Technical Note

Motion or Activity: their role in intra- and inter-subject variation in fMRI

Torben E. Lund [⋆,a], Minna D. Nørgaard [a], Egill Rostrup [a], James B. Rowe [a] and Olaf B. Paulson [a]

[a] *Danish Research Centre for MR, Copenhagen University Hospital, Hvidovre*

Abstract

Functional MRI (fMRI) carries the potential for non-invasive measurements of brain activity. Typically what are referred to as activation images are actually thresholded statistical parametric maps. These maps possess large inter-session variability. This is especially problematic when applying fMRI to pre-surgical planning because of a higher requirement for intra-subject precision. The purpose of this study was to investigate the impact of residual movement artefacts on intra-subject and inter-subject variability in the observed fMRI activation. Ten subjects were examined using three different word-generation tasks. Two of the subjects were examined 10 times on 10 different days using the same paradigms. We systematically investigated one approach of correcting for residual movement effects: the inclusion of regressors describing movement-related effects in the design matrix of a General Linear Model (GLM). The data were analysed with and without modeling the residual movement artefacts and the impact on inter-session variance was assessed using F-contrasts. Inclusion of motion parameters in the analysis significantly reduced both the intra-subject as well as the inter-subject-variance.

Key words: fMRI, inter-session variability, movement artefacts.

⋆ To whom correspondence should be addressed:
Torben E. Lund
Danish Research Centre for MR Copenhagen University Hospital
Kettegaard All 30, 2650 Hvidovre, Copenhagen, Denmark
Phone: +45 3632 3328
Fax: +45 3647 0302

Email address: torbenl@magnet.drcmr.dk.
URL: http://www.drcmr.dk.

Introduction

In a legendary study by McGonigle et al. (2000) the size of the activated area passing a fixed statistical threshold was shown to vary dramatically over 33 examinations of the same subject. This, and other studies, has lead to the common interpretation that fMRI has a large inter-session variance, not only when sessions span over different subjects, but even when the different sessions are examinations of the same subject. While McGonigle et al. (2000) emphasised that inter-session variation should not be assessed on thresholded images as the result will be highly threshold dependent, they did not provide information on the statistical significance of the task-by-session interaction. The presence of a significant task-by-session interaction is problematic in several ways. In the field of non-clinical brain mapping, using second level group analysis, there will be a decrease of sensitivity due to increased unexplained variance. When fMRI is used for presurgical planning. False negative activations might lead to inadvertant surgical removal of normal tissue, resulting in greater disability. Conversly, false positive activations could lead to incomplete resection of a tumour.

The potential contributors to the intra-subject variability are numerous (Genovese et al., 1997). Compared to the number of choices available during fMRI acquisition and preprocessing, only a limited number of studies have investigated the impact of acquisition parameters and pre-processing methods on the within-subject sessional variation in the fMRI signal.

Slice-orientation, as reported by Gustard et al. (2001), had a non-significant impact on the reproducibility of the fMRI signal when isotropic voxels and a simple motor paradigm were used. Spatial normalisation was shown to have a significant effect on the reproducibility of visual activation by some (Swallow et al., 2003), but not others (Miki et al., 2000). In two studies, the effect of including neighbourhood information on the intra-subject variability was investigated. Intra-subject variability was decreased when larger smoothing kernels were used (Rombouts et al., 1998) or when the four nearest in plane neighbours were included in the analysis (Yetkin et al., 1996). These results are in agreement with what has also been observed for inter-subject variability (White et al., 2001; Shaw et al., 2003). Unfortunately reduction of variability by smoothing comes at the cost of reduced spatial resolution. Physiological noise correction (Hu et al., 1995) was found to increase the test-retest reproducibility at 4T (Tegeler et al., 1999).

