TOWARDS EFFICIENT FMRI DATA RE-USE: CAN WE RUN BETWEEN-GROUP ANALYSES WITH DATASETS PROCESSED DIFFERENTLY WITH SPM?

Xavier Rolland[⋆] *Pierre Maurel*[⋆] *Camille Maumet*[⋆]

[⋆] Univ Rennes, Inria, CNRS, Inserm, IRISA UMR 6074, Empenn ERL U 1228, F-35000 Rennes, France

ABSTRACT

The increased amount of shared data creates an opportunity to reuse existing data to reach larger sample sizes and hence increase statistical power in neuroimaging studies. However, doing so may require to perform analyses using subject data processed differently. Here, we performed between-group analyses under the null hypothesis (making any detection a false positive), with data from the Human Connectome Project (HCP) (n=1080) processed with different pipelines. We compared the estimated false positive rates obtained to the theoretical false positive rate, to assess whether the variability in processing pipelines (called analytical variability) impacts the validity of the analyses. We found that some differences in parameter values caused invalidity, suggesting that analytical variability has to be taken into account before combining subject data processed with different pipelines.

***Index Terms*—** Brain Imaging, Analytical Variability, Reproducibility, Validity, Null Hypothesis, Pipeline

1. INTRODUCTION

Task-based functional Magnetic Resonance Imaging (fMRI) studies the activation of brain regions while a task is performed. Blood-Oxygen Level Dependent (BOLD) fMRI uses MRI to measure a BOLD signal (whose variations in time are related to brain activity), at each position of the brain. Multiple steps of processing are performed on the data: preprocessing, subject-level (first-level) and group-level (second-level) analysis. Series of steps performed for a complete analysis, or parts of it, are called pipelines (e.g. subject-level pipelines cover preprocessing and first-level analysis).

Many concerns have been raised over the last few years regarding the lack of reproducibility of fMRI studies [1, 2, 3]. One important cause of this issue is the overall low statistical power of fMRI studies induced by low sample sizes, which increases the likelihood that any positive result is false [4]. Larger sample sizes may be achieved by taking advantage of shared neuroimaging datasets, and combining subject data from multiple different sources for new studies. However, these datasets may include already processed subject data, and combining them may require to perform analyses using data processed differently. Multiple choices are possible at each processing step (different orders of operations, different parameter values, different software packages).

Multiple studies have explored how this variability in processing and analysis protocol (called analytical variability) may impact the reproducibility of neuroimaging results. For example, in fMRI, [5] showed substantial differences in results obtained across teams when completing a similar analysis with a different pipeline. Other studies have explored the variability resulting from specific processing and analysis parameters [6, 2, 1]. Frameworks for the optimization of pipelines by estimation of performance metrics associated with reproducibility have also been developed [7].

Here, we focused on how analytical variability in subject-level pipelines can impact the compatibility of subject data in between-group studies. If researchers want to use processed subject data coming from different sources, they must ensure that subject-level processing differences will not increase the probability of obtaining false positive results.

In order to assess the validity of between-group studies combining subject data processed differently, we processed raw data from the Human Connectome Project [8] with various pipelines which differed on a set of predefined parameters. After subject-level processing, we carried out a series of between-group analyses under the null hypothesis, with different pairs of subject-level pipelines. False positive rates were then used to assess the validity of these between-group analyses (with inflated false positive rates indicating the invalidity of the combination of pipelines).

2. MATERIAL

This study was performed using data from the HCP [8]. We used unprocessed fMRI data associated with the motor task and structural data for all available subjects ($n = 1080$). Multiple preprocessing and first-level analyses were performed (see section 3.2).

3. METHODS

In order to test the validity of between-group analyses using subject data processed differently, we performed analyses under null hypothesis assuming no difference in means across pipelines. We used the detection rate as an estimate of the

