# Same Data - Different Software - Different Results? Analytic Variability of Group fMRI Results

Alexander Bowring, Thomas E. Nichols, Camille Maumet

# Same Data - Different Software - Different Results? Analytic Variability of Group fMRI Results

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

## Introduction

A plethora 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 (Carp, 2012), alterations in software version (Glatard, 2015), and even changes in operating system (Gronenschild, 2012) 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 (Ioannidis, 2005), and compounded with a lack of data sharing, irreproducible research findings (Poldrack, 2017).

In this work, we seek to understand how choice of software package impacts analysis results. We reproduce the results of three published neuroimaging studies (Schonberg, 2012; Moran, 2012; Padmanabhan, 2011) with publicly available data within the three main neuroimaging software packages: AFNI, FSL and SPM, using parametric and nonparametric inference. All information for how to process, analyze, and model each dataset we obtain from the publication. We make a variety of comparisons to assess the similarity of our results across both software packages and choice of inference method.

## Methods

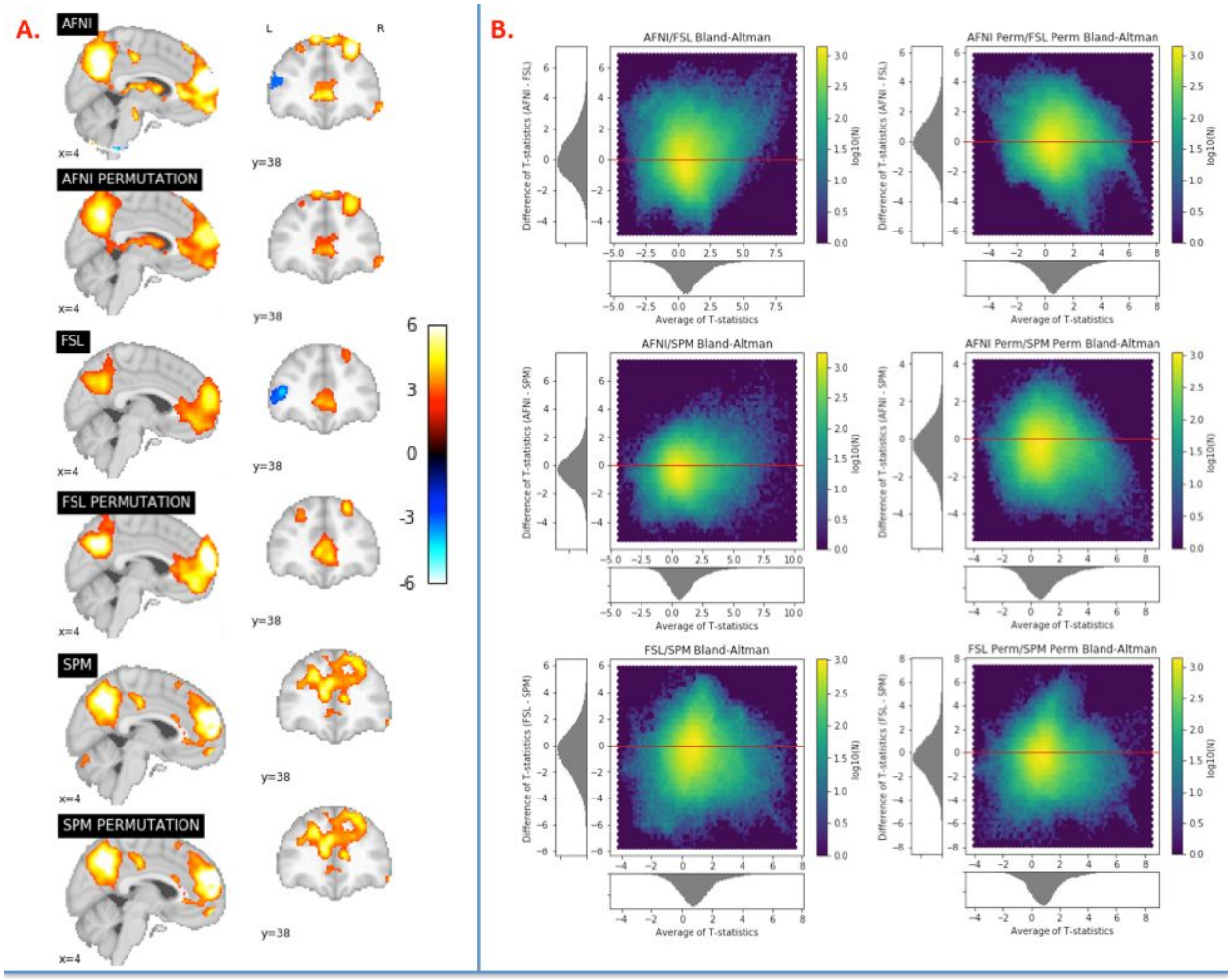
We reanalysed data from three published fMRI studies and attempted to replicate the result for the principal effect depicted in the main figure of each publication within the three packages. The dataset associated to each study was obtained from the OpenfMRI (Poldrack, 2015) database (ds000001, R: 2.0.4; ds000109, R:2.0.2; ds000120, R:2.0.4).

