



HAL
open science

On the validity of fMRI mega-analyses using data processed with different pipelines

Elodie Germani, Xavier Rolland, Pierre Maurel, Camille Maumet

► To cite this version:

Elodie Germani, Xavier Rolland, Pierre Maurel, Camille Maumet. On the validity of fMRI mega-analyses using data processed with different pipelines. 2025. inserm-04466478v3

HAL Id: inserm-04466478

<https://inserm.hal.science/inserm-04466478v3>

Preprint submitted on 16 Jan 2025


HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.





Distributed under a Creative Commons Attribution 4.0 International License

ON THE VALIDITY OF fMRI MEGA-ANALYSES USING DATA PROCESSED WITH DIFFERENT PIPELINES

Elodie Germani *
Univ Rennes, Inria, CNRS, Inserm
Rennes, France
elodie.germani@irisa.fr
 ORCID: 0000-0002-5786-9538

Xavier Rolland *
Univ Rennes, Inria, CNRS, Inserm
Rennes, France
xavier.rolland@ac-grenoble.fr

Pierre Maurel †
Univ Rennes, Inria, CNRS, Inserm
Rennes, France
 ORCID: 0000-0003-2539-7414

Camille Maumet ‡
Univ Rennes, Inria, CNRS, Inserm
Rennes, France
camille.maumet@inria.fr
 ORCID: 0000-0002-6290-553X

ABSTRACT

In neuroimaging and functional Magnetic Resonance Imaging (fMRI), many derived data are made openly available in public databases. These can be re-used to increase sample sizes in studies and thus, improve robustness. In fMRI studies, raw data are first preprocessed using a given analysis pipeline to obtain subject-level contrast maps, which are then combined into a group analysis. Typically, the subject-level analysis pipeline is identical for all participants. However, derived data shared on public databases often come from different workflows, which can lead to different results. Here, we investigate how this analytical variability, if not accounted for, can induce false positive detections in mega-analyses combining subject-level contrast maps processed with different pipelines. We use the HCP multi-pipeline dataset, containing contrast maps for N=1,080 participants of the HCP Young-Adult dataset, whose raw data were processed and analyzed with 24 different pipelines. We performed between-groups analyses with contrast maps from different pipelines in each group and estimated the rates of pipeline-induced detections. We show that, if not accounted for, analytical variability can lead to inflated false positive rates in studies combining data from different pipelines.

Keywords neuroimaging, analytical variability, pipelines, validity, data re-use

1 Introduction

Over the past few years, concerns have been raised regarding the lack of reproducibility of neuroimaging findings [1, 2, 3]. In particular, the low statistical power of studies was criticized, as effectively leading to low probabilities of identifying true effects but also to high probabilities of reporting false positive findings in the literature [1]. Researchers proposed different approaches to increase sample sizes, and thus statistical power, for instance with the development of large-scale studies [4, 5]. However, acquiring such an amount of data is costly and due to the challenge of finding participants, these studies often contain a few number of data per participant. In functional Magnetic Resonance Imaging (fMRI), a brain imaging technique in which brain activity is studied under different conditions, these datasets cover a limited subset of brain functions, limiting the flexibility of research questions that can be explored. A potential solution to increase sample size while avoiding these challenges is to reuse the data already acquired in other studies into meta- or mega-analyses [6].

* Co-first authors

† Joint senior authorship

‡ Corresponding author: camille.maumet@inria.fr

With the increased adoption of open science practices [7, 8, 9] and the development of dedicated research infrastructures [10, 11, 12], such as NeuroVault [10], OpenNeuro [11], more and more neuroimaging data from various studies have been made available to the scientific community. This includes raw data at the subject level, that can be re-analyzed using the same processing steps and combined in a mega-analysis, but also derived data (*i.e.* already processed) at the subject or group level. At the group level, derived data can be used in meta-analyses to build consensus results across multiple studies [6], but there are several limitations to this method due to publication bias [13].

At the subject level, individual contrast maps (after the subject-level processing) from different studies can be combined using mega-analyses (also known as *individual patient data (IPD) meta-analysis*). Their reuse is more optimal compared to raw data, not only because sharing of statistic maps is easier due to reduced privacy requirements, but also because it avoids having to perform costly re-computations. Indeed, fMRI studies require multiple processing steps on the data, both at the subject level (preprocessing of the raw fMRI data to prepare them for statistical analysis, and first-level analysis for each participant) and at the group level (second-level statistical analysis using the subject-level contrast maps resulting from first-level analysis). However, it is unlikely that derived data available on public databases come from the same pipeline. In addition, derived fMRI datasets can come from adaptable pipelines that apply different processing steps depending on which data is available (see for example fMRIPrep [14]). In practice, in those mega-analyses, confounds (such as differing pipelines) are typically accounted for by adding a nuisance covariate to the model [15]. This approach is useful to remove any detections that are induced by those confounds yet when it comes to the pipeline it is sometimes not straightforward to identify which aspect of the pipeline constitutes a nuisance factor and should be modeled.

Multiple studies have shown that different implementations of a processing pipeline can lead to different results in neuroimaging. These changes can arise from different levels of variations: different software packages [16] or software packages version [17], different algorithms and processing steps [18, 19, 20], different software environment [21], etc. In [19], 70 teams analyzed the same task-fMRI dataset, each with their usual pipeline, leading to 70 different analytical conditions. They found substantial differences in the results obtained across teams, in terms of statistic maps but also answers to binary hypotheses. This variability resulting from the processing and analysis protocol used on the data is also known as *analytical variability*.

Here, we systematically investigate how analytical variability impacts the results of fMRI mega-analyses with data from different pipelines (when no additional measures are taken to account for that variability). Previous studies have focused on how analytical variability affects the reproducibility of existing results in neuroimaging, by using different pipelines to complete a similar analysis in which the processing applied on all subject data is the same and comparing the results obtained across pipelines using different processing pipelines. In addition, dedicated frameworks for optimizing the choice of pipelines have been proposed based on an estimation of reproducibility performance [22, 23]. Notably, solutions to use different subject-level processing pipelines have been suggested in this context [24].

We explore the impact of pipeline-based differences on the results of between-group analyses that compare populations whose data were processed differently at the subject level. While, currently, this is not common practice, this setup could be useful in order to compare healthy versus pathological populations using processed data from different datasets (for example data of healthy participants from the minimally processed dataset of the Human Connectome Project (HCP) compared to patient data from another dataset). We carry out a series of between-groups analyses, with each group corresponding to subject-level contrast maps randomly sampled from the Human Connectome Project (HCP) Young Adult dataset [5] and processed with different pipelines (1 pipeline in each group). Since participants in all groups are sampled from the same dataset, we expect no population-based differences, and therefore all observed differences are attributable to the differing pipelines.

2 Material and Methods

The goal of this study is to test the validity of between-group analyses using subject-level contrast maps processed with different pipelines (when differences in pipelines are not accounted for in the statistical model). In the following sections, the term “pipeline” is used to refer to the subject-level pipeline.

The steps performed to estimate this validity are presented in Figure 1. First, we randomly sampled subject-level contrast maps processed through different pipelines from the HCP multi-pipeline dataset [25] (see section 2.1). Then, for each pair of pipelines, we performed a between-group analysis (see section 2.2). This group comparison was repeated 1,000 times in order to estimate the empirical rate of (pipeline-induced) significant differences. In the following, we denote this as the “false positive rate”, considering that those may be equivalently seen as false detections in a between-group analysis using a simple statistical model in which differences in pipelines are not accounted for.

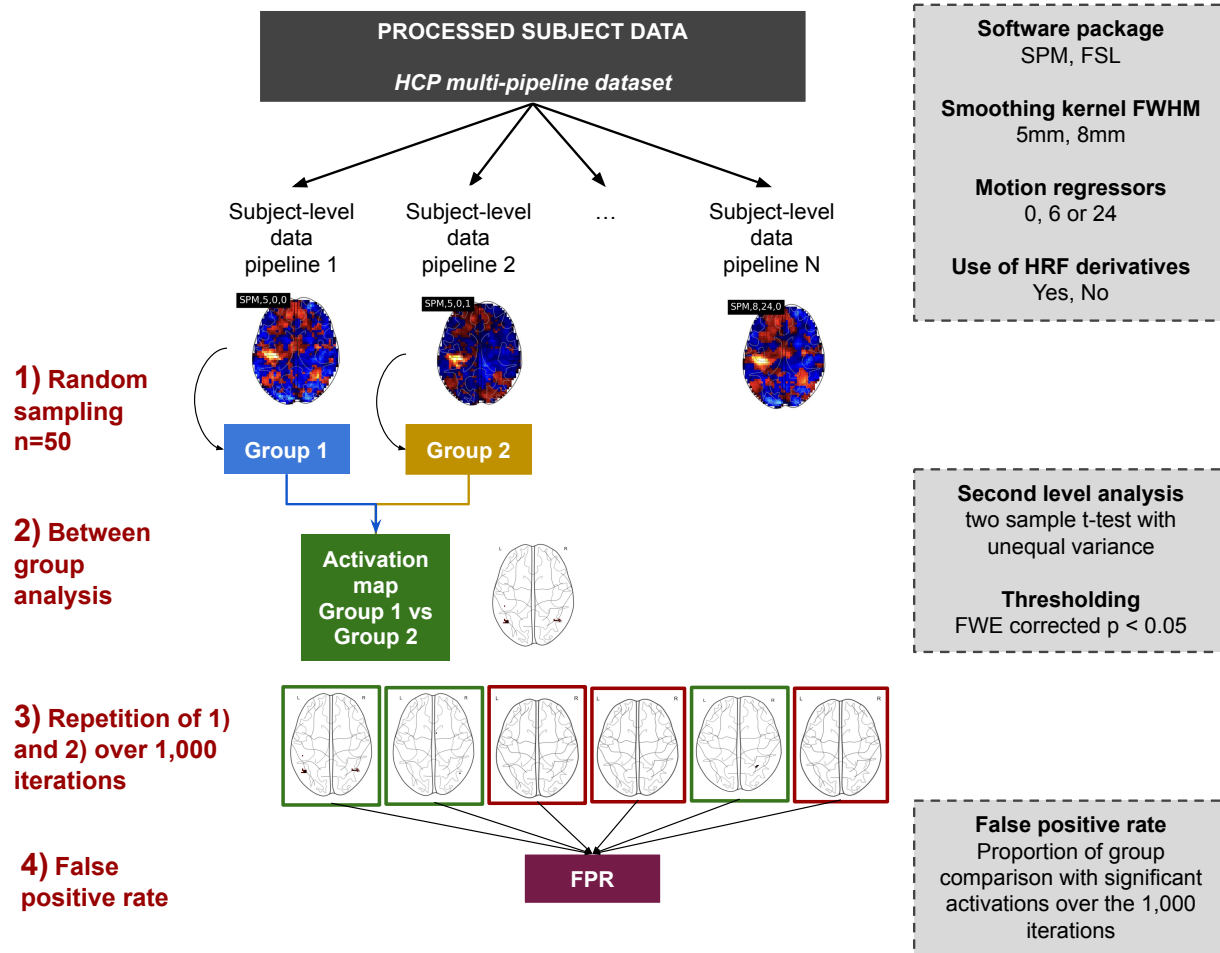


Figure 1: Overview of the method: 1) sampling of $n=50$ subject-level contrast maps for each group (i.e. one group = one pipeline) from the HCP multi-pipeline, 2) between-group analyses “Group 1 > or < Group 2”, 3) running 1,000 iterations of step 1 and step 2, and 4) estimation of the empirical rate of pipeline-induced significant differences referred to as the *false positive rate*.

All the scripts used to perform the study (group-level analysis and false positive rate estimation) are available on Software Heritage [26]: `swh:1:snp:585d3a0a3388a928ab3c6211c1826702aa618190`.

2.1 HCP multi-pipeline

This study was performed using derived data from the HCP Young Adult [5]. 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 Young Adult Open Access Data Use Terms available at: [27].

Subject-level contrast and statistic maps from 1,080 subjects of the HCP Young Adult S1200 release [5] for the motor task were obtained with 24 different pipelines. In brief, the pipelines implemented in the dataset varied on the following set of parameters:

- Software package: SPM (Statistical Parametric Mapping, RRID: SCR_007037) [28] or FSL (FMRIB Software Library, RRID: SCR_002823) [29].
- Smoothing kernel: Full-Width at Half-Maximum (FWHM) of 5 mm or 8 mm.
- Number of motion regressors included 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).

