Identifying sources of software-dependent differences in task fMRI analyses

Alexander Bowring, Thomas Nichols, Camille Maumet

Identifying sources of software-dependent differences in task fMRI analyses

Alexander Bowring, University of Oxford, UK;
Thomas Nichols, University of Oxford, UK;
Camille Maumet, Univ Rennes, Inria, CNRS, Inserm, IRISA, Rennes, France.

Introduction

A plethora of tools and techniques are available to process and model fMRI data. However, this flexibility comes with a drawback: the application of different analysis pipelines (Botvinik-Nezer, 2020), software versions (Glatard, 2015) and even operating systems (Gronenschild, 2012) can cause variation in the results of an fMRI study, increasing the risk of obtaining irreproducible research findings (Poldrack, 2017).

Recently, we discovered that the choice of software package used to conduct the analysis can also yield conflicting results (Bowring, Maumet & Nichols, 2019, *BMN*). We observed differences in the sizes and magnitudes of activated brain regions for three task fMRI datasets when the data were processed with AFNI, FSL and SPM. Here we revisit that work, seeking to find *where* in the analysis pipeline the greatest variation between software is induced. We run the same datasets through a series of hybrid analysis pipelines, mixing and matching the workflow steps from the three different packages. We apply quantitative comparisons to assess the similarity of our results and isolate the stages of the pipeline where the packages diverge.

Methods

In *BMN*, we reanalysed data from three published task fMRI studies (Schonberg, 2012; Moran, 2012; Padmanabhan, 2011), replicating the group-level result for the principal effect depicted in the main figure of each publication within the three packages. The datasets were obtained from the OpenfMRI (Poldrack, 2015) database (ds000001, R: 2.0.4; ds000109, R:2.0.2; ds000120, R:2.0.4).

Our first aim with this work was to identify whether the largest sources of software-variability are during the preprocessing or statistical modelling. To do this we repeated the *BMN* analyses, except this time applying a common fMRIPrep preprocessing strategy to each dataset before carrying out the rest of the analyses in the three packages. Subsequently, this led to a more in-depth exploration of variation between the softwares' modelling procedures. We partitioned the first-level data modelling into three components: the fMRI signal model (design matrix), noise model, and drift model. We then carried out further analyses, interchanging between the packages at these parts of the workflow (e.g. a pipeline using SPM's noise model and FSL's design matrix). Figure A shows all hybrid SPM/FSL pipelines implemented for the ds000001 dataset.
Finally, we applied quantitative methods to assess the similarity of our results: Pearson's *r*, assessing the correlation between the profile of statistical values obtained in the unthresholded maps, and Dice statistics, comparing the locations of activation in the FWE-thresholded maps.

All of the analysis code used for this submission has been made available via Github (Release SC2_0.1.0): https://github.com/AlexBowring/SC2/releases/tag/0.1.0

Results

Figure 1 presents comparisons of the ds000001 balloon analog risk task group-level results for combinations of SPM and FSL pipelines, where inference was on the '*reward_parametric* > *control_parametric'* effect (*t*-statistic). Group-level inference was conducted using a cluster-forming threshold $p < 0.01$, FWE-corrected clusterwise threshold $p < 0.05$.

Figure 2 presents comparisons of the ds000120 antisaccade task group-level results obtained for combinations of AFNI and SPM pipelines, where inference was on the main effect of time (*F*-statistic). Group-level inference was conducted using a cluster-forming threshold $p < 0.001$, FWE-corrected clusterwise threshold $p < 0.05$.

Conclusions

Our analyses have shed light on the main sources of pipeline-variability between AFNI, FSL and SPM. While differences a in low frequency drift models had negligible impact, we found substantially more variation from changes in the fMRI signal and group-level inference models. We hope that these results stimulate discussion in the community about inevitable variation ("agree to disagree") vs avoidable variation (e.g. consensus benchmarks available).

References

Botvinik-Nezer, R. (2020). 'Variability in the analysis of a single neuroimaging dataset by many teams.' Nature 582, 84–88.

