Statistical Inference for Same Data Meta-Analysis for Neuroimaging Multiverse Analyzes

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

*These authors contributed equally to this work

May 11, 2024

Abstract

Researchers using task-fMRI data have access to a wide range of analysis tools to model brain activity. This diversity of analytical approaches can lead to an inflated rate of false positives and contributes to the irreproducibility of neuroimaging findings. Multiverse analyses are a way to systematically explore 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 originating from a unique dataset. However, having multiple outputs for the same research question due to diverse analytical approaches may lack practicality and make it challenging to determine robust and consistent findings across studies. 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 here. In this work we consider a suite of methods to conduct meta-analysis in the multiverse setting, which we call same data meta-analysis (SDMA), accounting for inter-pipeline dependence among the results. First, we assessed the validity of theses methods in simulations. Then we tested them on the outputs of two real world multiverse analyses: "NARPS", a multiverse study originating from the same dataset analyzed by 70 different teams, and "HCP Yound Adult", a more homogeneous multiverse analysis originating from

the same Human Connectome Project dataset analyzed with 24 different pipelines built by the same team. Our findings demonstrate the validity of our proposed SDMA models under inter-pipeline dependence, and provide an array of options, with different levels of relevance, for the analysis of multiverse outputs.

Contents

1 Introduction **4**

2 Methods **5**

2.1 Theory 5

2.1.1 Input data 5

2.1.2 Notation 6

2.1.3 Conventional fixed effects meta-analysis model: Stouffer Method . 6

2.1.4 SDMA Stouffer 6

2.1.5 Consensus SDMA methods 7

2.1.6 SDMA GLS methods 9

2.2 Evaluations 10

2.2.1 Simulated outputs 10

2.2.2 Real-world multiverse analysis outputs sources 11

2.2.3 Assessment of Spatial Homogeneity of Correlation Q 12

2.2.4 Assessment of Validity 13

2.2.5 Interpretability of SDMA GLS Results 13

3 Results **16**

3.1 Q assumption 16

3.2 Results in simulations 16

3.3 Results using NARPS multiverse analysis outputs 18

3.4 Results using HCP Yound Adult multiverse analysis outputs 18

3.5 Comparison of Stouffer SDMA and SDMA GLS 19

4 Discussion **21**

5 Conclusion and future works **23**

6 equations for double check **28**

1 Introduction

The multiplicity of analytical methods available through a broad spectrum of tools for modeling brain activity can have a substantial impact on neuroimaging findings (Bowring et al., 2019; Botvinik-Nezer et al., 2020; Strother et al., 2004; Gronenschild et al., 2012; Glatard et al., 2015). This flexibility in analysis combined with selective reporting may result in an increased occurrence of false positives and hence contributes to the lack of reproducibility in neuroimaging results. In departure to traditional analyses in which a single method is used, a multiverse analysis can be used to generate multiple outputs from the same unique dataset (Steegen et al., 2016). These various output sets arise from executing an array of pipelines, each representing a different framework for Magnetic Resonance Imaging (MRI) analysis, which may include variations in both data processing and analytical steps. Multiverse analyses provide a systematic means to investigate the diversity of pipeline approaches applied to a specific dataset.

Each pipeline will produce an array of outputs, but in this work we focus only on the test statistic maps for a particular effect of interest, i.e. maps of T-scores or F-scores, both of which can be converted to Z-scores (While ideally we would work with parameter estimates and standard errors, across pipelines their units are often incompatible due to inconsistent scaling of the data, model and/or contrast, and thus we confine ourselves to Z-scores). The challenge is then how to combine these Z-scores to obtain valid and robust results from these output maps.

Meta-analysis is a natural approach to extract consensus inferences from these Z maps, yet a standard assumption of meta-analysis is independence amongst input datasets (Normand, 1999). However, when combining outputs from different analytical approaches applied to the same dataset (i.e., a multiverse approach), there will likely be very strong dependencies between the outputs. We thus propose a set of dependence-adjusted meta-analysis methods accounting for inter-pipeline dependence among the outputs. We identify two distinct stages of multiverse data analysis, exploratory data analysis, and statis-

tical inference. For exploratory data analysis, we propose a series of graphical methods to explore the dependency structure and assessing the potential spatial heterogeneity of this dependence. For inference, we propose a set of methods for combining a set of multiverse outputs which we call "same-data meta-analysis" (SDMA). We first assessed the validity of the proposed SDMA methods on simulated outputs. Subsequently, we examine their relevance on the outputs of two distinct real world multiverse analyses to gain deeper insights into the properties of each developed SDMA method. We conclude with a discussion of selecting the most suitable method based on specific use cases.

2 Methods

2.1 Theory

In the following we will develop four new same-data meta-analysis (SDMA) methods, two direct extensions of Stouffer combining method, and two based on Generalized Least Squares for the optimal combination of dependent outputs.

2.1.1 Input data

Broadly, there are three different types of outputs from a multiverse analysis: test statistics alone (e.g. only Z-score image), pairs of estimates and standard errors (e.g. as obtained from group task-fMRI analyses), and arbitrary values (e.g. correlations in connectome maps, or microstructural parameters from diffusion MRI). In this work, we consider only Z-scores, leaving the other two cases for future work. Further we assume all input maps take the form of Z-values, a reasonable starting point since other types of statistics can be converted to Z-scores. Throughout we assume Normality, the basic assumption that would be required for statistical inference on any individual pipeline. For the remainder of this manuscript, we will adopt the following standardized terminology for clarity: the term 'dataset' will consistently refer to the original task-fMRI dataset prior to analysis; 'outputs' will be used to describe the results of a multiverse analysis presented as Z

maps (one for each pipeline); and 'results' will denote the statistical maps derived from applying a meta-analysis model to the outputs of a multiverse analysis.

2.1.2 Notation

We assume each $\mathbf{Y_{kj}}$ is mean zero and variance one under the null hypothesis tested by the Z-score, but allow for inter-pipeline correlation with $K \times K$ correlation $\mathbf{Q}$. We develop all of these methods assuming spatial homogeneity of correlation, i.e. that all voxels share the same correlation $\mathbf{Q}$; this is a non-trivial assumption but below, we critically assess the assumption on the outputs of real world multiverse analyses.

2.1.3 Conventional fixed effects meta-analysis model: Stouffer Method

Stouffer method (Stouffer et al., 1949) is perhaps the most straighforward Z-score combining method, based on the sample mean of input Z-scores denoted $\bar{Y}_j = \frac{1}{K}\sum_k \mathbf{Y_{kj}}$:

$$Z_j^{\mathrm{S}} = \frac{\bar{Y}_j}{\sqrt{1/K}}$$

where Z_j^{S} is again a Z score and has mean zero and variance one under the null and magnified mean $\mu\sqrt{K}$ under the alternative. This traditional meta-analytic is a fixed effect method that is designed to powerfully combine evidence against the null. In essence, Stouffer combining creates an average map and then standardizes to account for $\mathrm{Var}(\bar{Y}_j) = 1/K$, producing a unit variance result.

