this effect does not appear in non-activated areas and/

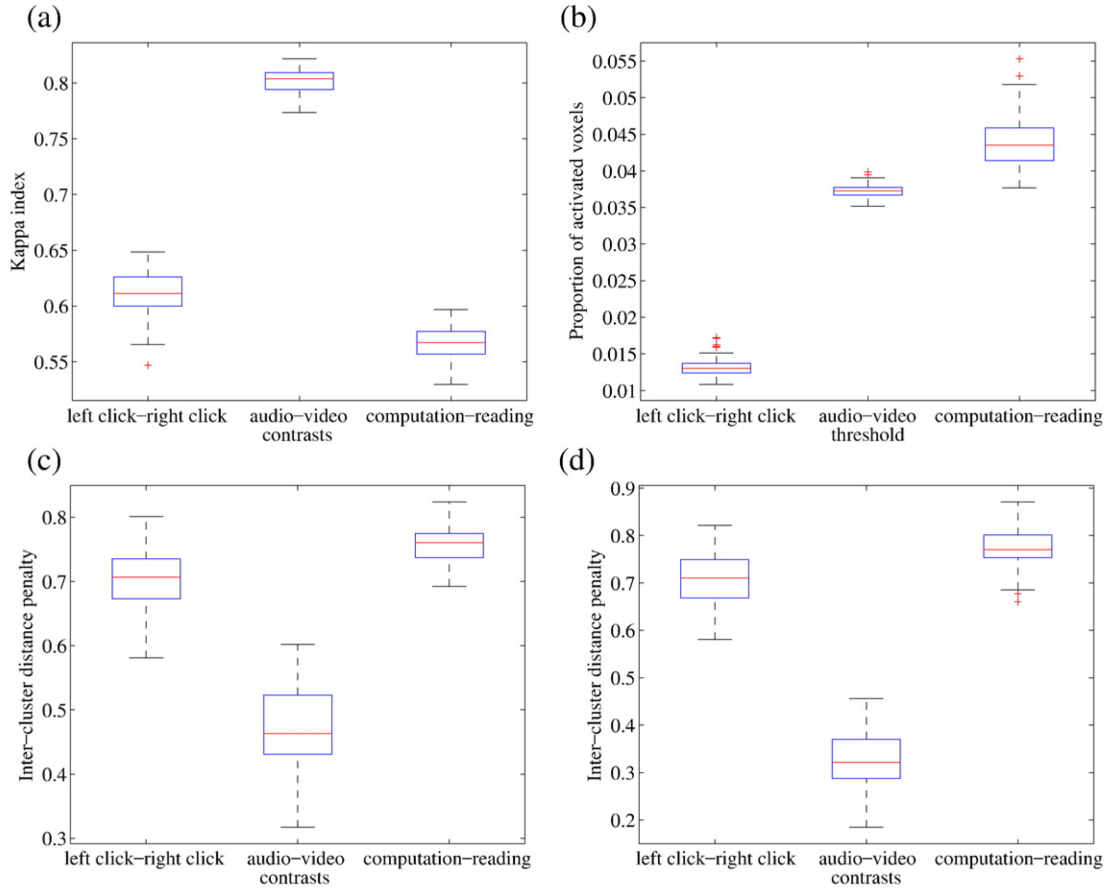

Fig. 6. Dependence of the reproducibility and of the sensitivity of the random effects analysis on the functional contrast under consideration. These results are obtained by drawing 5 disjoint groups of $S=16$ subjects in the population of 81 subjects, and applying the procedure described in the *Reliability measure at the voxel level* section. The threshold is $\theta=3.1$ ($p<0.001$). (a) Over 100 replications, the reliability index is higher for the *audio instructions–video instructions* contrast than for the *left–right button press* and *computation–reading* contrast. (b) However, the size of the putatively activated areas is greater for the contrast that shows regions involved in computation, and smaller for the contrast that shows the regions involved in motor activity. (c–d) The cluster variability penalty Φ is presented for clusters of more than $\eta=10$ (c) or $\eta=30$ (d) voxels (the lower the better). The behaviour is as expected, with the smallest value for the auditory-specific contrast.

or 2) spatial mis-registration[6] implies that at a given voxel, i.e. a given position in MNI space, some subjects have activity while other subjects do not, thus spatially widening the signal distribution. For simple contrasts such as those used (left or right button press, sentence listening), different cognitive strategies should be ruled out.

This inflated variance effect certainly deserves more investigation, given its prominent effect on statistics (sensitivity and reliability): for instance, the ΨFX statistic–that does not take into account the group variance, hence is simply a weighted average of the subject-based effects–seems more reliable than the MFX statistic, which is itself much more reliable than the RFX statistic (see Fig. 9). The effect of group variance is also an argument in favour of Bayesian analysis of fMRI data, if the reference signal level is not 0 (Friston and Penny, 2003).

