INTRODUCTION

Motivation and Synthesis of the FIAC Experiment: Reproducibility of fMRI Results Across Expert Analyses

Jean-Baptiste Poline,^{1*} Stephen C. Strother,²
Ghislaine Dehaene-Lambertz,³ Gary F. Egan,⁴ and Jack L. Lancaster⁵

¹UNAF, Service Hospitalier Frédéric Joliot-CEA, Orsay, France

²Rotman Research Institute, Baycrest Centre, Toronto, Ontario, Canada

³INSERM U562, Service Hospitalier Frédéric Joliot, CEA/DRM/DSV, Orsay, France

⁴Howard Florey Institute & Centre for Neuroscience, University of Melbourne, Victoria, Australia

⁵Research Imaging Center, University of Texas Health Science Center at San Antonio,
San Antonio, Texas, USA

Abstract: The Functional Imaging Analysis Contest (FIAC) culminated in the FIAC Workshop held at the 11th Annual Meeting of the Organization for Human Brain Mapping in Toronto in 2005. This special issue summarizes various analyses used by contestants with a single functional magnetic resonance imaging (fMRI) study, a cortical-language study using sentence repetition. The results from the cognitive neuroscientists who developed the test-base language study, and report their data analysis, are complemented by expert analyses of the same test-base data by most of the major groups actively developing fMRI software packages. Analyses include many variants of the general linear model (GLM), cutting-edge spatial- and temporal-wavelets, permutation-based, and ICA approaches. A number of authors also include surface-based approaches. Several articles describe the important emerging areas of diagnostics for GLM analysis, multivariate predictive modeling, and functional connectivity analysis. While the FIAC did not achieve all of its goals, it helped identify new activation regions in the test-base data, and more important, through this special issue it illustrates the significant methods-driven variability that potentially exists in the literature. Variable results from different methods reported here should provide a cautionary note and motivate the Human Brain Mapping community to explore more thoroughly the methodologies they use for analyzing fMRI data. *Hum Brain Mapp* 27:351–359, 2006.

© 2006 Wiley-Liss, Inc.

MOTIVATION FOR THE FIAC EXPERIMENT

The Functional Imaging Analysis Contest (FIAC) stemmed from an idea discussed within the Madic group (Orsay, UNAF-CEA, <http://www.madic.org>) during the 2004 Human Brain Mapping conference in Budapest. The feeling of the group, whose activity is mainly focused on developing techniques for functional image analysis, was that direct comparison of the various analysis methods is

*Correspondence to: Jean-Baptiste Poline, 4 Pl. General Leclerc, CEA/DRM/SHFJ, 91406 Orsay, France. E-mail: poline@shfj.cea.fr

DOI: 10.1002/hbm.20268

Published online 31 March 2006 in Wiley InterScience (www.interscience.wiley.com).

difficult, because key information is missing from many published articles and posters. To bring this information together in a single place, the group felt that it would be interesting, educational, and to some extent entertaining to have a single functional MRI (fMRI) dataset analyzed by most of the founding groups of currently used analysis techniques. When discussing the idea further it became apparent that a number of benefits could be gained from such an experiment. These are summarized below. After describing the motivation for the FIAC, we present a short synthesis of the results and conclusions about this experiment.

Comparing Techniques and Guiding Users

The most obvious goal was to determine which techniques would yield the “best” results. This is a legitimate question from neuroscientists or clinicians who typically assess the results in relation to their prior expectations, although this biases fMRI studies away from making new and unexpected discoveries. From a methodological point of view, “best” can be defined in terms of sensitivity and specificity for known signals and noise, and by the ability of the method to adapt to the data, i.e., the robustness of the method’s results to data that violate the underlying model assumptions. Because there is no gold standard for an actual brain imaging experiment, there is no easy answer to this question (and that can be seen as a disappointment—or a relief). It should be noted, however, that it is sometimes possible to estimate the quality of analyses without definite gold standards (see Chen et al. for some references).