- Presence (1) or absence (0) of the derivatives of the Hemodynamic Response Function (HRF) in the first-level GLM. The temporal derivative was added in FSL and both the temporal and dispersion derivatives in SPM.

These variations were chosen in particular due to the lack of consensus in the research community on their selection. In a many-analyst study [19], 70 teams analyzed the same task-fMRI dataset using their usual pipeline, and final statistic maps were compared to explore potential differences between results across pipelines and the source of these differences. A complementary evaluation of the source of pipeline-based differences was also performed across 3 reproductions of published fMRI studies [30]. Within the varying parameters, smoothing kernel size, number of motion regressors, and design of the statistical analysis were all identified as impactful. Since there is a lack of consensus on the values of these parameters, they can easily vary from one pipeline to another. Therefore shared derived data are likely obtained with pipelines for which these parameters are different.

In total, this led to 24 different subject-level pipelines (2 software packages \times 2 smoothing kernels \times 3 numbers of motion regressors \times 2 HRF). Together those contrast and statistic maps are referred to as the *HCP multi-pipeline dataset*. More details about the analysis and its implementation can be found in [25]. Briefly, the preprocessing steps included motion correction, coregistration of the functional images onto structural data, segmentation, registration into the MNI space, and smoothing. For subject-level statistical analysis, both SPM and FSL used a GLM to model experimental tasks, convolving event data with a Double Gamma HRF.

2.2 Between-group analyses

In this study, we explored the false positive rates induced by pipeline differences of between-group studies with subject-level contrast maps from different pipelines in three settings: within-pipeline (baseline), within-software (*i.e.* pipeline implemented in the same software package with different parameters) and between-software (*i.e.* pipeline implemented in different software packages with similar parameters).

2.2.1 Contrast post-processing

As FSL and SPM use different MNI templates [31] (*i.e.* MNI152Nlin6Sym for FSL, IXI549Space for SPM), subject-level contrast maps from different software packages had the same resolution (2 mm) but different dimensions. We used the following post-processing to harmonize the dimensions of the images. We used Nilearn [32] (RRID: SCR_001362) to resample all subject-level contrast maps to the dimensions of the MNI152Asym2009 brain template with a 2 mm resolution using third-order spline interpolation (continuous interpolation in Nilearn, and the default parameter of the resampling function in the library). We masked the contrast maps using the intersection of all subject-level brain masks (all pipelines).

FSL and SPM contrast maps are also scaled differently (see [33]). In both software packages, contrast maps are theoretically expressed in percent BOLD change but there are important differences in how this percent BOLD change is computed that effectively lead to scaling differences. Hence, in SPM, contrast maps units are closer to 2.5 times percent BOLD change due to the mask used to compute the global in-brain mean intensity. On the other hand, FSL contrast maps are scaled to 10,000 (*i.e.* 100 times percent BOLD change). We applied a factor to each contrast map to make them closer to percent BOLD change. Contrast maps in SPM and FSL were therefore rescaled by multiplying by $100/250 = 0.4$ and $100/10,000 = 0.01$ respectively.

All between-group analyses were performed on resampled, masked, and re-scaled subject-level contrast maps. As a sanity check, we also computed the between-group same-pipeline analyses on the original contrast maps (*i.e.* before post-processing). As expected, the estimated false positive rates were consistent with the results obtained on post-processed data (see Supplementary Table 1).

2.2.2 Analysis setup

For each between-group analysis, we randomly sampled 100 participants without replacement among the full set of 1,080 participants and split them into two groups ($N = 50$ in each group). This sample size is larger than typical sample size in fMRI studies (around 30 participants) [34]), but limiting atypical behaviors induced by small sample sizes. In each group, subject-level contrast maps were obtained using a different pipeline. This process was repeated for different groups and pairs of pipelines. We performed a one-tailed two-sample t-test with unequal variance and computed the statistic maps associated with H_0 : “no mean difference of activation between groups”. We used a voxelwise $p < 0.05$ FWE-corrected with Random Field Theory [35, 36], with approximately 130,000 comparisons (or 300-1000 resels, *i.e.*

independent comparisons) per between-group analysis. All between-group analyses were performed in SPM in order to have consistent conditions in all the second-level analyses.

2.3 False positive rates

For a given pair of pipelines, the between-group analysis was repeated 1,000 times with different sets of participants. The empirical false positive rate was estimated as the proportion of between-group analyses, across the repetitions (see section 2) with at least one significant detection (see Figure 1).

If the rate exceeds the nominal α -level of 0.05 (95% confidence interval [0.037; 0.064]), we can conclude that pipeline-based differences impact the validity of results (towards invalidity) and if the rate is lower we can conclude that the analysis is conservative. Of note, we use throughout the manuscript the term *validity* (resp. *invalidity*) only in relation to type I error (i.e. specificity) as per the definition of this term in Statistics. Measuring the effect of pipeline differences on type II error (i.e. sensitivity) is beyond the scope of the current manuscript.

2.4 Statistical distributions and P-P plots

P-P plots are usually used to observe how a given set of statistical values diverge from an expected distribution by plotting, for each k^{th} ordered statistical value, the expected associated p -value on the x-axis and the obtained p -value on the y-axis. Here, under the null hypothesis, p -values were expected to follow a uniform distribution $U(0, 1)$. Thus, for a set of N statistical values, the k^{th} ordered p -value follows a Beta distribution $\mathcal{B}(k, N - k + 1)$ with expected value $k/(N + 1)$ [37, 38]. Confidence bounds on the p -values were computed using the Beta distribution.

Here, we used a Bland-Altman [39] variant of P-P plots. Bland-Altman plots provide a visual representation of the difference between two measurements on the y-axis and the average of the two measurements on the x-axis. Here we adapted those plots to p -values, as follows:

- on the x-axis: the expected p -value in $-\log_{10}$
- on the y-axis: the difference between the $-\log_{10}$ obtained and the $-\log_{10}$ expected p -values.

This update made it easier to observe the behavior in the tails of the p -value distribution (which is of interest here). High statistical values (right tail of our sample) are associated to low p -values, *i.e.* to high $-\log_{10}$ p -values. We also looked at the distributions of the statistical values for multiple between-group analyses, and compared them with a Student distribution \mathcal{T}_{98} .

3 Results

3.1 Analyses using the same pipeline (baseline)

Table 1 shows the false positive rates obtained for all analyses with the same pipeline in both groups, separately for SPM and FSL. For all combinations, the false positive rates were below the expected value of 0.05, ranging between 0.012 and 0.028 for SPM and between 0.013 and 0.024 for FSL. These results, obtained with the same pipeline in both groups, are used as a baseline in the following.

3.2 Analyses using pipelines with different parameters

The following subsections present the results obtained with pipelines using different set of parameters (within software). In each case, we looked at the false positive rate (Figure 2), the statistical distributions (Supplementary Figures 3 and 11) and the associated P-P plots (Figure 4, Figure 5 and Supplementary Figures 6 and 9). To present the results, we chose a default value for each studied parameter – smoothing 5 mm FWHM, HRF with derivatives, and 24 motion regressors – and compared our results to those obtained with the default.

3.2.1 Different HRF

Adding derivatives to model the HRF was the most impactful of all three varying factors in both software packages. The false positive rates obtained with different HRF (*i.e.* *canonical HRF* > or < *HRF with derivatives*) are presented in Figure 2 (A) for the six analyses performed (*i.e.* with varying levels of smoothing and numbers of motion regressors – with the same setting in both pipelines).

SPM

	Smooth 5mm		Smooth 8mm	
	No derivatives	Derivatives	No derivatives	Derivatives
0 motion regressors	0.012	0.013	0.016	0.023
6 motion regressors	0.015	0.006	0.024	0.013
24 motion regressors	0.023	0.016	0.025	0.028

FSL

	Smooth 5mm		Smooth 8mm	
	No derivatives	Derivatives	No derivatives	Derivatives
0 motion regressors	0.014	0.013	0.015	0.023
6 motion regressors	0.018	0.014	0.018	0.018
24 motion regressors	0.015	0.013	0.016	0.024

Table 1: False positive rates for between-groups analyses with the same pipeline in both groups, with SPM and FSL and for all possible sets of parameters (number of motion regressors, smoothing kernel FWHM, and presence or absence of HRF temporal derivatives). The rates were always under 0.05.

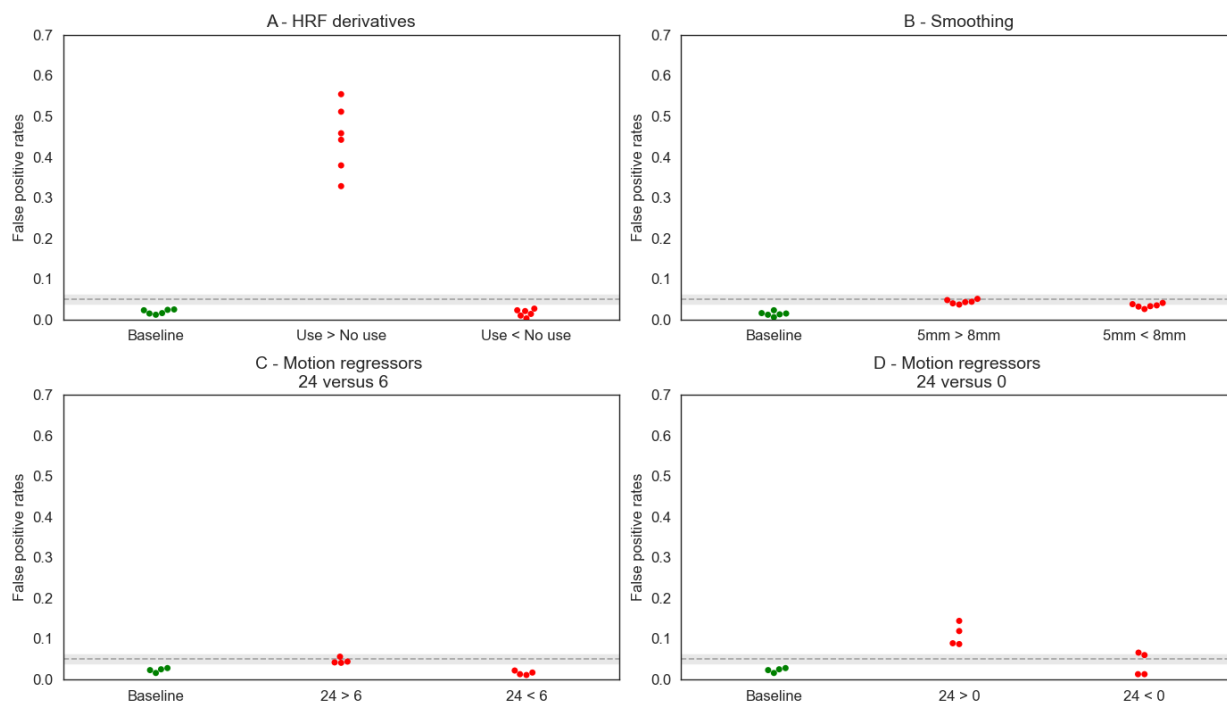


Figure 2: False positive rates for pipelines with a single differing parameter in SPM: A) HRF derivatives, B) smoothing and C and D) motion regressors. For each, we provide the false positive rates obtained for: 1/ Default > Variation and Default < Variation (Red) and 2/ baseline analysis with default parameters, used as a reference (Green, first column). The grey dashed line corresponds to the alpha level (0.05), and the grey band to the corresponding confidence interval at 95%.

