Same Data – Different Software – Different Results?

Analytic Variability of Group fMRI Results

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

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

Come and see the talk!
Informatics Session,
10:30 Tuesday,
Room 324 - 326

1. Introduction

A wealth of tools and techniques are now available to process and model fMRI data. However, this '**methodological plurality**' has come with a drawback.
Application of different analysis pipelines[1], alterations in software version[2], and even changes in operating system[3] have all been shown to cause **variation in the results** of a neuroimaging study. This high analytic flexibility has been pinpointed as a key factor that can lead to increased false-positives[4], and compounded with a lack of data sharing, irreproducible research findings[5].

Our Question

How does the choice of analysis software package impact the analysis results?

We **reproduce the results** of three published neuroimaging studies with publicly available data using **AFNI**, **FSL** and **SPM**, for parametric and nonparametric inference. We make a variety of comparisons to assess the similarity of both sets of results across software packages.

2. Method

We reanalyzed three published fMRI studies whose data have been made publicly available on the OpenNeuro repository: ds000001, ds000109, ds000120.
We aimed to recreate the main figure from each publication by replicating the original analysis pipeline within each software package. However, to maximize comparability across software we determined a number of processing steps to be included in all of our analyses.

Replicating the Study

- Analysis pipelines based on original study
- Info on how to model and analyze data obtained from publication
- Compare our results to publication figure

Maximizing Comparability

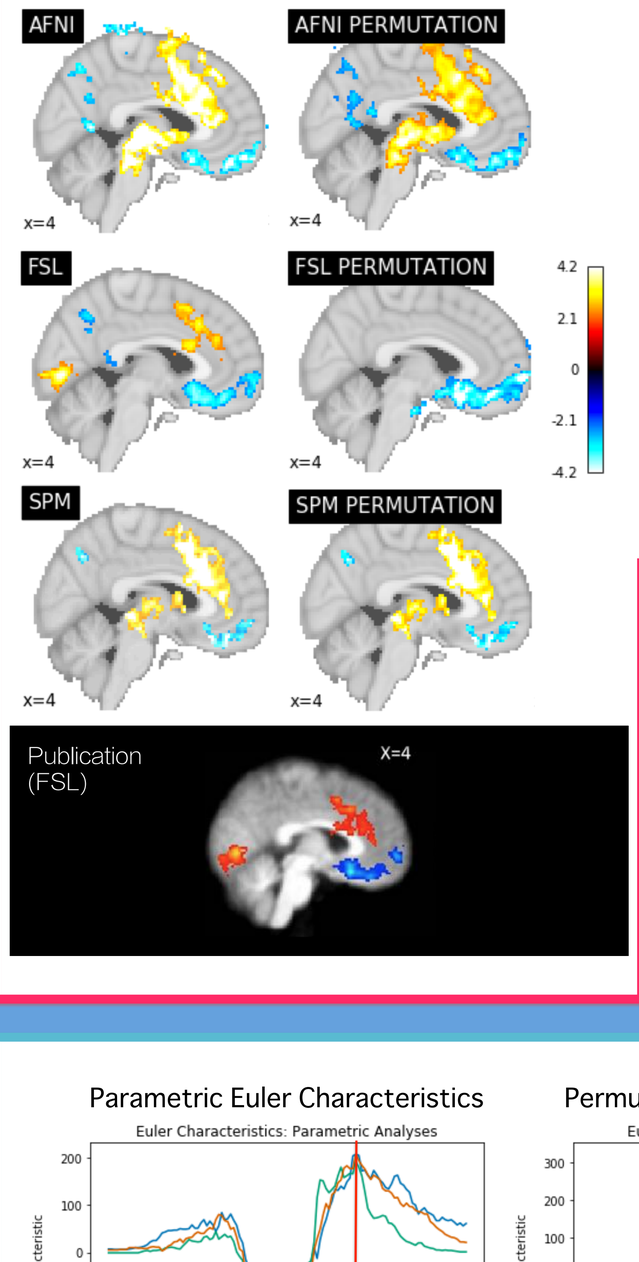
- Non-linear warping of functional data to MNI space
- 6 motion regressors added to GLM
- Data resampled to 2mm voxel sizes

To assess differences in the statistic maps obtained in each package we applied three quantitative comparison methods:

- Dice coefficients; differences in **locations** of activation clusters
- Euler characterizes; differences in **topological properties** of images
- Bland-Altman plots; assess differences in **magnitude** of stat values

