FMRI data analysis: How does analytical variability vary with sample size?

Elodie Germani, Camille Maumet

FMRI data analysis: How does analytical variability vary with sample size?

Elodie GERMANI[1], Camille Maumet[2] Institutions:

[1]Univ Rennes, Inria, CNRS, Inserm, Rennes, France, [2]Inria, Univ Rennes, CNRS, Inserm, Rennes, France

Introduction:

Neuroimaging workflows are highly flexible leaving researchers with many possible choices at each step of their analysis (Carp, 2012). Recents studies (Bhagwat et al., 2021; Botvinik-Nezer et al., 2020; Bowring et al., 2019) have demonstrated how different analytical choices can substantially impact neuroimaging results, effectively leading to a "vibration of effects" (Ioannidis, 2008). This observation is not limited to brain imaging and was also made across many scientific fields (Hoffmann et al., 2020). In psychology, (Klau et al., 2020) showed that the vibration of effects decreases and stabilizes as sample sizes increase.

We built on the results of (Botvinik-Nezer et al., 2020), in which a functional Magnetic Resonance Imaging (fMRI) "many analyst" study was conducted with 70 teams. In this study, each team used their favorite pipeline to analyze the same dataset and answer 9 pre-defined hypotheses. Here, after reproducing the pipelines used by 4 teams, we observe the impact of varying sample sizes on the vibration of effects.

Methods:

Using descriptions provided in the original study, we reproduced the SPM pipelines of 4 teams: 2T6S, V55J, Q6O0 and C88N. For each team, we assessed the quality of the reproduction by comparing our results with the statistic maps published on NeuroVault (Gorgolewski et al., 2015).

After this validation step, we focused our experiments on the results of Hypothesis 5 for which 80% of the teams agreed on a positive answer in the original paper. We randomly selected 20, 40, 60 and 80 participants (each set being a superset of the smaller ones) among the full dataset of 108 participants. We replicated the results of each pipeline with each subset of participants and with the full dataset.

Results:

1) Validation of the reproductions
For each team, we compared the reproduced and the original maps visually and quantitatively using Pearson's correlations (unthresholded maps) and Dice scores (thresholded maps).

Fig.1.A and B present the statistic maps for an example team (2T6S). We found similar activation patterns and the same observation was made for all reproduced teams. All correlations were above 0.98 (Fig.1.C) for all teams. There were differences in the number of activated voxels between the original and reproduced maps and sometimes many non-overlapping voxels with Dice scores ranging 0.63-0.81. Multiple attempts were necessary to obtain these results.

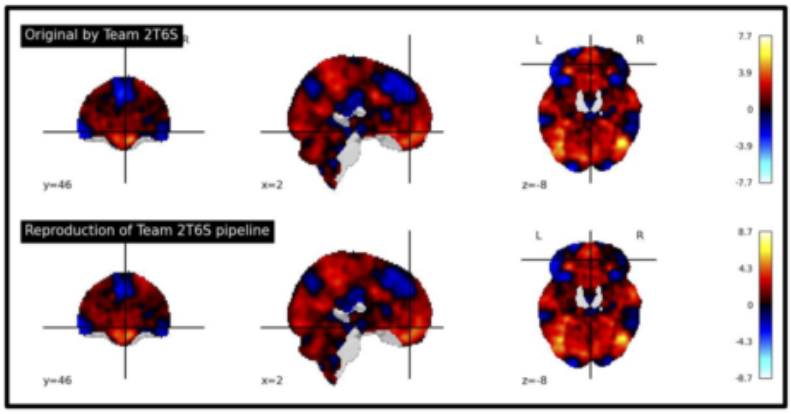
2) Analytical variability with varying sample sizes
The thresholded maps obtained for each sample size and each team are presented in Fig.2.A. Overall, the thresholded maps presented a lot of variability. For some pipelines, more participants were necessary to find activations in the ventro-medial prefrontal cortex (vmPFC) (which was the region of expected activation for H5). In particular, for the team Q6O0 with 60 participants there was no activation for H5 while all other teams had detected a significant activation.

For each result, we also looked at the mean activation value in the vmPFC. With smaller sample sizes (N=20 and N=40), for a majority of pipelines, there was no activation in the vmPFC (and hence a mean activation value of 0) (Fig 2.B). This value increased with larger sample size, except for 2T6S maps for which it was above 0 for all sizes, highlighting the higher sensibility of this pipeline, probably due to the absence of multiple testing correction. However, with growing sample sizes, mean activations seemed to converge, consistent with the previous observation that the vibration of effects reduces and stabilizes with larger sample sizes.

