Same Data - Different Software - Different Results?
Analytic Variability of Group fMRI Results
Alexander Bowring, Thomas E. Nichols, Camille Maumet

To cite this version:
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.
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