The take-home message of the FIAC experiment is therefore not the selection of “the best method,” but it provides details on the diversity of analytical approaches and qualitative information about variation of results across expertly performed analyses. In this sense, the FIAC experiment provides useful information for a more informed choice and use of the methods commonly or not so commonly used.

Why not provide a dataset with known signals and noise (and possibly no signal)? While this would be a worthwhile experiment (suggested as one of the possible follow-ups in the Appendix), it has two major drawbacks. First of all, it is extremely difficult to simulate a fully realistic functional neuroimaging dataset. An option would be to recreate fMRI noise from an actual experiment and linearly superimposed signals, although recent work on resting state networks makes it unclear how easy this is to accomplish (Beckmann et al., Goebel et al.). The efficiency of the methods could then be assessed, but the results would show that methods conforming to the correct model of the signal and noise chosen would perform best, and extrapolation of the results to actual brain signals would not be possible. This approach also limits the question one could ask of the data to that of signal detection. The second drawback is that it would not be enticing for neuroscientists (see below).

Establishing a Deeper Dialogue between Two Research Communities

We hoped that with an actual dataset we would attract scientists interested in the experimental paradigm and its domain. This could then lead to open and fruitful interactions in brain imaging between the neuroscience and data analysis communities. While the former would learn about the various analysis techniques available, the latter would better appreciate interesting neuroscience problems that need to be worked on. We therefore asked for the help of a number of well-established neuroscientists whose research interests encompass the brain mapping of language functions. We are thankful to Pascal Belin, Susan Bookheimer, Narly Golestani, Russ Poldrack, and Cathy Price for their participation. We also involved senior researchers well known for their expertise in methods development: Robert Cox, Karl Friston, Rainer Goebel, Tom Nichols, and Steve Smith, who are also sometimes involved in brain mapping research themselves.

In short, we hoped that the FIAC might lead to new analytical techniques, better adapted to the neuroimaging needs, but also show researchers the various analyses that they could use and examples of the results that can be obtained. Far from just a “contest,” we hoped to set up a forum for discussions between neuroscientists and the developers of fMRI analysis techniques and software.

Assessing a Component of the Neuroimaging Variability

There is a clear neuroscience interest in assessing the reproducibility of neuroimaging experiments. The field of functional neuroimaging is often considered by some scientists outside, but also within, the field as being seriously hampered by the variability of its results. Assessing the variability of the results across laboratories includes assessing the variability due to the analysis method employed. Because reproducibility is an important component of any scientific study (among others, such as interpretation within the context of the specific scientific field, explicit or implicit model used, etc.), an investigation of reproducibility within the FIAC dataset is aimed at helping the functional brain imaging field mature.

Different results may be obtained from a functional brain imaging experiment for the following reasons:

i) Analysis procedures:

- the general analytic approach, e.g., predictive coding vs. GLM, voxel-based vs. multivariate, voxel-based spatial normalization vs. individual regions of interest (ROIs), etc.;
- the analytical algorithms used and their implementation;
- the order of analytical procedures (e.g., slice timing vs. realignment etc.); and
- the analytical parameters chosen (order of the noise

model, smoothing, hemodynamic response function (HRF), etc.).

- ii) Variation of performance, attention, and other subject-specific factors, i.e., subject and scanner differences.
- iii) Differences in nominally similar experimental paradigms (including instructions, time of scanning, inclusion or exclusion rules) or implementation of the experimental paradigm in the scanner environment.

For the FIAC experimental dataset, many different analytical approaches were adopted. For example, in the test-base data (Dehaene-Lambertz et al.), repetition suppression was searched for either by comparing blocks or by computing the rate of signal drop within blocks and then comparing blocks. Some contributors modeled the first sentence of a block separately, some modeled the time by condition interactions, and some used data-driven, exploratory techniques. The specific impact of each analysis component was not addressed, and would be difficult to address since there are a large number of parameters with unknown interactions. However, an overall impression of the variability of results due to analysis can be gained when the variability due to subject, scanner, and paradigm implementation effects are held fixed. The FIAC experiment also allows us to address, in part, the variation of the experimental paradigm, since similar results were expected from the block and the event paradigms.