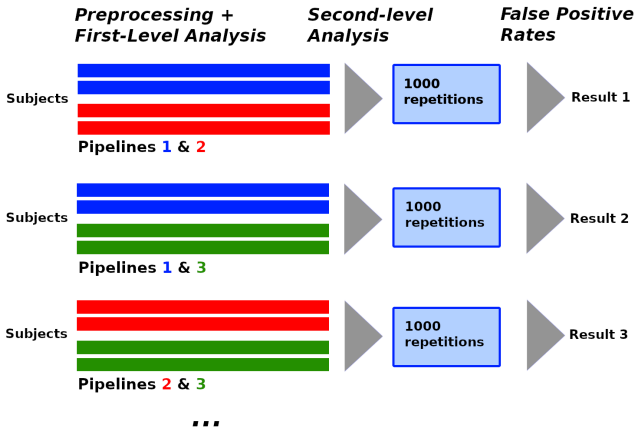

Fig. 1. Steps performed for the analysis: subject-level analysis on subject data with different pipelines, between-group analyses with subject data processed differently for multiple pairs of pipelines, and repetitions of each analysis 1000 times to estimate the false positive rates.

false positive rate and compared it to the expected false positive rate under the null hypothesis (Fig. 1). Analyses were performed with SPM12 r7771 (RRID: SCR_007037, [9]) with Octave 5.1.0 (RRID: SCR_014398), under Debian 10.6.

3.1. Subject-Level Pipelines

All subject data were processed through multiple pipelines, which carried out preprocessing and first-level analysis. The preprocessing steps were: spatial realignment of the functional data, coregistration of realigned data towards the structural data, segmentation of the structural data, nonlinear registration of the structural and realigned functional data towards a common space and smoothing of the normalized functional data. For each pipeline, we selected the first-level contrast corresponding to the right hand in the motor task.

Our pipelines varied on the following parameters:

- Smoothing kernel: Full-Width at Half-Maximum (FWHM) was 5mm or 8mm.
- Number of motion regressors in the General Linear Model (GLM) for the first-level analysis: 0, 6 (3 rotations + 3 translations) or 24 (3 rotations + 3 translations, 6 derivatives and the 12 corresponding squares of regressors).
- Presence or absence of the temporal derivatives of the Haemodynamic Response Function (HRF) in the GLM.

Apart from these parameters, each pipeline used the default settings. In total, those combinations provided a set of 12 different subject-level pipelines (2 FWHM $\times$ 3 numbers of motion regressors $\times$ 2 HRF). The steps performed and parameter values were chosen so as to represent typical pipelines found in the literature [1].

3.2. Between-group Analyses and False Positive Rates Estimation

In the remainder of the text, we will refer to as the "default pipeline", the subject-level pipeline with the following parameter values: 5mm FWHM for smoothing kernel, 24 motion regressors and no temporal derivatives for the HRF. We performed between-group analysis to compare two groups of 50 subjects, with each group processed respectively by the default pipeline and by one of the 11 alternative pipelines. The 100 subjects from the pair of groups were randomly sampled without replacement, uniformly among the 1080 subjects. For this reason, our analyses are under the null hypothesis of no difference between groups. In addition to comparisons with the default pipeline against the 11 different pipelines (referred to as "different pipeline analyses" in the following), to assess the compatibility between pipelines, we performed analysis with the default pipeline for both groups (referred to as "same pipeline analysis" in the following), as a sanity check.

We looked at the between-group difference in means for the right hand contrast. We performed two one-tailed t-test (for the comparisons default $>$ alternative and alternative $>$ default) with unequal variance using a voxelwise $p < 0.05$ FWE-corrected threshold. For conciseness, the results are presented below as one two-tailed t-test with a $p < 0.1$ FWE-corrected threshold.

For each pair of pipelines, the between-group analyses were repeated 1000 times with different groups of subjects to estimate the empirical false positive rate, which was the proportion of analyses over the 1000 repetitions with at least one significant voxel between both groups. This false positive rate is expected to be equal to 0.1 under the null hypothesis. The set of 1000 pairs of 50-subject groups used to estimate the empirical false positive rate was identical for all pairs of pipelines.

For same pipeline analysis, and for comparisons with the alternative pipeline varying from default on one or two parameter values, we also created variants of P-P plots to observe the behavior of the distribution of second-level statistical values. For each comparison, we created a subset of 1,000,000 statistical values chosen randomly over our 1,000 second-level analysis statistical maps. We obtained and sorted the p-values associated with these statistical values for the expected distribution under the null hypothesis (Student distribution with n=98 degrees of freedom). We made variants of P-P plots plotting the difference between obtained and expected -log(p-values), against the expected -log(p-values).

