Minimal Data Needed for Valid & Accurate Image-Based fMRI Meta-Analysis

Camille Maumet, Thomas Nichols

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

Statistics Department, University of Warwick, Coventry, UK.

Abstract

Meta-analysis provides a quantitative approach to summarise the rich functional Magnetic Resonance Imaging literature (fMRI). Due to the lack of availability of image data supportimg existing literature, the majority of fMRI meta-analysis are coordinate-based. However, when image data is available for each study, the optimal approach is to perform an image-based meta-analysis. A number of approaches have been proposed to perform such meta-analyses including combination of standardised statistics, just effect estimates or both effects estimates and their sampling variance. While the latter is the preferred approach in the statistical community, its properties are only guaranteed in large samples. Additionally, often only standardised estimates are shared, reducing the possible meta-analytic approaches. Finally, because the BOLD signal is non-quantitative care has to be taken in order to insure that effect estimates are expressed in the same units, especially when combining data from different software packages. Given the growing interest in data sharing in the neuroimaging community there is a need to identify what is the minimal data to be shared in order to allow for future image-based meta-analysis. In this paper, we compare the validity and the accuracy of nine meta-analytic approaches on simulated and real data.

Keywords: Meta-analysis, Neuroimaging, Mixed-effects

1. Introduction

Neuroscience studies have low power: true effects have low likelihood to be detected and, detections have a low likelihood to be true [4]. Many authors have advocated for increased sample sizes in neuroimaging (e.g. [31]) which would mechanically increase power and our community has started a shift towards the creation of large datasets. Large datasets are flourishing: created by consortia that collate datasets acquired at multiple sites (e.g. 1000FC/INDI [22]) or, by dedicated projects (e.g. UK Biobank [23]). Those efforts show promising results [7], but according to a recent study, sample sizes of neuroimaging studies remain low with a median just over 30 participants per study in 2015 [31]. Another approach to increase sample size is to combine the results of previous studies through meta-analyses. Meta-analysis, by providing inference based on the results of previously conducted studies, provides an essential tool to increase power while leveraging the existing literature. The mature field of task-based Functional Magnetic Resonance Imaging (fMRI) with its ample literature of more than 10 000 articles [32] is particularly well-suited for this approach.

Most fMRI meta-analyses are 'coordinate-based' (CBMA), i.e. they proceed by combining information about the locations of local maxima (and optionally their statistic value). A number of tools are resources are available for CBMA [9, 43]. But, the best practice method is an image-based meta-analysis (IBMA) that combines information from each study at every location in space [1], effectively enabling the discovery of consistent sub-threshold effects that would not be uncovered using coordinates only. For a long time, the standard in fMRI literature has been to report a table with coordinates and statistic values of each local maxima. The number of IBMAs was very limited as they required contacting the original authors of each study of interest to retrieve the summary statistic images, an approach that is both time-consuming and inefficient [2]. With the emergence of tools that facilitate sharing 3D volumes obtained as a result of a statistical analysis, such as NeuroVault [11], IBMAs becomes more and more easy to perform.

But the best approach to perform an IBMA is still unknown, with three different types of summary data under consideration:

1. the contrast estimates and contrast variance estimates (E+S).
2. the contrast estimates (E).
3. the standardized statistic maps (Z).

Figure 1: False positive rates of the meta-analytic estimators under the null hypothesis for $p < 0.05$.

The first option (E+S) is considered best practice and leads to more statistically efficient estimates [6] but, it requires the contrasts to be expressed with in the same units. In fMRI, units will depends on the field strength [5] as well as data, model and contrast vector scaling [25]. Also, while often neglected, the statistical methods for this first option rely on asymptotic, large sample results and may have poor performance for typical number of studies [12, 19, 13]. The second option (E) assumes homoscedasticity, i.e. constant variance of the study-level contrast estimates, an assumption that is difficult to respect in meta-analyses as the number of subjects is likely to vary across studies. Finally the third option (Z) mainly corresponds to historical methods that assume homogeneity, i.e. no between-study variance.