In SPM, the comparison *canonical HRF* > *HRF with derivatives* (Figure 2 - A) showed invalid false positive rates (above the theoretical 0.05 threshold) for all pipeline combinations. Similarly, in FSL, all combinations gave invalid results for

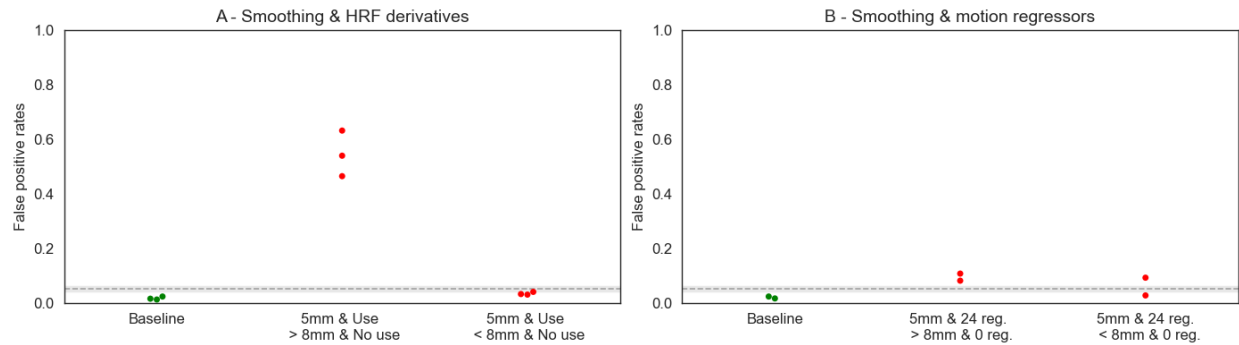


Figure 3: False positive rates for pipelines with two differing parameters in SPM: A) Smoothing and HRF, B) Smoothing and motion regressors. For each studied parameter, we provide the rates obtained for: 1/ Default $>$ Variation and Default $<$ Variation (Red) and 2/ baseline analysis with default parameters, used as a reference (Green, first column). The grey dashed line corresponds to the alpha level (0.05) and the grey band to the corresponding confidence interval at 95%.

this same comparison except two combinations: 5 mm or 8 mm smoothing $FWHM$ and 24 motion regressors. These two analyses led to values that were within the confidence bounds of the 0.05 threshold or slightly conservative (0.032 and 0.061 respectively). For the opposite comparison (*i.e.* canonical $HRF <$ HRF with derivatives) all combinations resulted in valid results with false positive rates under 0.05.

Figures 4 and Supplementary Figure 6 show the corresponding Bland-Altman P-P plots for comparisons with different HRF and otherwise default parameters. In both software packages, consistently with what we observed for the false positive rates, the comparison *canonical HRF > HRF with derivatives* led to values that were outside of the 95% confidence interval (grey area). In SPM, values were further away from the 95% confidence interval than in FSL.

The same observations could be made on the statistical distributions for both SPM and FSL (Supplementary Figures 3 and 11): both showed a shift in mean and variance, but this was smaller for FSL. The combination of pipeline parameters used in this Figure (*i.e.* pipelines with 5 mm $FWHM$ and 24 motion regressors, with different HRF derivatives) showed nearly valid false positive rates, as stated in the previous paragraph (see Figure 2), which could explain why the shift seemed smaller in FSL compared to SPM. We also observed the P-P plots for a different combination of FSL pipelines with other parameters (5 mm , 0 motion regressors) in Supplementary Figure 8 and found a similar shift as the one observed for SPM.

3.2.2 Different smoothing

The false positive rates obtained with different levels of smoothing (5 mm or 8 mm) in the pipelines are presented in Figure 2 (B) for the six analyses performed (*i.e.* with varying HRF models and number of motion regressors – with the same setting in both pipelines).

The false positive rates obtained with different levels of smoothing (5 mm $>$ or $<$ 8 mm) in the pipelines were above the 0.05 theoretical rate in FSL (ranging from 0.07 to 0.16) and within the confidence interval around the theoretical rate in SPM (ranging from 0.03 to 0.05). Compared to the baseline analyses using the same pipelines, the false positive rates were always inflated and were slightly higher for the tail $5\text{ mm} > 8\text{ mm}$.

The Bland-Altman P-P plots (Figure 4 and Supplementary Figure 6) are consistent with the observations made on the false positive rates. Between-group analyses using pipelines with different smoothing gave results outside of the 95% confidence interval in FSL and within the interval in SPM, with only a small positive difference in the direction $5\text{ mm} > 8\text{ mm}$.

The behaviors observed on the P-P plots can be explained by the positive shift in mean values and standard deviations observed on the statistical distribution for $5\text{ mm} > 8\text{ mm}$ for FSL (Supplementary Figure 11), which is less pronounced for SPM (Supplementary Figure 3).

3.2.3 Different number of motion regressors

The false positive rates obtained with different numbers of motion regressors (0, 6, and 24) are presented in Figure 2 (C - D) for the six analyses performed (*i.e.* with varying levels of smoothing and different HRF – with the same setting in

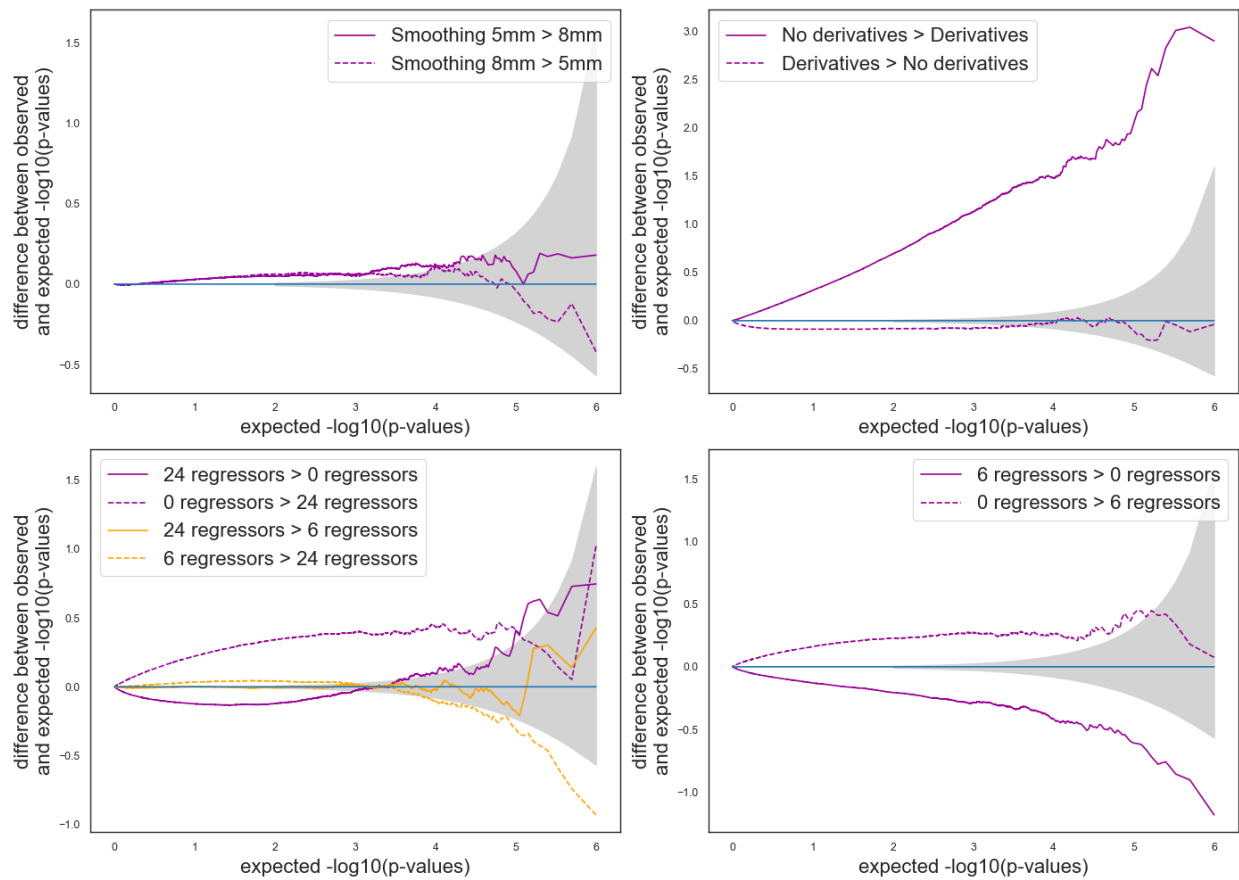


Figure 4: Bland-Altman P-P plots for pipelines with a single differing parameter in SPM. The grey shade corresponds to the 0.95 confidence interval. A curve above (respectively below) the confidence interval indicates invalidity (respectively conservativeness). Default parameters: 5 mm smoothing, 24 motion regressors and no HRF derivatives.

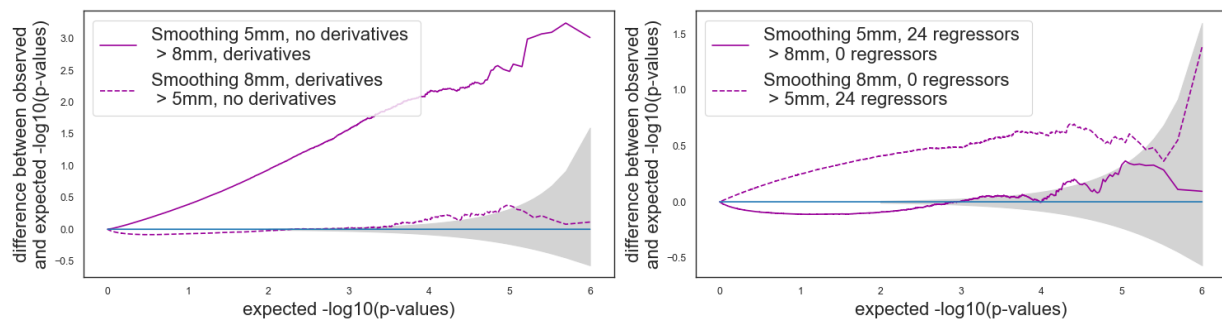


Figure 5: Bland-Altman P-P plots for pipelines with two differing parameters in SPM. The grey shade corresponds to the 0.95 confidence interval. A curve above (respectively below) the confidence interval indicates invalidity (respectively conservativeness). Default parameters: 5 mm smoothing, 24 motion regressors and no HRF derivatives.

both pipelines). We studied the combinations *24 motion regressors* > or < *6 motion regressors* (third column) and *24 motion regressors* > or < *0 motion regressors* (fourth column).

In SPM, false positive rates were below the 0.05 theoretical rate for all comparisons of *24 motion regressors* > or < *6 motion regressors*. For the comparison with no motion regressors, the false positive rates were higher and above 0.05 for *24 motion regressors* > *0 motion regressors* and slightly below for the opposite. In FSL, the results were dependent on the other pipeline parameters. All combinations led to invalid results (*i.e.* above the theoretical 0.05 threshold) except for *24 motion regressors* > *0/6 motion regressors* when using the canonical HRF (*i.e.* no HRF derivatives) in both pipelines.

In Section 3.2.1), we showed that all combinations of pipelines with varying HRF models led to invalid results except those with *5 mm* or *8 mm smoothing* and *24 motion regressors*. Here, we also observe invalid results for all combinations of pipelines with *24 motion regressors* > or < *0/6 motion regressors*, except those with *5 mm* or *8 mm smoothing* and *no HRF derivatives*. We can suppose that in FSL, when using *24 motion regressors*, the use of HRF derivatives in the GLM has a low impact on the results and similarly, when using the *canonical HRF*, using *0, 6 or 24 motion regressors* does not change the results much, and thus has a low impact on the validity of the mega-analyses combining subject-level data obtained from pipelines with different parameters.

In the Bland-Altman P-P plot for SPM (Figure 4), we observed more extreme values in the P-P plots for the comparisons “*24 motions regressors* > or < *0 motion regressors*” than for those of “*24 motions regressors* > or < *6 motion regressors*”, which is consistent with our observations on the false positive rates. The Bland-Altman P-P plot (Supplementary Figure 6) for FSL with *5 mm smoothing* and an HRF with derivatives, the comparison *24 motion regressors* > or < *0 motion regressors* were consistent with the invalid false positive rates found with such parameters: we found conservative results for the comparison *24 motion regressors* > *0 motion regressors* (plain line) and invalid results in the opposite direction (dashed line).

Statistical distributions (Supplementary Figures 3 and 11) also show a shift in mean and variance for the comparison “*24 motion regressors* > or < *0 motion regressors*”, for both SPM and FSL. This shift is not as important for the comparison “*24 motion regressors* > or < *6 motion regressors*”. The comparison “*6 motion regressors* > or < *0 motion regressors*” was also showed for comparison, and showed similar results as the “*24 motion regressors* > or < *0 motion regressors*” comparison.