2.1.4 SDMA Stouffer

The standard Stouffer result is based on an assumption of independent inputs. Given that this assumptions is not tenable in a multiverse setting, we propose a modification of the traditional Stouffer method, an SDMA version that accommodates an inter-pipeline correlation $\mathbf{Q}$. First, note that the variance of an average of K unit variance dependent variables with correlation $\mathbf{Q}$ is $\mathbf{1}^\top\mathbf{Q}\mathbf{1}/K^2$, where $\mathbf{1}$ is a vector of 1's. We thus propose

"SDMA Stouffer" Z^{SS} as the average with standardization to account for correlation $\mathbf{Q}$ among the inputs:

$$Z_j^{\mathrm{SS}} = \frac{\bar{Y}_j}{\sqrt{\mathbf{1}_k^\top \mathbf{Q} \mathbf{1}_k / K^2}}. \tag{1}$$

2.1.5 Consensus SDMA methods

While both original and SDMA Stouffer method scale the average to unit variance, it results in a scaling of the average. With independent datasets this is natural – when multiple studies all have evidence against the null, their combined evidence is yet stronger evidence than the mean Z, as reflected by the $\sqrt{K}$-amplification. With multiverse outputs, it is perhaps enigmatic: the original dataset is the same, but by combining similar but not identical versions of the outputs we can obtain results with amplified evidence against the null. Under the null hypothesis of mean zero signal everywhere there is no concern of signal amplification, but when signal is present it is impossible to scale a univariate (or single voxel j) average so that *both* mean and variance are preserved. However, for an image of statistics, we can shift the voxel-wise mean over voxels to have some target or "consensus" value.

In the following consensus methods, we propose two different ways to combine K test statistic images such that the output is based on an average yet the output is as similar as possible to the input as possible. In the following we denote μ_C and σ_C as the consensus mean and consensus standard deviation, respectively, we would like our final map to have. These could be arbitrary, but we assert that the most sensible values are the average over the K inputs

$$\mu_{\mathrm{C}} = \frac{1}{K} \sum_k \langle \mathbf{Y_k} \rangle \tag{2}$$

$$\sigma^2_{\mathrm{C}} = \frac{1}{K} \sum_k \langle \langle \mathbf{Y_k} \rangle \rangle \tag{3}$$

$$\tag{4}$$

of the respective voxel-wise statistics, where $\langle\cdot\rangle$ denotes image-wise average, i.e. $\langle \mathbf{Y_k} \rangle = \frac{1}{J}\sum_j \mathbf{Y_{kj}}$ is is the voxel-wise average for input k, and $\langle\langle\cdot\rangle\rangle$ is the voxel-wise variance, i.e. $\langle\langle \mathbf{Y_k} \rangle\rangle = (J-1)^{-1}\sum_j (Y_{jk} - \langle \mathbf{Y_k} \rangle)^2$

Consensus SDMA Stouffer Our first consensus method simply shifts the mean so that the image-wise mean of the output has the consensus mean:

$$Z_j^{\mathrm{CSS}} = Z_j^{\mathrm{SS}} - \langle Z^{\mathrm{SS}} \rangle + \mu_C,$$

that is, just the SDMA Stouffer value centered image-wise to have average μ_C. Of course, if the null hypothesis is true everywhere, then both $\langle Z^{\mathrm{SS}} \rangle$ and μ_C will be zero with high precision (since they are an average over many voxels) and this will have no impact. But if there is activation, it will.

In summary, the consensus SDMA Stouffer approach is accounting for inter-pipeline correlation, but then adjusting the resulting map so that it has the intervoxel average equal to the inter-pipeline average of the intervoxel average of each pipeline. so that it has the voxelwise average equal to the average overall all pipelines of the voxelwise averages.

Consensus Average The preceding SDMA Stouffer methods use statistical theory to account for the impact of dependence on the variability of the computed summary. However, alternatively, a less technical approach is to simply compute an average and use its own voxel-wise statistics to standardize before scaling and shifting to have the desired consensus mean and standard deviation.

$$Z_j^{\mathrm{CA}} = \frac{\bar{Y}_j - \langle \bar{Y} \rangle}{\sqrt{\langle\langle \bar{Y} \rangle\rangle}} \sigma_C + \mu_C$$

It is expected that the consensus average Z_j^{CA} will produce values very similar to consensus SDMA Stouffer Z_j^{CSS}. While Z_j^{CSS} uses statistical results to compute the impact of averaging K dependent inputs, Z_j^{CA} simply uses the naive Stouffer as a starting point, standardizing, scaling and shifting to desired consensus values.

2.1.6 SDMA GLS methods

SDMA Generalized Least Squares (GLS) When analyzing dependent outputs, the optimal, minimum variance estimates are obtained by generalized least squares (GLS), where both data and model are whitened. First consider the unwhitened case: For a regression of the K-vector of input data $\mathbf{Y}_j$ on a design matrix $\mathbf{X} = \mathbf{1}$, the least squares estimate is $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}_j = \mathbf{1}^\top\mathbf{Y}_j/K$ (the average) and the variance of the estimate is

$$(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \operatorname{Var}(\mathbf{Y}_j)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = \mathbf{1}^\top\mathbf{Q}\mathbf{1}/K^2, \tag{5}$$

exactly the variance found above, and the estimate divided by standard deviation is exactly the SDMA Stouffer (1).

So now instead consider whitening with $\mathbf{Q}^{-1/2}$, giving GLS mean estimate

$$\bar{Y}_j^{\mathrm{G}} = (\mathbf{X}^\top\mathbf{Q}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Q}^{-1}\mathbf{Y}_j = \frac{\mathbf{1}^\top\mathbf{Q}^{-1}\mathbf{Y}_j}{\mathbf{1}^\top\mathbf{Q}^{-1}\mathbf{1}} \tag{6}$$

and variance

$$(\mathbf{X}^\top\mathbf{Q}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Q}^{-1} \operatorname{Var}(\mathbf{Y}_j)\mathbf{Q}^{-1}\mathbf{X}(\mathbf{X}^\top\mathbf{Q}^{-1}\mathbf{X})^{-1} = (\mathbf{1}^\top\mathbf{Q}^{-1}\mathbf{1})^{-1}. \tag{7}$$

Thus our SDMA GLS is the GLS estimate divided by its standard deviation:

$$Z_j^{\mathrm{SG}} = \frac{\bar{Y}_j^{\mathrm{G}}}{\sqrt{(\mathbf{1}^\top\mathbf{Q}^{-1}\mathbf{1})^{-1}}} = \frac{\mathbf{1}^\top\mathbf{Q}^{-1}\mathbf{Y}_j}{\sqrt{\mathbf{1}^\top\mathbf{Q}^{-1}\mathbf{1}}} \tag{8}$$

The motivation behind using GLS is that, instead of weighting each output equally as in $\bar{Y}_j^{\mathrm{S}}$, we combine the K outputs according to $\mathbf{1}^\top\mathbf{Q}^{-1}$, which has the effect of down-weighting the influence of highly dependent pipelines. To illustrate, consider a scenario where the first half of outputs are derived from the same pipeline computed across various operating systems, resulting in virtually identical outputs. Conversely, the second half of pipelines produce nearly independent outputs. When calculating an unweighted average,

equal weight is assigned to all inputs. However, it might be more appropriate to assign greater weight to the second half of pipelines, as they provide more distinct and varied information.