Given the growing interest in data sharing in the neuroimaging community [31, 26], there is a need to identify what is the best approach for neuroimaging IBMAs and what is the minimal data needed. Here, we compared the use of IMBA using 9 meta-analytic approaches: 2 approaches use E+S's, 2 E's only and 5 Z's. We assessed and compared validity using simulated data under violation of each method's underlying assumptions: in small samples, under hetreroscedasticity, under heterogeneity or with contrast using mismatched units. Then, we estimated the accuracy of each method using a real dataset of 21 studies of pain.

2. Methods

2.1. Theory

For study $i = 1, \ldots, k$ we have contrast estimate $\hat{\beta}_i$, its contrast variance estimate s_i^2 (i.e. squared standard error), its equivalent Z-statistic map Z_i and its sample size n_i.

Combining contrast estimates and their standard error. The recommended approach is to fit contrast estimates and their standard error with a hierarchical general linear model (GLM) [6], creating a third-level (level 1: subject; level 2: study; level 3: meta-analysis). The general formulation for the study-level data is a meta-regression:

$$\hat{\boldsymbol{\beta}} = X\boldsymbol{\gamma} + \epsilon \tag{1}$$

where $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1 \ldots \hat{\beta}_k]^T$ is the vector of contrast estimates, X is the $k \times p$ study-level matrix. While the design matrix X can take any form, typically a one-sample model is used with $p == 1$ and X comprised of a column of ones. $\boldsymbol{\gamma} = [\gamma_1 \ldots \gamma_p]^T$ is a vector of meta-analytic parameters to estimate and $\epsilon \sim \mathcal{N}(0, W)$ is the residual error term.

In the most general case of a random-effects meta-analysis, i.e. assuming non-zero between-study variance, we have $W = \mathrm{diag}(\sigma_1^2/n_1 + \tau^2, \ldots, \sigma_k^2/n_k + \tau^2)$ where τ^2 denotes the between-study variance and σ_i^2/n_i denotes the variance of the contrast estimate for study i. Eq. (1) can be solved by weighted least squares giving:

$$\begin{aligned} \hat{\boldsymbol{\gamma}} &= (X^T W^{-1} X)^{-1} X^T W^{-1} \hat{\boldsymbol{\beta}} & (2) \\ \mathrm{Var}(\hat{\boldsymbol{\gamma}}) &= (X^T W^{-1} X)^{-1} & (3) \end{aligned}$$

Table S1 and S2 provides the WLS estimates and their sampling variances for one and two-sample tests. But in practice, the weight matrix W is unknown and has to be estimated from the data. Given $\hat{W}$ a consistent estimate of W, the feasible generalized least squares (FGLS) estimator is computed as:

$$\begin{aligned} \hat{\boldsymbol{\gamma}} &= (X^T \hat{W}^{-1} X)^{-1} X^T \hat{W}^{-1} \hat{\boldsymbol{\beta}} & (4) \\ \mathrm{Var}(\hat{\boldsymbol{\gamma}}) &= (X^T \hat{W}^{-1} X)^{-1} & (5) \end{aligned}$$

Here we use s_i^2 as an estimate of σ_i^2/n_i and estimate $\hat{\tau}^2$ from the data. FGLS is asymptotically efficient but its finite sample properties are unknown [12]. We used the 'FLAME 1' implementation available in FSL that provides inference by comparing the statistic to a Student distribution with $k - p$ degrees of freedom [42] as depicted in Table 1. This reference approach will be referred to as **Mixed-effects (MFX) GLM**.