The second part of this Introduction presents a short synthesis of the FIAC experiment and then proposes further uses of this experimental dataset. We also briefly list some recommendations for presenting the analysis methods' numerical results and visualization results of fMRI experiments.

FIAC: A SHORT SYNTHESIS

The FIAC was advertised on the OHBM website and was open to all HBM conference attendees (<http://www.madic.org>). Once the functional imaging data were made available on the web, the number of responses to the FIAC proposal was surprisingly high. The competitive aspect of the proposal had clearly generated considerable interest, and although this aspect was not the actual rationale for the "contest," it did produce a sufficient number of responses for a reasonable sampling of the popular methods. The editors of the special issue and the organizers of the FIAC 2005 would like to express their sincere thanks for the large amount of work done in a short time by the different contributors to the issue. An additional difficulty arose from the need to analyze data from another laboratory, which is a challenge even for experts. Various contributors reported problems with identifying the left and right brain hemispheres after data processing steps. We would also like to emphasize that most of the contributors are primarily statisticians, engineers, and physicists, but many went to con-

siderable lengths to discuss the experimental paradigm and its interpretation in their articles.

Summary of the Analyses Performed

Participants consisted of active groups in fMRI data analysis methods. Many groups are associated with neuroscience laboratories that have applied statistics, computer science, or mathematics research groups. This is due to the characteristic of the OHBM meeting that aims at gathering the different communities. Packages used in the analyses include SPM5, FSL, BrainVoyager (the only commercial package among the contributors), BAMB, fMRI/BrainStat, and AFNI, to name a few.

The methods employed and the datasets used are summarized in Table I. The type of analyses performed, although quite variable, almost all include linear modeling of the blood oxygen level-dependent (BOLD) response. The different models used produced large differences in some instances, but produced enough similarities for comparisons between the models. Our goal was not to assess the variation of sensitivity or specificity of the techniques, but we focused on evaluating the variations among the results presented for the block data and the GLM analyses, which provides the largest number of similar analyses. The three latter contributions to this special issue propose less common analytical approaches. Zhang et al. demonstrated the application of a Diagnosis toolbox, Chen et al. studied how the experimental conditions can be predicted by some regional activity, and Patel et al. presented a functional connectivity analysis involving a hierarchy across regions.

Although the main feature of the FIAC was to use a common dataset, part of the analysis procedure consisted of choosing the data to be included in the study and the data selection criteria. For instance, should subjects with excess motion artifact be excluded, and how much is too much? There are numerous "rules of thumb" or criteria that are seldom documented or validated outside individual laboratories, and known only to the researchers at those laboratories. Several examples of differences in analysis results are given in Table II. For instance, one or two of the following five subjects were excluded from most analyses (Subjects 0, 2, 7, 10, 11), while Subjects 8 and 12 were excluded from two analyses and Subject 5 from 8 analyses. Subject 5 was not included in the list of potential outliers in the diagnosis analysis, but was excluded by most because the data were not complete. Since group sizes are typically small in fMRI analyses, results can be very sensitive to the specific subjects included in an experiment. This is an important issue that requires further research.

Summary of the Reported Results

We extracted the results described for the block data for the three main contrasts of experimental conditions: effect of sentence ($DSt.DSp + DSt.SSp - (SSt.DSp + SSt.SSp)$), effect of speaker ($DSt.DSp + SSt.DSp - (DSt.SSp + SSt.SSp)$), and interaction (positive or negative), where D = different, S = same, St = sentence, and Sp = speaker (Dehaene-

TABLE I. Summary of the data analyzed and the main types of analyses performed by each contributor