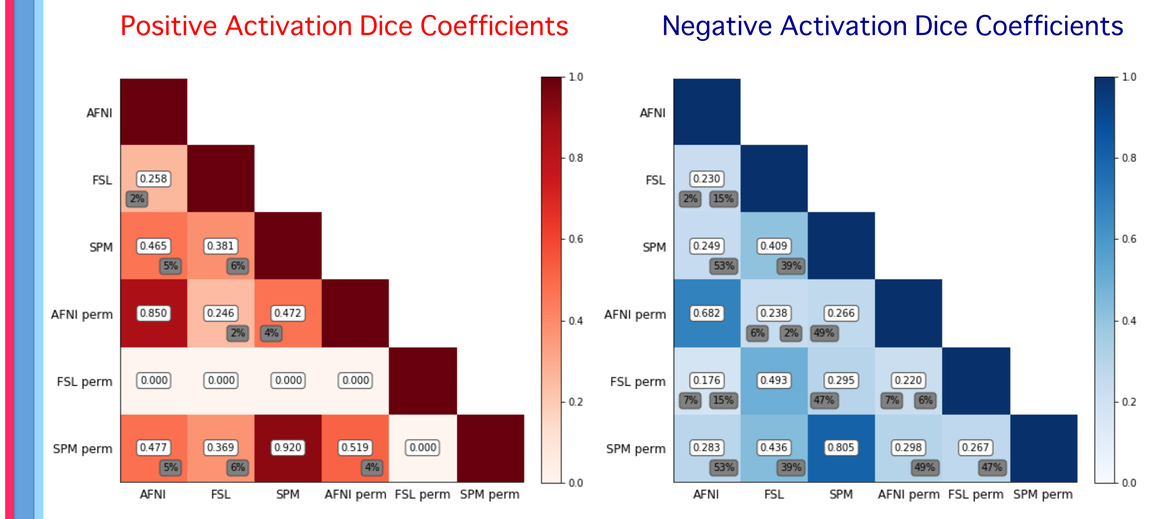
3. Results

Here we present comparisons of the main contrast 'pumps control vs pumps demean' from the ds000001 study,
See our bioRxiv preprint for full results! https://doi.org/10.1101/285585