Consensus SDMA GLS We can likewise define a consensus SDMA GLS, Z^{CSG}, which is shifted to have a consensus image-wise average:

$$Z_j^{\mathrm{CSG}} = Z_j^{\mathrm{SG}} - \langle Z^{\mathrm{SG}} \rangle + \mu_C \tag{9}$$

2.2 Evaluations

2.2.1 Simulated outputs

Null outputs generation We simulated a set of Z-statistic maps Z_j according to a K-dimensional normal:

$$Y_j \sim \mathcal{N}(0, \mathbf{Q})$$

In a first scenario of null, independent pipelines, we generate outputs with $\mu = 0$ and $\mathbf{Q} = \mathbf{I}$.

In a second scenario of null, correlated pipelines, we set $\mu = 0$ and considered different levels of dependence, specifically using compound symmetric correlation structures where all correlations are equal. The correlation was set to one of three possible values (0.2, 0.5, and 0.8).

In a third scenario of null, three pipelines were independent and the others were correlated pipelines. We varied the correlation according to the three values described above.

For each scenario, the number of pipelines and voxels were respectivelly varied from a grid of K in {20,50,100} and J in {5.000, 10.000, 20.000}. Simulations were implemented in Python (3.11.6). Summary heatmaps for each of the main scenario can be found in Figure 1. All analysis scripts of the present study are readily accessible to the reader

online https://github.com/JLefortBesnard/SDMA.

2.2.2 Real-world multiverse analysis outputs sources

NARPS outputs description The Neuroimaging Analysis Replication and Prediction Study Botvinik-Nezer et al. (2020), also called the NARPS study, recently evaluated the degree and impact of analytic flexibility on task-fMRI results. They assessed the real-world variability of outputs across independent teams analyzing the same dataset. The dataset included task-fMRI data from 108 individuals, each performing one of two versions of a task previously used to study decision-making under risk Tom et al. (2007); Canessa et al. (2013). This dataset is available at: https://openneuro.org/datasets/ds001734/versions/1.0.4. 70 teams were provided with the raw data and an optional preprocessed data, and were asked to analyze the data to test nine hypotheses, each consisting of a yes/no question regarding significant activity in a specific brain region in relation to a particular feature of the task. Among other outputs, each team submitted the unthresholded statistic maps supporting each hypothesis test. In our study, we used the unthresholded statistic maps of 55 from the 70 teams included in NARPS. 15 statistic maps were removed from the image-based analysis (see Supplementary Table 1 for details). We assessed the sensibility of each meta-analytic estimator on this set of 55 unthreshold maps. These maps are available at https://github.com/poldrack/narps/tree/master/ImageAnalyses. In this paper, we present our results within the first NARPS hypothesis. Results in the remaining hypotheses can be found in the supplementary.

HCP Yound Adult outputs description The Human Connectome Project (HCP) is an ambitious 5-year effort to characterize brain connectivity and function and their variability in healthy adults (Van Essen et al., 2012). In particular, the HCP Young Adult provides task-fMRI data for different tasks and cognitive processes. Using the motor task-fMRI data from the HCP Yound Adult, Germani et al. (2023) evaluated the

variability of outputs across 6 different contrasts and 24 different preprocessing and first-level analyses from the same dataset (1,080 participants from the HCP Young Adults S1200 release). The 24 different pipelines differed in 4 parameters: software package (SPM or FSL), smoothing kernel (5 or 8mm), number of motion regressors (0, 6 or 24) included in the General Linear Model (GLM) for the first-level analysis, and presence or absence of the derivatives of the Hemodynamic Response Function (HRF) in the GLM for the first-level analysis. Unthresholded statistic maps were obtained for each pipeline, resulting in 24 maps per contrast. In our work, we assessed the sensibility of each meta-analytic estimator for the 24 unthresholded maps obtained for the right-hand contrast. The data paper presenting these maps is available in Germani et al. (2023). Note that these multiverse analysis outputs were generated within a single laboratory using only two software tools, in contrast to the NARPS outputs which involved 70 different teams and multiple software applications. As a consequence, the unthresholded maps of these HCP Yound Adult outputs are more homogeneous than the 70 unthresholded maps of the NARPS mutliverse analysis, enabling us to examine the effects of heterogeneity.

2.2.3 Assessment of Spatial Homogeneity of Correlation Q

Given that these SDMA methods assume that the inter-pipeline correlation is the same across the brain, we measured heterogeneity within the outputs of the NARPS and HCP Yound Adult multiverse analysis. We thus computed the magnitude of the difference between the inter-pipeline correlation across the whole brain and across brain regions.

First, we calculate the difference between correlation matrix of the whole brain and of a set of 5 brain regions (frontal, parietal, temporal, insular, and occipital) derived from the Harvard-Oxford atlas (Caviness Jr et al., 1996; Rademacher et al., 1992; Jenkinson et al., 2012). This similarity matrix highlights where and how much the matrices differ element-wise.

$$\mathbf{Q}_{Si} = \mathbf{Q}_i - \mathbf{Q}_b$$

with $\mathbf{Q}_i$ the correlation matrix within one of the 5 brain regions and $\mathbf{Q}_b$ the correlation

matrix within the whole brain.

Then, the Frobenius norm of these similarity matrices $\mathbf{Q}_{Si}$ is computed, as the square root of the sum of the squares of its elements, representing the magnitude of the difference between the two matrices.

$$\|\mathbf{Q}_{Si}\|_F = \mathrm{Tr}(\mathbf{Q}_{Si}\mathbf{Q}_{Si}^T)$$

2.2.4 Assessment of Validity

The false positive rate of each meta-analytic estimator was assessed using simulated outputs. T-values were calculated for each SDMA methods using the SDMA equations described previously. Subsequently, P-values were derived from these T-maps using the cumulative distribution function of a standard normal distribution. These P-values were left uncorrected for multiple comparisons to facilitate comparison between SDMA methods. The real world multiverse analysis outputs were likewise analyzed and qualitative comparisons were made between results from different SDMA methods.

2.2.5 Interpretability of SDMA GLS Results

We found that results were quite similar among methods based on the sample mean (SDMA Stouffer, consensus SDMA Stouffer, and Consensus Average), and quite different from the methods involving whitening (SDMA GLS and consensus SDMA GLS). Also, Given that SDMA GLS downweights the contribution of highly dependent pipelines, comparing the weight and contribution of various sets of pipelines might help visualizing SDMA GLS method behavior. Thus we developed an approach to measure the influence of each study on each of the two types of methods, SDMA Stouffer with the SDMA GLS.