Motion artefacts are some of the most important contributors to fMRI signal, unrelated to neural activity (Hajnal et al., 1995). Much of their effects can be removed by realignment but residual movement artefacts that are not accounted for by standard rigid-body realignment still exist. A commonly used

method (Rowe & Passingham, 2001; Salek-Haddadi et al., 2003) for correcting these residual movement artefacts is to include movement-parameters in the design matrix of a general linear model. If only the raw movement parameters (translations and rotations) are included, it is assumed that the effects are linear, and that movement in opposite directions result in opposite signal changes. This is not always the case. Consider for example the 1- dimensional case of a grey matter voxel lying between two white matter voxels. In this case movement of the voxel in either direction will lead to a signal increase. Using a Volterra expansion of the movement parameters, higher order and differential effects can also be modeled, including spin history effects (Friston et al., 1996). In the case of stimulus locked motion the inclusion of movement parameters in the design matrix is likely to remove not only residual movement artefacts but true activation as well, since these effects are no longer uniquely associated with the paradigm regressor.

The purpose of the present study was to investigate the extent to which the inter-session variability of the task-related fMRI activation could be improved by modeling residual motion, by inclusion of a Volterra expansion of movement parameters in the general linear model. In contrast to previous studies which have used coincidence maps, multi-panel displays or other strongly threshold dependent measures, we formally assessed the significance of the task-by-session interaction using F-contrasts.

Materials and methods

Experimental setup

Ten healthy volunteers (labelled A-J) were examined with three tasks of overt word generation: Categorical (generation of words from a specific category), Alphabetical (generation of words starting with a specific letter) and Semantic (generation of verbs associated with a specific noun). Two of the subjects (A&B) were examined 10 times. The alphabetical and categorical paradigms were presented in a boxcar design with active and baseline condition lasting 44s each. In the semantic paradigm, baseline and activation conditions each lasted 20s. Using a 1.5T Siemens Vision scanner and a gradient echo EPI, a set of 104 volumes (20 slices (interleaved acquisition), matrix:128×128, resolution (xyz): 1.56mm×1.56mm×5mm, TR=5.5s, TE=66ms, flipangle=90∘) was acquired in each of the three paradigms.

Data analysis

Pre-processing and data analysis was carried out using the SPM package (http://www.fil.ion.ucl.ac.uk/spm/). Data were realigned (6-parameter rigid body) (Friston et al., 1995), spatially normalised (Ashburner & Friston, 1999) and resampled to the EPI template and spatially smoothed using a 3D Gaussian kernel (FWHM=5mm). Residual movement effects were then modeled as described in Friston et al. (1996) by including a Volterra expansion of the 6 rigid-body motion parameters as nuisance covariates in the design matrix of a GLM (Worsley & Friston, 1995) implemented in SPM2. The Volterra expansion consisted of linear and quadratic effects of the 6 movement parameters belonging to each volume, and also included spin-history effects as linear and quadratic effects of motion parameters in the previous volume, giving a total of 24 regressors in addition to those describing the paradigm and baseline.

Nine different sets of images (Subject A: Categorical, Alphabetical and Semantic; Subject B: Categorical, Alphabetical and Semantic; 10 Subjects: Categorical, Alphabetical and Semantic) were analysed with two types of models each, giving 18 analyses in all. Both types of models were implemented as multi-session design matrices, and both modelled serial correlations as a first order auto regressive (AR(1)) process (Friston et al., 2002), and low frequency drifts as a discrete cosine set (128s cut-off).

In the first type of model, only the paradigm regressor and session specific baseline (mean value) were included in the design matrix (size $1040\times(10\times2)$). In the second type of model, the Volterra expanded motion parameters (24 regressors per session) were also included in the design matrix (size $1040\times(10\times(2+24))$). The expanded motion parameters were specified as covariates of no interest in the design matrix in SPM2. This was necessary because SPM2 estimates a global AR(1) process within a mask determined by the effects of interest. As the temporal autocorrelation has spatial structure (Worsley et al., 2002, Figure 2) with much higher temporal correlation in grey matter than in white matter, specifying all regressors as being of interest could bias the global AR(1) estimate toward less correlation, as white matter voxels will also show correlation with the movement parameters.

