Statistical inference for Same Data Meta-Analysis for neuroimaging multiverse analyzes

Lefort-Besnard Jeremy[1], Nichols E. Thomas[2*], Maumet Camille[1*]

[1] Inria, Univ Rennes, CNRS, Inserm, IRISA UMR 6074, Empenn ERL U 1228, Rennes, France

[2] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Oxford, UK

[*] Sharing last authorship

Problem statement

Researchers using fMRI data have a wide range of analysis tools to model brain activity. This diversity of analytical approaches means there are many possible variations of the same imaging result (Bowring et al., 2019). Analyzing a dataset with a single approach can thus be misleading.

Alternatively, a multiverse analysis can be used, where multiple sets of results are obtained from running different pipelines on the same dataset. Such a setting produces multiple outputs in the form of a statistical map. Meta-analysis approaches can then help to effectively extract a unique result from these maps.

A required assumption for traditional meta-analyses is the independence among input datasets (Fig. 1, a). This assumption is no longer true in a multiverse setting (Fig. 1, b), thus treating dependent studies as independent may lead to invalidity and hence inflated false positive findings.

Here, we present a variety of same data meta-analysis (SDMA) models. The validity and accuracy of these models were assessed in a set of simulations as well as on two real-world multiverse outputs originating from the same data: the "NARPS" multiverse analysis (Botvinik-Nezer et al., 2020) and the "HCP Young Adult" multiverse analysis (Germani et al., 2023) which generated respectively 70 and 24 different statistical maps.