Note that the SDMA Stouffer method (Eq. (1)) can be re-written

$$Z_j^{\mathrm{SS}} = \sum_{k=1}^{K} w^{\mathrm{Q}} Y_j = w^{\mathrm{Q}} \left(\sum_{k=1}^{K} Y_j \right) \tag{10}$$

where

$$w^{\mathrm{Q}} = (\mathbf{1}^\top \mathbf{Q} \mathbf{1})^{-1/2} \tag{11}$$

showing that every study $k = 1, ..., K$ has equal influence on the resulting statistic Z_j^{SS}.

Now consider rewriting SDMA GLS (Eq. (8)), as

$$Z^{\mathrm{CGS}} = \sum_{k=1}^{K} w_k^{\mathrm{QGC}} Y_k = w^{\mathrm{Q}} \left(\sum_{k=1}^{K} \frac{w_k^{\mathrm{QGC}}}{w^{\mathrm{Q}}} Y_k \right) \tag{12}$$

where

$$w_k^{\mathrm{QGC}} = (\mathbf{1}^\top \mathbf{Q}^{-1} \mathbf{1})^{-1/2} \sum_{k'} ((\mathbf{Q}^{-1}))_{k'k}, \tag{13}$$

which shows that each pipeline has an unequal contribution.

These expressions show that the weight of each pipeline is constant in SDMA Stouffer and equal to w^{Q}, while in SDMA GLS it varies and is w_k^{QGC}. In terms of the actual contribution to statistic (i.e., each element of the summand), the k-th contribution for SDMA Stouffer is $w^{\mathrm{Q}} Y_k$, but if we divide it by w^{Q} we obtain the original measurement.

If we wanted to compare the relative impact of each study we would examine each element of the summand in the versions above. For GLS, this motivates considering the k-th contribution as $w_k^{\mathrm{QGC}} Y_k / w^{\mathrm{Q}}$, the k-th element of the summand adjusted so that is as a comparable to the original measurement as possible. Specifically this is

$$\frac{w_k^{\mathrm{QGC}}}{w^{\mathrm{Q}}} Y_k = \frac{(\mathbf{1}^\top \mathbf{Q}^{-1} \mathbf{1})^{-1/2}}{(\mathbf{1}^\top \mathbf{Q} \mathbf{1})^{-1/2}} \sum_{k'} ((\mathbf{Q}^{-1}))_{k'k} Y_k. \tag{14}$$

As a sanity check, note that in the case that the pipelines were actually independent, $\mathbf{Q} = \mathbf{I}$, this expression would just be Y_k.

In the following, our aim was to evaluate the behavior of both the SDMA Stouffer and SDMA GLS Approaches. We wanted to compare weights and contributions between highly dependent pipelines and the rest. We thus first defined subgroups based on their similarities (see subgroups definition below). Then, we calculated the SDMA Stouffer and SDMA GLS weights assigned to each pipeline and computed their contribution, as

detailed above. We then looked at two indicators: the contribution of each subgroup (as the sum of contributions of all pipelines from a subgroup) and the average weight (over all pipelines in a subgroup).

Subgroups within the NARPS outputs The authors of the NARPS study Botvinik-Nezer et al. (2020) calculated Spearman correlations between whole-brain unthresholded statistic maps between each team and then clustered the pipelines based on similarities. The authors performed this clustering analysis for the nine hypotheses tested in NARPS. To assess and directly compare the performance of both the SDMA Stouffer and the SDMA GLS methods, we utilized their three subgroup solutions obtained within the first hypothesis, encompassing majority (highly correlated pipelines), opposite (anti-correlated pipelines), and unrelated (independent pipelines) subgroups (Supplementary Table 2 and Supplementary Figure 5 and 6). Given that SDMA GLS downweights the contribution of highly dependent pipelines, comparing the weight and contribution of various sets of pipelines might help visualizing SDMA GLS method behavior.

Subgroups within the HCP Yound Adult Similarly to the approach taken in NARPS, we computed Spearman correlations among whole-brain unthresholded statistical maps from each of the 24 pipelines from (Germani et al., 2023), revealing highly correlated maps. Subsequently, we performed pipeline clustering based on these similarities and adopted the 2-cluster solution (Supplementary Figure 8). We thus divided the 24 pipelines into into 2 subgroups, namely FSL and SPM. Again, the weight and contribution of these two sets of pipelines were computed.

3 Results

3.1 Q assumption

We found that the Frobenius norm score was very small, below 10, over the five different regions on the seven different contrasts of the NARPS outputs. On the HCP Yound Adult outputs, the Frobenius norm score was even smaller, around 1 (Table 1 and Supplementary Figure 5 and 6).

Brain region	**Frobenius norm score**							
	NARPS							**HCP**
	Hyp 1	**Hyp 2**	**Hyp 5**	**Hyp 6**	**Hyp 7**	**Hyp 8**	**Hyp 9**	
Frontal	5.2	4.4	7.8	7.6	7.7	7.2	6	1.2
Occipital	4.7	6	5.5	5.9	5.6	5.6	5.2	0.7
Parietal	5.8	5.5	5.1	4.8	5.1	4.9	5.1	0.6
Temporal	6.2	5.2	8.9	6	9.8	5.9	4.9	0.7
Insular	8.4	10.3	7.9	7.6	8.3	7.7	8.4	0.7

Table 1: Testing spatial homogeneity in NARPS and HCP outputs

3.2 Results in simulations

We evaluated the sensibility of each meta-analytic estimator in a set of simulations. Summary heatmaps for each of the main scenario can be found in Figure 1. In simulations under the null scenario, where no effect was present, we find that when pipelines are independent (Figure 2, upper row), all meta-analysis methods performed well (i.e. within the confidence bounds). However, in the correlated settings, we find that Stouffer method has a dramatically inflated false-positive rate whereas the SDMA estimators worked as expected (Figure 2, middle row). This observation remains consistent even in the scenario where a few independent pipelines were included in the correlated outputs (Figure 2, bottom row). These results were found in the 3 main simulations, where K is set to 20 pipelines, J to 20.000 voxels, and the correlation value to 0.8. Results were essentially identical for other combinations of J, K, and correlation values (Supplementary Figure 1, 2, 3, 4).

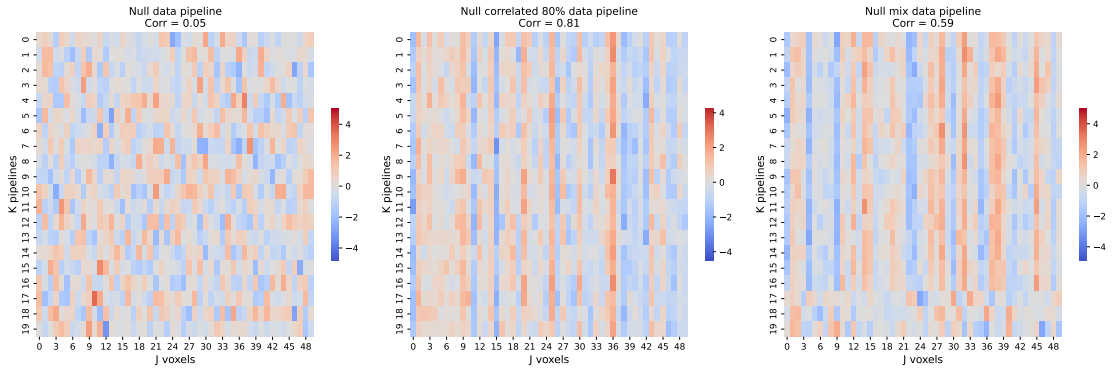

Figure 1: The Z-statistic values were displayed for the independent (scenario 1) and correlated (scenario 2) pipeline outputs null case simulations, as well as for the null case including correlated and independent pipeline outputs (scenario 3). For the sake of visualization, only the Z-values for the 20 first teams (pipelines) and the 50 first voxels were displayed. Note that we illustrated only one simulation per scenario, even though several simulations were explored for each scenario.