Contribution first author	Single subject	Group block	Group Event	Main analyses performed
Aston et al., Phiwave	X	X		GLM3D
Beckmann et al., FSL	X	X	X	GLM3D, HRFestim, MixMod, Mixed FX, Tensor-ICA
Dehaene-Lambertz et al., SPM2		X		GLM3D
Goebel et al., BrainVoyager	X	X	X	GLM3D, GLMsurf, group-ICA,
Meriaux et al., Madic		X		GLM3D, Mixed FX
Noppeney and Penny, SPM5	X			GLM3D, Global "Conjunction"
Saad et al., AFNI	X	X	X	GLM3D, GLMsurf, HRFestim, conjunction
Suckling et al., BAMM		X		GLM3D
Taylor et al., BrainStat	X	X	X	GLM3D, HRFestim, Delay analysis
Zhang et al., SPMd	X	X		Artifacts detection, model and assumption checking
Chen et al., Predictive modeling	X			Predictive modeling, reproducibility analysis, GLM3D
Patel et al., Ascend. Network		X		Functional connectivity, "ascendant" networks

GLM3D: General Linear Model on volumetric data.

GLMsurf: General Linear Model on the cortical surface, after cortical alignment.

Delay Analysis: estimating the delay of the hemodynamic function and contrasting it for different experimental conditions.

Group ICA /Tensor ICA: Two different group ICA techniques. Group ICA uses the clustering of the ICA components across subjects; Tensor ICA extracts subject, space, and condition components (three-dimensional decomposition).

MixMod: Mixture modeling of the null and alternative hypotheses, allowing quantification of the probability of accepting the null hypothesis (i.e., the probability that these two effects are equal).

HRFestim: a procedure to estimate or partly estimate the hemodynamic response function (HRF).

Lambertz et al.). Some authors report other contrasts, such as sentence or speaker effects restricted to one condition of the other factor, e.g., DSt.DSp-DSt.SSp. Often, two analyses were performed on the block data. Most commonly, the

activity in the blocks was compared, with or without removing the estimated effect of the first sentence. Some analyses modeled the decrease of activity within a block, and compared this decrease between blocks, introducing time as an

TABLE II. Example of the variation of analysis parameters and multiple comparison correction methods (these are not necessarily the greatest causes of variation in the reported results)

Contributors	Subjects excluded	Smooth. spatial filter FWHM (mm)	Distortion correction with field maps	Modeled first sentence	Multiple comparisons correction method
Aston et al., Phiwave	0 5 11 (Fieldmap) 8 (spikes)	Space dependant	Yes	Yes	No correction
Beckmann et al., FSL	5	5	Yes	No	Gaussianized random fields and spatial mixture modeling
Dehaene-Lambertz et al., SPM2	5	8	No	Yes	Random fields (SPM2)
Goebel et al., BrainVoyager	5 7 8 and 12	0 (surf) 5 (3D)	No	NR	3D Monte Carlo simulations based on Forman MRM 1995
Meriaux et al., Madic	5	5	No	Yes	Based on permutations
Noppeney and Penny, SPM5	Single subj. analysis	8	Yes	Yes (factor time)	Random fields (SPM5)
Saad et al., AFNI	5 10 12	4 (block) 8 (event)	No	Yes	Monte Carlo simulations based on Forman MRM 1995
Suckling et al., BAMM	5	5	No	No	Time reshuffling, test of the median across subjects
Taylor et al., BrainStat	2 5	None; Var. ratio smoothing 7 to 10	No	Yes	Random Fields and Discrete local maxima (BrainStat, fMRISat)
Zhang et al., SPMd	Show 1 9 8 or 4 as potential outliers	None	No	No	Not applicable
Chen et al., Predictive modeling	Single subj. analysis	6	No	No	Not applicable
Patel et al., Ascend. Network	4 subjects are excluded	None	No	No	Not applicable

additional factor. This second analysis was much less sensitive (probably due to the increased degrees of freedom or number of tests required to estimate the shape of the within block decrease), but the results were reported when significant. We have chosen not to report the event-related data results, although these do show more sensitivity according to several authors, because the low number of analyses (four) was insufficient to provide a reasonable estimate of the variability of results.

Extracting the needed information from the contributed articles was often difficult. The reader is strongly recommended to use the summary only as an adjunct to reading the complete articles, as the summary cannot adequately describe the wide variety of analytical issues reported. The summary is intended to convey meta-information about the variability and reproducibility of the reported results for a common paradigm.