Bowring, A. (2019). 'Exploring the impact of analysis software on task fMRI results.' Human Brain Mapping. 40(11), 3362-3384.

Glatard, T. (2015). 'Reproducibility of neuroimaging analyses across operating systems.' Frontiers in neuroinformatics, 9.

Gronenschild, E. H. (2012). 'The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements.' PloS one, 7(6), e38234.

Moran, J. M. (2012). 'Social-cognitive deficits in normal aging.' Journal of neuroscience, 32(16), 5553-5561.

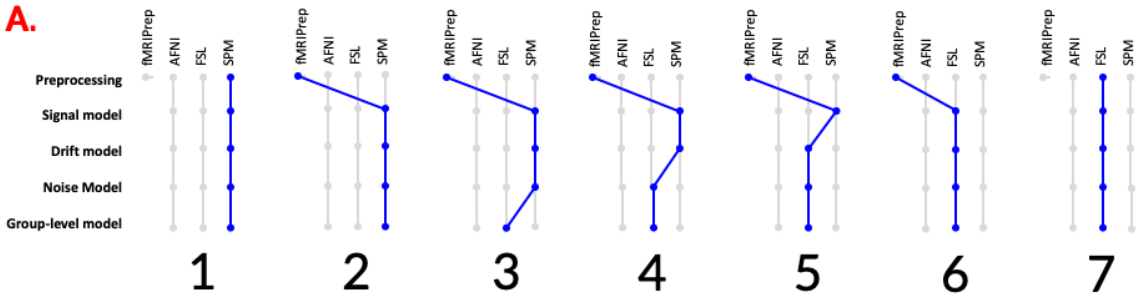
Padmanabhan, A. (2011). 'Developmental changes in brain function underlying the influence of reward processing on inhibitory control.' Developmental cognitive neuroscience, 1(4), 517-529.

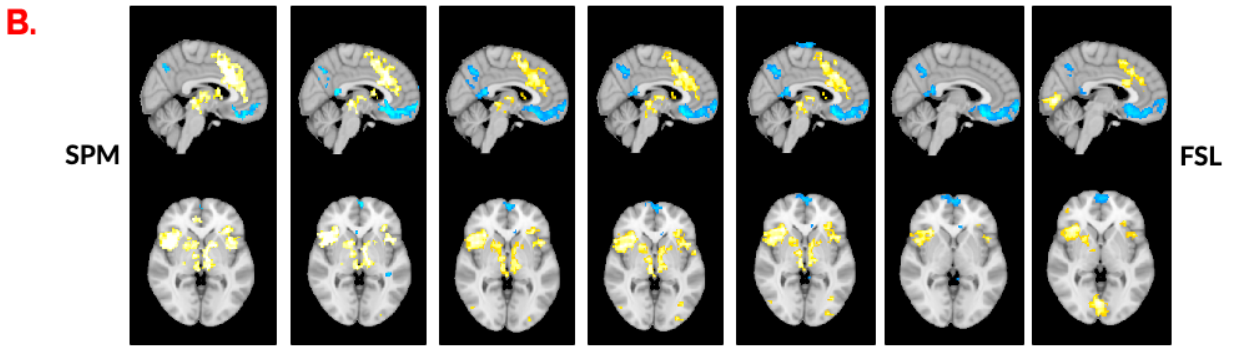
Poldrack, R. A. (2015). 'OpenfMRI: open sharing of task fMRI data.' NeuroImage. vol. 144, part B, pp. 259-261.