In a fixed-effects meta-analysis, i.e. assuming no or negligible between-study variance, we have $W = \mathrm{diag}(\sigma_1^2/n_1, \ldots, \sigma_k^2/n_k)$. We obtain a feasible generalied least squares estimate by approximating σ_i^2/n_i by s_i^2. We used the 'Simple OLS' implementation available in FSL that provides inference by comparing the statistic to a Student distribution with $(\sum_{i=1}^{k}(n_i - 1)) - p$ degrees of freedom [37] as depicted in Table 1. This reference approach will be referred to as **Fixed-effects (FFX) GLM**.

	Meta-analysic statistic	Nominal H_0 distrib.	Inputs	Assumptions
MFX GLM	$\left(\sum \kappa_i \hat{\beta}_i\right) / \sqrt{\sum_{i=1}^{k} \kappa_i}$ with $\kappa_i = 1/(\hat{\tau}^2 + s_i^2)$	$\mathcal{T}_{k-1}$	$\hat{\beta}_i, s_i^2$	IGE, large sample.
RFX GLM	$\left(\sum_{i=1}^{k} \frac{\hat{\beta}_i}{\sqrt{k}}\right) / \widehat{\sigma_C^2}$	$\mathcal{T}_{k-1}$	$\hat{\beta}_i$	IGE; $\tau^2 + \sigma_i^2 = \sigma_C^2 \ \forall i$
Contrast Perm.	$\left(\sum_{i=1}^{k} \frac{\hat{\beta}_i}{\sqrt{k}}\right) / \widehat{\sigma_C^2}$	Empirical	$\hat{\beta}_i$	ISE.
FFX GLM	$\left(\sum_{i=1}^{k} \frac{\hat{\beta}_i}{s_i^2}\right) / \sqrt{\sum_{i=1}^{k} 1/s_i^2}$	$\mathcal{T}_{(\sum_{i=1}^{k} n_i - 1) - 1}$	$\hat{\beta}_i, s_i^2$	$\tau^2 = 0$, large sample.
Fisher	$-2 \sum_i \ln P_i$	$\chi^2_{(2k)}$	Z_i	$\tau^2 = 0$
Stouffer	$\sqrt{k} \times \frac{1}{k} \sum_i Z_i$	$\mathcal{N}(0,1)$	Z_i	$\tau^2 = 0$
Weighted Z	$\frac{1}{\sqrt{\sum_i n_i}} \sum_i \sqrt{n_i} Z_i$	$\mathcal{N}(0,1)$	Z_i, n_i	$\tau^2 = 0$
Z RFX	$\left(\sum_{i=1}^{k} Z_i\right) / \sqrt{k} \hat{\sigma}$	$\mathcal{T}_{k-1}$	Z_i	IGE; $1 + \tau^2/\sigma_i^2$ cst.
Z Perm.	$\left(\sum_{i=1}^{k} Z_i\right) / \sqrt{k}$	Empirical	Z_i	ISE.

Table 1: Statistics for one-sample meta-analysis tests and their sampling distributions under the null hypothesis H_0. Empirical null distributions are determined using permutations with sign flipping. IGE=Independent Gaussian Errors, ISE=Independent Symmetric Errors. Note: $P_i = \Phi(-Z_i)$, $\widehat{\sigma_C^2}$ is the unbiased sample variance.

Combining contrast estimates. If the s_i^2 are unavailable, the contrast estimates $\hat{\beta}_i$ can be combined by assuming that the within-study contrast variance σ_i^2/n_i is constant ($\sigma_i^2/n_i = \sigma^2 \ \ \forall i$) or negligible in comparison to the between-study variance ($\sigma_i^2/n_i \ll \tau^2$). Then $W = \text{diag}(\sigma_C^2, \ldots, \sigma_C^2)$ where σ_C^2 combines the within and between-study variances, i.e. $\sigma_C^2 \simeq \tau^2$ or $\sigma_C^2 \simeq \tau^2 + \sigma^2$ (note, however, in this setting we do not separately estimate τ^2 or σ^2). Under these assumptions, Eq. (1) can be solved by ordinary least squares giving:

$$\hat{\gamma} = (X^T X)^{-1} X^T \hat{\beta} \quad (6)$$

$$\text{Var}(\hat{\gamma}) = (X^T X)^{-1} \sigma_C^2 \quad (7)$$

Given $\hat{\sigma}_C^2$ the unbiased sample variance, we obtain the statistics presented in Table 1 for one sample tests. This approach will be referred to **Random-effects (RFX) GLM** in the following. Inference can be carried out by comparing the RFX GLM statistic to a Student distribution with $k-1$ degrees of freedom, this result holds asymptotically as well as in small samples [12].

As an alternative to parametric approaches, non-parametric inference [15, 27] can be performed by comparing the one-sample RFX GLM T-statistic to the distribution obtained with "sign flipping", i.e. randomly multiplying each study's data by 1 or -1, justified by an assumption of independent studies and symmetrically distributed random error. For two-sample tests, the non-parameteric distribution can be obtained by random permutation of the group labels. This approach will be referred to as **Contrast permutation**.

Combining standardised statistics. When only test statistic images are available there are a several alternative approaches available. **Fisher**'s meta-analysis provide a statistic to combine the associated p-values [8]. **Stouffer**'s approach combines directly the standardised statistic [36]. In [44] following [20], the author proposed a weighted method that weights each study's Z_i by the square root of its sample size [3,7]. This approach will be referred to as **Weighted Z**. All these meta-analytic statistics assumes no or negligible between-study variance and are only suited for one-sample tests. The corresponding statistics and nominal null distributions are presented in Table 1.

As suggested in [1], to get an RFX-like method based on Stouffer's approach, the standardised statistical estimates Z_i can be combined in an OLS analysis. The corresponding estimate, referred as **Z RFX** is also provided in 1. Non-parametric inference [15, 27] can also be obtained by sign flipping on the Z_i's. This approach will be referred to as **Z permutation**.

2.2. Approximations

In practice, all the methods are either approximate in small samples or rely on assumptions that might not be verified in the context of neuroimaging meta-analyses. Methods based on FGLS (MFX GLM and FFX GLM) have approximate

parametric null distributions. The nominal distributions of RFX GLM and two-sample contrast permutations are under the (unrealistic) assumption of homogeneous standard errors over studies; even if all studies are 'clean' and conducted at the same center, variation in sample size will induce differences in s_i^2's. The fixed-effects methods (Fisher, Stouffer, wieghted Z and FFX GLM) assume homogeneity across studies, i.e. zero between-study variance. All contrast methods (MFX GLM, RFX GLM, Contrast permutation and FFX GLM) require the contrasts to be expressed with in the same units. Finally, RFX Stouffer and Z RFX are ad-hoc methods and their validity has to be investigated.

2.3. Experiments

2.3.1. Validity on null data

We used Monte Carlo simulations to empirically investigate the validity of each estimator. We simulated a set of contrast estimates $\hat{\beta}_i$ and contrast variance estimates s_i^2 according to:

$$\hat{\beta}_i \sim \mathcal{N}(0, \frac{\sigma_i^2}{n} + \tau^2) \tag{8}$$

$$s_i^2 \sim \frac{\sigma_i^2}{n} \frac{\chi^2_{(n-1)}}{n-1} \tag{9}$$

where σ_i^2 was either constant across studies (homoscedasticity) and taken from $\{n \times [0.25, 0.5, 1, 2, 4]\}$ or varying across studies (heteroscedasticity) such as $\max(\sigma_i^2) = \alpha \min(\sigma_i^2)$ with $\alpha \in \{2, 4, 8, 16\}$ and $\text{mean}(\sigma_i^2) = 1$, allowing for 5 different values of σ_i^2 that were linearly spaced and repeated as many times as needed for the specified number of studies k. The between-study variance τ^2 was set to zero (homogeneity) or 1 (heterogeneity). We looked at four different number of studies per meta-analysis $k \in \{5, 10, 25, 50\}$ and two number of subjects $n \in \{20, 100\}$. We set the default number of subjects per studies to $n = 20$ which is a common sample size in existing neuroimaging studies [30]. A total of 144 parameter sets ($9\sigma_i^2 \times 2\tau^2 \times 4k \times 2n_i$) was therefore tested and a total of 1 026 000 realisations was performed for each parameter set.