Non-normality is another important factor. To our knowledge, this has not been investigated before, since it requires a high number of subjects. Interestingly, the importance of non-normality is reduced when considering normalized effects $\tau(s, v)$ instead of raw effects $\hat{\beta}(s, v)$. This shows that first-level statistics can play an important role in group statistics. In particular, the difference observed between the normality of τ and $\hat{\beta}$ maps possibly indicates that the current way of normalizing signal magnitude with respect to the mean signal may not be optimal for inter-subject comparison (this is also an open question for inter-session variability). However, the normalization with respect to first-level variance might not be satisfactory, since it could in turn be highly dependent on acquisition artifacts, motion and physiology, whether these are modelled of not. We are not aware of any successful signal calibration strategy, but mixed-effects model may solve part of the problem. Interestingly, several areas with significant non-normality are found at the periphery of activation maxima, confirming the impact of spatial shifts on group statistics. Once again, further investigations on non-normality may be performed, e.g. searching different groups of subjects in the population or outlier subjects (see Kherif et al., 2004). Robust statistics might also be used for inference (Wager et al., 2005), but at the risk of a weaker control on specificity. Moreover, such inference schemes raise the difficult question of the generalizability of group results to other groups of subjects (given that the concept of *outlier* is ill-defined when considering a small group). In general, it is advisable to use non-

[6] Spatial mis-registration may be artefactual (incorrect normalization) or not (intrinsically different functional anatomy).

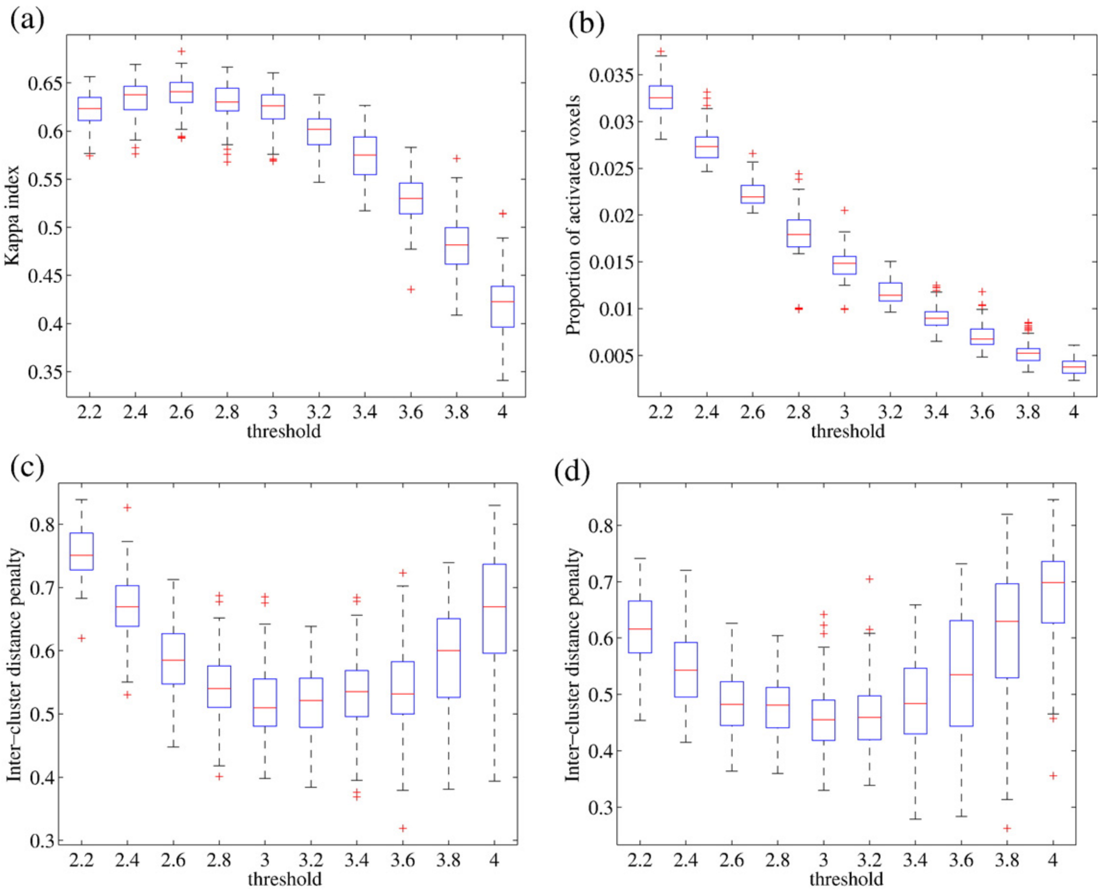

Fig. 7. Dependence of the reproducibility, the sensitivity, and the distance between supra-threshold clusters of the group random effects analysis on the threshold chosen to binarize the statistic maps. These results are obtained by drawing 5 disjoint groups of $S=16$ subjects in the population of 81 subjects, and applying the procedure described in the *Reliability measure at the voxel level* section. This is performed on the images of the *left–right button press* contrast, with 100 resamplings. (a) The reproducibility index κ shows is maximized for $\theta \sim 2.7$. (b) The sensitivity decreases when θ increases. (c,d) The average distance between supra-threshold clusters of more than 10 (c) or 30 (d) voxels across groups has a minimum around $\theta \sim 3$.

parametric assessment to obtain reliable thresholds (Mériaux et al., 2006a). However, the choice of robust statistics (statistics that adapt to non-normal data) is not necessarily advantageous: for instance, the Wilcoxon statistic did not perform better than other statistics in our experiments (see Fig. 9).