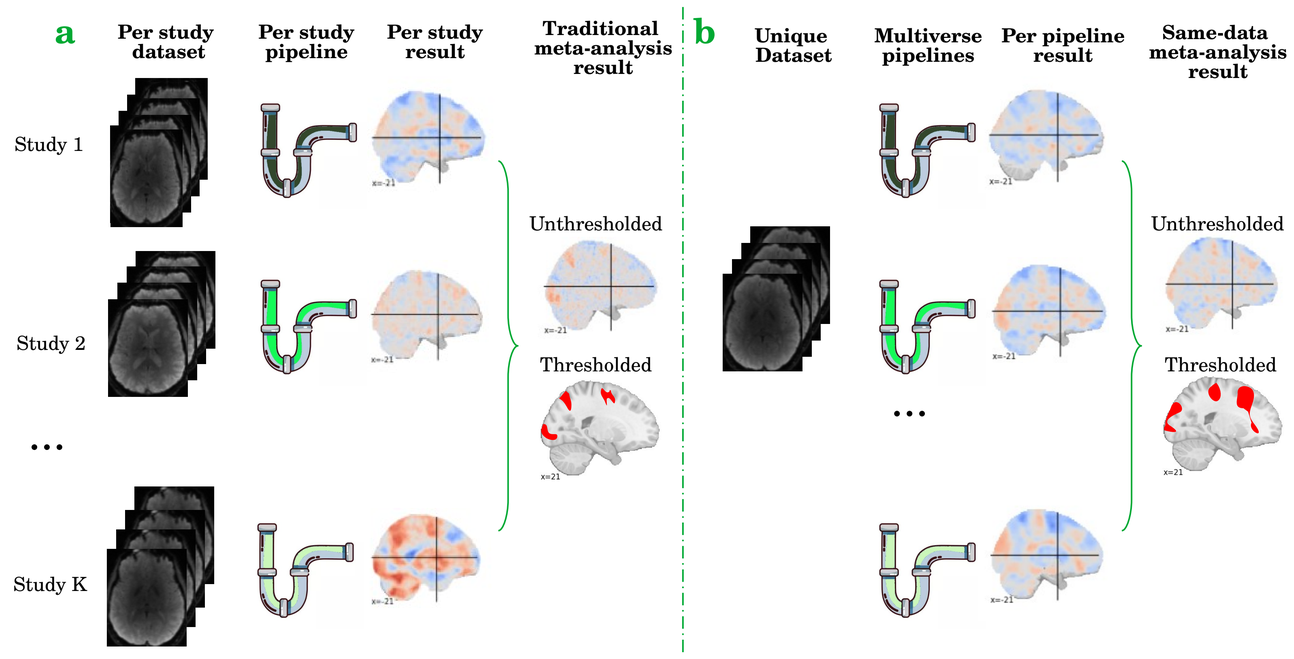


Fig 1: *Traditional meta-analyses setting (a) and multiverse setting (b). In the traditional approach, there are K independent datasets analyzed (typically with K distinct pipelines) and used to estimate an effect. In a multiverse setting a single dataset is run through K different pipelines to estimate an effect. Dependence from using a single dataset invalidates traditional meta-analysis methods when applied to multiverse data.*

Meta-analyses models

Traditional versus same data meta-analysis models

Stouffer

$$\bar{Y}_j^{\mathrm{S}} = \sqrt{K}\,\bar{Y}_j$$

SDMA Stouffer

$$\bar{Y}_j^{\mathrm{SC}} = \frac{\bar{Y}_j}{\sqrt{\mathbf{1}^\top \mathbf{Q}\mathbf{1}/K^2}}$$

Consensus SDMA Stouffer

$$\bar{Y}_j^{\mathrm{CS}} = \frac{\bar{Y}_j - \mu_C}{\sqrt{\mathbf{1}^\top \mathbf{Q}\mathbf{1}/K^2}} + \mu_C$$

Consensus Average

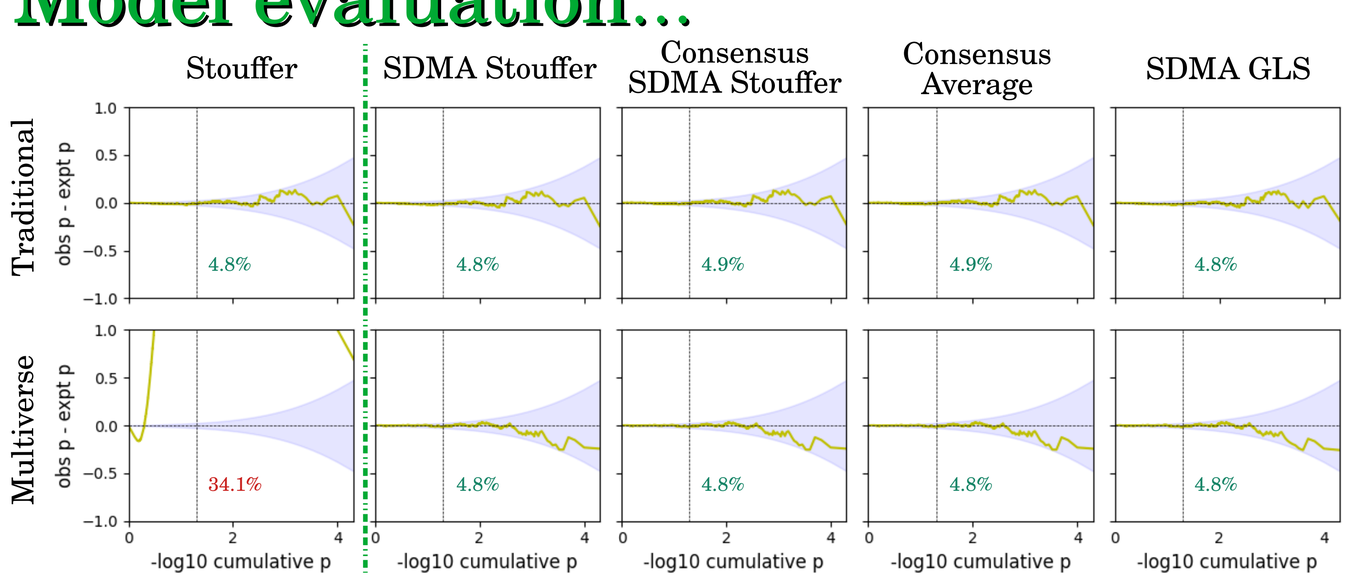
$$\bar{Y}_j^{\mathrm{CA}} = \frac{\bar{Y}_j - \mu_C}{\sigma^2_{\bar{Y}}}\sigma_C + \mu_C$$