4. RESULTS

4.1. False positive rates

False positive rates for comparisons between the default pipeline and each of the 11 alternative pipelines are presented in Table 1. Same pipeline analysis (using the default pipeline

	Smoothing, 5 mm		Smoothing, 8 mm	
	No der.	Der.	No der.	Der.
0 motion reg.	**0.109**	0.097	**0.237**	**0.245**
6 motion reg.	0.044	0.046	**0.139**	**0.145**
24 motion reg.		0.035	**0.113**	**0.131**

Table 1. **Empirical false positive rates for analyses comparing the 11 alternative pipelines to the default pipeline** (smoothing=5mm, no der. and 24 motion reg.) at FWE-corrected $p < 0.1$ two-tailed. Invalid results (> 0.1) are in bold. The false positive rate obtained for same pipeline analysis (default pipeline in both groups, corresponding to the grey cell) was equal to 0.040. reg.= regressors, der.=derivatives.

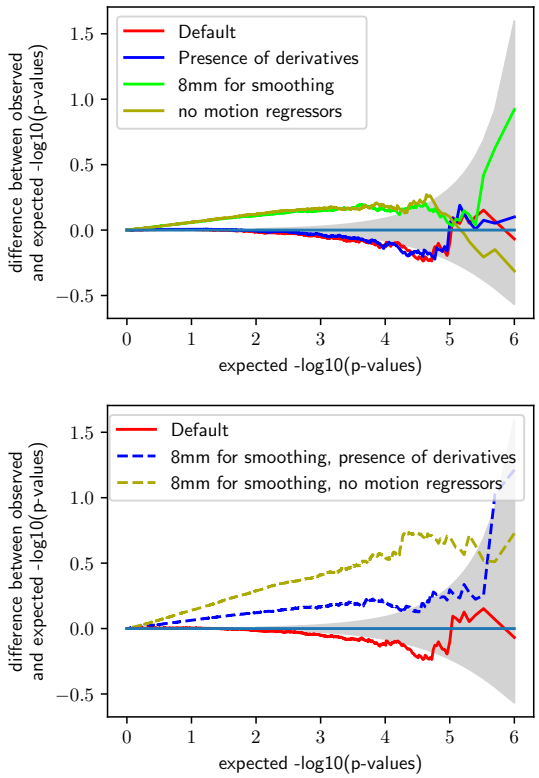

Fig. 2. Variants of P-P plots for distributions obtained with various analyses between the default pipeline and pipelines with one (top) or two (bottom) varying parameter values (indicated in legend), against the expected distribution, with 0.95 confidence interval (in grey). The curve for same pipeline analysis (alternative pipeline equal to default) is plotted in red.

in both groups) led to conservative results with a false positive rate of 0.040, which is consistent with the literature [10].

For analyses combining pipelines with a single varying factor, the temporal derivatives of the HRF was the least impacting of all three factors (temporal derivative, smoothing or motion regressors) with a false positive rate of 0.035 (third row, second column). Smoothing was the most impacting factor with a false positive rate of 0.113 (third row, third column). For motion regressors, the two possible alternative values led to false positive rates of 0.044 (motion reg = 6, second row, first column) and 0.109 (motion reg=0, first row, first column).

For analyses combining pipelines with multiple varying factors, all comparisons with different smoothing kernel (FWHM 8mm in the alternative pipeline, third and fourth column) led to invalid results, with false positive rates above 0.1. Other parameters being equal, comparisons with same smoothing always led to smaller false positive rates (column 2 versus 4 and column 1 versus 3). Similarly, comparisons with different numbers of motion regressors (especially with no motion regressors in the alternative pipeline) gave higher false positive rates than comparisons with the same number of motion regressors (line 1 versus line 3). More generally, multiple varying factors always led to higher false positive rates, suggesting that the combination of differences between pipelines leads to a combination of effects.

4.2. P-P plots

P-P plots are shown on Figure 2. Positive differences outside the confidence interval indicate invalidity (-log(p-values) higher than expected) whereas negative differences indicate conservativeness.

For same pipeline analysis (using the default pipeline for both groups), the results observed were conservative. The effects of parameter differences on P-P plots are similar to those observed on false positive rates: no clear effect of the modeling of the HRF, effect of differences in smoothing and numbers of motion regressors, and combination of effects with combination of differences. The P-P plots show that the results observed with a specific thresholding in Table 1, regarding the effect of parameter differences, are representative of general tendencies for statistical values overall.