Measuring the reliability of group studies

The reliability of an activation pattern measures how systematically a given voxel or region will be found when performing a group study in a particular group of subjects. Taking advantage of the great number of subjects, we have used a bootstrap procedure and two measures for assessing the reliability of the group studies: one models the activated/non-activated state of voxel as a mixture of binomial distributions, and quantifies the difference between the null and the active mode, while the other defines how well clusters of supra-threshold activity match across groups.

The first criterion has already been proposed in the literature; it has the advantage of yielding very stable results across splits (see Figs. 6 and 9). One reason is that all the R groups are used in each single computation of the parameters, while the cluster-based measure is based on pairwise comparisons. However, care must be taken because the estimation may come trapped in local minima (although we have never observed convergence problems in our experiments), or because the joint estimation of the different parameters may imply some non-trivial interaction between the parameters (e.g. the sensitivity λ might not be independent from κ; across splits there is on average a negative correlation of around -0.3 between κ and λ, which is significant). More importantly, results at the voxel level are not as important as the presence of a strong local maximum or a significant cluster, which deserve being reported.

This has incited us to develop a second measure [see Eq. (9)], which takes into account only extended clusters and compares the position of their centres of mass, a measure related to the study of (Murphy and Garavan, 2004). Note that the penalty function Φ stabilizes to $\Phi \simeq 1$ as soon as the distance exceeds 12 mm. Cluster centres that are separated by 20 mm are no more likely homologous than clusters whose centres are separated by 50 mm. (this is true because we are reporting group results; when reporting individual results, greater variability might be allowed). Averaging across supra-threshold clusters yields an idea of how frequently close clusters will be obtained across groups of subjects. This pairwise measure is somewhat more variable than the voxel-based indexes, but it yields an independent confirmation of possible differences in reliability.

As reported, the bootstrap dispersion depends strongly on the contrast studied, confirming the appropriateness of these measures