We found that the reported results do vary significantly for the most similar analytical approaches. Results were clearly consistent when signals were very strong (e.g., listening to sentences vs. rest), which is very reliably reported as activating the bilateral superior temporal sulci and gyri. However, brain imaging protocols are often concerned with signals that are weak or variable, or both, and barely detectable.

A summary of the results is given in Table III. Figure 1 provides a summary of the reported Talairach activation locations across all conditions from all articles in this issue that provided such coordinates. It is clear that many analyses found two strong loci of effects in the left temporal gyrus and weaker, more scattered effects in the right temporal gyrus. The repetition of sentence (sentence effect) was consistently found in the left temporal gyrus (all contributors), but the location of the maxima varied up to several centimeters when using the Talairach coordinates (Fig. 2). The volume of the significant activation was also quite variable if we compare the four studies that gave this information (1,701 mm³, Goebel et al.; 3,592 mm³, Beckmann et al.; 3,904 mm³, Dehaene-Lambertz et al.; and 24,192 mm³, Meriaux et al.). A significant number of regions were reported by only one or two contributors (right anterior superior temporal sulcus (STS) (twice), right inferior precentral, right inferior frontal, anterior insula, left inferior frontal gyrus (IFG), left inferior parietal, right middle temporal gyrus (MTG)).

The main effect of speaker was found less often, sometimes by restricting the analysis to the same sentence conditions. Regions reported significant were most often the left posterior STS/MTG, left and right superior temporal, right middle STS (using correction for multiple comparison in the temporal lobe), right Broca area. *Notice that almost half of the analyses did not report a significant speaker effect*, and that the distance between local maxima found in the left temporal region exceeded 2 cm (Fig. 2). To summarize, there was no strong consensus between individual reported results across analyses.

Interaction between speaker and sentence effects was reported in right Broca and left posterior STS/MTG, left and

right temporal-occipital, right middle STS, and left middle temporal. Several authors noted that the interaction was generally due to less activity for the same sentence, same speaker condition compared to others. Once again, localization of activity produced local maxima in the same gyrus or cortical lobe that could be greater than 2 cm apart.

While it is possible that the observed variability results from thresholding effects and that the underlying maps are in fact similar to each other, this is yet to be confirmed. It is also possible that some procedures do not appropriately control the risk of type I error (false alarm), although what is appropriate in terms of choosing a threshold for a particular *P* value varies with the approach (e.g., false discovery rate, random field theory, simulations) and the increasing sophistication of the analytic corrections (Taylor and Worsley). If the differences are mainly due to the choice of subjects excluded, then this confirms the sensitivity of the results to this choice. Some procedures vary in the null hypothesis they test. For instance, permutations of the experimental conditions (or time/wavelet coefficients) within subjects lead to a null hypothesis of the kind “none of the subjects have activated,” i.e., a fixed-effects null hypothesis. In theory it is therefore enough that one of the subjects activates to reject a subsequent group test. This is not the usual group analysis hypothesis, although this may be somewhat ameliorated, as in BAMB (Suckling et al.), by using a group median that ensures that at least half of the subjects show a significant fixed-effects response.

Other statistical procedures rely on simulations to perform multiple comparison correction, and therefore assume hypotheses on the distribution of the data and its covariance that are not necessarily tenable, as shown by diagnostic procedures. However, this criticism also partly applies to model-based techniques such as random field theory. Some statistical procedures may benefit from a clearer descriptive term. The so-called global conjunction analysis performs an operation that is closer to a logical disjunction, and this may still confuse some researchers even if the information can be found on discussion lists or in the literature. However, this is well described (see Noppeney and Penny). On the other hand, for most analyses the risk of a false-negative might be very high, and modeling the alternative hypothesis should lead to more appropriate trade-offs between type I and II errors. Reporting type II error risk would provide analysis sensitivity, a most useful piece of information. As for non-inferential analyses, it is interesting to note that both ICA techniques, (Beckmann et al., Goebel et al.), although based on different principles for grouping the data across subjects, do show some consistency in reporting the so-called resting-state network.