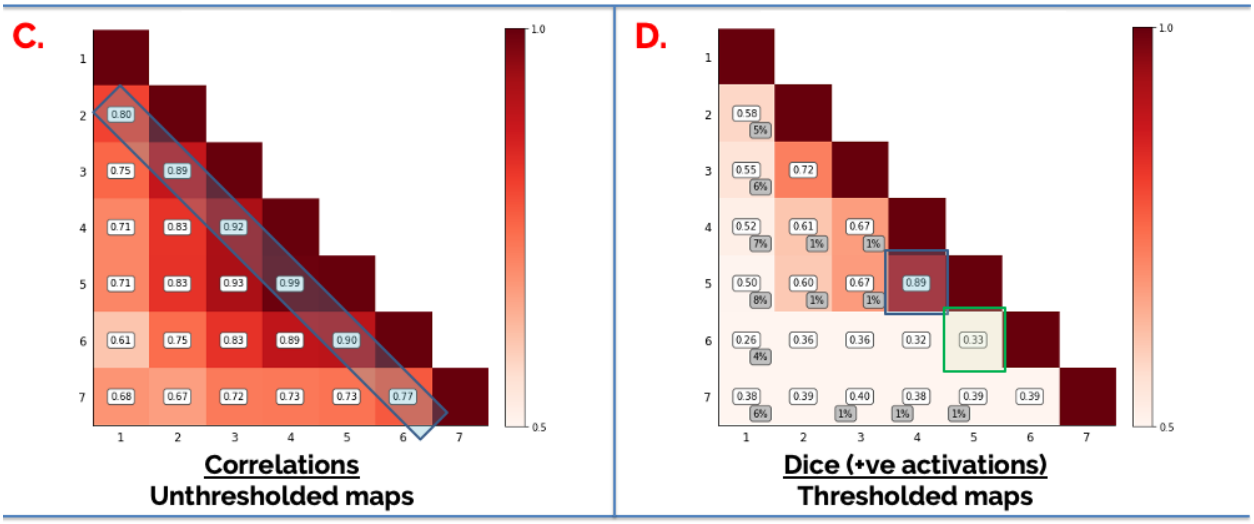
Poldrack, R. A. (2017). 'Scanning the horizon: towards transparent and reproducible neuroimaging research.' Nature Reviews Neuroscience, 18(2), 115-126.

Schonberg, T. (2012). 'Decreasing ventromedial prefrontal cortex activity during sequential risk-taking: an fMRI investigation of the balloon analog risk task.' Frontiers in neuroscience, 6.

ds000001 SPM/FSL Pipelines & Results







The ds000001 group-level results for all pipelines used to create these figures have been made available in a Neurovault respitory: https://neurovault.org/collections/8381/