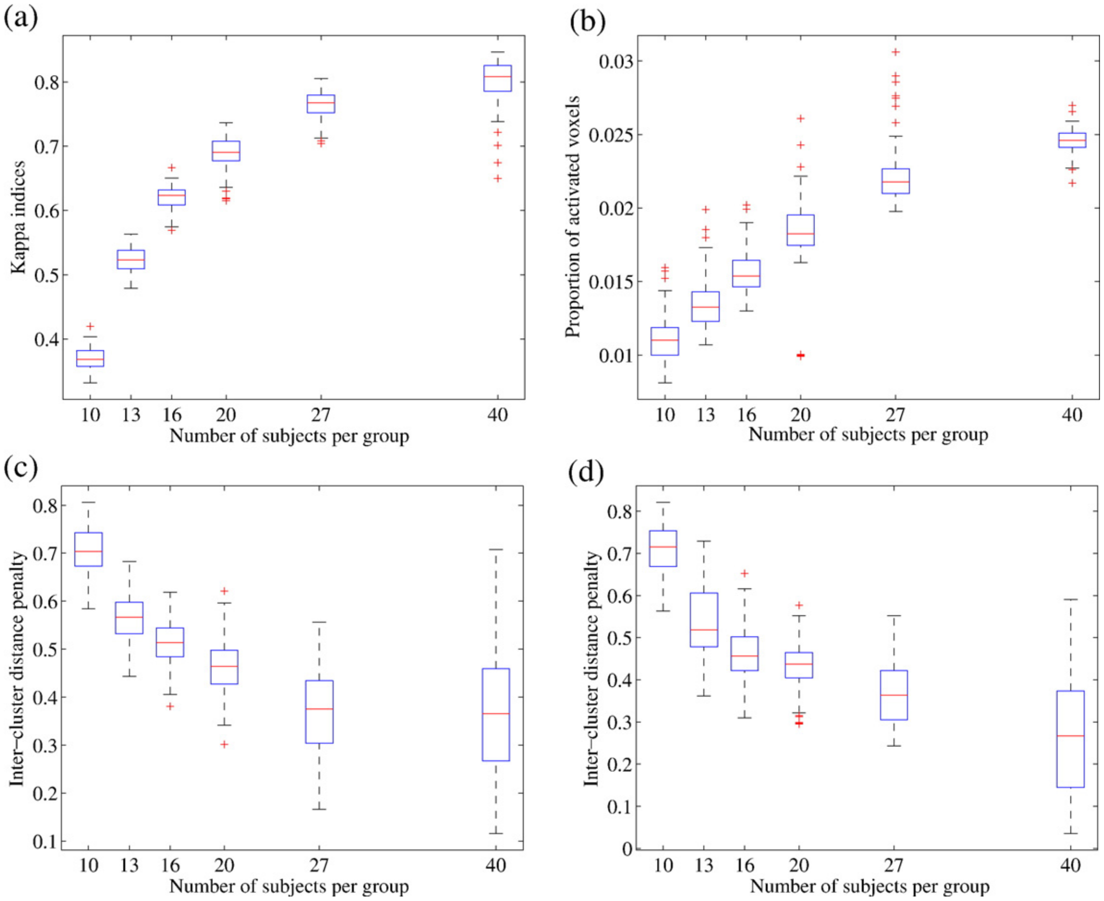

Fig. 8. Effect of the RFX group size on reproducibility κ (a), sensitivity λ (b) and the average distance between supra-threshold cluster centroids Φ (c–d). The reliability is assessed considering disjoint groups of size $S=10$, 13, 16, 20, 27, 40 within the population of 81 subjects. This is performed on the images of the *left–right button press* contrast, with 100 resamplings. (a) The reproducibility index increases with S and reaches a plateau for $S>20$. (b) The size of putatively activated areas steadily increases with S. (c–d) The average intra-cluster distance decreases with S; it reaches a plateau for $S>20$ when $\eta=10$ (c), whereas it further decreases when $\eta=30$ (d).

(Fig. 6). In general terms, κ and Φ have a similar behaviour (κ is high when Φ is low and vice versa). This was not obvious, given that the two measures are independent and based on completely different approaches. It suggests that our results reflect intrinsic features of data.

Our setting for the study of the reliability may also be used to compare competing pre-processing techniques or analysis frameworks, in addition to previous contributions based on cross-validation (Strother et al., 2002) and information theory (Kjems et al., 2002). The main difference between our approach and the cross-validation scheme from (Strother et al., 2002) is that:

- The analysis is univariate (based on one map) in our case, while it was multivariate in Strother et al. (2002), with a dimension reduction of the data. Though the interpretation of univariate results is conceptually simpler, it may not generalize to parametric designs such as those used in LaConte et al. (2003), Shaw et al. (2003), and Strother et al. (2004).
- The reproducibility measure used in Strother et al. (2004) is map-based correlation, whereas we compare supra-threshold areas. This is an advantage, since only the supra-threshold areas are of interest, but introduces an artefactual dependence on the threshold.
- In Strother et al. (2002), two-fold reproducibility is considered, while we need an R-fold splitting of the group with $R>2$ (although the estimation procedure still converges for $R=2$). Our method requires a large database of subjects, but is more general.

Is there a best threshold?

The fundamental question of finding an optimal threshold to label areas as activated has rarely been addressed, since it requires the modelling at the voxel level of both the null and the alternative hypothesis to control both the false positive and false negative rates. This is possible here, thanks to the large number of subjects.

Interestingly, we find a relatively low value for the optimal threshold ($\theta^* \sim 2.7$ when considering κ, $\theta^* \sim 3$ when considering Φ; note that these two measures are independent). The corresponding p-values (0.0035–0.001, without correction) are not conservative, so that such thresholds do not allow a very strict control of the rate of false positives. Interestingly, such thresholds are often used in the literature. It is possible that researchers through trial and error converged to this value.

Family-wise error control procedures such as Bonferroni, Random Field Theory (Ashburner et al., 2004), and, to a lesser