Prior to the analyses we determined a number of processing steps to be included in all of our reproductions, for example, inclusion of six motion regressors in the analysis design matrix to remove motion-related artefacts. Although this meant deviating from an exact reproduction of a publication's analysis, these steps were included to maximise comparability. Excluding these procedures, we endeavoured to choose the analysis pipeline within each package most consistent with the publication given the limitations of the software. Scripts were written to carry out the analyses in each package, and for FSL and SPM, export the group-level results as NIDM-Results packs.

For each study, the activation maps were uploaded to Neurovault (Gorgolewski, 2015). We applied three quantitative comparison methods: Bland-Altman plots, assessing differences in the magnitudes of activations between the unthresholded group T-statistic maps; Dice statistics, comparing the locations of activation in the FWE-thresholded maps. Finally, Euler Characteristics were computed for each software's group T-statistic map characterizing differences in the topological properties of the thresholded images. Comparisons were made both between software, as well as within software for the parametric and nonparametric inference results.

## Results

Figures A-E present comparisons of the group-level results in each package for reproductions of the main contrast '*false belief* > *false photograph'* from the publication associated to the ds000109 dataset. Group-level inference was conducted using a cluster-forming threshold $p < 0.005$, FWE-corrected clusterwise threshold $p < 0.05$. While qualitatively the regions of activation determined in the thresholded images are similar, the comparisons display striking differences across software, as well as between parametric and nonparametric inferences within FSL.




*The group-level results for all software packages used to create figures A-E are available on Neurovault at https://neurovault.org/collections/2238/.*

A. Slice comparison of the thresholded T-statistic maps for all packages using parametric and nonparametric inference with cluster-forming threshold $p < 0.005$, FWE-corrected clusterwise threshold $p < 0.05$. In general, regions of the brain determined as showing significant activation were similar across packages and inference methods. However, the results reveal substantial discordance in the magnitude of T-statistic values, as well as the precise locations of activation within a region. Notably, the AFNI and FSL thresholded maps obtained using parametric inference are the only results to have displayed significant deactivation. For AFNI and FSL, the T-statistic values of significant voxels was also dependent on if parametric or nonparametric inference methods were applied. Conversely, the only difference between the SPM thresholded maps is that parametric inference determined a slightly larger number of significant clusters than permutation inference.

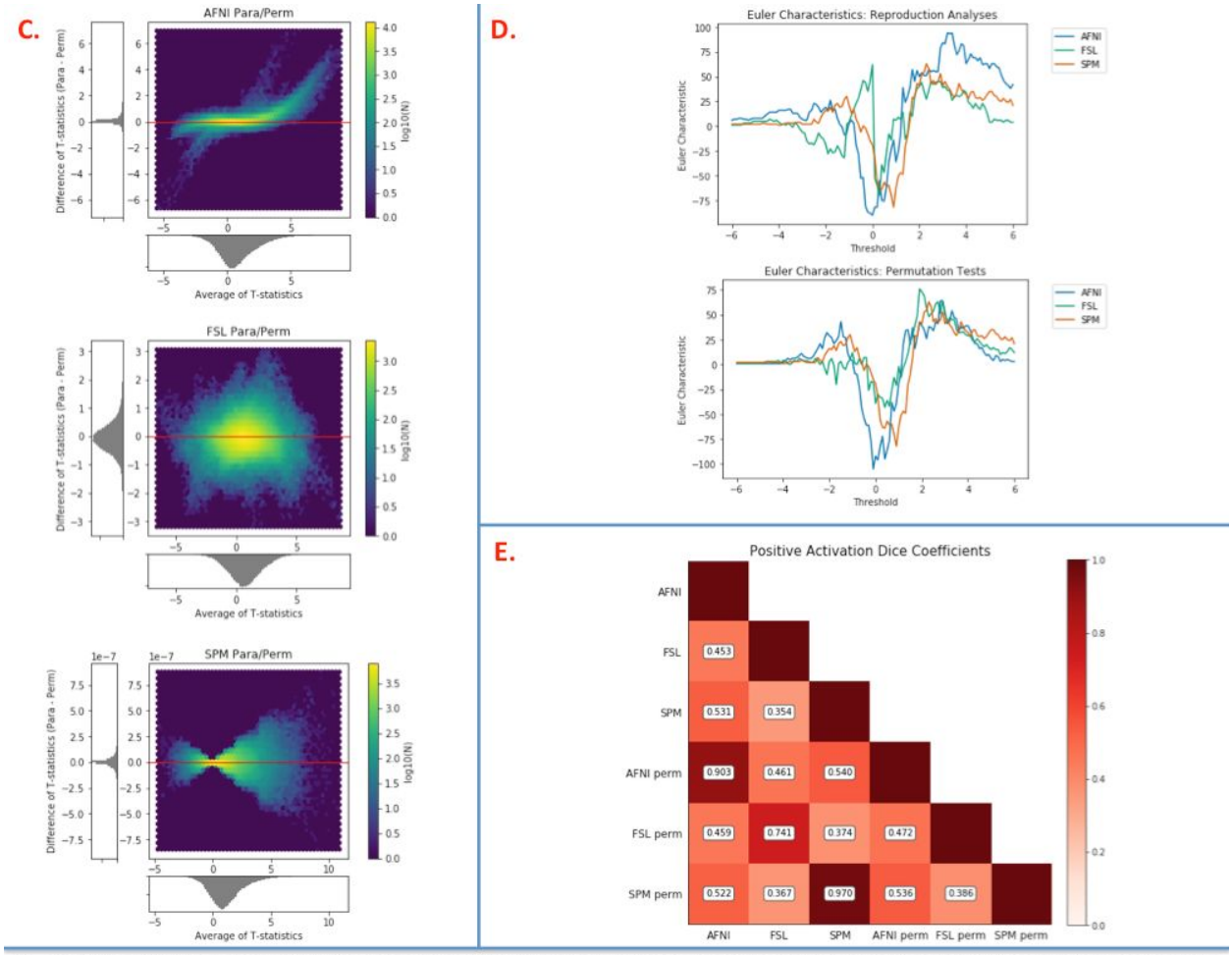
B. Bland-Altman 2D histograms comparing the unthresholded T-statistic maps between software for each type of inference method. Common to all plots are a cloud of densities that disperse away from the origin. High density counts either side of the x-axis show sizeable disagreement for all of the comparisons; differences in computed T-statistic values extend as far as four for a large quantity of voxels in all of the plots. In general, densities are distributed evenly either side of the x-axis for all average activation values, suggesting that although there is variability between software, overall no package consistently determined larger activation than the other two. Notably, the marginal distribution of the difference of T-statistics is approximately bell-shaped around a mean of zero in all of the plots. This is comparable to the distribution of densities expected from observations of the difference of two independent Gaussian random variables, characterizing the stochastic disposition between software on the size of the T-statistic at a given voxel.

