Identifying Sources of Software-dependent Differences in Task-fMRI Analyses

Alex Bowring[1], Camille Maumet[2], Thomas E. Nichols[1]

[1.] The University of Oxford, UK [2.] Inria, Univ Rennes, CNRS, Inserm, France

Introduction

A plethora of tools and techniques are available to process and model fMRI data. However, this **analytical flexibility** comes with a drawback: the application of different analysis pipelines, software versions and even operating systems can cause **variation in the results** of an fMRI study. When combined with **selective reporting practices**, where only methods that report a favorable outcome are likely to be published, the consequences of this can lead to **overstated and irreproducible research findings**.

Previous Research

In ***Bowring, Maumet, Nichols, 2019 (Human Brain Mapping)***, we investigated the question:

How does the choice of analysis software package impact task-fMRI analysis results?

We **reproduced the results** of three published task-fMRI studies using **AFNI, FSL and SPM**, and then applied a range of quantitative comparison methods to assess the similarity between the group-level statistical results maps.

We observed considerable differences in the shapes and magnitudes of activated brain regions for AFNI, FSL and SPM.

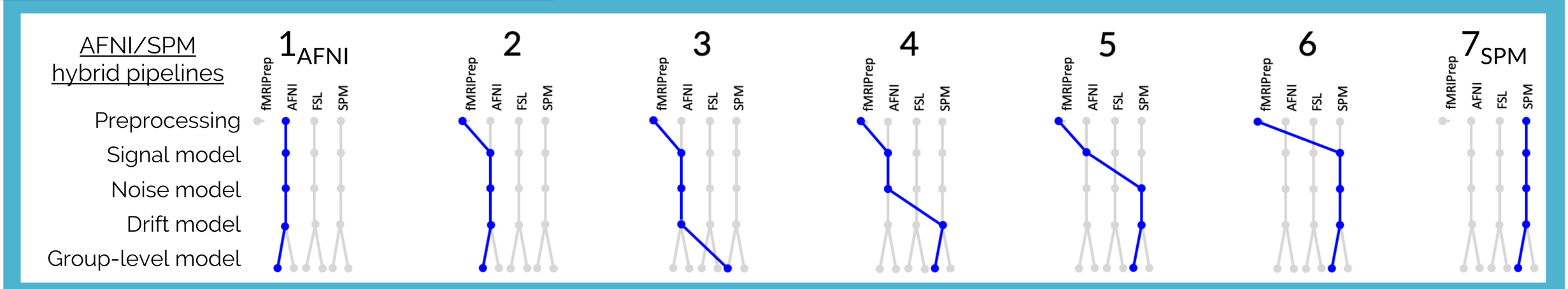
This Work

While our previous research showed that AFNI, FSL, and SPM can produce conflicting analysis results, the question remains:

Where in the analysis workflow does variation between software occur?

To investigate this we revisit our previous work, running the same three datasets through a **collection of hybrid pipelines** that **interchange procedures** from AFNI, FSL, SPM at **different stages** of the analysis workflow, as well as utilizing a common **fMRIPrep preprocessing** pipeline. We then compare the unthresholded and thresholded group-level statistic maps to **isolate** the parts of the analysis workflow **where the three packages diverge**.

Pipeline Generation



Partition analysis pipeline into 5 stages: Preprocessing, signal model, noise model, drift model, Group-level model

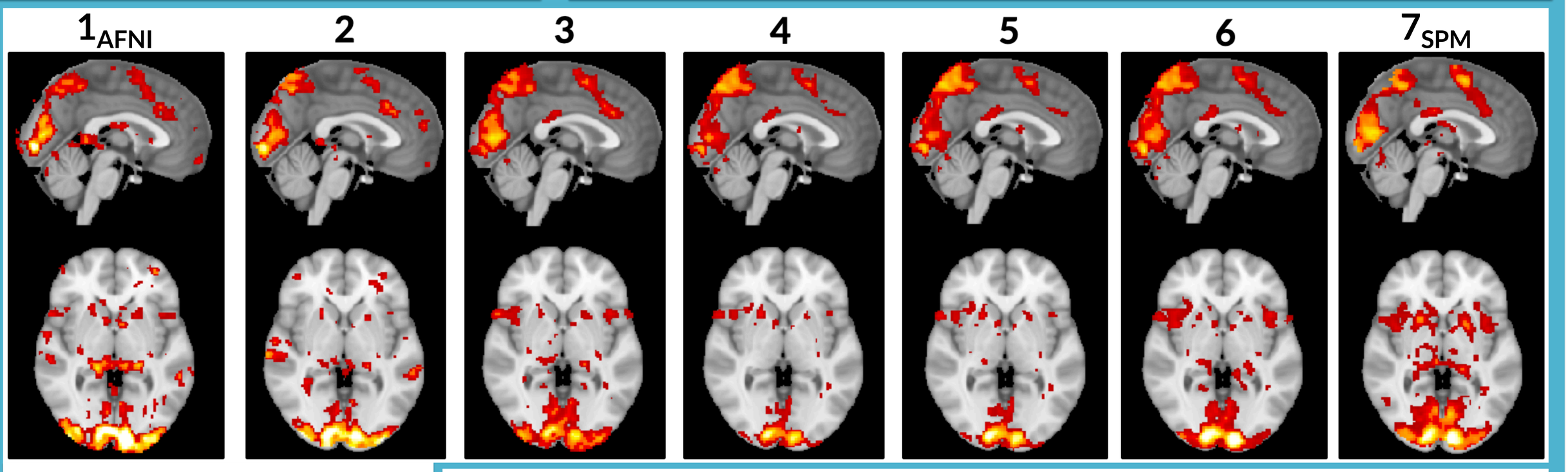
Carry out analyses, interchanging between software at the 5 analysis stages. **Figure shows 7 hybrid AFNI/SPM pipelines**