For each of the 18 analyses a t-contrast per session was used to test for the effect of the paradigm, and an F-contrast was used to test for the intra- or inter-subject variation. The F-test was constructed so that each of the 10 rows (SPM notation) in the F-contrast tested for the deviation of a specific session (or subject) from the mean of the other 9 sessions/subjects. For example, the first 10 columns (`spm_conman` notation) of the third row in the F-contrast of a model without motion parameters reads $[-1,-1,+9,-1,-1,-1,-1,-1,-1,-1]$. This contrast can be interpreted as a test for significant task-by-session interaction.

To determine the typical activation for each of the 18 analyses a mixed effect analysis was performed. As opposed to a typical two-level random effects analysis (Holmes & Friston, 1998), the mixed effect analysis acknowledges that the design may be unbalanced. This would happen if the correlation of movement parameters with the paradigm differs between sessions. The mixed effect analysis is implemented in SPM2 with the program `spm_mfx` and is described in (Friston et al., 2005).

The variance maps (F-tests) were thresholded at p=0.05, the family-wise error rate (FWE) was controlled using Gaussian Random Field Theory (GRF) (Worsley et al., 1996). The t-maps for the individual sessions were thresholded at p=0.05 corrected using false discovery rate (FDR) (Benjamini & Hochberg, 1995) which has adaptive features, and does not depend on heavy smoothing. This choice was motivated by the demands of presurgical planning, a typical example of single subject fMRI, in which false negatives are of major concern. As the FWE thresholding based on GRF may be too conservative for typical second-level analysis with low degrees of freedom (Nichols & Hayasaka, 2003), we used FDR thresholding only for the mixed effect analysis.

Results

This study produced 18 separate analyses. We are therefore only able to show detailed results from a subset of these analyses, and the remainder will only be commented on.

Intra-subject variance

In 6 of 6 intra-subject analyses (the three tasks in subjects A and B separately) inclusion of motion parameters reduced the inter-session variation over 10 sessions (see Figure 1 row 1 and 2). The resulting thresholded activation maps were correspondingly clearer with more focal activation patterns. For subject B in particular, inter-session variability was still visible after the correction, but this was restricted to areas in the prefrontal and temporal lobe that are typically involved in word generation.

Inter-subject variance

In all inter-subject analyses, inclusion of the motion parameters also reduced the inter-session variance significantly (see Figure 1 third row). Remaining

voxels with significant inter-session variability were mainly located in areas typically involved in word generation and also found active in the mixed effects activation map (not shown).

Impact on activation maps

As a representative example of activation maps from individual sessions we show in Figure 2a how the activation maps from the individual sessions of subject A, alphabetical paradigm, change when movement parameters are included in the design matrix. These maps should be compared with typical activation patterns (over 10 sessions of the Alphabetical task for Subject A) found in the maps from the mixed effect analysis shown in Figure 2b. From Figure 2a (simple) it is clear that examinations 5 and 7 contribute substantially to the inter-session variance map in Figure 1 first row, with areas that neither reflect typical language areas nor activation typical to the specific subject. This widespread "activation" is clearly removed when movement effects are modelled. In examinations 3 and 8 (Figure 2a) the language-related frontal and temporal cortex activation seen in the FDR thresholded maps disappear when movement effects are modelled. This indicates stimulus locked motion. Language areas in the maps from the remaining sessions are relatively unchanged when movement parameters are included. After modeling of movement related effects, the language related areas seen in FDR thresholded images from single sessions are very similar to the results of the mixed-effect analysis over multiple sessions in Figure 2b.

Discussion

McGonigle et al. (2000) highlighted the problem of large intra-subject variability in fMRI. We have here shown that residual movement artefacts are indeed a large part of the problem. Additionally, we have presented a method for explicitly assessing the significance of inter-session variance by the use of an F-contrast.