## Conclusions

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 54% (intra-software, parametric-vs-nonparametric, were better, e.g. 97% for SPM). This work supports the need for open sharing of data, and the importance of understanding the fragility of one's results under the choice of software used.

## 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. Gorgolewski, K. J. (2015). NeuroVault. org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in neuroinformatics*, *9*.
4. 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.




C. Bland-Altman 2D histograms comparing the unthresholded T-statistic maps between the parametric and nonparametric results within each software package. For AFNI, the concentration of densities along the x-axis signals harmonization between the two methods. However, the dissipation of these densities above the x-axis for large positive T-statistic values, respectively below the x-axis for large negative T-statistics, exemplify that parametric inference methods were more liberal overall. Noting the scale of the y-axis in the SPM plot, this histogram shows that the unthresholded T-statistic images obtained for both inference methods were identical up to numerical error, and that the size of this error scaled with the absolute size of average activation. In comparison, the cloud of densities displayed in the FSL histogram - similar to the between-software plots in Figure B – suggest much greater incoherence between the two methods within this package.

D. Euler Characteristics (ECs) for each software's T-statistic map thresholded using T-values between -6 and 6. A schematic characterization of the EC is that it computes the number of clusters minus the numbers of 'handles' plus the number of 'holes' in a thresholded image; for large T-values, this will simply be the number of clusters. In this respect, differences in the Euler Characteristic values for large thresholds seen in both plots suggest a high variability in the topological properties of thresholded activation images dependent on choice of software package and inference method used.

E. Dice statistic values for pairwise comparisons between the positive thresholded T-statistic maps. Dice is the size of the overlapping region divided by the average size of each region. Dice coefficients computed between software packages are substantially smaller than 1, suggesting high spatial variability in the location of significant activations. Coefficients for the within-software comparisons of the parametric and nonparametric results are much improved by comparison, insinuating spatial coherence between the two methods within each software package.


5. Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.
6. Moran, J. M. (2012). Social-cognitive deficits in normal aging. *Journal of neuroscience*, *32*(16), 5553-5561.
7. Padmanabhan, A. (2011). Developmental changes in brain function underlying the influence of reward processing on inhibitory control. *Developmental cognitive neuroscience*, *1*(4), 517-529.
8. Poldrack, R. A. (2015). OpenfMRI: open sharing of task fMRI data. *NeuroImage*.
9. Poldrack, R. A. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, *18*(2), 115-126.
10. 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*.
11. Padmanabhan, A. (2011). Developmental changes in brain function underlying the influence of reward processing on inhibitory control. *Developmental cognitive neuroscience*, *1*(4), 517-529.