Systematic review and evaluation of meta-analysis methods for same data meta-analyses in a multiverse setting

Jeremy Lefort-Besnard, Thomas E. Nichols, Camille Maumet

Systematic review and evaluation of meta-analysis methods for same data meta-analyses in a multiverse setting

Lefort-Besnard Jeremy[1], Nichols 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

Keywords: multiverse | meta-analysis | neuroimaging | open science | reproducibility

Introduction

Researchers using task-fMRI data have access to a wide range of analysis tools to model brain activity. This diversity of analytical approaches has been shown to have substantial effects on neuroimaging results (Botvinik-Nezer et al., 2020; Bowring et al., 2018; Carp, 2012; Glatard et al., 2015). Combined with selective reporting, this analytical flexibility can lead to an inflated rate of false positives and contributes to the irreproducibility of neuroimaging findings (Poldrack et al., 2017). Multiverse analyses are a way to systematically explore and integrate pipeline variation on a given dataset. We focus on the setting where multiple statistic maps are produced as an output of a set of analyses. Meta-analysis is a natural approach to extract consensus inferences from these maps, yet the traditional assumption of independence amongst input datasets does not hold. In this work we consider a suite of methods to conduct meta-analysis in the multiverse setting, accounting for inter-pipeline dependence among the results..

Methods

We propose several same data meta-analysis (SDMA) methods based on the traditional 'Stouffer' fixed-effects meta-analysis (Stouffer, 1949):

- *SDMA Stouffer* in which correlation across pipelines is taken into account,
- *Consensus SDMA Stouffer* and *Consensus Average* methods, where the combined inference is calibrated to be as similar to the input pipelines as possible, and
- *General Least Squares (GLS) SDMA*, where inter-pipeline correlation is used to find the statistically optimal combination of pipeline results.

The validity of these models were assessed in a set of simulations. Here, we focus on false positive control in two scenarios: 1/ independent pipelines with no significant results (*null case*), and 2/ correlated pipelines with no significant results (*null correlated case*). These meta-analysis models were and (Botvinik-Nezer et al., 2020), a multiverse analysis with 70 different statistic maps originating from the same data. Finally, given that these SDMA methods assume that the inter-pipeline

correlation is the same across the brain, we measured heterogeneity with the Frobenius norm between the whole brain and a set of several brain regions derived from the AAL atlas.

Results

Simulation results under the null setting of no effect and independent pipelines show that all the tested meta-analysis estimators are valid (Fig 1 top row). However, when data are correlated, as typically observed in a multiverse setting, only the SDMA estimators had valid inferences (Fig 1 lower row), while the conventional meta-analysis approach (Stouffer) dramatically overestimated the number of false-positives. On the real world dataset (Fig 2), the GLS method finds more significant voxels while the 3 other methods all have similar sensitivity. Finally, we found that the Frobenius norm was 0.01% across brain regions, supporting the validity of the consistent correlation assumption.

Discussion

We compared several methods for combining multiverse results that account for the dependence among inputs. Our findings demonstrated the validity of the SDMA models under inter-pipeline dependence. As expected, the (traditional) Stouffer's method is liberal while the SDMA methods are all valid and present different levels of significance. These different levels illustrated different types of inference that practitioners can choose based on the assumptions of their analyses. As an illustration, in their work, Botvinik-Nezer and colleagues (2020) implemented the Consensus Average model to combine inferences, striving to align them closely with each of the input pipelines, under the assumption that all pipelines were equally valid and thus contributed equally relevant information. This assumption may not hold true in other multiverse settings.

References

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., & Benoit, R. G. (2020). *Variability in the analysis of a single neuroimaging dataset by many teams*.

Bowring, A., Nichols, T. E., & Maumet, C. (2018). *Same data-different software-different results? Analytic variability of group fmri results*. 1–3.

Carp, J. (2012). On the plurality of (methodological) worlds: Estimating the analytic flexibility

of FMRI experiments. *Frontiers in Neuroscience*, *6*, 149.

Glatard, T., Lewis, L. B., Ferreira da Silva, R., Adalat, R., Beck, N., Lepage, C., Rioux, P., Rousseau, M.-E., Sherif, T., & Deelman, E. (2015). Reproducibility of neuroimaging analyses across operating systems. *Frontiers in Neuroinformatics*, *9*, 12.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., & Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, *18*(2), 115–126.