3.2.4 Combined effects of parameters

We observed the combined effects of:

- differences in smoothing and in HRF model
- differences in smoothing and in motion regressors

The false positive rates obtained with different smoothing and different HRF models or different motion regressors in the pipelines are presented in Figure 3 for the different analyses performed.

In both SPM and FSL, the first set of between-group analyses (*5 mm smoothing, canonical HRF*) > (*8 mm smoothing, HRF with derivatives*) led to invalid results with false positive rates largely above the 0.05 theoretical threshold (around 0.60). The opposite test provided conservative results.

In SPM, the results for (*5 mm smoothing, canonical HRF*) > (*8 mm smoothing, HRF with derivatives*) were close to those obtained for the analyses with a single varying parameter *canonical HRF* > *HRF with derivatives* (from 0.46 to 0.63 in the combined effect analysis and from 0.32 to 0.52 in the exploration of HRF derivatives effect only, see Figure 2 - A). In the isolated analyses, the effect of changing the smoothing kernel FWHM was not very important in SPM (“*5mm vs 8mm smoothing kernel FWHM*”), which might explain why the false positive rates did not increase much in the combined effect analyses.

Under FSL, the previous analyses on the effect of each of these parameters separately (changing smoothing kernel FWHM and changing HRF model separately) both gave inflated false positive rates, and their combined effect largely increased the false positive rates (up to 0.77) compared to the effect of changing the use HRF derivatives alone (up to 0.49).

Similar observations can be made on the P-P plots on Figure 5 and Supplementary Figure 9.

In both SPM and FSL, the second set of analyses (*5 mm smoothing, 24 motion regressors*) > or < (*8 mm smoothing, 0 motion regressors*), we found invalid results for nearly all combinations. In SPM, the false positive rates were only slightly above the theoretical threshold of 0.05 (0.081 and 0.11), which is consistent with our previous observation:

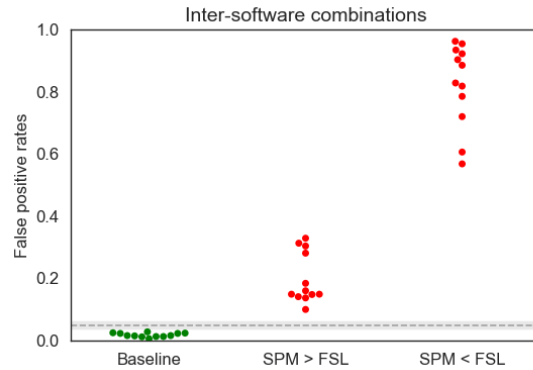


Figure 6: False positive rates for pipelines with different software packages. We provide the false positive rates obtained for: “FSL > SPM” and “FSL < SPM” (Red), and for the corresponding analyses within-pipelines, *i.e.* the baseline (Green, first column). The grey dashed line corresponds to the alpha level (0.05) and grey band to the corresponding confidence interval at 95%..

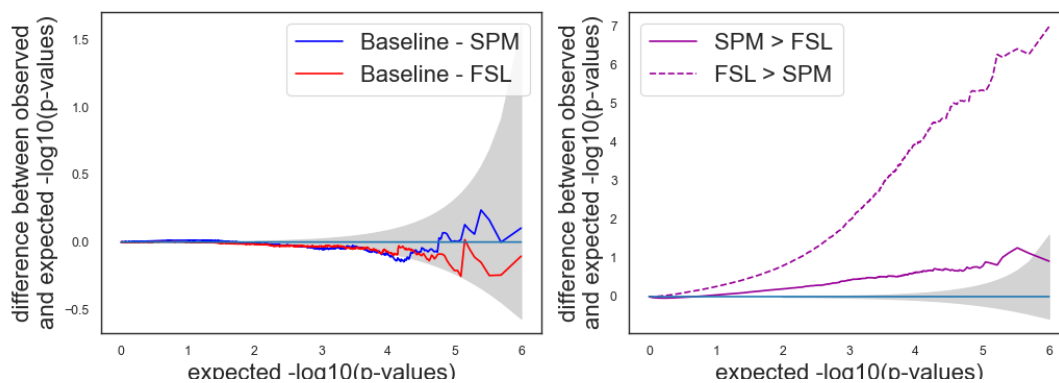


Figure 7: Bland-Altman P-P plots for pipelines with different software packages. The grey shade corresponds to the 0.95 confidence interval. A curve above (respectively below) the confidence interval indicates invalidity (respectively conservativeness). Default parameters: 5 mm smoothing, 24 motion regressors, and no HRF derivatives.

initially, changing smoothing kernel FWHM and number of motion regressors separately led to false positive rates close to 0.05, consistently, their combination led to rates that were only slightly invalid.

For both SPM and FSL, we observed shifts in the distributions of statistical values (Supplementary Figures 3 and 11). These shifts were similar to those obtained for changes in motion regressors only.

3.3 Analyses using pipelines with different software packages

We also explored the ability to use in a same between-group analysis subject-level data obtained with different software packages (here FSL and SPM). We performed the analyses for all possible combinations SPM > or < FSL: 2 smoothing kernels \times 3 numbers of motion regressors \times 2 HRF models, corresponding to 12 between-software comparisons – with the same setting for both SPM and FSL pipelines. The false positive rates are displayed in Figure 6. For all between-software analyses, the false positive rates were above 0.05. We obtained lower values for *SPM > FSL* (between 0.10 to 0.32), than for the opposite test (between 0.56 to 0.95). In all cases, false positive rates were largely increased compared to the reference analyses (*i.e.* using the same software in both groups). This observation was consistent with the P-P plot, which showed a large deviation from the 95% confidence interval for the direction *SPM < FSL* (Figure 7). Figure 8 shows the distribution of statistical values for the between-software comparison with all other parameters set with default values (*i.e.* 5mm smoothing kernel, 24 motion regressors and no HRF derivatives). We can see a shift in

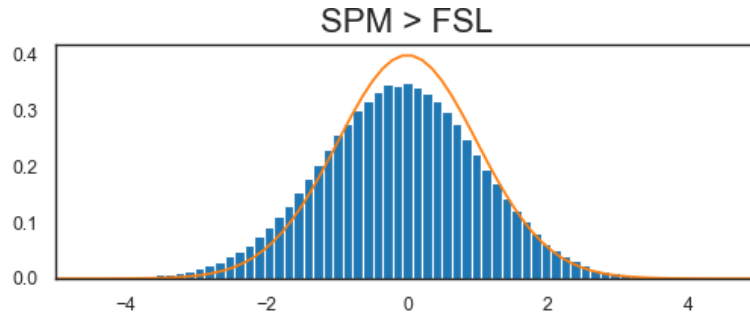


Figure 8: Distribution of statistical values for between-software analyses, compared to the expected distribution.

terms of mean and standard deviation of values. This shift was larger than those observed, for instance, for the effect of HRF derivatives, which was the most impacting factor on within-software comparisons.

4 Discussions

In this study, we showed that between-group analyses that use data generated by different pipelines can lead to invalidity when differences in pipelines are not properly accounted for. In almost all cases, combining data processed with different pipelines led to false positive rates above the theoretical 0.05 threshold. These results, obtained when combining subject-level contrast maps processed differently, suggest that it is necessary to consider how analytical variability may affect the results when combining data.

When performing analyses using the same pipeline on all participants' data (as traditionally done in the literature) results were valid for all analyses. Although the false positive rates obtained were lower than the 5% rate, the results were similar to those obtained in [40]. The level of smoothing, combined with the thresholding method that we chose (*i.e.* voxelwise FWE-corrected based on random field theory in SPM), may be responsible for these lower rates [41].

Our results for different pipeline analyses suggest that some factors have a larger impact than others. We saw that for differences regarding the size of the smoothing kernel and number of motion regressors (6 > or < 24 motion regressors) within SPM software package, results were similar to those obtained with identical pipeline analyses, suggesting that participant data can be combined without having to consider the differences in pipelines, if this is the only difference. This is not the case for differences in the use of HRF derivatives and use of motion regressors (0 motion regressors > or < 6 or 24 motion regressors), which gave invalid results.

We also saw that combining multiple differences in parameters could result in bigger effects, depending on the effect of each parameter alone. The combination of two parameters that both have a high effect on the results led in our case to inflated false positive rates, while the combination of parameters that had a limited effect did not lead to higher false positive rates (*e.g.* smoothing and motion regressors in SPM). This suggests that it may be possible to model the effect caused by specific variations in the subject-level pipelines. To enable this in the future, it is essential that the pipelines used is shared with enough details to allow a reproduction of the exact processing applied on the data.

However, the ability to model the effect of parameters is limited to specific variations. For example, for each variation of parameter, we saw different effects across the two software packages under study (SPM and FSL). Overall, observations were similar, but false positive rates were often increased in FSL compared to SPM for the same comparison. This suggests that some parameters values are more robust to changes when combined together, here, in FSL, when using 24 motion regressors, combining data with different use of HRF derivatives led to false positive rates close to the baseline analysis (*i.e.* same pipeline in both groups).

The most important source of invalidity was found when studying the effect of differences in software packages. SPM and FSL both implement similar pipeline steps with different settings. While we tried to align some parameters between the two software packages by changing the software package default values (*e.g.* smoothing kernel, type of HRF, etc.), some steps are specific to each software and cannot be changed by the user, causing potential differences between the results. We tried to correct some of these differences, in particular for the unit scale of subject-level contrast maps. But, even with these corrections, we still found highly inflated false positive rates when comparing pipelines with the same values for the parameters under study and different software packages. We suppose that differences in how software packages scale the data were not compensated by our simple rescaling approach and that more work will be needed to be able to combine subject-level data from two different software packages in the same analysis.

In this study, we focused on between-group analyses in which each group of participants was processed with a different pipeline. While this an extreme setup (in which the effect of interest is perfectly confounded with differences in pipelines), in practice, other combinations may be observed, for example with multiple pipelines used within a group. Our setup – in which processing pipelines varied depending on the group – was justified by the use-case in which data from various public datasets are used in the same analysis. For example, specific datasets have been created to study various neurological disorders, usually associated with a minimal processing pipeline dedicated to the study, and the corresponding minimally processed data (Alzheimer’s Disease Neuroimaging Initiative (ADNI) [42] for Alzheimer’s disease, Autism Brain Imaging Data Exchange (ABIDE) [43] for autism, etc). Researchers may want to use these minimally processed data and compare groups of participants with one group composed of participants with a specific conditions from one of these processed datasets, and the other group composed of healthy participants from another processed dataset (e.g. from the Human Connectome Project [5]).

We chose to study variations induced by 4 types of parameters (software package, HRF, smoothing and number of motion regressors), within each software package based on their widespread use in the neuroimaging community [44]. Yet, in practice, there are many more variations: researchers might use different software versions, perform or not specific sub-steps in the analysis (for example, the use or not of slice-timing correction), use different HRF models etc. Therefore, in real conditions, the differences observed between pipelines will likely be more important. In future works, other analyses may be done for other varying parameters using the same framework.

For other types of confounds such as imaging site, scanner effect or age and sex, harmonization or mitigation methods have been proposed to take these into account. Several studies [45, 46, 47, 15] proposed to remove confounds by incorporating them as additional regressors in the analysis, or by estimating batch-specific parameters (such as mean and variance) and then using these to standardize the data with frameworks such as ComBat [46]. Alternative approaches have also been proposed, such as restriction, where the study is limited to participants with specific characteristics. In [48], authors employed this method in a cohort study focused on males of the same age and nationality. In practice, pipeline-based differences may be accounted for by adding a confound in the statistical analysis (see for instance [15]), or using harmonization techniques such as ComBat [46]. However, the specific impact of pipeline-based differences remained unexplored. Our results can be used to understand in which cases differences in analytical pipelines must be accounted for.

Recently, deep learning frameworks, and in particular generative models used for style transfer [49], showed their potential for such task in converting data between different domains (e.g. acquisition site) [50, 51]. Currently, the most widespread practice is to include a covariate in the statistical model for pipeline-based confounds. We envision that other methods such as style transfer may provide additional solutions to mitigate analytical variability in such analyses (see for instance [52]).

5 Conclusion

Our study shows that between-group analysis using subject-level data which have been processed differently can be affected by pipeline-based differences. While some parameters did not have significant effects, others produced invalid results, suggesting that it is necessary to model those pipeline-based confounds.