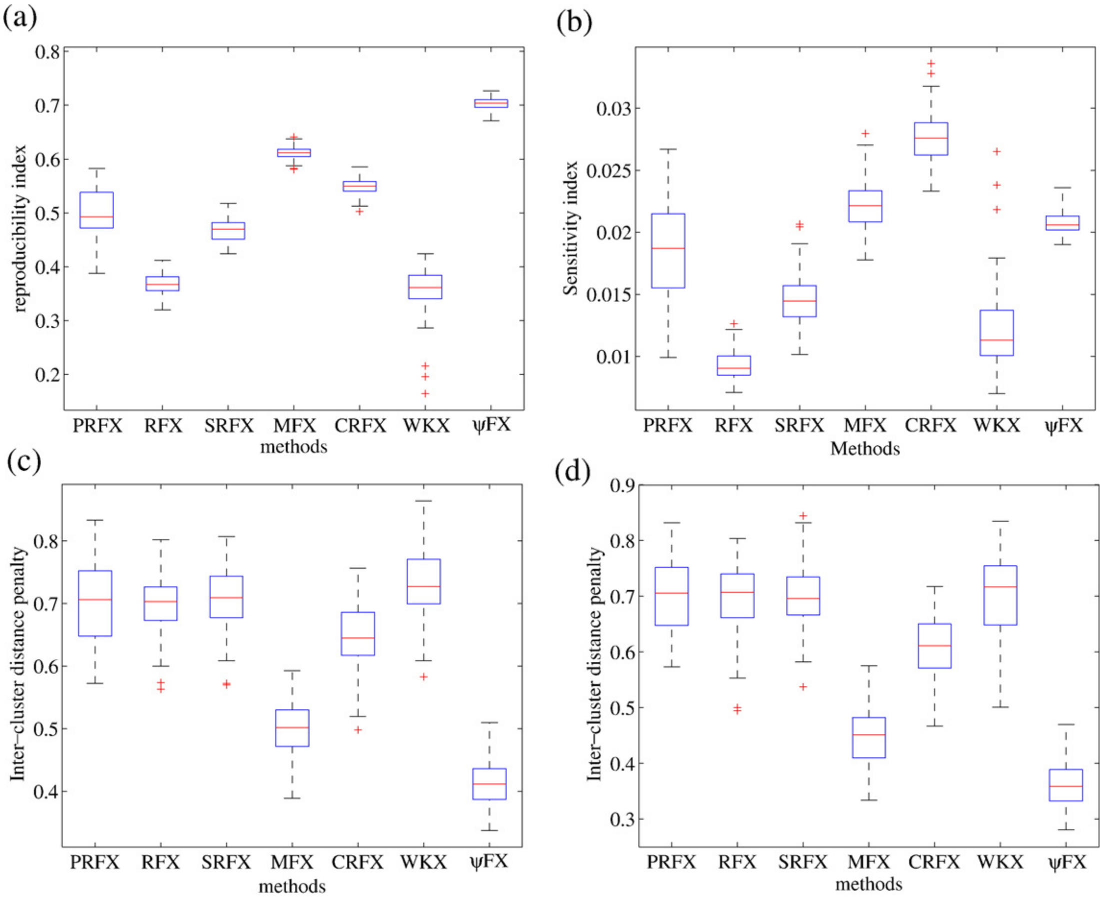

Fig. 9. Dependence of the reliability κ (a), the sensitivity λ (b), inter-supra-threshold cluster distance penalty Φ (c–d) of the statistical analysis on the group statistic used. Φ is based on clusters of size greater than $\eta=10$ (c) or $\eta=30$ (d). These quantities assessed considering $R=8$ disjoint groups of size $S=10$ within the population of 81 subjects, using the *left–right button press* contrast, and 100 resamplings.

extent, False Discovery Rate (Genovese et al., 2002), typically require the use of much higher thresholds. In this study, we have chosen a relatively lenient threshold $p<10^{-3}$ uncorrected because specificity control was not our main point. However, a good compromise between the control of false positives and the reliability may be the use of cluster-level or parcel-level inference. It is important that those will control for the number of brain regions reported, and not only for the number of voxels.

We obtained very similar results with higher SNR functional contrasts such as the auditory contrast. The optimal threshold was slightly higher, between 3 and 3.5 (in a z scale). However, it is not obvious that our results generalize to datasets with different structure and our point is certainly not to justify lenient thresholding procedures. Nevertheless, the question of an optimized threshold for reproducibility (accounting for both false positive and false negative) could be addressed more systematically in neuroimaging studies.

How does the sample size affect the reliability of the results?

Another fundamental question concerns the number of subjects that should be included in a study. The point here is not only the sensitivity (Desmond and Glover, 2002; McNamee and Lazar, 2004), but also the reliability (Murphy and Garavan, 2004) of the results. Our results clearly indicate that $S=20$ is a minimum if one wants to have acceptable reliability, and preferably $S=27$. Most studies currently do not have this number of subjects, and one can therefore be legitimately concerned with the reliability of many *findings* from neuroimaging studies: activation detection is the result of a relatively arbitrary thresholding procedure, while true activation configurations show a complex picture (Jernigan et al., 2003). While the specificity of detection procedures is strongly controlled, some activated areas might not be reported due to the lack of power. Increasing S should somewhat reduce the false negative rate, and thus increase the reproducibility of the studies.

One might object that in our case, only one session was available for each subject, and that the quick event-related design might yield poor results in terms of detection. However, the impact of this problem is limited for the following reasons:

- The results that we describe are related to very basic contrasts (auditory and motor activity) for which we could check that most of the subjects (motor contrast) or even every subject (auditory contrast) had significant functional activity in expected regions, which has to be compared with the subtle functional contrasts that are often under investigation.
- The relatively limited number (20 to 60) of trials is perfectly taken into account in mixed-effects models, in which the first-order variance reflects the uncertainty about the activation value

related to the effect. This is in particular the case in Figs. 3–5, where *group* variance, estimated within a mixed-effects framework (see Appendix A), is shown. In models that neglect first-level variance, such as random effects analyses, unmodelled first-level variance should simply yield an inflated group variance. In particular, it does not explain deviation from normality in the data, observed in Figs. 3–5.
- Our finding is consistent with earlier simulations and studies (Desmond and Glover, 2002; Murphy and Garavan, 2004).

For these reasons, we hope that this paper will promote the use of larger cohorts in neuroimaging studies.

Reliability of the different statistical tests

One of the most important practical questions is to describe or design the most efficient ways to perform group studies in neuroimaging. Based on this first study we can suggest some guidelines.