Three types of analyses were computed: a one-sample meta-analysis testing significance of the mean effect in a group of kstudies, a two-sample meta-analysis testing significance in mean differences between two groups of k studies each and an unbalanced two-sample meta-analysis testing significance in mean differences between two groups of $2/5 \times k$ and $8/5 \times k$ respectively. Two-sample analysis asuumed a common between-study variance across groups.

We conducted those simulations to evaluate the validity of each estimator in small samples and under violations of their assumptions, such as inhomogeneity of contrast variances s_i^2 or presence of non-negligible between-study variance. Furthermore, we studied the robustness of contrast-based methods to the presence of mismatched units across studies. To simulate units mismatch, each contrast estimates $\hat{\beta}_i$ and contrast variance estimates s_i^2 was replaced by a rescaled version: $\hat{\beta}_i^* = \hat{\beta}_i a_i$ and $s_i^{2*} = s_i^2 a_i^2$. 2 types of unit mismatched were investigated:

- Mismatched contrast vectors: a_i linearly sampled between 0.4 and 1.6 (mean 1).
- Mismatched data scaling algorithms (simulating data from different analysis sofware): either 20% or 50% of the studies were rescaled with a factor $a_i = 2$ other studies keeping the original scaling.

In the case of two-sample tests, the same mismatch was applied to both groups.

The simulations were implemented in Matlab R2016b [34] on a High Permformance Computing cluster. Code is available at: `https://github.com/cmaumet/zmeta_buster` and depends on the IBMA toolbox cb5aee [21], SPM12 revision 6196 (RRID:SCR_007037) [10], FSL 5.0.10 (RRID:SCR_002823) [17] and SnPM f254ff (RRID:SCR_002092) [27]. Figures were computed in R [33] jupyter notebooks [29] and depends on ggplot2 [40] and cowplot [41]. The code used to compute the figures displayed in this manuscript is available at: `https://github.com/cmaumet/zmeta`.

2.3.2. Accuracy on real data

We then used Receiver-Operating-Characteristics (ROC) curves to assess the sensibility of each of the nine meta-analytic estimators on real data. We used a dataset of 21 studies investigating pain of which 10 were computed with SPM and 11 with FSL, this dataset is available at: http://neurovault.org/collections/1425/.

Comparability of contrast estimates depends on equivalent scaling of the data, models, and contrast vectors. Data scaling in FSL is performed by setting the median brain intensity to 10,000 while in SPM the mean brain intensity is set to 100. Furthermore, due to imprections of SPM data scaling algorithm, gray matter values tends to be closer to 200 than 100. We therefore resclaled SPM contrast estimates and standard errors by 2 and FSL contrast estimates and standard errors by 100. Furthermore when contrasts were not constructed to preserve units, with sum of positive elements equal to 1, sum of negative elements equal to -1, we further rescale the data to take into account these undesired effects.

We used Neurosynth [43]'s automated large scale coordinate-based meta-analysis of pain as ground truth of areas that should present an effect (cf. http://neurosynth.org/analyses/terms/pain/). For each meta-analytic approach, we estimated the true positive rate over a range of thresholds and combined these values to the false positive rates computed on simulated data to create the ROC curves.

The real data analyses were implemented in Matlab R2016b on a Mac Book Pro laptop. Code is available at: `https://github.com/cmaumet/zmeta_rocs`. Software dependencies and repositories for figures are identical as for simulated analyses. Real data figures were computed in Python 3.6.0 [38] jupyter notebooks and depends on nibabel [3], Scipy [18], numpy [39], scikit-learn [28] and matplotlib [16].