In the present study we modelled residual movement effects by including a Volterra expansion of motion-parameters in the design matrix of a GLM. Using this method we were able to assign a large proportion of the inter-session variation observed in fMRI to differences in movement during scanning. This effect was large both within and between subjects. This means that all types of inter-session variance in activation patterns are affected by residual movement artefacts, even though one might expect the effect to be more pronounced in the same subject. When the movement parameters are included in the model,

the activation image from a single examination resembles the mixed effect activation image from a single subject examined 10 times. This suggest that a more typical response for a given subject from a single session may be achieved by including movement parameters in the design matrix.

The difference in movement patterns between sessions means that the study is no longer balanced at the first level. A typical two-level random effects analysis is therefore in principle not optimal. Our results therefore suggest the importance of a true mixed effect analysis in multi-subject studies.

Our results also show the importance of careful model selection in single subject analyses such as presurgical mapping. At the risk of biased model selection, it may be justified to analyse the data with several different models e.g. with and without movement parameters included in the design matrix. If correction for movement effects abolishes what was thought to be task-related activation then this suggests the presence of stimulus locked motion, reducing the validity of the conclusions. The best solution would be to repeat the experiment. If, on the other hand, modeling of movement effects reduces spurious activations or leaves the activation map unchanged, it increases confidence in the activation images.

The inclusion of movement parameters in the design matrix is only one way to correct for residual movement artefacts. The method has been used in many papers already (Friston et al., 1996; Rowe & Passingham, 2001; Salek-Haddadi et al., 2003), but the impact of these regressors has, however, so far not been studied systematically in the context of inter-session variability.

In this study the images were smoothed with a narrow Gaussian kernel. This could lead to over-conservative estimates of FWE correction using GRF, and larger inter-session variance. However, it is the preferred strategy when fMRI is performed in the context of pre-surgical mapping to preserve spatial resolution. While the FDR approach for thresholding in a simple model can be liberal, we find that the combination of FDR thresholding and inclusion of movement parameters in the design matrix led to more consistent and representative activation images for a given subject.

One must consider other sources of inter-session variation. For instance, we did not in the present study formally assess the effects of session-to-session variation in task performance, which could in part explain the inter-session variation observed in the language areas that remained even after modeling residual movement effects. Furthermore, differences in cardiac and respiratory induced noise and true differences in the BOLD signal e.g. due to different levels of hormones or drugs such as caffeine (Mulderink et al., 2002) may also contribute to the inter-session variance in fMRI.

In conclusion, a significant proportion of the inter-session variability of task-

related activation can be explained by differences in movement patterns. The inclusion of movement parameters is therefore highly recommended. In the context of single-subject studies such as pre-surgical mapping, the validity of an fMRI activation can be assessed by analysis of the data with and without inclusion of movement parameters in the design matrix. Activation that persists after modeling movement effects can be considered more valid than activation that disappears after modeling movement.

References

Ashburner, J. & Friston, K. J. (1999). Nonlinear spatial normalization using basis functions. *Hum Brain Mapp*, **7**, 254–66.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Friston, K. J., Ashburner, J., Frith, C., Poline, J.-B., Heather, J. D., & Frackowiak, R. S. J. (1995). Spatial registration and normalization of images. *Human Brain Mapping*, **2**, 165–189.

Friston, K. J., Williams, S., Howard, R., Franckowiak, R. S. J., & Turner, R. (1996). Movement-Relatet effects in fMRI time series. *Magn Reson Med*, **35**, 346–355.

Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., & Ashburner, J. (2002). Classical and bayesian inference in neuroimaging: theory. *NeuroImage*, **16**, 465–83.

Friston, K. J., Stephan, K. E., Lund, T. E., Morcom, A., & Kiebel, S. (2005). Mixed-effects and fMRI studies. *NeuroImage*, **25**(1), 244–252.

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

Gustard, S., Fadili, J., Williams, E. J., Hall, L. D., Carpenter, T. A., Brett, M., & Bullmore, E. T. (2001). Effect of slice orientation on reproducibility of fMRI motor activation at 3 Tesla. *Magn Reson Imaging*, **19**, 1323–31.