6 Code availability

All analysis pipelines were executed in Python v3.8. The executions require the installation of SPM and FSL software packages. To facilitate reproducibility, we provide a NeuroDocker image that can be pulled from Dockerhub and that contains all necessary software packages. The Docker image is available at: https://hub.docker.com/r/elodiegermani/open_pipeline.

6.1 HCP Multi-pipeline

Python scripts to run the pipelines and create the dataset were made available publicly in the Software Heritage public archive: `swh:1:snp:17870c3d782aa25a7ffdd6165fe27ce6eac6c90b` [53].

6.2 Between-group analyses, figures and tables

Python and Matlab scripts to run the experiments and to create the figures and tables of this article are available in the Software Heritage public archive: `swh:1:dir:4381210db83c93bca14cf685be0ec293128412c8` [26].

- Programming language: Python3.8, Matlab
- Licence: MIT
- Requirements: multiple Python libraries, available in the Docker container `open_pipeline`

7 Data availability

This study was performed using derived data from the HCP Young Adult [5], publicly available at ConnectomeDB. Data usage requires registration and agreement to the HCP Young Adult Open Access Data Use Terms available at: [27].

The HCP multi-pipeline dataset [25] is in the process of being made publicly available on Public-nEUro [54] (we are currently pending approval from Data Protection Officers at our institute).

8 Ethics

This study was performed using derived data from the HCP Young Adult [5]. No experimental activity involving the human participants was made by the authors. Only publicly released data were used.

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 Young Adult Open Access Data Use Terms available at: [27].

9 Authors contributions

All authors participated to the conceptualization of the project. E.G. and X.R. developed the project (wrote the code, performed the experiments, and analyzed results) and wrote the original draft of the manuscript. C.M. and P.M. supervised the project and provided feedback on the manuscript (writing: review and editing).

10 Competing interests

The authors declare that they have no competing interests.

11 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). Elodie Germani was supported by Region Bretagne (ARED MAPIS) and Agence Nationale pour la Recherche for the programm of doctoral contracts in artificial intelligence (project ANR-20-THIA-0018).

References

- [1] Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, 14(5):365–376, 2013.
- [2] Russell A Poldrack, Chris I Baker, Joke Durnez, Krzysztof J Gorgolewski, Paul M Matthews, Marcus R Munafò, Thomas E Nichols, Jean-Baptiste Poline, Edward Vul, and Tal Yarkoni. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature reviews neuroscience*, 18(2):115, 2017.

- [3] Rotem Botvinik-Nezer and Tor D. Wager. Reproducibility in Neuroimaging Analysis: Challenges and Solutions. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 8(8):780–788, 2023. ISSN 2451-9022. doi:10.1016/j.bpsc.2022.12.006.
- [4] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12, 2015. doi:https://www.doi.org/10.1371/journal.pmed.1001779.
- [5] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.
- [6] Gholamreza Salimi-Khorshidi, Stephen M Smith, John R Keltner, Tor D Wager, and Thomas E Nichols. Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. *Neuroimage*, 45(3):810–823, 2009.
- [7] Jean-Baptiste Poline, Janis L. Breeze, Satrajit S. Ghosh, Krzysztof Gorgolewski, Yaroslav O. Halchenko, Michael Hanke, Karl G. Helmer, Daniel S. Marcus, Russell A. Poldrack, Yannick Schwartz, John Ashburner, and David N. Kennedy. Data sharing in neuroimaging research. *Frontiers in Neuroinformatics*, 6, 2012. ISSN 1662-5196. doi:10.3389/fninf.2012.00009.
- [8] Russell A Poldrack and Krzysztof J Gorgolewski. Making big data open: data sharing in neuroimaging. *Nature neuroscience*, 17(11):1510, 2014.
- [9] Guiomar Niso, Rotem Botvinik-Nezer, Stefan Appelhoff, Alejandro De La Vega, Oscar Esteban, Joset A. Etzel, Karolina Finc, Melanie Ganz, Rémi Gau, Yaroslav O. Halchenko, Peer Herholz, Agah Karakuzu, David B. Keator, Christopher J. Markiewicz, Camille Maumet, Cyril R. Pernet, Franco Pestilli, Nazek Queder, Tina Schmitt, Weronika Sójka, Adina S. Wagner, Kirstie J. Whitaker, and Jochem W. Rieger. Open and reproducible neuroimaging: From study inception to publication. *NeuroImage*, 263:119623, 2022. ISSN 1095-9572. doi:10.1016/j.neuroimage.2022.119623.
- [10] Krzysztof J Gorgolewski, Gael Varoquaux, Gabriel Rivera, Yannick Schwarz, Satrajit S Ghosh, Camille Maumet, Vanessa V Sochat, Thomas E Nichols, Russell A Poldrack, Jean-Baptiste Poline, et al. Neurovault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in neuroinformatics*, 9:8, 2015.
- [11] Christopher J Markiewicz, Krzysztof J Gorgolewski, Franklin Feingold, Ross Blair, Yaroslav O Halchenko, Eric Miller, Nell Hardcastle, Joe Wexler, Oscar Esteban, Mathias Goncalves, Anita Jwa, and Russell Poldrack. The OpenNeuro resource for sharing of neuroscience data. *eLife*, 10:e71774, 2021. ISSN 2050-084X. doi:10.7554/eLife.71774.
- [12] Christian Barillot, Elise Bannier, Olivier Commowick, Isabelle Corouge, Anthony Baire, Ines Fackfack, Justine Guillaumont, Yao Yao, and Michael Kain. Shanoir: Applying the Software as a Service Distribution Model to Manage Brain Imaging Research Repositories. *Frontiers in information and communication technologies*, 2016. doi:10.3389/fict.2016.00025.
- [13] John PA Ioannidis, Marcus R Munafo, Paolo Fusar-Poli, Brian A Nosek, and Sean P David. Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in cognitive sciences*, 18(5): 235–241, 2014.
- [14] Oscar Esteban, Christopher J. Markiewicz, Ross W. Blair, Craig A. Moodie, A. Ilkay Isik, Asier Erramuzpe, James D. Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, Hiroyuki Oya, Satrajit S. Ghosh, Jesse Wright, Joke Durnez, Russell A. Poldrack, and Krzysztof J. Gorgolewski. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1):111–116, 2019. ISSN 1548-7105. doi:10.1038/s41592-018-0235-4.
- [15] Fidel Alfaró-Almagro, Paul McCarthy, Soroosh Afyouni, Jesper L. R. Andersson, Matteo Bastiani, Karla L. Miller, Thomas E. Nichols, and Stephen M. Smith. Confound modelling in UK Biobank brain imaging. *NeuroImage*, 224:117002, 2021. ISSN 1053-8119. doi:10.1016/j.neuroimage.2020.117002.
- [16] Alexander Bowering, Camille Maumet, and Thomas E Nichols. Exploring the impact of analysis software on task fmri results. *Human brain mapping*, 40(11):3362–3384, 2019.
- [17] 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.

- [18] Xinhui Li, Lei Ai, Steve Giavasis, Hecheng Jin, Eric Feczko, Ting Xu, Jon Clucas, Alexandre Franco, Anibal Sólón Heinsfeld, Azeez Adebimpe, et al. Moving beyond processing and analysis-related variation in neuroscience, 2021.
- [19] Rotem Botvinik-Nezer, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A Mumford, R Alison Adcock, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, pages 1–7, 2020.
- [20] Stephen C. Strother. Evaluating fMRI preprocessing pipelines. *IEEE engineering in medicine and biology magazine: the quarterly magazine of the Engineering in Medicine & Biology Society*, 25(2):27–41, 2006. ISSN 0739-5175. doi:10.1109/memb.2006.1607667.
- [21] 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.
- [22] Stephen LaConte, Jon Anderson, Suraj Muley, James Ashe, Sally Frutiger, Kelly Rehm, Lars Kai Hansen, Essa Yacoub, Xiaoping Hu, David Rottenberg, and Stephen Strother. The Evaluation of Preprocessing Choices in Single-Subject BOLD fMRI Using NPAIRS Performance Metrics. *NeuroImage*, 18(1):10–27, January 2003. ISSN 1053-8119. doi:10.1006/nimg.2002.1300. URL <https://www.sciencedirect.com/science/article/pii/S1053811902913005>.
- [23] 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 Suppl 1:S196–207, 2004. ISSN 1053-8119. doi:10.1016/j.neuroimage.2004.07.022.
- [24] Nathan W. Churchill, Robyn Spring, Babak Afshin-Pour, Fan Dong, and Stephen C. Strother. An Automated, Adaptive Framework for Optimizing Preprocessing Pipelines in Task-Based Functional MRI. *PLoS ONE*, 10(7): e0131520, July 2015. ISSN 1932-6203. doi:10.1371/journal.pone.0131520. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4498698/>.
- [25] Elodie Germani, Elisa Fromont, Pierre Maurel, and Camille Maumet. The HCP multi-pipeline dataset: an opportunity to investigate analytical variability in fMRI data analysis. working paper or preprint, December 2023. URL <https://inserm.hal.science/inserm-04356768>.
- [26] Elodie Germani, Xavier Rolland, Pierre Maurel, and Camille Maumet. Software heritage archive for the gitlab repository "hcp_pipelines_compatibility", 2024. URL https://archive.softwareheritage.org/swh:1:dir:4381210db83c93bca14cf685be0ec293128412c8;origin=https://gitlab.inria.fr/egermani/hcp_pipelines_compatibility;visit=swh:1:snp:579fb7e69702ce1f9f7192b5e73772a213a35c29;anchor=swh:1:rev:7979bf2d392a0c37c22615c1a4c826735c0b49e8.
- [27] Human connectome project: Data usage agreement. <https://www.humanconnectome.org/study/hcp-young-adult/document/wu-minn-hcp-consortium-open-access-data-use-terms>, 2013.
- [28] William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E Nichols. *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.
- [29] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. *Fsl. Neuroimage*, 62(2):782–790, 2012.
- [30] Alexander Bowring, Thomas E. Nichols, and Camille Maumet. Isolating the sources of pipeline-variability in group-level task-fMRI results. *Human Brain Mapping*, 43(3):1112–1128, 2022. ISSN 1097-0193. doi:10.1002/hbm.25713. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.25713>.
- [31] Alan C. Evans, Andrew L. Janke, D. Louis Collins, and Sylvain Baillet. Brain templates and atlases. *NeuroImage*, 62(2):911–922, 2012. ISSN 10538119. doi:10.1016/j.neuroimage.2012.01.024.
- [32] Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gael Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 2014. doi:10.3389/fninf.2014.00014.
- [33] Thomas Nichols. SPM plot units, 2012. URL <https://web.archive.org/web/20230606094719/https://blog.nisox.org/2012/07/31/spm-plot-units>.
- [34] Jérôme Dockès, Kendra Oudyk, Mohammad Torabi, Alejandro I de la Vega, and Jean-Baptiste Poline. Mining the neuroimaging literature. *eLife Sciences Publications, Ltd*, 2024. doi:10.7554/elife.94909.1. URL <http://dx.doi.org/10.7554/elife.94909.1>.
- [35] Matthew Brett, Will Penny, and Stefan Kiebel. An Introduction to Random Field Theory. 2003.