CONCLUSION

An evaluation of the variability and similarity of the results of the FIAC dataset showed that the analytical methods and/or their settings and implementation details do produce different results. These differences relate to the size and number of activations and their reported Talairach coordi-

TABLE III. Summary of results for the block data group analysis for each contribution

	Main effect of all vs. rest	Main effect of sentence	Main effect of speaker	Interaction effect	Notes: general	Notes: specific to methods
Aston et al., Phiwave	Large activity in temporal lobes bilat. Signif. not computed	Left temporal gyrus Significance not computed	Weak signals in bilat. temporal gyri, significance not computed	ventral visual stream, SMA, ventral and medial prefrontal	Excluded subjects 8 for spikes, 0 5 and 11 for lack of fieldmaps.	Applied wavelet transform for greater MSE estimates. Results are not quantified in terms of type I error.
Beckmann et al., FSL	HG bilat., planum temporale, planum polare, Lat. Sup. temp. STS, MTG	Left Ant STS (simple main FX for DSp)	NS Contrast 13: Right Broca and left post. STS/MTG Mixture model: DStSSp < DStDSp: HG Bilat	Right Broca and left post. STS/MTG	Results suggest that both effects are processed in a single system without asymmetries for content vs. voice processing	3 ways mixed effects model. Estimation of the session effects. Gaussian assumption for cluster test. Show improvements with H1 modeling and HRF estimation.
Dehaene-Lambertz et al., SPM2	Bilateral temporal	Left temporal	NS	NS	Analysis of hemispheric asymmetries	First sentence modeled separately.
Goebel et al., Brain-Voyager	Bilateral temporal	Left Ant STG	NS	Left n Right temporal-occipital	Excluded subj 5, 7, 8, 12; ica + clustering : left STG and IFG for suppress of SStSSp; Performed a within block analyses.	Control of false positive based on simulations that require assumptions (Forman et al., 95)
Meriaux et al., Madic	NR	Left middle STS Right anterior STS With SVC and restricted to DSp: left post STS	Left post STS With SVC: left post and right middle STS	With SVC: Right middle STS	SVC: encompasses the perisylvian areas (about 3000 voxels), all subjects analyzed	Mixed effects model with permutation techniques.
Noppeney and Penny, SPM5	Not reported	Bilat. fronto-temp.; Right inf. frontal gyrus, middle temporal regions.	NS; Conjunction: left post sup temp.	NS	Single subject analysis	“Conjunction” analysis : interpretation or naming may cause confusion. The text however is clear about the issue.
Saad et al., AFNI	STG/STS, Brodmann areas 21, 22, and 42.	Not tested	Not tested	Not tested	Used two analyses: fixed response or variable shape for each condition Used conjunction Removed subjects 5 10 12 In Meriaux et al :interaction found in right STS/STG. Here: bilateral Ant. MTG, left post. MTG and left IFG.	Control of false positive based on simulations that require assumptions (Forman et al., 95)

TABLE III. (continued).