First of all, given the results on normality tests, non-parametric assessment of functional activity should be preferred to analytical tests, which may rely on incorrect hypotheses. This can be done using adapted toolboxes e.g. SnPM (Hayasaka and Nichols, 2003) or Distance (Mériaux et al., 2006a), www.madic.org). It is worthwhile to note that C implementation of the tests reduces computation time to a reasonable level (e.g. cluster-level *p*-values can be computed in less than 1 min on a 10-subject dataset). Non-parametric estimation of the significance improves both the study sensitivity and reproducibility.

Second, mixed-effects models should systematically be preferred to mere random effects analyses: there is some information in the first level of the data that improves the estimation of the group effects/variance and statistic.

Third, cluster- and parcel-based inference should be preferred to voxel-based thresholding. Cluster-level inference is of frequent use, which benefits the sensitivity and the reliability of group analyses. However, it is based on the assumption that activated regions are large, which is not necessarily true. Parcel-based inference may thus be an interesting alternative, since it further allows some spatial relaxation in the subject-to-subject correspondence. The price to pay is a larger variability of the results due to a less stable decision function (activated vs non-activated). We recommend the combination of one of these techniques together with MFX. By contrast, stronger smoothing (12 mm) did not increase significantly the results reliability.

Fourth, to our surprise, ΨFX was found to be the most reliable technique. Although the statistic function does not take into account the group variance–as argued earlier, this is probably the reason for its higher performance–its distribution under the null hypothesis is tabulated by random swaps of the effects signs, so that it is indeed a valid group inference technique. However, it should be used with care because first the thresholds have to be computed voxel per voxel (i.e. are not spatially stationary), and second the statistic value itself has no obvious interpretation, in contrast to the RFX and MFX statistics.

Conclusion

This analysis is also a starting point for developing new strategies in brain mapping data analysis. Several directions could be considered in the future.

- First, one could relate fMRI inter-subject variability to behavioural differences and individual or psychological characteristics of the subjects. Once again, such investigation may be undertaken only on large databases of subjects.
- Second, further efforts should be made to relate spatial functional variability to anatomical variability. While some cortex-based analysis reports have indicated a greater sensitivity than standard volume-based mappings (Fischl et al., 1999), statistical evidence is still lacking, and it is not clear at all how much can be gained when taking into account macroanatomical features, e.g. sulco-gyral anatomy. Similarly, diffusion-based imaging may add useful information to improve cross-subject brain cartography (Behrens et al., 2006).
- Third, at a statistical level, we think that intermediate levels of descriptions could be used more systematically between the subjects and the group level. Identification of outlier subjects, possible subgroups and so on can be investigated (Kherif et al., 2004; Thirion et al., 2005; Thirion et al., 2006b), though finding meaningful distance and separation criteria is not straightforward. For instance, it would be interesting to know what proportion of subjects had a significant activity in a given region; such a simple question requires solving issues in across-subjects correspondences and in statistical thresholding (how can one be sure that two foci of activity in two subjects are homologous?).

Finally, we hope that these results will help establish useful guidelines when planning acquisition and analyzing group functional neuroimaging datasets.

Appendix A. Estimation of the group variance in a mixed-effects model

The joint estimation of the group effect and the group variance proceeds from Eqs. (1) and (2). At a voxel *v*, *S* values of estimated effects $\hat{\beta}$ are available, together with *S* estimates of the associated variances $\hat{\sigma}^2$ (we drop the voxel index *v* for simplicity). We also assume that the estimated variance is correct, so that $\sigma^2=\hat{\sigma}^2$ (note that the estimator relies on $\nu=100$ degrees of freedom).

For this model, the log-likelihood of the data is written as:

$$\mathcal{L}\left(\hat{\beta} \mid \bar{\beta}, v_g\right) = cst - \frac{1}{2}\left(\sum_{s=1}^{S} \log\left(\sigma^2(s)+v_g\right) + \sum_{s=1}^{S} \frac{(\bar{\beta}-\hat{\beta}(s))^2}{\sigma^2(s)+v_g}\right) \tag{10}$$

maximizing $\mathcal{L}$ with respect to $\bar{\beta}$ while keeping v_g fixed yields:

$$\bar{\beta} = \sum_{s=1}^{S} \frac{\hat{\beta}(s)}{\sigma^2(s)+v_g}\left(\sum_{s=1}^{S} \frac{1}{\sigma^2(s)+v_g}\right)^{-1} \tag{11}$$

while the minimization of $\mathcal{L}$ with respect to v_g, while β^- is fixed yields

$$\sum_{s=1}^{S} \frac{(\bar{\beta}-\hat{\beta}(s))^2}{(\sigma^2(s)+v_g)^2} = \sum_{s=1}^{S} \frac{1}{\sigma^2(s)+v_g} \tag{12}$$

Let $L(v_g)$ and $R(v_g)$ be the left and right hand side terms in Eq. (12). We solve it by iterating the solution of $L(v_g)=R$ in under the constraint $v_g > 0$ using Newton's method, then updating the right hand side term. This procedure always converges in a few iterations.