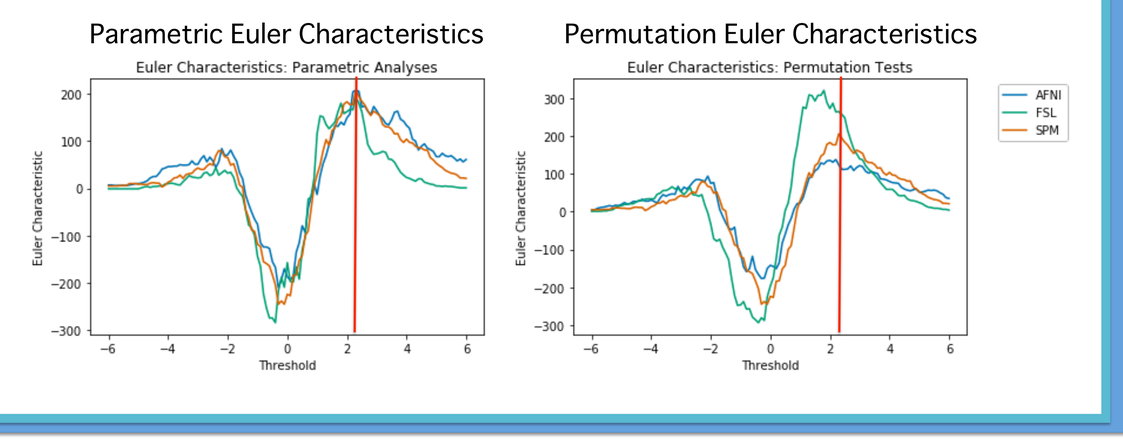
Thresholded Statistic Maps

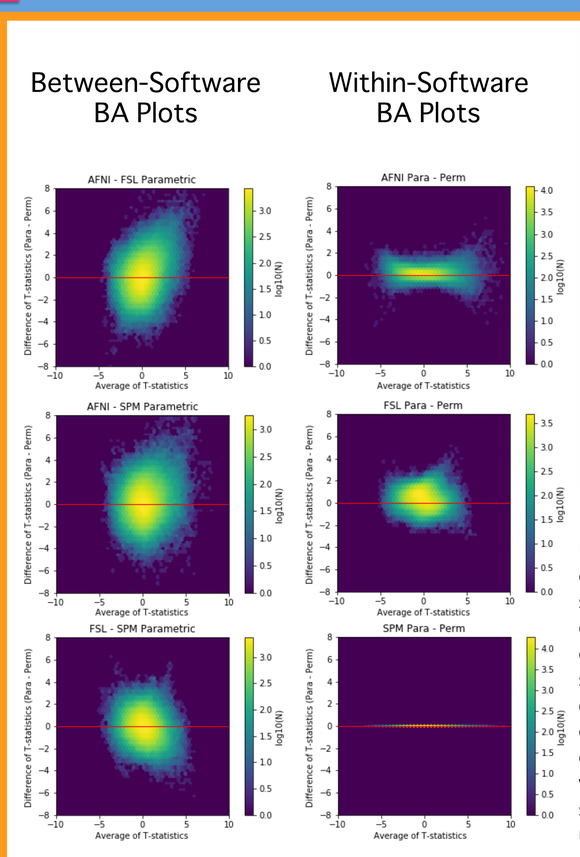
Thresholded T-statistic maps for all packages using parametric and non-parametric inferences, FWE clusterwise threshold $p < 0.05$, with a cluster- forming threshold $p < 0.01$ uncorrected.
Overall, the locations of significant clusters are broadly similar across packages and inference methods. However, the results show discordance in the magnitude of T-statistic values, as well as differences in the sizes of corresponding clusters.
The FSL parametric result is most similar to the published figure (bottom), which is to be expected since FSL was used in the original study. AFNI and SPM generally find larger clusters but lack the primary visual response. All permutation methods use ordinary least squares (no first-level variance information), as does SPM parametric; this explains the close match between the two SPM results. The significant cingulate cluster is large and covers multiple anatomical regions; in FSL permutation this cluster happened to break up and lose significance.



Euler Characteristics

The Euler Characteristic (EC) summarizes the topology of a thresholded image. It is the number of clusters minus the number of 'handles' plus the number of 'holes' in a thresholded image; for extreme positive or negative thresholds this is essentially the number of clusters, however for moderate thresholds it provides a signature of the image. The EC is therefore informative over a range of thresholds. For the cluster-forming threshold (red line, $p = 0.01$, $z = 2.32$), the EC is larger for FSL permutation compared to parametric. This suggests permutation found more smaller, fragmented clusters, consistent with above findings.





Dice Coefficients

Dice coefficient values for pairwise comparisons between the positive (left) and negative (right) thresholded T-statistic maps. Dice is the size of the overlapping region divided by the average size the regions. Between-software dice coefficients are substantially smaller than 1 (perfect agreement), suggesting high spatial variability in the location of activations. As FSL permutation inference found no significant positive activations, the dice values for the corresponding row in the left triangle are all zero. Within-software coefficients are better by comparison, suggesting spatial coherence across the two inference methods within each package. The percentage of activation in one software's thresholded map that fell outside of the analysis mask of the other software is displayed in grey. These values indicate the variability in the size of the analysis masks used by each package.

Bland–Altman Plots

Bland-Altman (BA) plots (here presented as histograms) compare the unthresholded T-statistic maps between software for parametric inference (left column), and comparing both inference methods within software (right column). For between-software comparisons, we see a great spread in the vertical axis, indicating substantial disagreement. In general, densities are distributed evenly each side of the x-axis, suggesting that no package consistently greater activation than the other two. For the within-software comparisons, since SPM uses ordinary least squares for both methods, we see no differences in the unthreshold T-statistic maps. AFNI and FSL both show some disagreement, however the scale of these differences is far less relative to the inter-software comparisons.

4. References

1. Carp, J. (2012), On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments. Frontiers in Neuroscience, 6.
2. Glatard, T. (2015). Reproducibility of neuroimaging analyses across operating systems. Frontiers in neuroinformatics, 9.
3. 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.
4. Ioannidis, J. P. (2005). Why most published research findings are false. PLoS medicine, 2(8), e124.
5. Poldrack, R. A. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. Nature Reviews Neuroscience, 18(2), 115-126.

5. Conclusion

We have found a disappointing level of agreement between software packages. While the general pattern of activations found was similar, the best inter-software Dice overlap was 52% (AFNI permutation - SPM permutation). This work supports the need for open sharing of data, and the metrics and measurements introduced here encourage further exploration into software validation.