5. DISCUSSIONS

We observed that differences between subject-level pipelines on specific parameters (smoothing, number of motion regressors) had an effect which caused invalid results when combined in between-group analyses. Therefore, pairs of pipelines with different values for these parameters cannot be combined for analyses with differences between groups: it would be impossible to know if the between-group differences have an effect, as it would be confounded with the effect of pipeline differences.

Our results suggest that some combinations of pipelines should be avoided, while other could be used. Differences in modeling of the HRF had no observable effect in our experiments. This is not the case for smoothing and motion

regressors, which gave invalid results.

We performed analyses with different subject-level processing pipelines, where the pipelines are confounded with the groups. This situation may happen in practice if we want to do group comparisons where each group is associated with a specific dataset, which is itself associated with a specific subject-level pipeline. Other situations may also happen in which there are multiple pipelines used on the subjects within each group. Such situations may be investigated in future work.

Since there are many more parameters which may vary between subject-level pipelines in practice (different models for the HRF, performing or not specific substeps such as slice-timing correction, software packages, etc), effects of pipeline differences would likely be more important in real conditions.

Here, we focused our investigations on measuring deviations from the theoretical false positive rate under the null. While this is important in order to assess the validity of the statistical approaches, the lack of sensibility is also an important methodological issue that will require further investigations in future work.

5.1. Conclusion

Our study shows that processing applied to data must be taken into consideration when combining them, due to the potential invalidity of the results. The framework that we defined may be used with other variations of parameters or other paradigms in future work, and also using different software packages, to assess the generalizability of our results. Also, we may try to create methods to model and correct the effect of analytical variability, which would allow us to combine data without having to consider the differences in terms of processing.

6. COMPLIANCE WITH ETHICAL STANDARDS

This study was performed using data from the Human Connectome Project (HCP). Written informed consent was obtained from participants and the original study was approved by the Washington University Institutional Review Board. We agreed to the HCP Open Access Data Use Terms.

7. ACKNOWLEDGEMENTS

Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

Xavier Rolland was supported by Region Bretagne (ARED Varanasi) and by EU H2020 project OpenAIRE-Connect (Grant agreement ID: 731011).

8. REFERENCES

[1] J. Carp, "On the plurality of (methodological) worlds: estimating the analytic flexibility of fmri experiments," *Frontiers in neuroscience*, vol. 6, pp. 149, 2012.

[2] A. Bowring, C. Maumet, and T. E. Nichols, "Exploring the impact of analysis software on task fmri results," *Human brain mapping*, vol. 40, no. 11, pp. 3362–3384, 2019.

[3] S. C. Strother, "Evaluating fmri preprocessing pipelines," *IEEE Engineering in Medicine and Biology Magazine*, vol. 25, no. 2, pp. 27–41, 2006.

[4] K. S. Button, J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò, "Power failure: why small sample size undermines the reliability of neuroscience," *Nature reviews neuroscience*, vol. 14, no. 5, pp. 365–376, 2013.

[5] R. Botvinik-Nezer, F. Holzmeister, C. F. Camerer, A. Dreber, J. Huber, M. Johannesson, M. Kirchler, R. Iwanir, J. A. Mumford, R. A. Adcock, et al., "Variability in the analysis of a single neuroimaging dataset by many teams," *Nature*, pp. 1–7, 2020.

[6] T. Glatard, L. B. Lewis, R. Ferreira da Silva, R. Adalat, N. Beck, C. Lepage, P. Rioux, M.-E. Rousseau, T. Sherif, E. Deelman, et al., "Reproducibility of neuroimaging analyses across operating systems," *Frontiers in neuroinformatics*, vol. 9, pp. 12, 2015.

[7] N. W. Churchill, R. Spring, B. Afshin-Pour, F. Dong, and S. C. Strother, "An automated, adaptive framework for optimizing preprocessing pipelines in task-based functional mri," *PloS one*, vol. 10, no. 7, pp. e0131520, 2015.

[8] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. J. Behrens, E. Yacoub, K. Ugurbil, Wu-Minn HCP Consortium, et al., "The wu-minn human connectome project: an overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

[9] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical parametric mapping: the analysis of functional brain images*, Elsevier, 2011.

[10] Anders Eklund, Thomas E Nichols, and Hans Knutsson, "Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates," *Proceedings of the national academy of sciences*, vol. 113, no. 28, pp. 7900–7905, 2016.