- [36] K. J. Worsley, S. Marrett, P. Neelin, A. C. Vandal, K. J. Friston, and A. C. Evans. A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4(1):58–73, 1996. ISSN 1065-9471. doi:10.1002/(SICI)1097-0193(1996)4:1<58::AID-HBM4>3.0.CO;2-O.
- [37] Tian Ge, Jianfeng Feng, Derrek P. Hibar, Paul M. Thompson, and Thomas E. Nichols. Increasing power for voxel-wise genome-wide association studies: The random field theory, least square kernel machines and fast permutation procedures. *NeuroImage*, 63(2):858–873, November 2012. ISSN 1053-8119. doi:10.1016/j.neuroimage.2012.07.012. Publisher: Elsevier BV.
- [38] Herbert A. David and Haikady N. Nagaraja. *Order statistics*. John Wiley, Hoboken, N.J, 3rd ed edition, 2003. ISBN 978-0-471-72216-8 978-0-471-65401-8 978-0-471-38926-2. doi:10.1002/0471722162.
- [39] Davide Giavarina. Understanding Bland Altman analysis. *Biochemia Medica*, 25(2):141–151, 2015. ISSN 18467482. doi:10.11613/BM.2015.015.
- [40] 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*, 113(28):7900–7905, 2016.
- [41] Thomas Nichols and Satoru Hayasaka. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, 12(5):419–446, 2003. ISSN 0962-2802. doi:10.1191/0962280203sm341ra.
- [42] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer’s disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.
- [43] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.
- [44] Joshua Carp. On the plurality of (methodological) worlds: estimating the analytic flexibility of fmri experiments. *Frontiers in neuroscience*, 6:149, 2012.
- [45] Anil Rao, Joao M. Monteiro, and Janaina Mourao-Miranda. Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage*, 150:23–49, 2017. ISSN 1053-8119. doi:https://doi.org/10.1016/j.neuroimage.2017.01.066.
- [46] Jean-Philippe Fortin, Elizabeth M Sweeney, John Muschelli, Ciprian M Crainiceanu, Russell T Shinohara, Alzheimer’s Disease Neuroimaging Initiative, et al. Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage*, 132:198–212, 2016.
- [47] Joanne C Beer, Nicholas J Tustison, Philip A Cook, Christos Davatzikos, Yvette I Sheline, Russell T Shinohara, Kristin A Linn, Alzheimer’s Disease Neuroimaging Initiative, et al. Longitudinal combat: A method for harmonizing longitudinal multi-scanner imaging data. *NeuroImage*, 220:117129, 2020.
- [48] Kiyana Zarnani, Thomas E. Nichols, Fidel Alfaro-Almagro, Birgitte Fagerlund, Martin Lauritzen, Egill Rostrup, and Stephen M. Smith. Discovering markers of healthy aging: a prospective study in a Danish male birth cohort. *Aging*, 11(16):5943–5974, 2019. ISSN 1945-4589. doi:10.18632/aging.102151.
- [49] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image Style Transfer Using Convolutional Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. doi:10.1109/CVPR.2016.265.
- [50] Harrison Nguyen, Richard W. Morris, Anthony W. Harris, Mayuresh S. Korgoankar, and Fabio Ramos. Correcting differences in multi-site neuroimaging data using Generative Adversarial Networks, 2018. URL <http://arxiv.org/abs/1803.09375>.
- [51] Mengting Liu, Piyush Maiti, Sophia Thomopoulos, Alyssa Zhu, Yaqiong Chai, Hosung Kim, and Neda Jahanshad. Style Transfer Using Generative Adversarial Networks for Multi-site MRI Harmonization. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 313–322. Springer International Publishing, 2021. ISBN 978-3-030-87199-4. doi:10.1007/978-3-030-87199-4_30.
- [52] Elodie Germani, Camille Maumet, and Elisa Fromont. Mitigating analytical variability in fMRI results with style transfer, 2024. URL <http://arxiv.org/abs/2404.03703>.
- [53] Elodie Germani, Elisa Fromont, Pierre Maurel, and Camille Maumet. Software heritage archive for the gitlab repository "hcp_pipelines", 2023. URL <https://archive.softwareheritage.org/swh:1:dir:>

67ce4a985abc2206169943486b91db7acb998a54;origin=https://gitlab.inria.fr/egermani/hcp_pipelines;visit=swh:1:snp:17870c3d782aa25a7ffdd6165fe27ce6eac6c90b;anchor=swh:1:rev:3cd5ecce2bbc7d5a38c89878435b0b526541b24d.

[54] Public nEUro, 2023. URL <https://public-neuro.github.io/index.html>.

Supplementary materials

Supplementary methods

Bland-Altman P-P plots For a given pair of pipelines, we have 1,000 group analyses, which makes a total of more than 150M voxel values. Since the resulting list of voxels obtained is very large and using it for further observations can be very time-consuming, for each between-group analysis using two given pipelines, on which we wanted to make observations, we only took a random sample of 1,000,000 values from the concatenation of statistical values over the 1,000 corresponding group analyses.

To compute the p-values, we transformed the statistic values using the survival function of the Student's t-distribution with 98 degrees of freedom (50 participants + 50 participants - 2). This corresponds to 1-CDF (cumulative distribution function) of the t-distribution. The confidence intervals were computed using a beta distribution for each kth value with the lower bound being k and the upper bound being 1, 000, 000 - k + 1. After conversion to logarithmic scale, this gave us the confidence intervals for the distribution of p-values.

Supplementary table

SPM

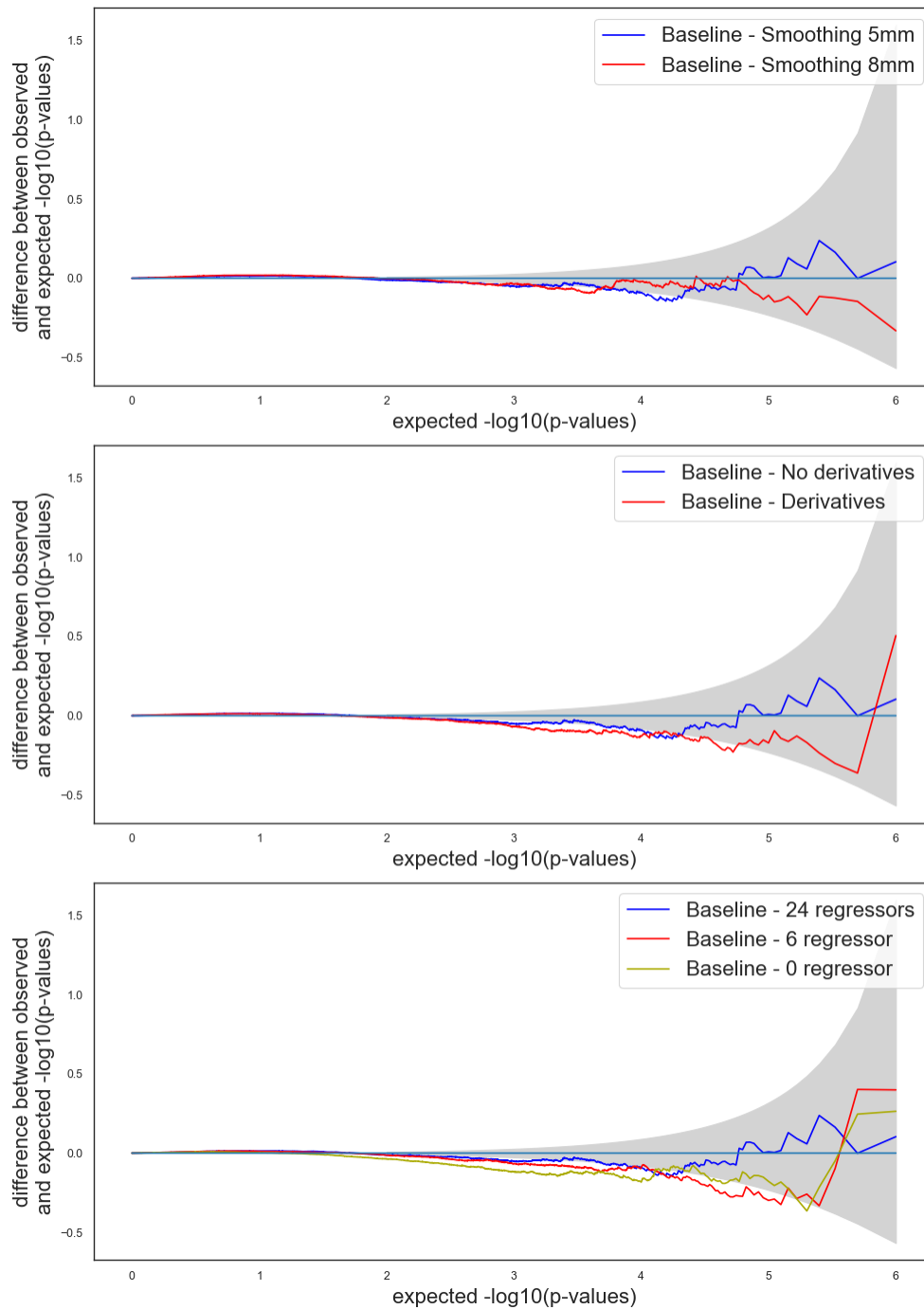
	Smooth 5 mm		Smooth 8 mm	
	No derivatives	Derivatives	No derivatives	Derivatives
0 motion regressors	0.014	0.019	0.025	0.019
6 motion regressors	0.013	0.015	0.021	0.025
24 motion regressors	0.021	0.015	0.018	0.019

FSL

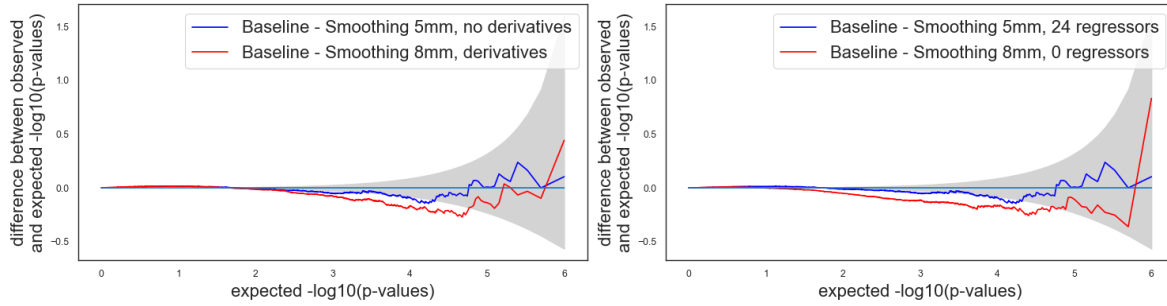
	Smooth 5 mm		Smooth 8 mm	
	No derivatives	Derivatives	No derivatives	Derivatives
No motion regressors	0.01	0.013	0.014	0.014
6 motion regressors	0.015	0.017	0.017	0.022
24 motion regressors	0.017	0.02	0.014	0.012

Supplementary Table 1: False positive rates for between-groups analyses with the same pipeline in both groups, using contrast maps without post-processing with SPM and FSL and for all possible sets of parameters (number of motion regressors, smoothing kernel FWHM and presence or absence of HRF temporal derivatives). The rates were always under 0.05.

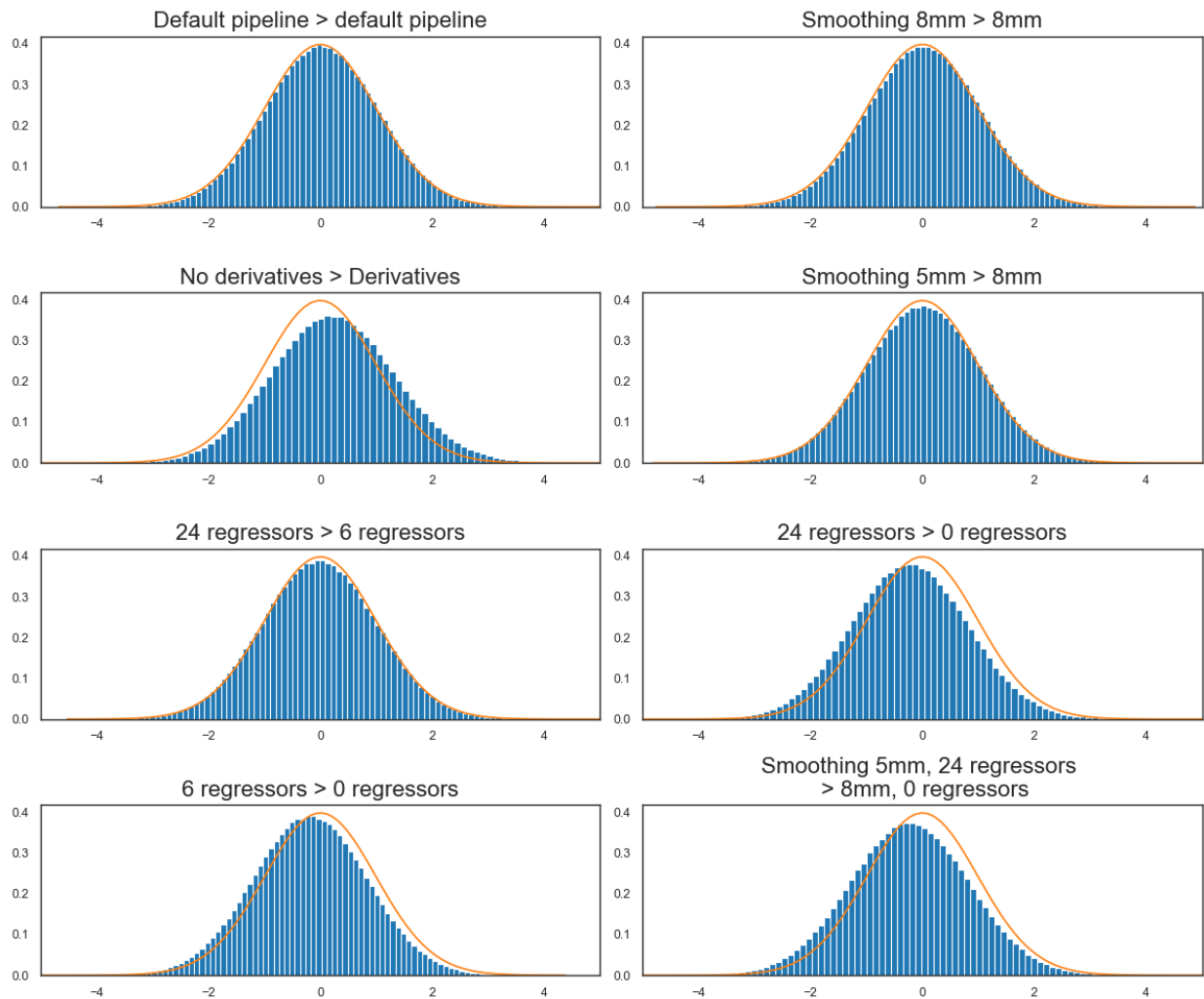
Supplementary figures for analyses within SPM