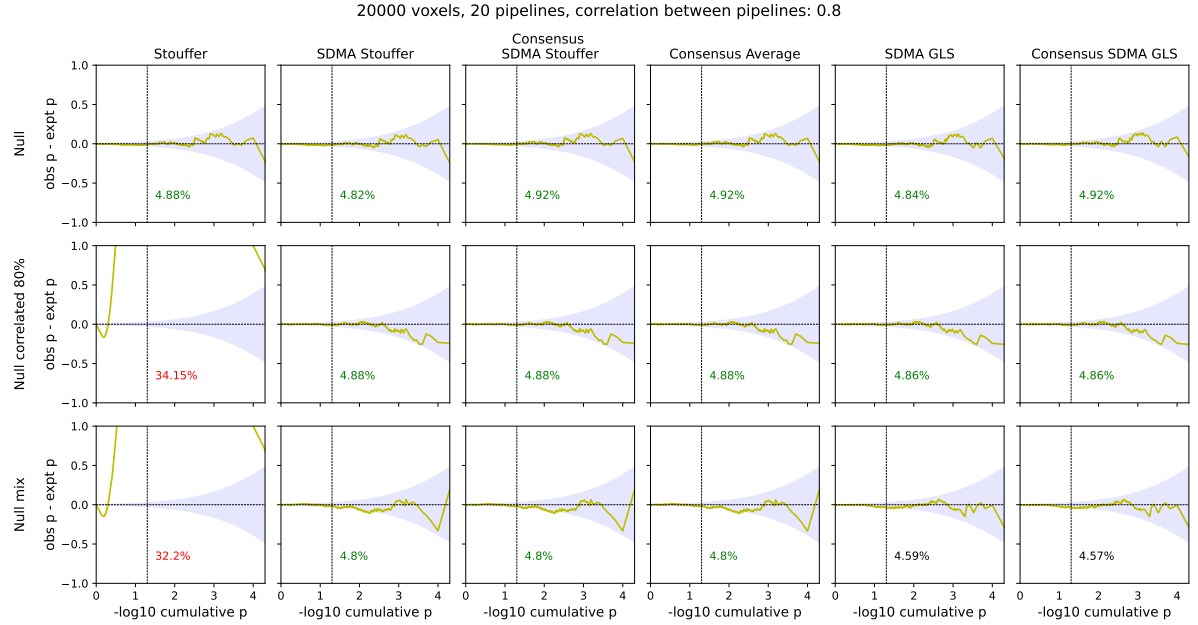

Figure 2: Comparative P-P plots for each meta-analysis estimator in the independent (upper row), correlated pipelines (middle row), and mix (bottom row) simulations, where the y-axis is the difference in observed and expected -log10 ordered P-value, and the x-axis is the sorted expected -log10 ordered P-value. The blue shading depicts the nominal 95% confidence interval for each expected ordered P-value. At the bottom of each plot is the false positive rate for $\alpha = 5\%$, displayed in red when significantly different from nominal and green otherwise. As expected, only the SDMA methods (all methods on the right of the "Stouffer") performed well in the dependent multiverse setting.

3.3 Results using NARPS multiverse analysis outputs

The meta-analysis estimators were computed using the statistic map from each of the 55 NARPS teams, producing a map of P-value. Significant T-values (P<0.05 uncorrected) are plotted in MNI space, where all results have been masked with an atlas generated from the intersection of the MNI and pipelines brain maps to exclude voxels that were not present in all teams. Figure 3 (left) shows the first NARPS hypothesis; see Supplementary Figure 7 for other hypotheses.

Areas of significance of the P-values were plotted with a value equal to the corresponding T-value, and then plotted within an atlas generated from the intersection of the MNI and pipelines brain maps (Figure 3, left column). The percentage of significant P-values was similar in the SDMA Stouffer (9.13%), in the consensus SDMA Stouffer (11.07%), and in the consensus average (14.16%). However, the GLS methods exhibited divergent outcomes, exhibiting a substantially higher proportion of significant voxels (48.39% and 45.79%). Note that unlike the outcomes observed in the simulation scenario, we are no longer operating under the null hypothesis. Consequently, we should not anticipate a significance level of 5% anymore, and moreover, there is no definitive truth available (i.e., we lack knowledge of which method yields the most optimal outcome).

3.4 Results using HCP Yound Adult multiverse analysis outputs

Combining the statistic maps from each of the 24 pipelines created a P-value map. Significant T-values (P<0.05 uncorrected) are plotted in the same MNI space as NARPS (Figure 3 right). In contrast to the findings of NARPS, all estimators yielded comparable results, ranging from 28.29% to 32.46% of significant voxels.