Hajnal, J. V., Bydder, G. M., & Young, I. R. (1995). fMRI: does correlation imply activation? *NMR Biomed*, **8**, 97–100.

Holmes, A. P. & Friston, K. J. (1998). Generalisability, Random Effects and Population Inference. *NeuroImage*, **7**(4), s754.

Hu, X., Le, T. H., Parrish, T., & Erhard, P. (1995). Retrospective Estimation and Correction of Physiological Fluctuation in Functional MRI. *Magn Reson Med*, **34**, 201–212.

McGonigle, D., Howseman, A., Athwal, B., Friston, K., Frackowiak, R., & Holmes, A. (2000). Variability in fMRI: an examination of intersession differences. *NeuroImage*, **11**(6 Pt 1), 708–34.

Miki, A., Raz, J., van Erp, T. G., Liu, C. S., Haselgrove, J. C., & Liu, G. T.

(2000). Reproducibility of visual activation in functional MR imaging and effects of postprocessing. *AJNR*, **21**, 910–5.

Mulderink, T. A., Gitelman, D. R., Mesulam, M. M., & Parrish, T. B. (2002). On the use of caffeine as a contrast booster for BOLD fMRI studies. *NeuroImage*, **15**, 37–44.

Nichols, T. & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Midical Research*, **12**, 419–446.

Rombouts, S., Barkhof, F., Hoogenraad, F., Sprenger, M., & Scheltens, P. (1998). Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. *Magn Reson Imaging*, **16**(2), 105–13.

Rowe, J. B. & Passingham, R. E. (2001). Working memory for location and time: activity in prefrontal area 46 relates to selection rather than maintenance in memory. *NeuroImage*, **14**, 77–86.

Salek-Haddadi, A., Lemieux, L., Merschhemke, M., Friston, K. J., Duncan, J. S., & Fish, D. R. (2003). Functional magnetic resonance imaging of human absence seizures. *Ann Neurol*, **53**, 663–7.

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

Swallow, K. M., Braver, T. S., Snyder, A. Z., Speer, N. K., & Zacks, J. M. (2003). Reliability of functional localization using fMRI. *NeuroImage*, **20**, 1561–77.

Tegeler, C., Strother, S. C., Anderson, J. R., & Kim, S. G. (1999). Reproducibility of bold-based functional MRI obtained at 4 T. *Human Brain Mapping*, **7**(4), 267–83.

White, T., O'Leary, D., Magnotta, V., Arndt, S., Flaum, M., & Andreasen, N. (2001). Anatomic and functional variability: the effects of filter size in group fMRI data analysis. *NeuroImage*, **13**(4), 577–88.

Worsley, K. J. & Friston, K. J. (1995). Analysis of fMRI time-series revisited–again. *NeuroImage*, **2**, 173–81.

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

Worsley, K. J.and Marrett, S., P. Neelin, P., Vandal, A. C., Friston, K. J., & Evans, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, **4**(1), 58–73.

Yetkin, F. Z., McAuliffe, T. L., Cox, R., & Haughton, V. M. (1996). Test-retest precision of functional MR in sensory and motor task activation. *AJNR*, **17**, 95–8.