Finally, the joint estimation of $\overline{\beta}$ and v_g proceeds by successive re-estimation of both terms, and always converges in practice. Finally, this joint estimation based on a C implementation is fairly quick, even on a large dataset.

References

Ashburner, J., Friston, K., Penny, W. (Eds.), 2004. Human Brain Function, 2nd edition. Academic Press.

Beckmann, C., Jenkinson, M., Smith, S., 2003. General multi-level linear modelling for group analysis in fMRI. NeuroImage 20, 1052–1063.

Behrens, T.E.J., Jenkinson, M., Robson, M.D., Smith, S.M., Johansen-Berg, H., 2006. A consistent relationship between local white matter architecture and functional specialisation in medial frontal cortex. NeuroImage 30 (1), 220–227.

Brammer, M., Bullmore, E., Simmons, A., Grasby, P., Howard, R., Woodruff, P., Rabe-Hesketh, S., 1997. Generic brain activation mapping in functional magnetic resonance imaging: a nonparametric approach. Magn. Reson. Imaging 15 (7), 763–770.

Bullmore, E., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., Brammer, M., 1999. Global, voxel, and cluster tests, by theory and permutation, for difference between two groups of structural MR images of the brain. IEEE Trans. Med. Imag. 18, 32–42.

Collins, D.L., G., L.G., Evans, A.C. (1998). Non-linear cerebral registration with sulcal constraints. In MICCAI'98, LNCS-1496, pages 974–984.

Desmond, J.E., Glover, G.H., 2002. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. J. Neurosci. Methods 118 (2), 115–128.

Fischl, B., Sereno, M.I., Tootell, R.B., Dale, A.M., 1999. High-resolution intersubject averaging and a coordinate system for the cortical surface. Hum. Brain Mapp. 8 (4), 272–284.

Flandin, G. (2004). Utilisation d'informations géométriques pour l'analyse statistique des données d'IRM fonctionnelle. PhD thesis, Université de Nice-Sophia Antipolis.

Flandin, G., Kherif, F., Pennec, X., Malandain, G., Ayache, N., Poline, J.-B., 2002. Improved detection sensitivity of functional MRI data using a brain parcellation technique. Proc. 5th MICCAI, LNCS 2488 (Part I). Springer Verlag, Tokyo, Japan, pp. 467–474.

Fox, M.D., Snyder, A.Z., Zacks, J.M., Raichle, M.E., 2006. Coherent spontaneous activity accounts for trial-to-trial variability in human evoked brain responses. Nat. Neurosci. 9 (1), 23–25.

Friston, K.J., Penny, W., 2003. Posterior probability maps and SPMs. NeuroImage 19 (3), 1240–1249.

Friston, K.J., Holmes, A.P., Worsley, K.J., 1999. How many subjects constitute a study? NeuroImage 10 (1), 1–5.

Friston, K., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002. Classical and Bayesian inference in neuroimaging: theory. NeuroImage 16 (2), 465–483.

Genovese, C.R., Noll, D.C., Eddy, W.F., 1997. Estimating test–retest reliability in functional MR imaging. I: Statistical methodology. Magn. Reson. Med. 38 (3), 497–507.

Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15 (4), 870–878.

Hayasaka, S., Nichols, T., 2003. Validating cluster size inference: random field and permutation methods. NeuroImage 20 (4), 2343–2356.

Hellier, P., Barillot, C., Corouge, I., Gibaud, B., Le Goualher, G., Collins, D.L., Evans, A., Malandain, G., Ayache, N., Christensen, G.E., Johnson, H.J., 2003. Retrospective evaluation of intersubject brain registration. IEEE Trans. Med. Imag. 22 (9), 1120–1130.

Hollander, M., Wolfe, D., 1999. Nonparametric Statistical Inference, 2nd edition. John Wiley and Sons, New York, USA.

Holmes, A., Blair, R., Watson, J., Ford, I., 1996. Nonparametric analysis of statistic images from functional mapping experiments. J. Cereb. Blood Flow Metab. 16, 7–22.

Jernigan, T.L., Gamst, A.C., Fennema-Notestine, C., Ostergaard, A.L., 2003. More "mapping" in brain mapping: statistical comparison of effects. Hum. Brain Mapp. 19 (2), 90–95.

Kherif, F., Poline, J.-B., Mériaux, S., Benali, H., Flandin, G., Brett, M., 2004. Group analysis in functional neuroimaging: selecting subjects using similarity measures. NeuroImage 20 (4), 2197–2208.

Kjems, U., Hansen, L.K., Anderson, J., Frutiger, S., Muley, S., Sidtis, J., Rottenberg, D., Strother, S.C., 2002. The quantitative evaluation of functional neuroimaging experiments: mutual information learning curves. NeuroImage 15 (4), 772–786.