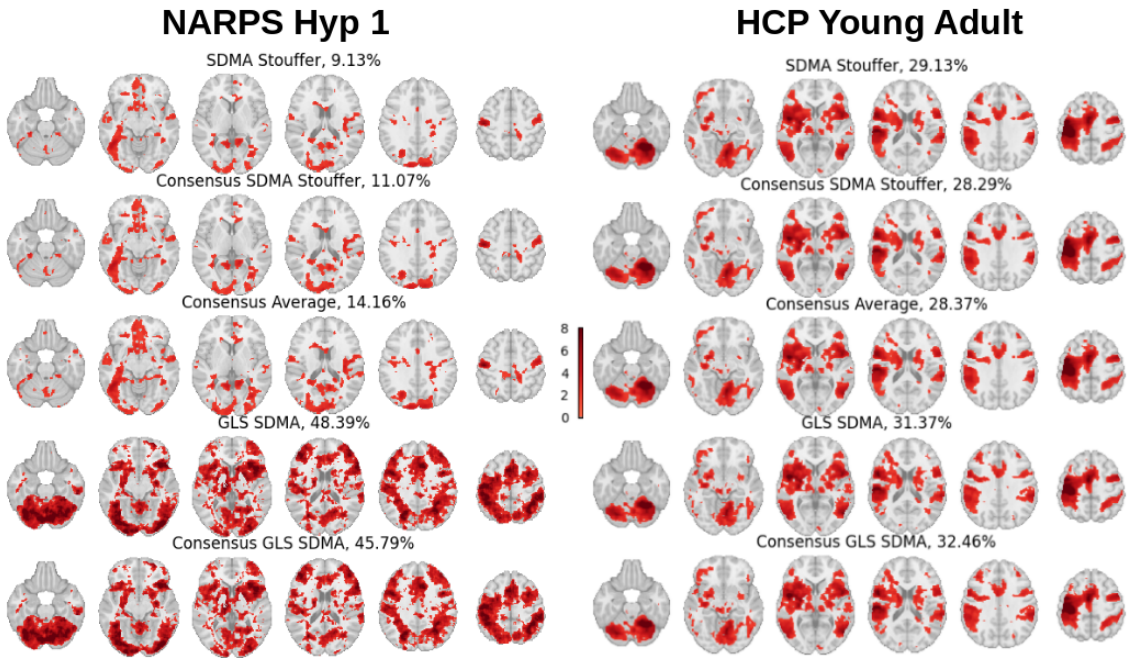

Figure 3: Uncorrected significant P-values (indicated by their corresponding T-values) for each meta-analysis estimator. Demonstration of different SDMA methods using the statistic maps from the NARPS study (first hypothesis) and using the statistic maps from Germani et al. (2023) (*HCP Yound Adult*). Maps were thresholded at $p \leq 0.05$ uncorrected to allow for direct comparison. Name of the SDMA model and percentage of significant voxels are displayed on each map.

3.5 Comparison of Stouffer SDMA and SDMA GLS

Motivated by the differences observed in the results in the NARPS outputs, between equally weighted and whitened SDMA methods, we examined the weight and contribution assigned by Stouffer SDMA and SDMA GLS across three distinct pipeline subgroups in the NARPS outputs: majority, opposite, and unrelated subgroups. Our results showed that using the SDMA Stouffer method, the final significance map closely resembles the contribution map of the majority subgroup, which contains most of the pipelines (Figure 4, left section). Equal weighting is allocated to every pipeline and consequently to each subgroup, resulting in the majority group exerting the greatest influence. Examination of weights and contributions per pipeline subgroup reveals that GLS attributed greater importance to the unrelated and opposite subgroups (figure 4, right section), with the majority of significant voxels originating from the opposite subgroup, a surprising re-

sult as the significant effects are in fact in the opposite direction of the largest collection of studies.

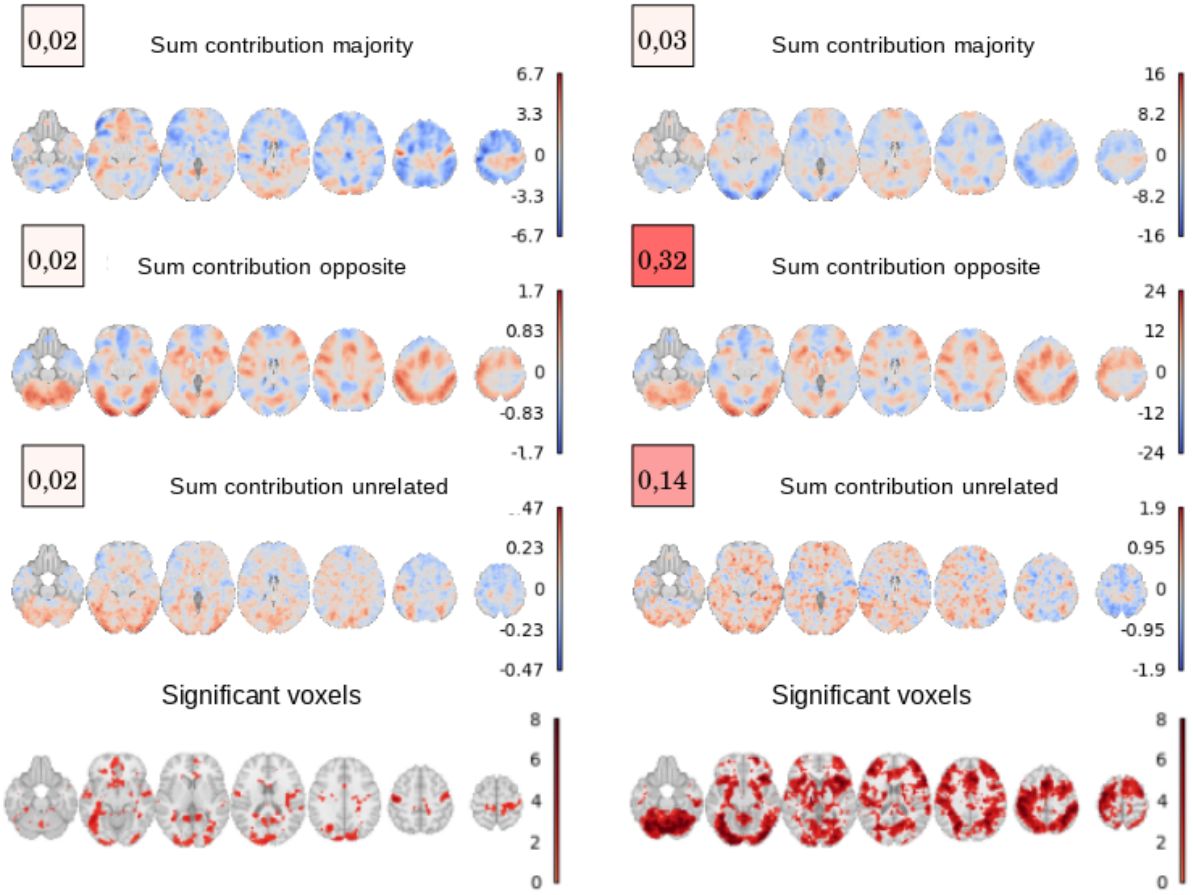

Figure 4: Characteristics of the SDMA Stouffer and the SDMA GLS methods illustrated on the first NARPS hypothesis outputs. The authors of the NARPS study computed Spearman correlations between whole-brain unthresholded statistic maps for each team and clustered them accordingly based on their similarities. In our work, we employed their three subgroup solutions, which included majority, opposite, and unrelated subgroups. The left panel illustrates the aggregated SDMA Stouffer contributions within each subgroup depicted on an MNI atlas, along with the mean SDMA Stouffer weight per subgroup (colored square). Likewise, the right panel showcased the aggregated contributions and average weights per subgroup, assigned by the SDMA GLS estimator. The bottom row displays the significance levels for each method

4 Discussion

The primary objective of this paper is to introduce and assess several same-data meta-analysis (SDMA) techniques. Our results can be summarized in four main findings.

As expected, we find that the traditional Stouffer method produces dramatically inflated false positive rates in presence of collinearity among pipelines. Conversely, we show that our SDMA Stouffer and SDMA Stouffer consensus methods are valid, robust and suitable for multiverse settings. However, while the SDMA GLS methods are valid in simulation, our findings with the NARPS outputs show that complex and negative dependencies among pipelines can lead to unexpected results.

The conventional Stouffer method fails to address the dependency structure within the multiverse outputs The conventional Stouffer method is grounded in an assumption of independent inputs, and as expected we found greatly inflated false positive rates in the presence of dependence. This inadequacy motivated the creation of the five different SDMA methods for combining multiverse outputs.