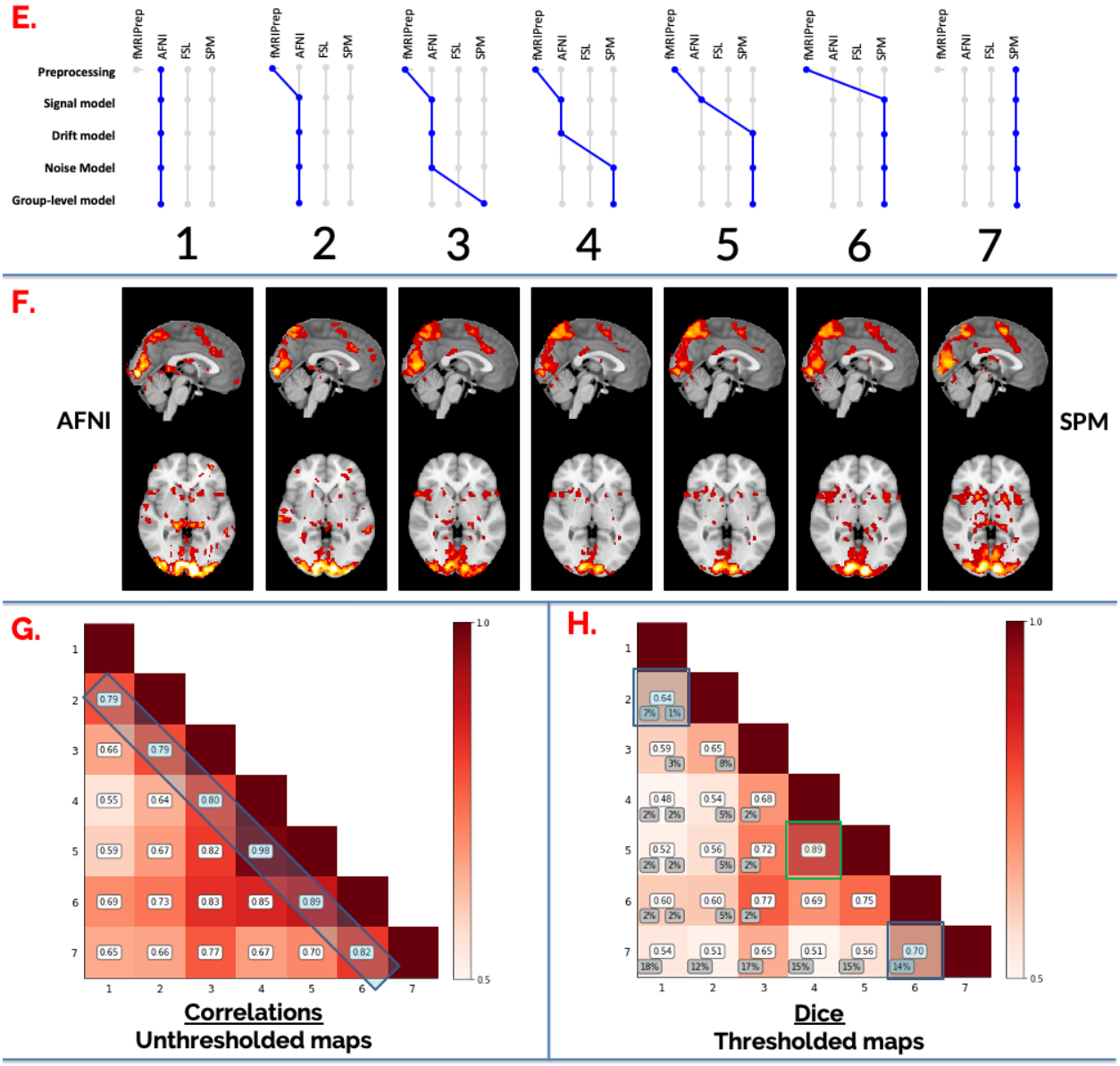
A. Diagrams to represent the 7 pipelines carried out for the ds000001 dataset that included analysis procedures from either SPM or FSL. Pipelines **1** and **7** are the workflows where the entire analysis was carried out in SPM and FSL, respectively. Pipelines **2** and **6** are similar to **1** and **7** except that fMRIPrep was used to preprocess the data rather than each package's respective preprocessing workflow. For pipelines **3** to **5**, SPM and FSL were interchanged at various stages of the statistical modelling. For example, in pipeline **4** SPM's first-level signal and drift model were applied to the data alongside FSL's first-level noise model and group-level model. Notably, the ordering of the pipelines is such that only one specific analysis procedure is changed between adjacent pipelines. For instance, the only difference between pipelines **3** and **4** was whether SPM's or FSL's noise model was applied, and therefore any discrepancies between the group-level results for these pipelines are wholly attributable to differences between the two softwares' noise models.

B. The thresholded *t*-statistic maps for all pipelines using parametric inference with a cluster-forming threshold $p < 0.01$, FWE-corrected clusterwise threshold $p < 0.05$. For most sets of results, positive activation was identified in the anterior cingulate, the anterior insula (bilateral) and the thalamus (bilateral), and negative activation was identified in the dorsomedial prefrontal cortex. The main exception to this was pipeline **6**, which did not determine activation in either the anterior cingulate or thalamus. On further investigation of the unthresholded maps, we found that this was due to a facet of the clusterwise inference performed here: while the other pipelines obtained a single large activation cluster in the anterior cingulate, for pipeline **6** this broke up into smaller, disconnected clusters, causing the activation to be 'thresholded out' after the FWE clusterwise correction. Finally, it is notable that the activation clusters for pipelines **1** and **2** (using SPM's group-level inference model) appear to be brighter than the other sets of results (which use FSL's group-level model). This may be because the mixed-effects model implemented by FSL's FLAME deweights the most variable subjects, consequentially leading to smaller statistic values than SPM's OLS group-level model.