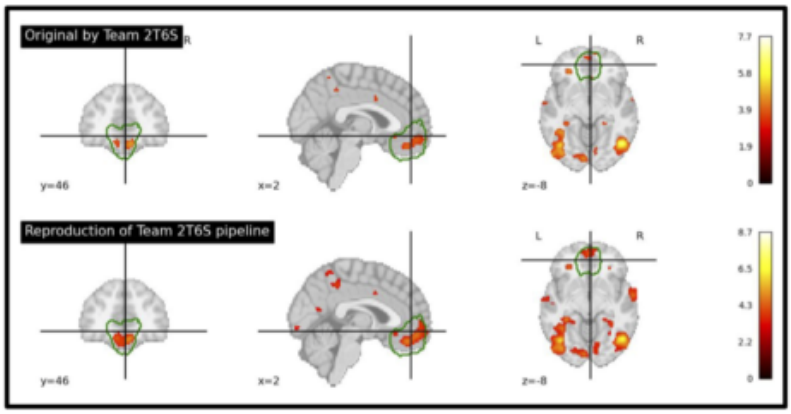
Conclusions:

With this work, our objective was to show the impact of sample size on analytical variability. Our findings show that, in fMRI data analysis, the vibration of effects decreases with sample size. Our results also suggest that some variability remains even for large sample sizes. Further work will be needed in order to include more pipelines and investigate which part of the pipelines are the most impactful.

A. Original and reproduced statistic maps for H5 with an example team (2T6S).



B. Original and reproduced thresholded maps for H5 with an example team (2T6S).

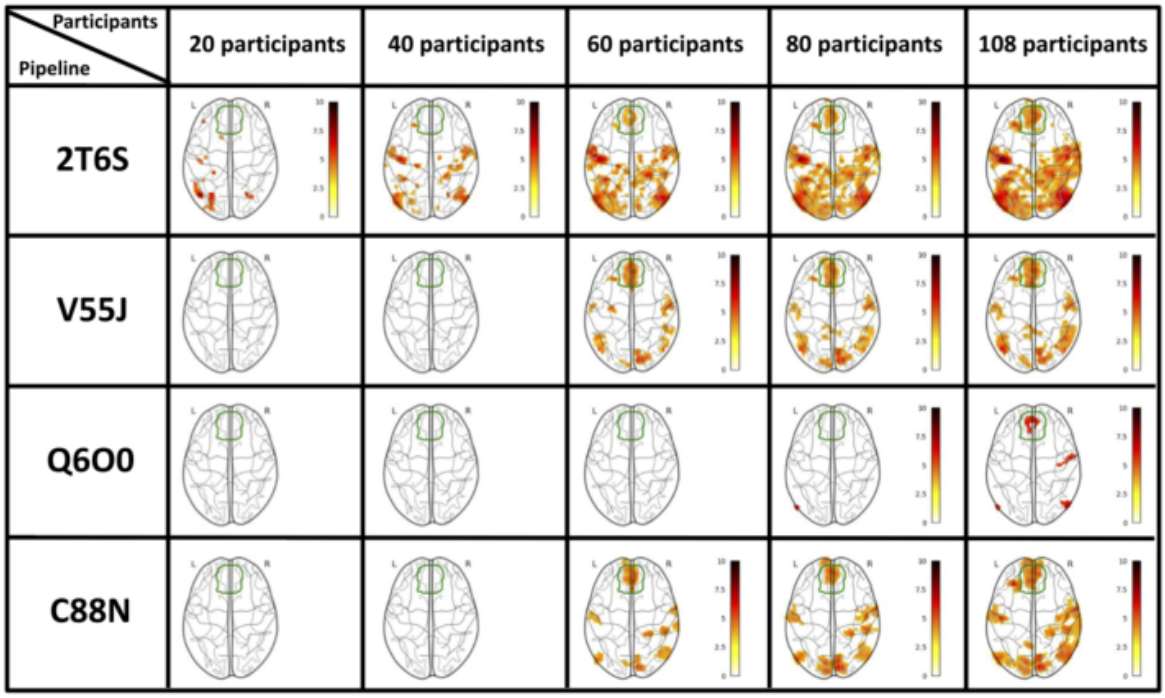


C. Comparison of original and reproduced maps for H5 with all reproduced teams.

	2T6S	**C88N**	**Q6O0**	**V55J**
Pearson's correlation coefficient between original and reproduced statistic maps	0.98	0.99	0.99	0.98
Mean Squared Error between original and reproduced statistic maps	0.14	0.05	0.11	0.15
Number of activated voxel: reproduced / original	20375 / 13195	10077 / 8490	516 / 922	5641 / 4799
Number of non overlapping activated voxels between original and reproduced maps	7654	3479	528	263
Dice similarity score between thresholded original and reproduced maps	0.77	0.81	0.63	0.75

Figure 1. Quality check of the pipelines reproductions. Statistic maps **(A & B)** with the example of team 2T6S and quantitative metrics for all reproduced pipelines **(C)**.