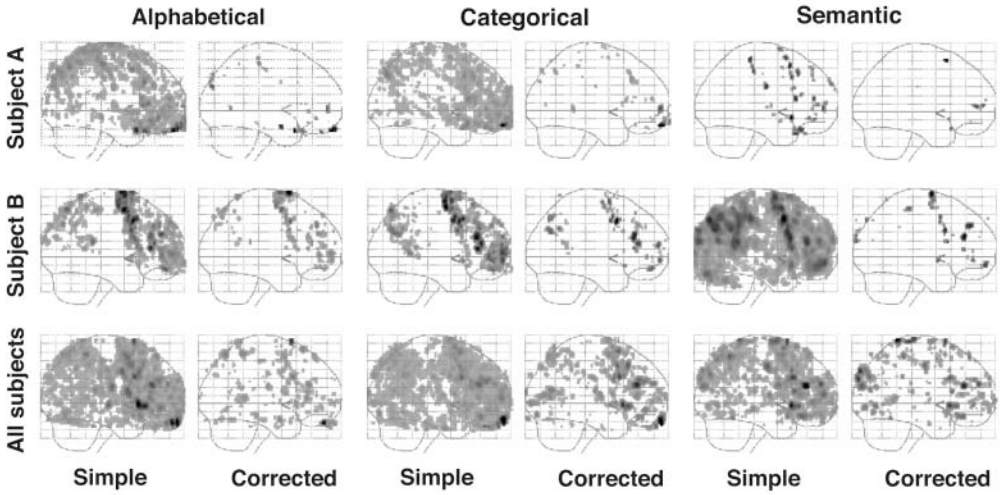

Fig. 1. F-test for inter-session variance in fMRI activation. The multi panel display shows maximum intensity saggital projections of the inter-session variation in task-related activity observed during the three different language paradigms (Alphabetical, Categorical and Semantic) across the 10 repeated examinations of subject A and B and across all 10 subjects. For each task the left hand column (Simple) shows results from the model without movement parameters, and the right hand column (Corrected) shows results from the model with movement parameters included. All maps are thresholded at p=0.05, corrected for multiple testing, controlling FWE using GRF.

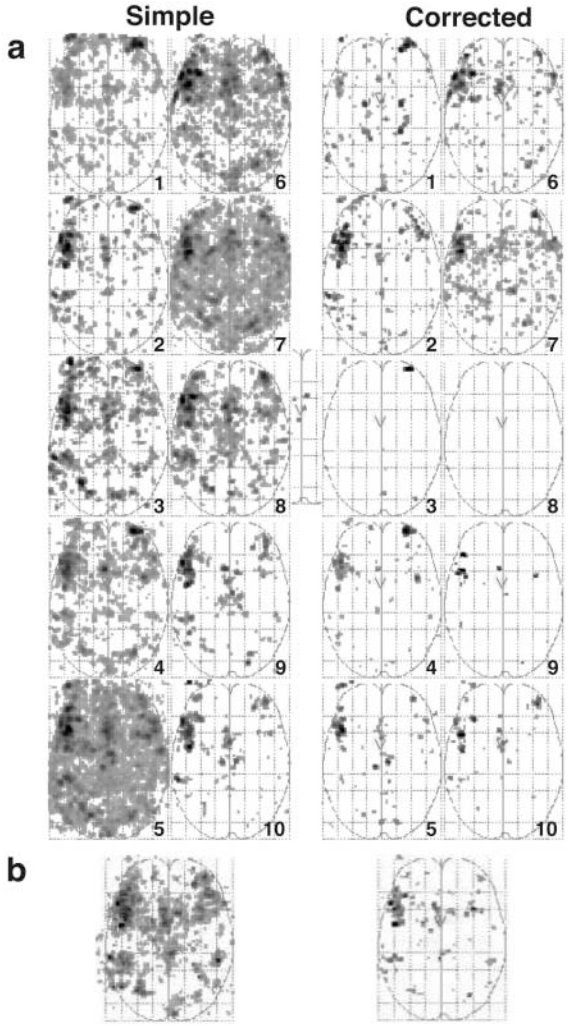

Fig. 2. t-test for effect of alphabetical word generation. The multi panel display **(a)** shows the thresholded activation images (transversal maximum intensity projections) from subject A through the 10 different sessions of the Alphabetical paradigm, without (simple) and with (corrected) modeling of residual movement artefacts. Figure **(b)** shows the results of the the mixed effect analysis of subject A, alphabetical task, without (Simple) and with (Corrected) correction for residual movement artefacts. All maps are thresholded adaptively at p=0.05, corrected for multiple testing, controlling FWE weakly with FDR.