SDMA methods demonstrated effectiveness in both independent and multiverse simulations. Every SDMA methods developed in this work worked as expected in simulations of both independent and dependent outputs under the null scenario, producing nominal false positive rates. The correlation degree among pipelines did not influence these findings, nor did the number of voxels and pipelines included in the analysis. Our simulation results indicate that the developed SDMA methods are suitable in the context of a multiverse setting.

Application of SDMA methods in homogeneous outputs In our analysis using outputs of real world multiverse analysis, all proposed SDMA methods produced nearly identical results in homogeneous outputs. The HCP Yound Adult multiverse outputs were carried out by a single team. As a result, these multiverse outputs are relatively

homogeneous, and we found all five of our methods produced nearly identical results. Notably, the methods that should be theoretically optimal (using GLS whitening instead of equally weighted average) were most sensitive, detecting more voxels than the other methods (Figure 3 right). Overall, these results indicate that the five developed methods are robust and consistent across scenarios with minimal variability. We also note that the motivation for the Consensus SDMA Stouffer method is to reduce the magnification of the significance from combining distinct information across the different pipelines. However, there was not a substantial difference between Consensus and SDMA Stouffer regarding the results obtained using HCP Young Adult outputs.

Application of SDMA methods in heterogeneous outputs The NARPS analyses were carried out by 70 different teams, and exhibit appreciable variability with some teams exhibiting negligible or even negative correlation with the main subgroup of teams. As a result, these multiverse outputs are quite heterogeneous, and while the SDMA Stouffer, the consensus SDMA Stouffer and the consensus average methods yield virtually identical results, the GLS-based SDMA methods produced quite different combined maps. We investigated the source of these differences and found they can be attributed to the presence of anticorrelated pipelines (opposite subgroup) in the NARPS outputs. Our examination of weights and contributions within each subgroup of pipelines indicates that GLS assigns more weight to the unrelated and opposite subgroups, while diminishing the impact of pipelines from the majority subgroup. In instances involving highly heterogeneous pipelines, interpreting the resulting outcomes can be difficult and could be unstable in the presence of anticorrelated or otherwise outlier pipelines.

Overall Recommendations Among these five methods we recommend the SDMA Stouffer as the basic go-to method that is robust and easy to interpret. If there is a concern that effects are being amplified, either Consensus SDMA Stouffer or Consensus Average can be used. Finally, if one has a relatively homogeneous set of outputs and wants to maximize the statistical power, SDMA GLS should produce the optimal inference.

5 Conclusion and future works

Multiverse analyses offer a systematic approach to practically address analytical variability, an important driver of irreproducibility in neuroimaging research, by exploring and integrating variation across different analysis pipelines applied to the same dataset. In this study, our emphasis was on meta-analysis methods designed specifically for the multiverse setting, which considers inter-pipeline dependence among outputs. Through simulations and assessments on two real-world multiverse analysis outputs, we verified the effectiveness of our proposed SDMA models. Furthermore, our findings underscored that GLS methods in scenarios with high heterogeneity may result in unclear and difficult-to-interpret outcomes, suggesting they may not be appropriate for application in a multi-expert context like NARPS.

References

A Bowring, C Maumet, and TE Nichols. Exploring the impact of analysis software on task fMRI result. *Human Brain Mapping*, 2019. doi: 10.0.1002/hbm.24603. URL https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/hbm.24603.

Rotem Botvinik-Nezer, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A Mumford, R. Alison Adcock, Paolo Avesani, Blazej M. Baczkowski, Aahana Bajracharya, Leah Bakst, Sheryl Ball, Marco Barilari, Nadège Bault, Derek Beaton, Julia Beitner, Roland G. Benoit, Ruud M. W. J. Berkers, Jamil P. Bhanji, Bharat B. Biswal, Sebastian Bobadilla-Suarez, Tiago Bortolini, Katherine L. Bottenhorn, Alexander Bowring, Senne Braem, Hayley R. Brooks, Emily G. Brudner, Cristian B. Calderon, Julia A. Camilleri, Jaime J. Castrellon, Luca Cecchetti, Edna C. Cieslik, Zachary J. Cole, Olivier Collignon, Robert W. Cox, William A. Cunningham, Stefan Czoschke, Kamalaker Dadi, Charles P. Davis, Alberto De Luca, Mauricio R. Delgado, Lysia Demetriou, Jeffrey B. Dennison, Xin Di, Erin W. Dickie, Ekaterina Dobryakova, Claire L. Donnat, Juergen Dukart, Niall W. Duncan, Joke Durnez, Amr Eed, Simon B. Eickhoff, Andrew Erhart, Laura Fontanesi, G. Matthew Fricke, Shiguang Fu, Adriana Galván, Remi Gau, Sarah Genon, Tristan Glatard, Enrico Glerean, Jelle J. Goeman, Sergej A. E. Golowin, Carlos González-García, Krzysztof J. Gorgolewski, Cheryl L. Grady, Mikella A. Green, João F. Guassi Moreira, Olivia Guest, Shabnam Hakimi, J. Paul Hamilton, Roeland Hancock, Giacomo Handjaras, Bronson B. Harry, Colin Hawco, Peer Herholz, Gabrielle Herman, Stephan Heunis, Felix Hoffstaedter, Jeremy Hogeveen, Susan Holmes, Chuan-Peng Hu, Scott A. Huettel, Matthew E. Hughes, Vittorio Iacovella, Alexandru D. Iordan, Peder M. Isager, Ayse I. Isik, Andrew Jahn, Matthew R. Johnson, Tom Johnstone, Michael J. E. Joseph, Anthony C. Juliano, Joseph W. Kable, Michalis Kassinopoulos, Cemal Koba, Xiang-Zhen Kong, Timothy R. Koscik, Nuri Erkut Kucukboyaci, Brice A. Kuhl, Sebastian Kupek, An-

gela R. Laird, Claus Lamm, Robert Langner, Nina Lauharatanahirun, Hongmi Lee, Sangil Lee, Alexander Leemans, Andrea Leo, Elise Lesage, Flora Li, Monica Y. C. Li, Phui Cheng Lim, Evan N. Lintz, Schuyler W. Liphardt, Annabel B. Losecaat Vermeer, Bradley C. Love, Michael L. Mack, Norberto Malpica, Theo Marins, Camille Maumet, Kelsey McDonald, Joseph T. McGuire, Helena Melero, Adriana S. Méndez Leal, Benjamin Meyer, Kristin N. Meyer, Glad Mihai, Georgios D. Mitsis, Jorge Moll, Dylan M. Nielson, Gustav Nilsonne, Michael P. Notter, Emanuele Olivetti, Adrian I. Onicas, Paolo Papale, Kaustubh R. Patil, Jonathan E. Peelle, Alexandre Pérez, Doris Pischedda, Jean-Baptiste Poline, Yanina Prystauka, Shruti Ray, Patricia A. Reuter-Lorenz, Richard C. Reynolds, Emiliano Ricciardi, Jenny R. Rieck, Anais M. Rodriguez-Thompson, Anthony Romyn, Taylor Salo, Gregory R. Samanez-Larkin, Emilio Sanz-Morales, Margaret L. Schlichting, Douglas H. Schultz, Qiang Shen, Margaret A. Sheridan, Jennifer A. Silvers, Kenny Skagerlund, Alec Smith, David V. Smith, Peter Sokol-Hessner, Simon R. Steinkamp, Sarah M. Tashjian, Bertrand Thirion, John N. Thorp, Gustav Tinghög, Loreen Tisdall, Steven H. Tompson, Claudio Toro-Serey, Juan Jesus Torre Tresols, Leonardo Tozzi, Vuong Truong, Luca Turella, Anna E. van 't Veer, Tom Verguts, Jean M. Vettel, Sagana Vijayarajah, Khoi Vo, Matthew B. Wall, Wouter D. Weeda, Susanne Weis, David J. White, David Wisniewski, Alba Xifra-Porxas, Emily A. Yearling, Sangsuk Yoon, Rui Yuan, Kenneth S. L. Yuen, Lei Zhang, Xu Zhang, Joshua E. Zosky, Thomas E. Nichols, Russell A. Poldrack, and Tom Schonberg. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):84–88, jun 2020. ISSN 0028-0836. doi: 10.1038/s41586-020-2314-9. URL http://www.nature.com/articles/s41586-020-2314-9.