For preprocessing we also utilized a common **fMRIPrep** workflow

At the group-level we considered both **parametric** and **non-parametric** inference (**2 branches in bottom row of figure**)

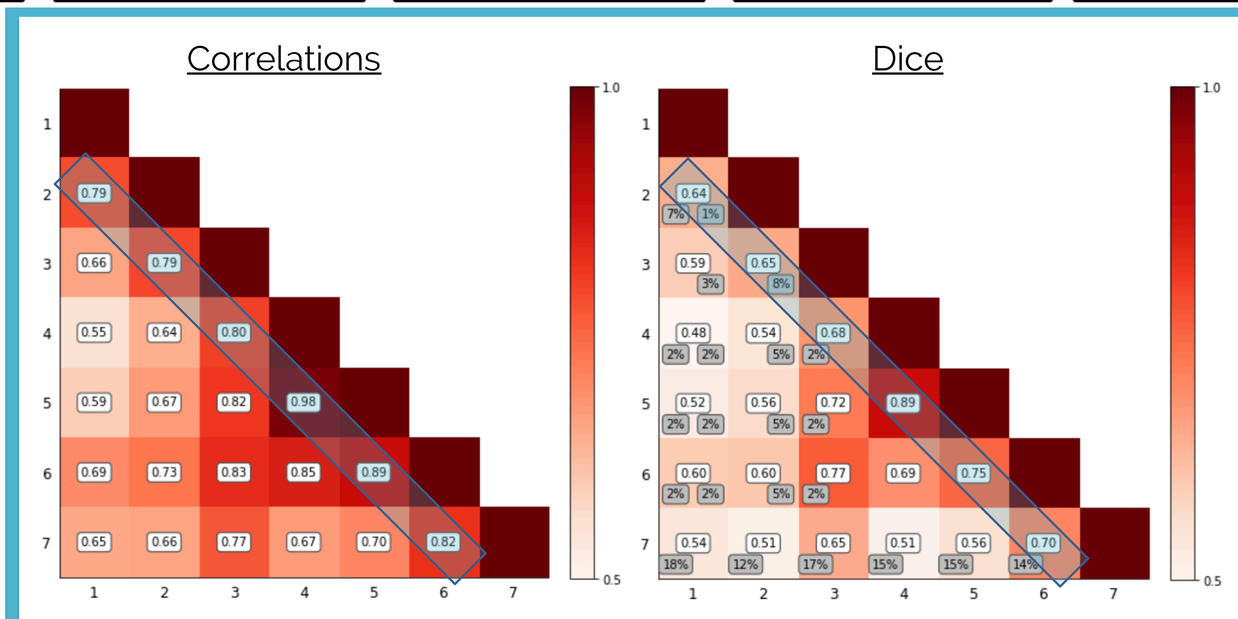
Results

Here we present parametric group-level results obtained from AFNI/SPM hybrid pipelines reproducing the results of *Padmanabhan et al., 2011* (OpenNeuro dataset ds000120), the main effect of an antisaccade task fit with a flexible Fourier HRF basis (F-test).



Thresholded Statistic Maps

Slice comparisons of the thresholded F-statistic maps obtained for the seven AFNI/SPM hybrid pipelines displayed in the 'Pipeline Generation' figure. Parametric inference was carried out 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; activation was 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 activation clusters for the two pipelines using AFNI's group-level model (pipelines **1** and **2**) generally reported larger statistic values, particularly in the occipital lobe, where SPM's activations were more concentrated.



Quantitative Comparisons

Correlations provide pairwise comparisons of the unthresholded F-statistic maps, and Dice coefficients provide comparisons of the thresholded maps (grey values show % of activation for one pipeline's thresholded map that fell outside the other pipeline's analysis mask). In both plots, values on the off-diagonal (blue windows) show the correlation/dice between statistic maps when the software package was changed at a single analysis step. It is notable that the correlation/Dice values for pipelines **4/5** were the highest here, suggesting a strong similarity between AFNI's and SPM's drift model. Conversely, comparisons of pipelines **1/2** the lowest, indicating larger differences between the preprocessing workflows carried out in AFNI and fMRIPrep.

Conclusion

Our results have revealed an intricate picture regarding areas of the the pipeline where a change of software can manifest as variability in task-fMRI results. While the final regions of activation for *Padmanabhan et al.* were shown to be contingent on the software package used for the group-level model and preprocessing, for the two other datasets we reanalyzed (not presented here) we observed that differences in the first-level signal and noise models contributed most to analytic variability. However, it is notable that *all* our analysis results have indicated harmonization as to how AFNI, FSL, and SPM model the low-frequency fMRI drifts.

References

Carp J. (2012), On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments, Frontiers in Neuroinformatics

Bowring A. (2019), Exploring the impact of analysis software on task fMRI results, Human Brain Mapping

Esteban O. (2018), fMRIPrep: a robust preprocessing pipeline for functional MRI, Nature Methods

Padmanabhan A. (2011), Developmental changes in brain function underlying the influence of reward processing on inhibitory control, Developmental Cognitive Neuroscience