A. Thresholded maps (glass brains) replicated with different pipelines and sample sizes.



B. Mean activation value inside the ventromedial prefrontal cortex (extracted from Harvard/Oxford atlas) with different pipelines and sample sizes.

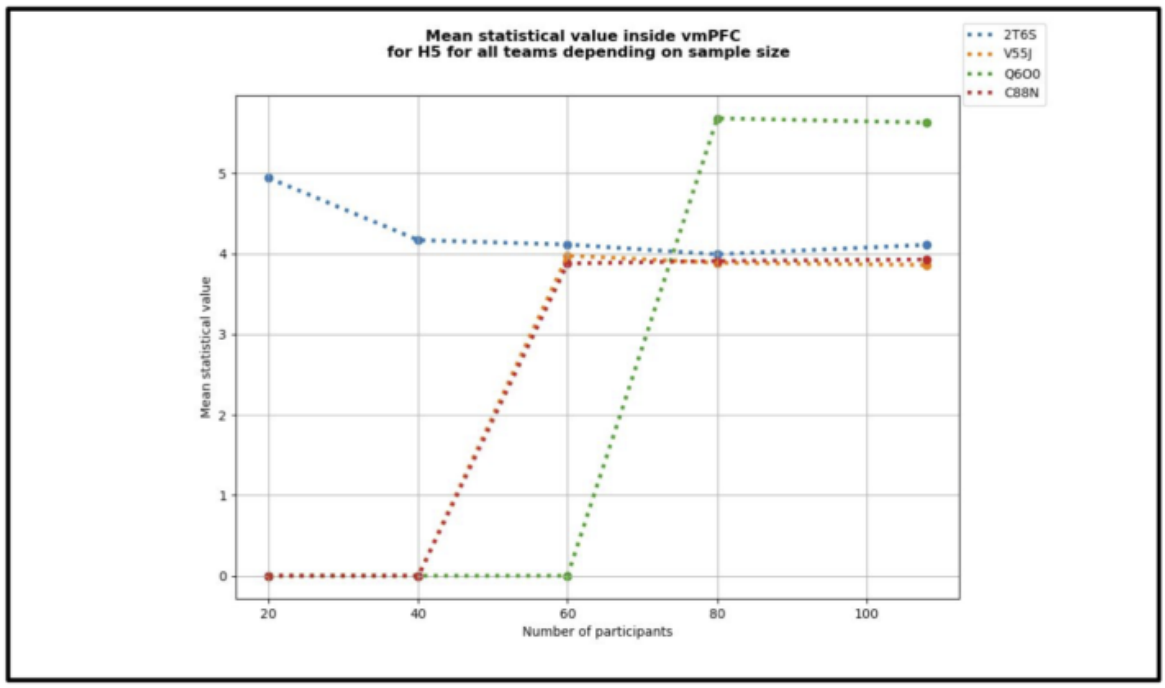


Figure 2. Impact of sample sizes on the vibration of effects through visual comparison **(A)** and metric computation **(B)**.

References

Bhagwat, N. (2021), ‘Understanding the impact of preprocessing pipelines on neuroimaging cortical surface analyses’, GigaScience, vol. 10, no. 1.

Botvinik-Nezer, R. (2020), ‘Variability in the analysis of a single neuroimaging dataset by many teams’, Nature, vol. 582, pp. 84–88.

Bowring, A. (2019), ‘Exploring the impact of analysis software on task fMRI results’, Human Brain Mapping, vol. 40, no. 11, pp. 3362– 3384.

Carp, J. (2012), ‘On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments’, Frontiers in Neuroscience, vol. 6, no. 149.

Gorgolewski, K.J. (2015), ‘NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain’ Frontiers in Neuroinformatics, vol. 9, no. 8.

Hoffmann S. (2021), ‘The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines’, Royal Society of Open Science, vol. 8, no. 4.

Ioannidis, J.P.A. (2008), ‘Why Most Discovered True Associations Are Inflated’, Epidemiology, vol. 19, no. 5, pp. 640-648.

Klau, S. (2020), ‘Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology’, Department of Statistics: Technical Reports, vol. 232.