LaConte, S., Anderson, J., Muley, S., Ashe, J., Frutiger, S., Rehm, K., Hansen, L.K., Yacoub, E., Hu, X., Rottenberg, D., Strother, S., 2003. The evaluation of preprocessing choices in single-subject BOLD fMRI using NPAIRS performance metrics. NeuroImage 18 (1), 10–27.

Liou, M., Su, H.-R., Lee, J.-D., Cheng, P.E., C.-C., H., Tsai, H., 2003. Bridging functional MR images and scientific inference: reproducibility maps. J. Cogn. Neurosci. 15 (7), 935–945.

Liou, M., Su, H.-R., Lee, J.-D., Aston, J.A.D., Tsai, A.C., Cheng, P.E., 2005. A method for generating reproducible evidence in fMRI studies. NeuroImage.

Lund, T.E., Madsen, K.H., Sidaros, K., Luo, W.-L., Nichols, T.E., 2006. Non-white noise in fMRI: does modelling have an impact? NeuroImage 29 (1), 54–66.

Maitra, R., Roys, S.R., Gullapalli, R.P., 2002. Test–retest reliability estimation of functional MRI data. MRM 48 (1), 62–70.

McNamee, R.L., Lazar, N.A., 2004. Assessing the sensitivity of fMRI group maps. NeuroImage 22 (2), 920–931.

Mériaux, S., Roche, A., Dehaene-Lambertz, G., Thirion, B., Poline, J.-B., 2006a. Combined permutation test and mixed-effect model for group average analysis in fMRI. Hum. Brain Mapp. 402–410.

Mériaux, S., Roche, A., Thirion, B., Dehaene-Lambertz, G., 2006b. Robust statistics for nonparametric group analysis in fMRI. Proc. 3th Proc. IEEE ISBI, Arlington, VA.

Murphy, K., Garavan, H., 2004. An empirical investigation into the number of subjects required for an event-related fMRI study. NeuroImage 22 (2), 879–885.

Neumann, J., Lohmann, G., 2003. Bayesian second-level analysis of functional magnetic resonance images. NeuroImage 20 (2), 1346–1355.

Nichols, T., Holmes, A., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15, 1–25.

Shaw, M.E., Strother, S.C., Gavrilescu, M., Podzebenko, K., Waites, A., Watson, J., Anderson, J., Jackson, G., Egan, G., 2003. Evaluating subject specific pre-processing choices in multisubject fMRI data sets using data-driven performance metrics. NeuroImage 19 (3), 988–1001.

Smith, S.M., Beckmann, C.F., Ramnani, N., Woolrich, M.W., Bannister, P.R., Jenkinson, M., Matthews, P.M., McGonigle, D.J., 2005. Variability in fMRI: a re-examination of inter-session differences. Hum. Brain Mapp. 24 (3), 248–257.

Stiers, P., Peeters, R., Lagae, L., Hecke, P.V., Sunaert, S., 2006. Mapping multiple visual areas in the human brain with a short fMRI sequence. NeuroImage 29 (1), 74–89.

Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., Rottenberg, D., 2002. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. NeuroImage 15 (4), 747–771.

Strother, S., Conte, S.L., Hansen, L.K., Anderson, J., Zhang, J., Pulapura, S., Rottenberg, D., 2004. Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis. NeuroImage 23 (Suppl. 1), S196–S207.

Thirion, B., Pinel, P., Poline, J.-B., 2005. Finding landmarks in the functional brain: detection and use for group characterization. Proc. MICCAI2005, Palm Springs, USA.

Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., Poline, J.-B., 2006a. Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets. Hum. Brain Mapp. 27 (8), 678–693.

Thirion, B., Roche, A., Ciuciu, P., Poline, J.-B. (2006b). Improving sensitivity and reliability of fmri group studies through high level combination of individual subjects results. In Proc. MMBIA2006, New York, USA.

Thyreau, B., Thirion, B., Flandin, G., Poline, J.-B., 2006. Anatomo-functional description of the brain: a probabilistic approach. Proc. 31st IEEE ICASSP, vol. V. IEEE, Toulouse, France, pp. 1109–1112.

Wager, T., Keller, M., Lacey, S., Jonides, J., 2005. Increased sensitivity in neuroimaging analyses using robust regression. NeuroImage 26 (1), 99–113.

Wei, X., Yoo, S.-S., Dickey, C.C., Zou, K.H., Guttmann, C.R.G., Panych, L.P., 2004. Functional MRI of auditory verbal working memory: long-term reproducibility analysis. NeuroImage 21 (3), 1000–1008.

Woolrich, M., Behrens, T., Beckmann, C., Jenkinson, M., Smith, S., 2004. Multi-level linear modelling for fMRI group analysis using Bayesian inference. NeuroImage 21 (4), 1732–1747.

Worsley, K.J., 2005. An improved theoretical *P* value for SPMs based on discrete local maxima. NeuroImage 28 (4), 1056–1062.

Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., Evans, A., 2002. A general statistical analysis for fMRI data. NeuroImage 15 (1), 1–15.

Zar, J.H., 1999. Biostatistical Analysis. Prentice-Hall, Inc., Upper Saddle River, NJ.