C. Correlation coefficients (Pearson's *r*) for pairwise comparisons of the unthresholded *t*-statistic maps obtained with all pipelines. Values on the off-diagonal of the matrix (highlighted by the blue window) show the correlations between the final group-level statistic maps when the software package is changed at one particular analysis step. In this regard, the fact that these values are all relatively high (all correlations > 0.77) suggests that differences in the overall activation profiles obtained by SPM and FSL are not due to substantial divergence between the two packages at one specific analysis procedure, but rather an accumulation of smaller differences across the entire pipeline.

D. Dice statistic values for pairwise comparisons between the positive threshold *t*-statistic maps obtained with all pipelines. Dice measures the overlap of voxels between two sets of thresholded maps as a proportion of the total spatial extent covered by both maps activation's (i.e. a Dice coefficient of 1 indicates the areas of activation are identical in both maps, while 0 indicates complete disagreement). Grey values show the percentage of 'spill-over' activation, that is, the percentage of activation in one pipeline's thresholded map that fell outside the analysis mask of the others. The Dice coefficient of 0.33 for pipelines **5** and **6** (green window) signifies a dramatic change in the regions of activation identified when using FSL's signal model rather than SPM's, as already discussed in the caption to Fig. B. On the other hand, the Dice coefficient of 0.89 for pipelines **4** and **5** (blue window) suggests a strong similarity between SPM's and FSL's drift models, backed up by the corresponding correlation value of 0.99 in Fig. C. It is notable that the Dice values here are on-the-whole worse than the correlations in Fig. C., suggesting that the regions with higher brain activation were also the most variable across pipelines.

ds000120 AFNI/SPM Pipelines & Results



The ds000120 group-level results for all pipelines used to create these figures have been made available in a Neurovault respitory: https://neurovault.org/collections/9324/

E. Diagrams to represent the 7 pipelines carried out for the ds000120 dataset that included analysis procedures from either AFNI or SPM. The interpretation here is the same as discussed in Fig. A, except here we consider pipelines using procedures from AFNI and SPM rather than SPM and FSL.

F. The thresholded *F*-statistic maps for all pipelines using parametric inference with a cluster-forming threshold $p < 0.001$, FWE-corrected clusterwise threshold $p < 0.05$. Qualitatively there were similarities between all the sets of results here, with activation determined in the occipital pole (bilateral), lateral occipital cortex (bilateral) and occipital fusiform gyrus, as well as the supplementary motor cortex and middle frontal gyrus (bilateral). However, there was greater variation between the pipelines in areas where weaker effects were present, as can be seen by the different scatterings of smaller clusters in the axial slice displayed (bottom row). It is notable that the activations in the occipital lobe for pipelines applying SPM's group-level inference model seem to be more concentrated than the two pipelines (**1** and **2**) which used AFNI's group-level model.

G. Correlation coefficients for pairwise comparisons of the unthresholded *F*-statistic maps obtained with all pipelines. The overall distribution of values seen here is remarkably similar to the corresponding correlations displayed for ds000001 in Fig. C, suggesting that the main sources of variability between software are independent of the dataset being analyzed. Focusing on the values on the off-diagonal of the matrix (blue window), we once again observe that the correlations between maps is relatively high when the software package is changed at only one individual analysis stage. In particular, the correlation of 0.98 for pipelines **4** and **5** suggests that the differences between AFNI's and SPM's modelling of the low-frequency fMRI drifts are minimal.

H. Dice statistic values for pairwise comparisons between the thresholded *F*-statistic maps obtained with all pipelines. The distribution of values here is similar to the correlations presented in Fig. G, although the Dice values are generally smaller. Once again, the Dice value of 0.89 for the two pipelines (**4** and **5**) that differed by the choice of drift model was the largest (green window), while the relatively smaller Dice values for the pipelines where fMRIPrep was used instead of each software's preprocessing workflow (**1** and **2**, **6** and **7**) indicate a more substantial change in the final regions of activation determined.