Supplementary Figure 1: Baseline for Figure 4. Bland-Altman P-P plots for pipelines with no differing parameters in SPM. The grey shade corresponds to the 0.95 confidence interval. A curve above (respectively below) the confidence interval indicates invalidity (respectively conservativeness). Default parameters: 5 mm smoothing, 24 motion regressors and no HRF derivatives.

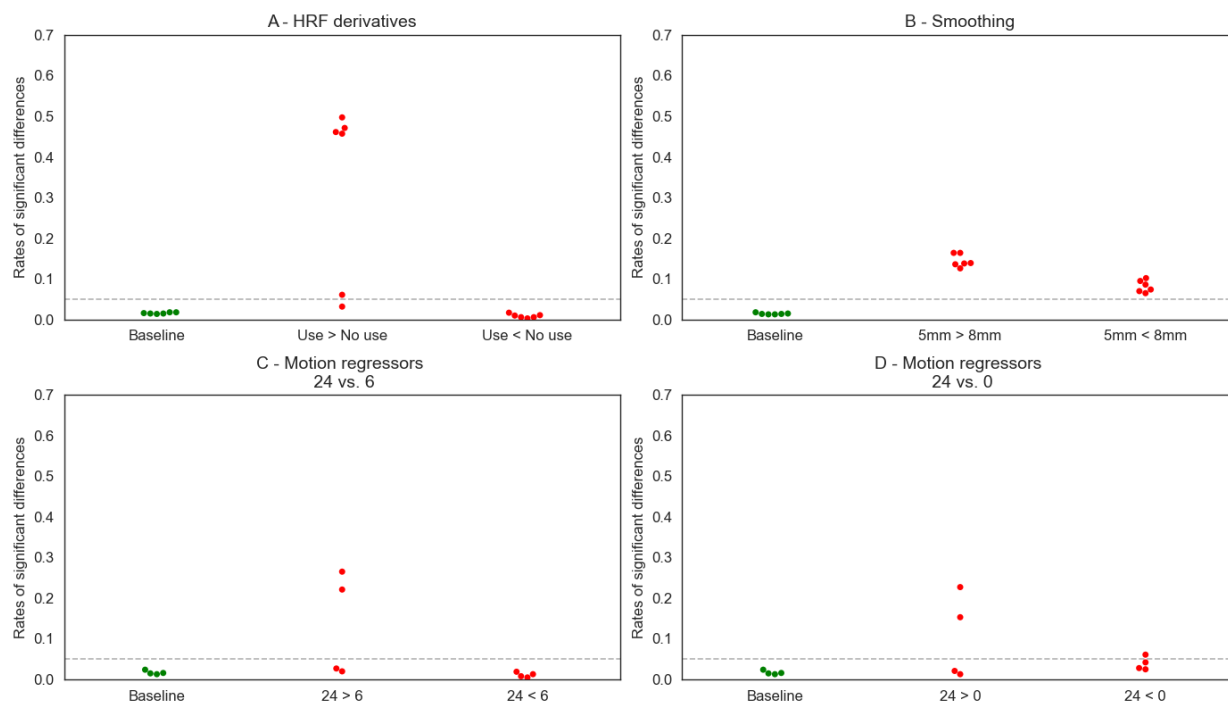


Supplementary Figure 2: Baseline for Figure 5. Bland-Altman P-P plots for pipelines with no differing parameters within SPM. The grey shade corresponds to the 0.95 confidence interval. A curve above (respectively below) the confidence interval indicates invalidity (respectively conservativeness). Default parameters: 5 mm smoothing, 24 motion regressors and no HRF derivatives.

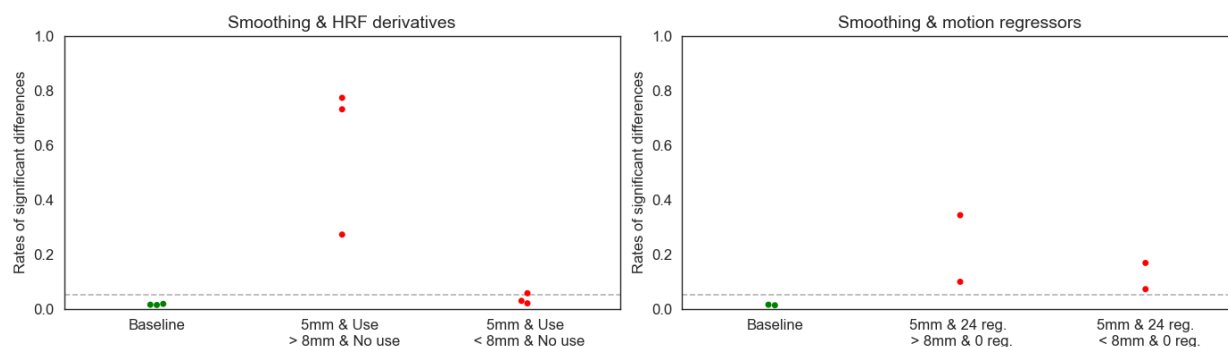


Supplementary Figure 3: Distribution of statistical values for multiple between-group analyses under SPM, compared to the expected distribution. Default parameters: 5 mm smoothing, 24 motion regressors and no HRF derivatives. Pipelines which differ from the default pipeline are put in bold. The orange curve represents the Student distribution with 98 degrees of freedom, which is the expected distribution in our case under null hypothesis.

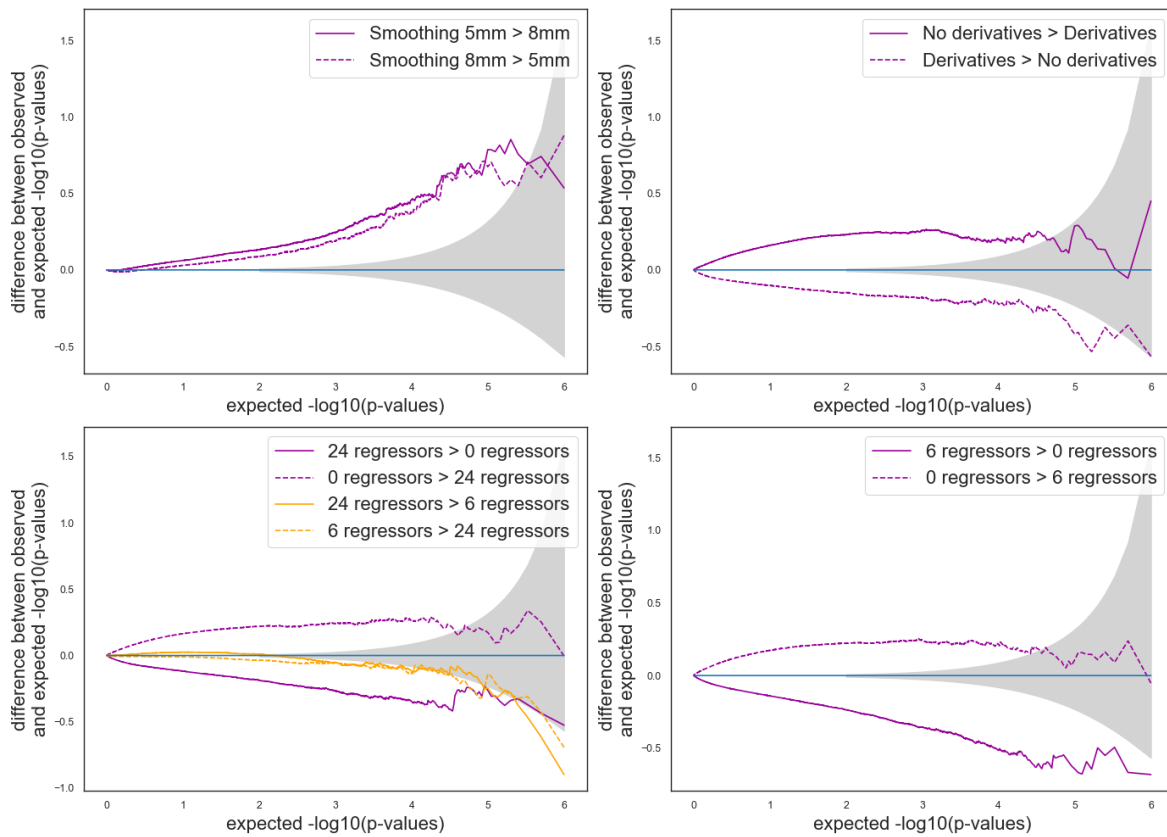
Supplementary figures for analyses within FSL