	Main effect of all vs. rest	Main effect of sentence	Main effect of speaker	Interaction effect	Notes: general	Notes: specific to methods
Suckling et al., BAMM	Bilateral temp. and prefrontal. Deactivation ant. and post. Cing. Parietal.	Ant temp bilat, left IFG, ant insula	Bilateral superior temporal	In regions found for main effects, left middle temp. due to drop of SStSSp	Weak evidence of within-block suppression of activation during repetition of sentences.	The test is based on the time reshuffling in the wavelet space; the null hypothesis seems to be that none of the subject has activated.
Taylor et al., BrainStat	Not tested	Left MTG, Right MTG Left inf parietal lobule	Trend ($P \sim 0.1$) LMTG (B21)	NS	Excluded subjects 2 and 5. Analyzed the delay of the HRF: increased for sentence effects. Combination of events and blocks gives more results (esp right inf parietal lobule)	Smoothing : from 7 to 10 mm to get 40 DF. First sentence modeled separately.
Zhang et al., SPMd	Standard tests not applicable. Finds non “normal” regions in left STG or STS for the group study				Show that the autocorrelation correction remains problematic (shown on 1 subject) Finds subj 1, 9 and possibly 4 as outliers	Normality tests voxels per voxels.
Chen et al., Predictive modeling	Standard tests not applicable. A selection of regions involved: RSTG, LSTG, LMFG, R Insula, RIFG, L precentral gyrus.				Single subject, The good accuracy of prediction does not translate to reliable selected features; artifacts in fMRI makes the problem hard to solve	Multivariate Z-scored activation maps produced by split-half resampling
Patel et al., Ascend. Network	Standard tests not applicable. Baseline networks: (BA41 BA42, BA22 (STG)), (BA22 posterior, STG, and Broca’s BA44). Ascendant connection is found from BA 41 to post. cingulate (BA 30) and BA44, and a stronger connection from Wernicke’s to STG and the posterior cingulate in the sentence effect contrast.				Functional connectivity with 3 regions: primary auditory cortex (AI), Wernicke (BA22) and BA39 is presented	“Ascendant” is used in the sense of “active in a super set of the time,” a possibly confusing name, but this is clearly stated in the text.

NS: not significant; NR: not reported; FX: Effect; Signif: test significance; SVC: small volume correction; STS: superior temporal sulcus; STG: superior temporal gyrus; MTG: middle temporal gyrus; IFG: inferior frontal gyrus; HG: Heschl gyrus; Post: Posterior; Sup: Superior; Ant: Anterior; Bilat : Bilateral.

nate locations. We did not attempt to evaluate which components in the analytic algorithms produced the largest variability in the results: e.g., intrasubject processing such as realignment, temporal interpolation if performed, spatial normalization (leading to potential localization bias, but also to loss of sensitivity), smoothing, temporal model chosen, statistical tests, and the correction for multiple comparisons procedures. Again, components may interact (such as the model chosen and the realignment procedure, the echoplanar imaging (EPI) distortion correction, the spatial normalization, etc.), making component-by-component optimization and comparison only a partial solution. Specific

hierarchical experiments controlling for all except two or more components should be set up for this purpose.

One goal of the FIAC experiment was to identify new techniques and improved use of standard techniques in order to provide neuroscientists with information to enhance their usual analytical approach. This goal has already been partly achieved, in the form of new results from the FIAC, including the sentence effect in the right temporal lobe, and the possible implication of Broca’s area. However, the variable results of the different methods used should provide a cautionary note and motivate neuroscientists to explore more thoroughly the methodology for analyzing

Figure 1.

The reported Talairach activation locations across all conditions from all articles in this issue that provided such coordinates.

their data. The lack of understanding of complex analysis techniques by members of the neuroscience community has, in the past, induced suspicion, misuse, and/or misinterpretation, a situation that needs to be dealt with by 1) increasing the technical education before using new methods with courses, didactic articles, and educational websites; and 2) improved understanding of the consequences of choosing different analysis techniques in fMRI.

The FIAC experiment also had frustrating aspects. The major one lies in the difficulty in comparing results of various analyses, while not knowing the ground truth. We suggest that this experiment should be complemented by a set of benchmark simulated data, although this would have its own limitations. Such data might be helpful in quantifying the capabilities of different methodological approaches with respect to the simulations, but their extrapolation to any particular real fMRI dataset will remain unknown. Data-driven approaches focusing on empirical ROC curves based on reproducibility and prediction metrics may be able to

bridge this gap between simulations and real data, but more research is needed. Another limitation encountered was that the dialogue between the methods and neuroscience communities did not develop as hoped, mainly because of time and organizational constraints.

How do we choose the fMRI analysis technique and software? As the number of acquisition and analysis methods and available software implementations increases, it becomes more and more difficult to choose the best analysis tools in an even semi-informed way for a particular experimental design. From our experience, the main deciding factors in choosing analysis methods are:

- **Publicity.** Very good neuroscience articles have been published with the method/package, and therefore reviewers will not question the technique, so it is easier to publish results.