SDMA GLS

$$\bar{Y}_j^{\mathrm{SG}} = \frac{\bar{Y}_j^{\mathrm{G}}}{\sqrt{(\mathbf{1}^\top \mathbf{Q}^{-1}\mathbf{1})^{-1}}}$$

Notation: Y_{kj} is the test statistic for pipeline $k = 1, \ldots, K$ and voxel $j = 1, \ldots, J$; $\bar{Y}_j$ is the average voxel value across pipelines; $\boldsymbol{Q}$ the interpipeline correlation matrix; μ_c and σ_c^2 respectively the average of the voxel-wise means and the average of the voxel-wise variances; $\sigma_{\bar{Y}}^2$ the voxel-wise variance of the $\bar{Y}_j$. Stouffer's method is a traditionnal meta-analysis model. The GLS estimate $\bar{Y}_j^G$ down-weights the influence of highly dependent pipeline and is defined as $\frac{\mathbf{1}^\top \mathbf{Q}^{-1}\mathbf{Y}_j}{\mathbf{1}^\top \mathbf{Q}^{-1}\mathbf{1}}$.

Model evaluation...



...On simulated data:

Simulations of results under the null scenario, where no effect is present, show that when pipelines are independent, all meta-analysis estimators are valid (Fig. 2, top). However, when data are correlated, as in the multiverse setting, only the same data meta-analysis estimators are valid (Fig. 2, bottom). These results demonstrate the validity of the same data meta-analysis models in those simulated settings.

Fig 2: *Comparative P-P plots for each meta-analysis estimator on traditional (upper row) or multiverse (lower row) simulated data, where the y-axis is the difference in observed and expected sorted p-value, and the x-axis is the expected sorted p-value. The blue shadow depicts the theoretical confidence interval, and the observed false-positive rate at p<0.05 is written on each plot, and appears in green when the results are valid (i.e. within the confidence bounds).*