Stouffer, S. A. (1949). Adjustment during army life. *Studies in social psychology in World War II.* Princeton Univ. Press.

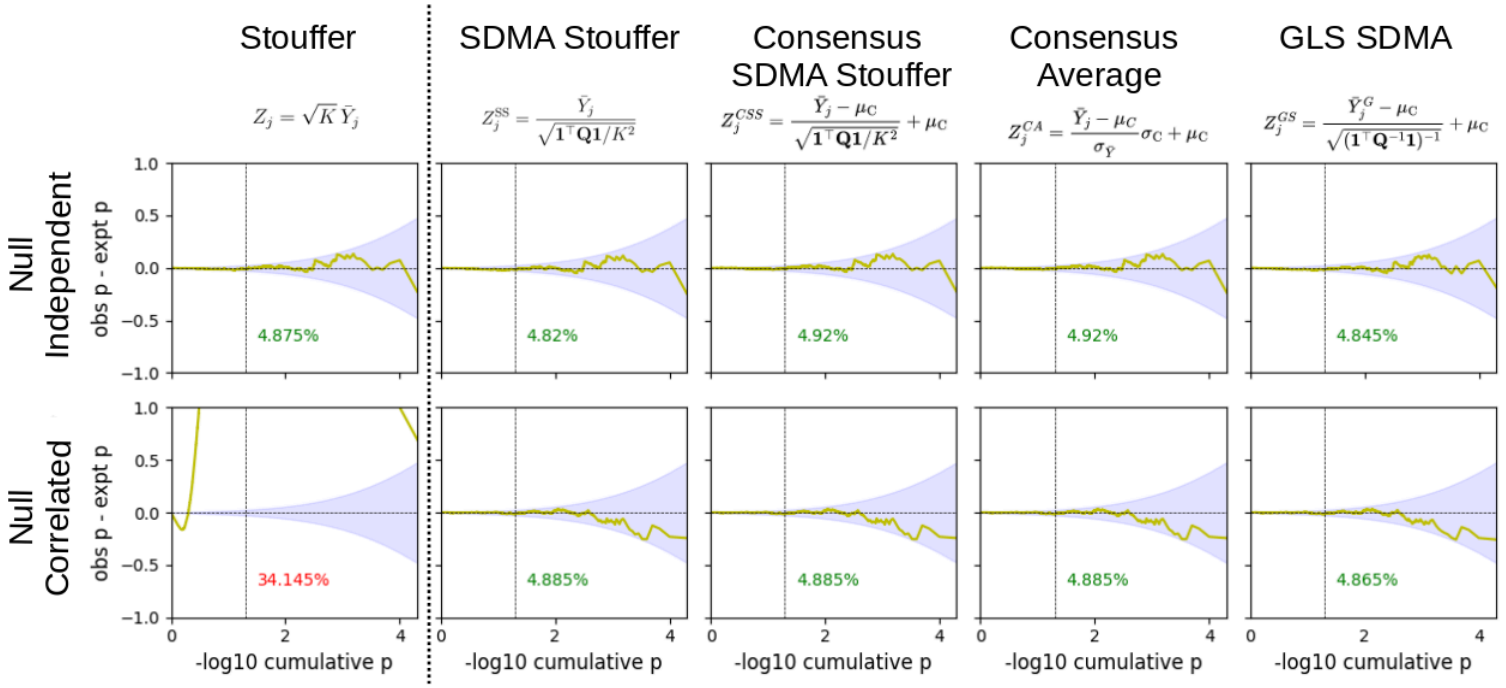

Figure 1: comparative P-P plots of the meta-analysis estimators in traditional and multiverse settings.

Comparative P-P plots for each meta-analysis estimator in the independent (*upper row*) and correlated data (*lower row*) simulations, where the y-axis is the difference in observed and expected sorted -log10 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). As expected, only the SDMA methods (right side of the dotted line) perform well in the multiverse setting.