- **History.** As it is hard to learn how to use specific tools, let alone a new package, there is considerable inertia, since researchers would rather spend time doing experiments than

Figure 2.

The peaks of activation reported in four articles that analyzed the block design and reported the Talairach coordinates of the activation maxima for the main effects of sentence repetition (in red) and of speaker repetition (in blue), projected on axial slices of an individual subject. 1: Meriaux et al., Madic; 2: Goebel et al., BrainVoyager; 3: Beckman et al., FSL; and 4: Dehaene-Lambertz et al., SPM2.

learning a new processing environment unless a specific analysis is required and only available in a given package.

- **Help and local availability.** A statistician or a methods group is involved in the project, and questions or specific demands can be directed to these collaborators. Active discussion lists (such as the SPM, AFNI, or FSL lists) are also often a good way to get much-needed help. Alternatively, the software is locally available and other colleagues in your laboratory have experience using it.

- **Technique.** The package has demonstrated good technical performance in the literature—often outside the brain mapping literature. It also should have good data visualization and verification capacities. One challenge is that packages are constantly being revised and developed, so that the technical literature is not always a reliable guide to the currently available tools and packages.

The choice of analysis tools and their associated packages influences the manner and type of results that are reported, and we propose a short set of recommendations in the Appendix. A full set of recommendations should be established and maintained in the future.

To conclude, we would like to encourage initiatives that would allow the various package components to be tested easily in the same software environment. In other words, if all the analyses presented in this issue could be easily performed by a single operator, functional brain imaging should gain in reproducibility and operability. This would help to choose the analysis and/or create consensus activation maps, and data would be more easily explored and better interpreted. A number of systems have been proposed but their use is still limited. On the methods side, it is important that this is accompanied by a set of well-documented program interfaces that make high-level methods development and testing much faster. The emphasis would then turn from developing the best package (and re-implementing already developed code) to improving or developing new components able to build on already existing components from various laboratories.

ACKNOWLEDGMENTS

The authors thank the Marseille fMRI center with JL Anton, A. Campagne, B. Nazarian, and M. Roth for their excellent work and their friendship. We had great support from the IFR49 (<http://www.ifr49.org>), with special thanks to I. Denghien and N. Terrien (IFR49), S. Dodel (IFR49 and

Madic), the other members of Madic and the UNAF laboratory in Orsay. Our thanks also to C. Pallier (Unicog) for his efficient help.

APPENDIX

Short Set of Recommendations for Reporting Results Learned from Reading the FIAC Articles

1. Use nonthresholded maps or very-low threshold maps on the cortical surface, or in a space that allows full visualization. Report nonsignificant results if they show a trend, since the $P < 0.05$ threshold is arbitrary.
2. Provide a clear description of the temporal model and the test performed (statistic, degrees of freedom, etc.). Consider best practice in statistics applied to medical studies or biology.
3. Show the effect size and its standard deviation, and show individual subject values whenever possible.
4. Report localization referring to the template used for normalization, report the smoothing used, show individual localization as far as possible.
5. Run diagnosis (model fit, group homogeneity), especially on regions reported to be significant.
6. Graph the BOLD response at the maxima of activations for the different conditions.

Future Proposals for the FIAC

1. Conduct a more precise evaluation procedure that may involve using simulated datasets.
2. Conduct a prediction contest (proposed in HBM 2006, Florence).
3. Construct a database of the results and the programs.
4. Use the results for educational purposes: the dataset can be disseminated together with the results of many analyses.
5. Use benchmark data for co-registration, realignment, spatial normalization, EEG analysis and source reconstruction, etc.
6. Conduct an interpretation contest for neuroscientists with the results of one analysis (proposal of C. Beckmann).
7. Integrate data from other modalities, such as EEG for source localization, anatomical data for segmentation, etc.
8. Allow more than 3 months for completing data exploration and analysis.