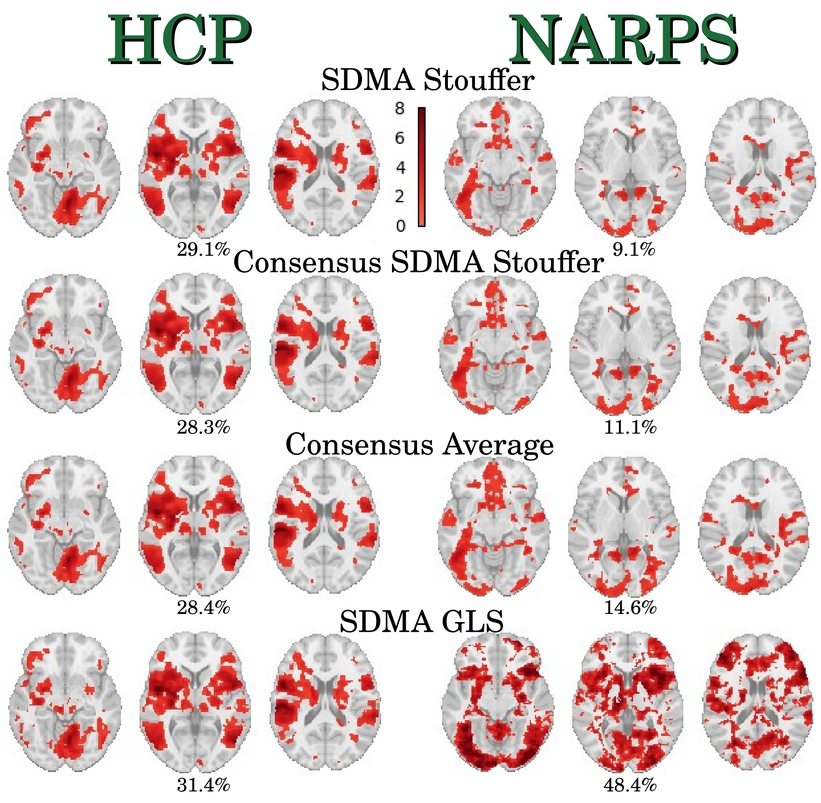


Fig 3: *Demonstration of the SDMA methods using the statistical maps from the first hypothesis in the NARPS study (parametric effect of gains in a mixed gambles task - left panel) and from the HCP young adult right hand contrast (right panel). Maps were thresholded at P<0.05 uncorrected to allow for direct comparison. Percentages of significant voxels are displayed on each map.*

...On real data:

The same data meta-analysis methods exhibit varying levels of significance depending on their particular characteristics. The SDMA Stouffer, the Consensus SDMA Stouffer and the Consensus Average methods put an equal weight on each pipeline, while the SDMA GLS method decreases weights of correlated pipelines.

In the HCP multiverse outputs, generated within a single laboratory using only two software tools, all SDMA estimators yield comparable results (Fig. 3, left panel).

In the NARPS multiverse outputs, which encompass 70 different teams and involve multiple software applications, the GLS method exhibits divergent outcomes (Fig. 3, right panel). In NARPS, a subset of pipelines is anticorrelated with others. GLS attributes greater importance to the anticorrelated pipelines (Fig. 4), with the majority of significant voxels originating from the opposite subgroups.

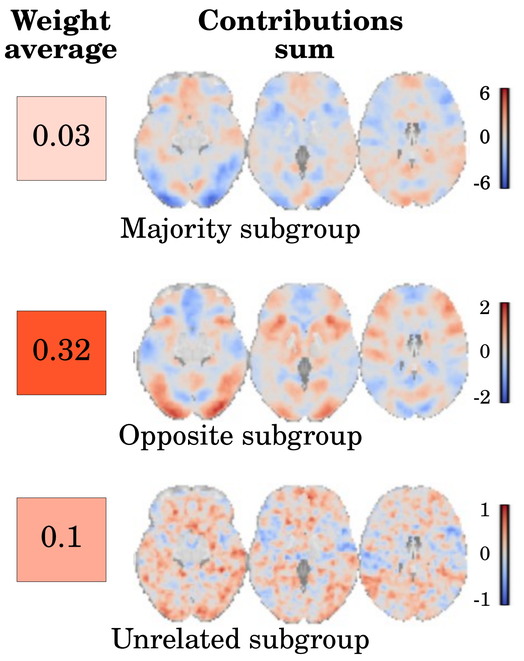


Fig 4: *Average weights (left) and aggregated contributions (right) per subgroup, assigned by the SDMA GLS estimator. Investigators of the NARPS study computed Spearman correlations between whole-brain unthresholded statistical maps for each team and clustered them accordingly based on their similarities. In our work, we employed their three cluster solutions, which included majority, opposite, and unrelated subgroups.*

Conclusion

Our findings underscored that the SDMA GLS method in scenarios with high heterogeneity may result in unclear and difficult-to-interpret outcomes, suggesting that it may not be the best option for application in a multi-expert context like NARPS.

Bibliography

Bowring A, Maumet C, Nichols TE. *Exploring the impact of analysis software on task fMRI results*. Hum Brain Mapp. 2019

Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, ... & Rieck JR. *Variability in the analysis of a single neuroimaging dataset by many teams*. Nature. 2020

Germany E, Fromont E, Maurel P, Maumet C. *The HCP multi-pipeline dataset: an opportunity to investigate analytical variability in fMRI data analysis*. Preprint. 2023

Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RMJ. *The American soldier: Adjustment during army life*. Princeton University Press. 1949