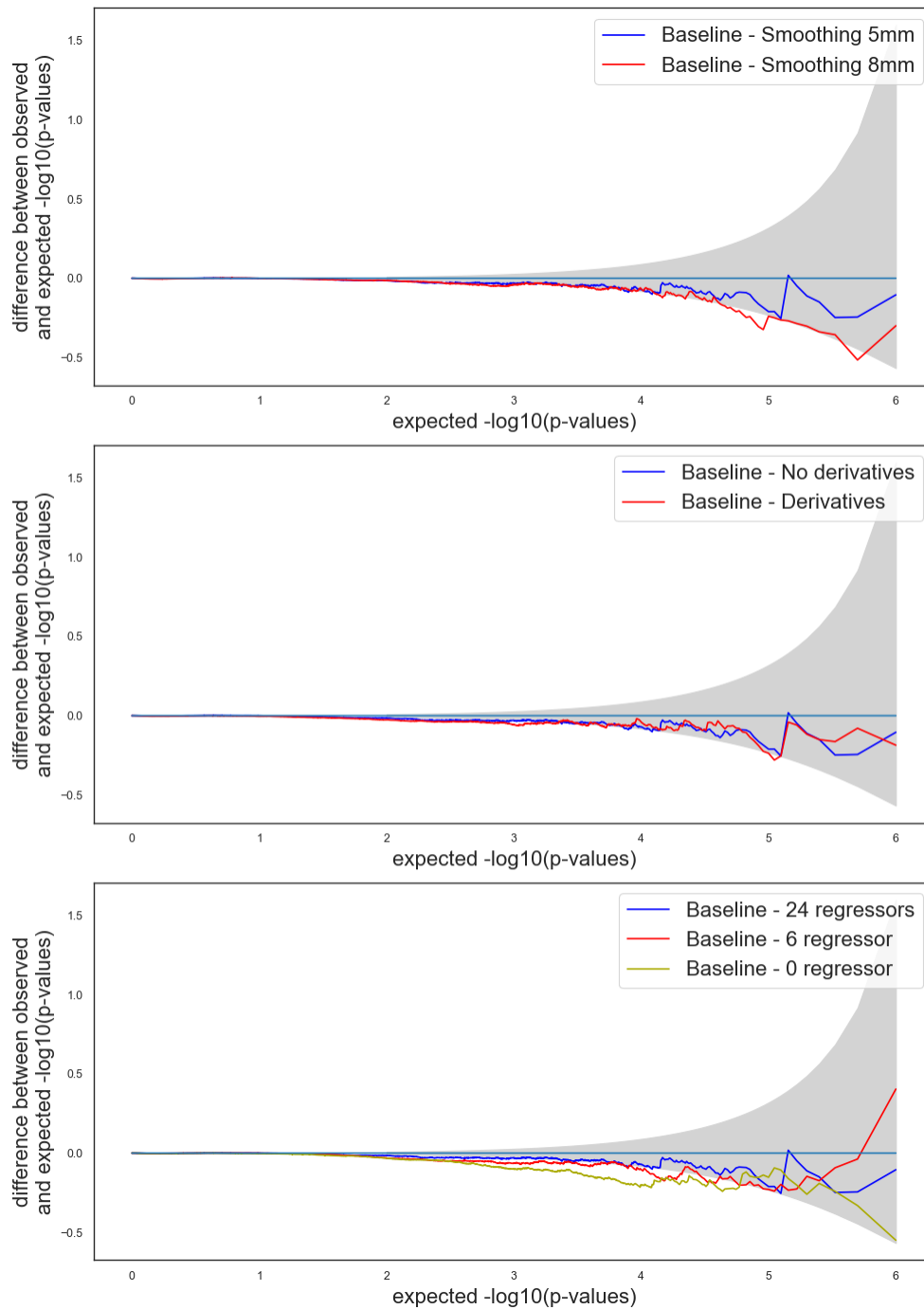
Supplementary Figure 4: False positive rates for pipelines with a single differing parameter in FSL: A) HRF derivatives, B) smoothing and C and D) motion regressors. For each, we provide the false positive rates obtained for: 1/ Default > Variation and Default < Variation (Red) and 2/ baseline analysis with default parameters, used as a reference (Green, first column). The grey dashed line corresponds to the alpha level (0.05), and the grey band to the corresponding confidence interval at 95%.



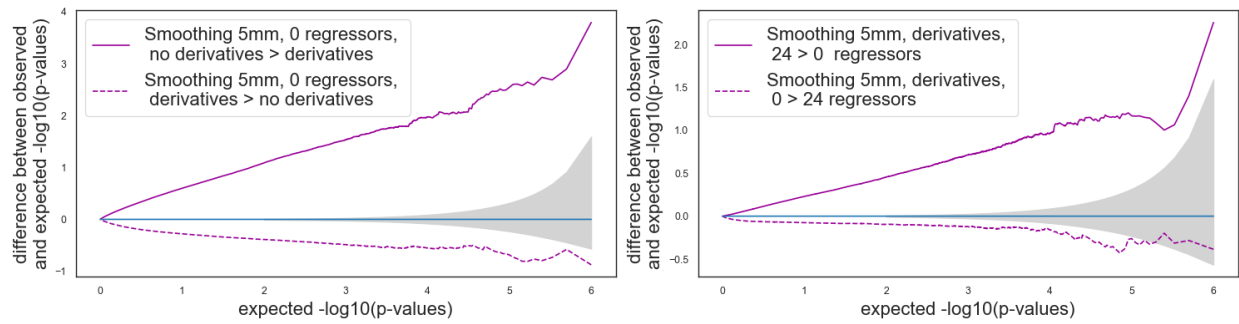
Supplementary Figure 5: False positive rates for pipelines with two differing parameters in FSL: A) Smoothing and HRF, B) Smoothing and motion regressors. For each studied parameter, we provide the rates obtained for: 1/ Default > Variation and Default < Variation (Red) and 2/ baseline analysis with default parameters, used as a reference (Green, first column). The grey dashed line corresponds to the alpha level (0.05) and grey band to the corresponding confidence interval at 95%.



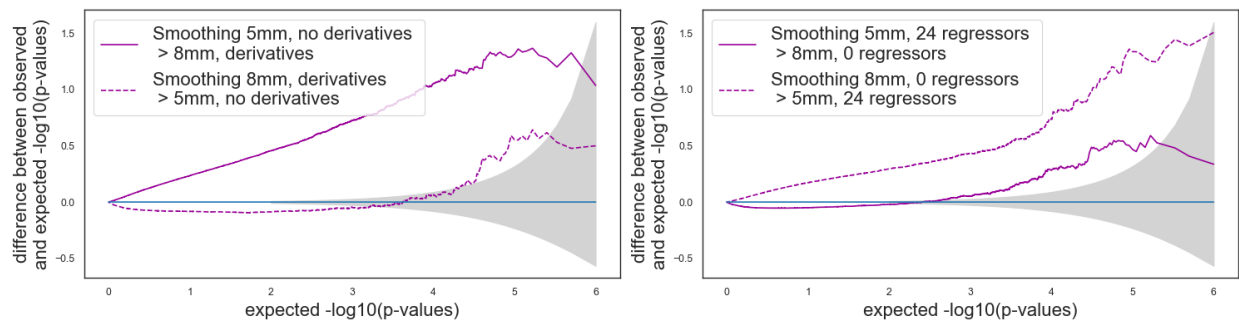
Supplementary Figure 6: Bland-Altman P-P plots for pipelines with a single differing parameter in FSL. The grey shade corresponds to the 0.95 confidence interval. A curve above (respectively below) the confidence interval indicates invalidity (respectively conservativeness). Default parameters: 5 mm smoothing, 24 motion regressors and no HRF derivatives.



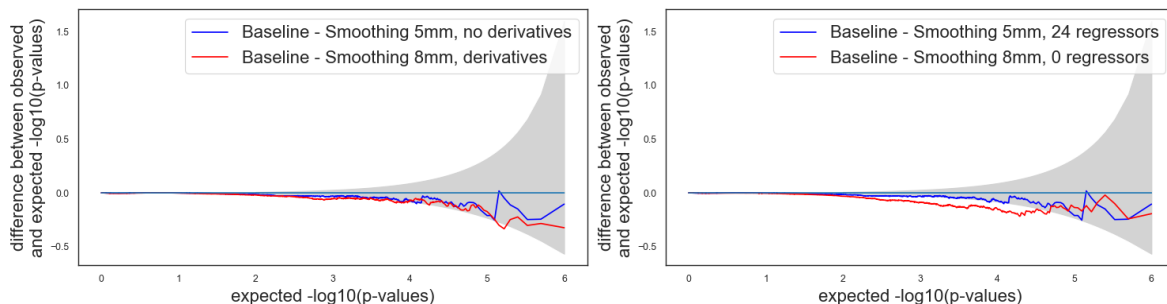
Supplementary Figure 7: Baseline for Supplementary Figure 6. Bland-Altman P-P plots for pipelines with no differing parameters in FSL. The grey shade corresponds to the 0.95 confidence interval. A curve above (respectively below) the confidence interval indicates invalidity (respectively conservativeness). Default parameters: 5 mm smoothing, 24 motion regressors and no HRF derivatives.



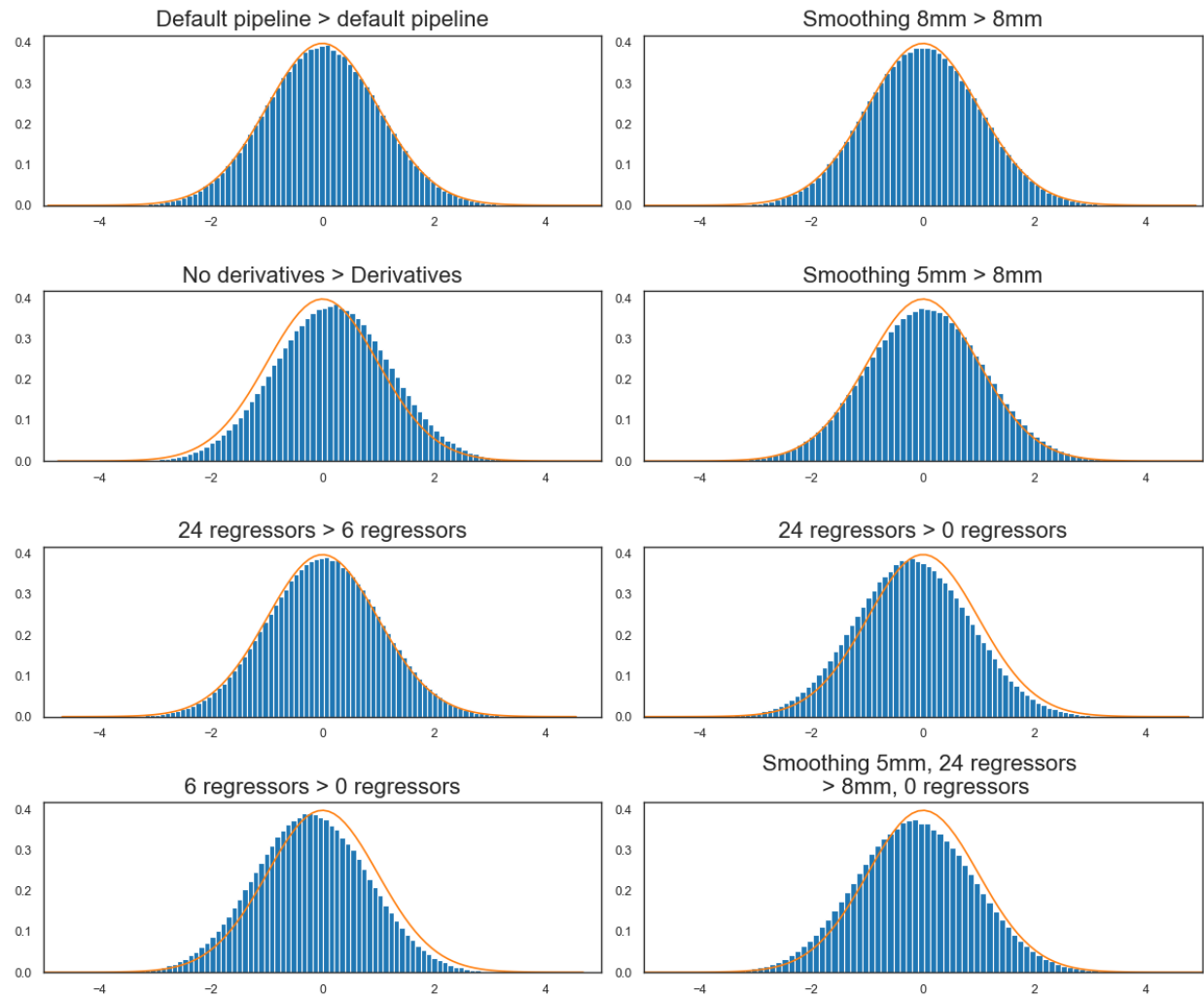
Supplementary Figure 8: Bland-Altman P-P plots for pipelines with a single differing parameter in SPM. The grey shade corresponds to the 0.95 confidence interval. A curve above (respectively below) the confidence interval indicates invalidity (respectively conservativeness). Default parameters values were modified to 5 mm smoothing, 0 motion regressors and no HRF derivatives to explore the impact of fixed parameters on the validity of analyses.



Supplementary Figure 9: Bland-Altman P-P plots for pipelines with two differing parameters in FSL. The grey shade corresponds to the 0.95 confidence interval. A curve above (respectively below) the confidence interval indicates invalidity (respectively conservativeness). Default parameters: 5 mm smoothing, 24 motion regressors and no HRF derivatives.



Supplementary Figure 10: Baseline for Figure 9. Bland-Altman P-P plots for pipelines with no differing parameters in SPM. The grey shade corresponds to the 0.95 confidence interval. A curve above (respectively below) the confidence interval indicates invalidity (respectively conservativeness). Default parameters: 5 mm smoothing, 24 motion regressors and no HRF derivatives.



Supplementary Figure 11: Distribution of statistical values for multiple between-group analyses under FSL, compared to the expected distribution. Default parameters: 5 mm smoothing, 24 motion regressors and no HRF derivatives. Pipelines which differ from the default pipeline are put in bold. The orange curve represents the Student distribution with 98 degrees of freedom, which is the expected distribution in our case under null hypothesis.