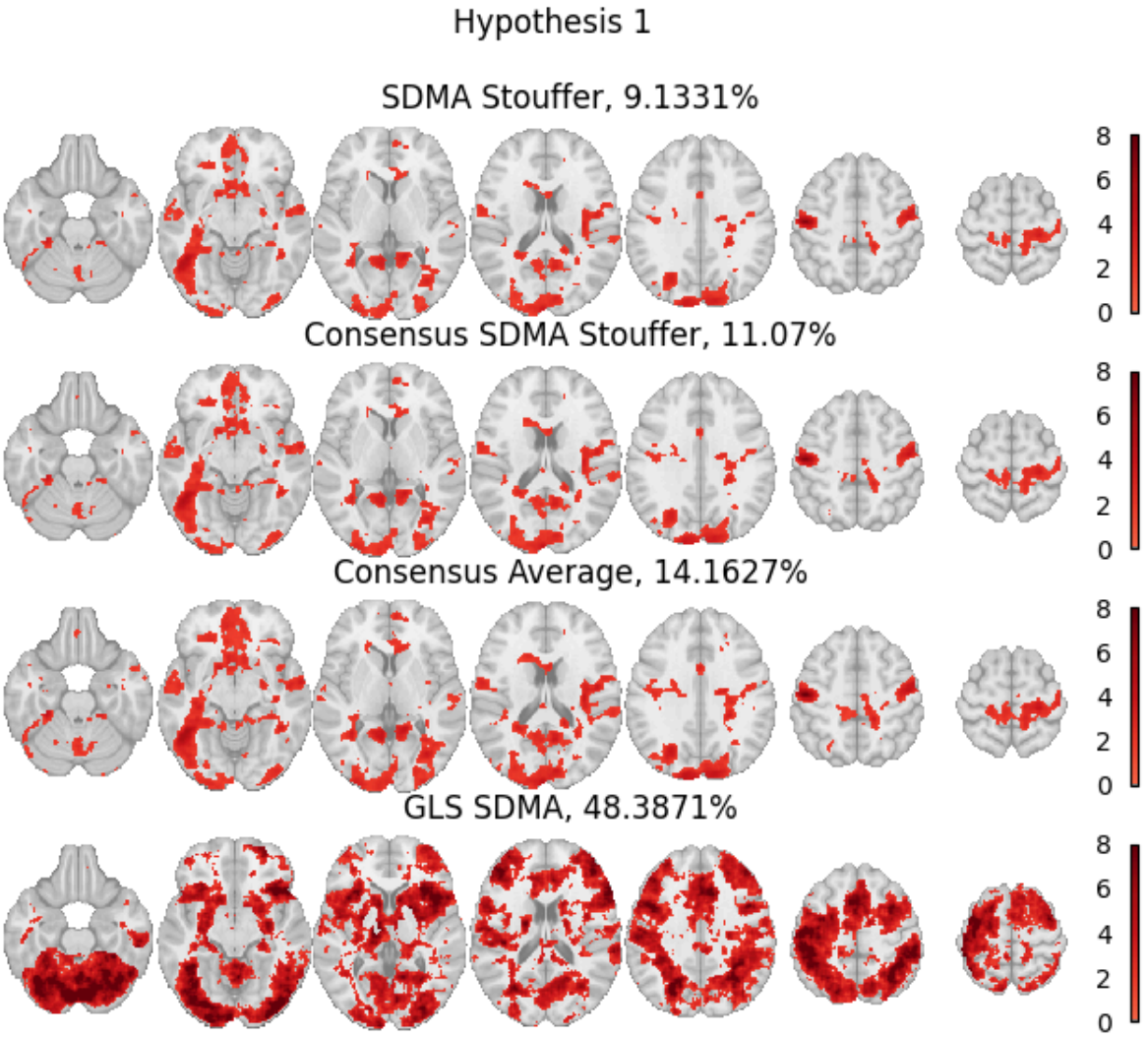

Figure 2: significant p-values for each meta-analysis estimator.
Demonstration of different SDMA methods using the statistical maps from the first hypothesis in the NARPS study (parametric effect of gains in a mixed gambles task). Maps were thresholded at P<0.05 uncorrected to allow for direct comparison. Name of the MA model and percentage of significant voxels are displayed on each map.