Stephen Strother, Stephen La Conte, Lars Kai Hansen, Jon Anderson, Jin Zhang, Sujit Pulapura, and David Rottenberg. Optimizing the fmri data-processing pipeline using prediction and reproducibility performance metrics: I. a preliminary group analysis. *Neuroimage*, 23:S196–S207, 2004.

Ed HBM Gronenschild, Petra Habets, Heidi IL Jacobs, Ron Mengelers, Nico Rozendaal, Jim Van Os, and Machteld Marcelis. The effects of freesurfer version, workstation type, and macintosh operating system version on anatomical volume and cortical thickness measurements. *PloS one*, 7(6):e38234, 2012.

Tristan Glatard, Lindsay B Lewis, Rafael Ferreira da Silva, Reza Adalat, Natacha Beck, Claude Lepage, Pierre Rioux, Marc-Etienne Rousseau, Tarek Sherif, Ewa Deelman, et al. Reproducibility of neuroimaging analyses across operating systems. *Frontiers in neuroinformatics*, 9:12, 2015.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11 (5):702–712, 2016. doi: 10.1177/1745691616658637. URL https://doi.org/10.1177/1745691616658637. PMID: 27694465.

Sharon-lise T Normand. Tutorial in Biostatistics Meta-Analysis: Formulating, Evaluating, Combining, and Reporting. *Statistics in medicine*, 18:321–359, 1999.

Samuel A Stouffer, Edward A Suchman, Leland C Devinney, Shirley A Star, and Robin M Williams Jr. *The American soldier: Adjustment during army life. (Studies in social psychology in World War II), Vol. 1.* Princeton Univ. Press, Oxford, England, 1949.

SM Tom, CR Fox, C Trepel, and RA Poldrack. The neural basis of loss aversion in decision-making under risk. *Science*, 2007. doi: 10.1126/science.1134239. URL https://pubmed.ncbi.nlm.nih.gov/17255512/.

Nicola Canessa, Chiara Crespi, Matteo Motterlini, Gabriel Baud-Bovy, Gabriele Chierchia, Giuseppe Pantaleo, Marco Tettamanti, and Stefano F Cappa. The functional and structural neural basis of individual differences in loss aversion. *Journal of Neuroscience*, 33(36):14307–14317, 2013.

David C Van Essen, Kamil Ugurbil, Edward Auerbach, Deanna Barch, Timothy EJ Behrens, Richard Bucholz, Acer Chang, Liyong Chen, Maurizio Corbetta, Sandra W

Curtiss, et al. The human connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.

Elodie Germani, Elisa Fromont, Pierre Maurel, and Camille Maumet. The hcp multi-pipeline dataset: an opportunity to investigate analytical variability in fmri data analysis. *arXiv preprint arXiv:2312.14493*, 2023.

Verne S Caviness Jr, James Meyer, Nikos Makris, and David N Kennedy. Mri-based topographic parcellation of human neocortex: an anatomically specified method with estimate of reliability. *Journal of cognitive neuroscience*, 8(6):566–587, 1996.

J Rademacher, AM Galaburda, DN Kennedy, PA Filipek, and VS Caviness Jr. Human cerebral cortex: localization, parcellation, and morphometry with magnetic resonance imaging. *Journal of cognitive neuroscience*, 4(4):352–374, 1992.

Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

6 equations for double check

$$Z_j^{\mathrm{S}} = \frac{\bar{Y}_j}{\sqrt{1/K}}$$

$$Z_j^{\mathrm{SS}} = \frac{\bar{Y}_j}{\sqrt{\mathbf{1}_k^\top \mathbf{Q} \mathbf{1}_k / K^2}}.$$

$$\mu_{\mathrm{C}} = \frac{1}{K} \sum_k \langle \mathbf{Y_k} \rangle$$

$$\sigma_{\mathrm{C}}^2 = \frac{1}{K} \sum_k \langle\langle \mathbf{Y_k} \rangle\rangle$$

$$Z_j^{CSS} = Z_j^{\mathrm{SS}} - \langle Z^{\mathrm{SS}} \rangle + \mu_C = \frac{\bar{Y}_j - \mu_{\mathrm{C}}}{\sqrt{\mathbf{1}^\top \mathbf{Q} \mathbf{1} / K^2}} + \mu_{\mathrm{C}}$$

$$Z_j^{CA} = \frac{\bar{Y}_j - \langle \bar{Y} \rangle}{\sqrt{\langle\langle \bar{Y} \rangle\rangle}} \sigma_C + \mu_C = \frac{\bar{Y}_j - \mu_{\bar{Y}}}{\sigma_{\bar{Y}}} \sigma_{\mathrm{C}} + \mu_{\mathrm{C}}$$

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

$$Z_j^{CGS} = Z_j^{\mathrm{SG}} - \langle Z^{\mathrm{SG}} \rangle + \mu_C = \frac{\bar{Y}_j^G - \mu_{\mathrm{C}}}{\sqrt{(\mathbf{1}^\top \mathbf{Q}^{-1} \mathbf{1})^{-1}}} + \mu_{\mathrm{C}}$$

NEW:

$$Z_j^{CGS} = \frac{Z_j^{\mathrm{SG}} - \langle Z^{\mathrm{SG}} \rangle}{\sqrt{\langle\langle Z^{\mathrm{SG}} \rangle\rangle}} * \sigma_{\mathrm{C}} + \mu_C$$

with

$$\langle Z^{\mathrm{SG}} \rangle = \frac{1}{J} \sum_j Z_j^{\mathrm{SG}}$$

and

$$\langle\langle Z^{\mathrm{SG}} \rangle\rangle == \frac{1}{J} \sum_j \langle\langle Z_j^{\mathrm{SG}} \rangle\rangle$$

with

$$\langle\langle Z_j^{\mathrm{SG}} \rangle\rangle = (J-1)^{-1} \sum_j (Z_j^{\mathrm{SG}} - \langle Z^{\mathrm{SG}} \rangle)^2$$