2.3.3. Presence of heterogeneity and heteroscedasticity in real data

To investigate the presence of between-study variations we used Cochran's Q test of heterogeneity [14]:

$$Q = \sum_{i=1}^{i=k} w_i(\hat{\beta}_i - \hat{\theta})^2 \tag{10}$$

where $\hat{\theta}$ is the fixed-effects GLM meta-analytic estimate such that $\hat{\theta} = \frac{\sum_{i=1}^{i=k} w_i \hat{\beta}_i}{\sum_{i=1}^{i=k} w_i}$. The theoretical weights are $w_i = n_i/\sigma_i^2$, but in practice we use $w_i = 1/s_i^2$. The nominal null distribution of Q is χ^2_{k-1}

We also computed the I^2 statistics:

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2} \tag{11}$$

We used the between-study variance estimated by FSL's FLAME [35] and computed $\hat{\sigma}^2$ was computed as the average within-study contrast variance across all studies. Using this metric, voxels with values close to 0 present negligible between-study variance and values close to 1 outline appreciable study heterogeneity and the importance of random-effects models.

3. Results

3.1. Robustness to violation of model assumptions

Fig. 2A presents the one-sample simulation results in small samples, i.e. under small number of studies or small number of subjects. We focus here on methods fow which valitidy is only guaranteed in large samples: FFX GLM and MFX GLM, under ideal conditions otherwise (i.e. τ^2=0 for FFX GLM and τ^2=1 for MFX GLM). When the number of subjects is small, FFX GLM is invalid for all within-study variances investigated, regardless of the number of studies included in the meta-analysis. On the other hand MFX GLM is conservative for small number of studies and constant within-study variance. More surprinsingly, while MFX GLM is valid for constant within-study variances it is invalid in the presence of large variations in the within-study variances, regardless of the number of subjects included in each study.

Fig. 2B presents the one-sample simulation results under heteroscedasticity (σ_i^2 varying across studies). We focus here on methods fow which valitidy is only guaranteed under homoscedasticity: RFX and Contrast permutation, in a sample of 25 studies with 20 subjects each under ideal conditions otherwise (i.e. τ^2=1). RFX GLM and contrast permutation are robust to heteroscedasticity for all settings studied. RFX GLM is closer to nominal. For small P-values, Contrast Permutation is conservative as expected due to the discrete nature of its distribution.

Fig. 2C presents the one-sample simulation results under heterogeneity ($\tau^2 > 0$). We focus here on methods fow which valitidy is only guaranteed under homogeneity: Fisher, Stouffer, Weighted Z and FFX GLM, with a sample of $k = 25$ studies with $n_i = 20$ subjects each. All fixed-effects methods are invalid under heterogeneity.

Similar behaviours are observed for two-sample tests (cf. Supplementary Fig. S1 and Fig. S2).

3.2. Robustness to units mismatch

Fig. 3 presents the simulation results under unit mismatches for one-sample tests. When different scaling algorithm are used (Fig. 3, 2nd and 3rd columns), e.g. with different neuroimaging software packages (provided that differences in scaling targets have been accounted for), Contrast Permutation has a behaviour that is close to nominal. RFX GLM is valid but conservative. MFX GLM is robust to the presence of unit mimatches when the studies are homoscedastic. In the presence of strong heteroscedasticity, MFX GLM remains invalid as when the units are matched due to finite sample inaccuracies (cf. previous section). In the presence of slight heteroscedasticity, unit mismatchs can cause invalidity of othewise valid MFX GLM. When the contrast are scaled differently (Fig. 3, 4th column), we observe a very similar pattern than for different scaling algorithm. Similar behaviours are observed for two-sample tests (cf. Supplementary Fig. S3 and Fig. S4).

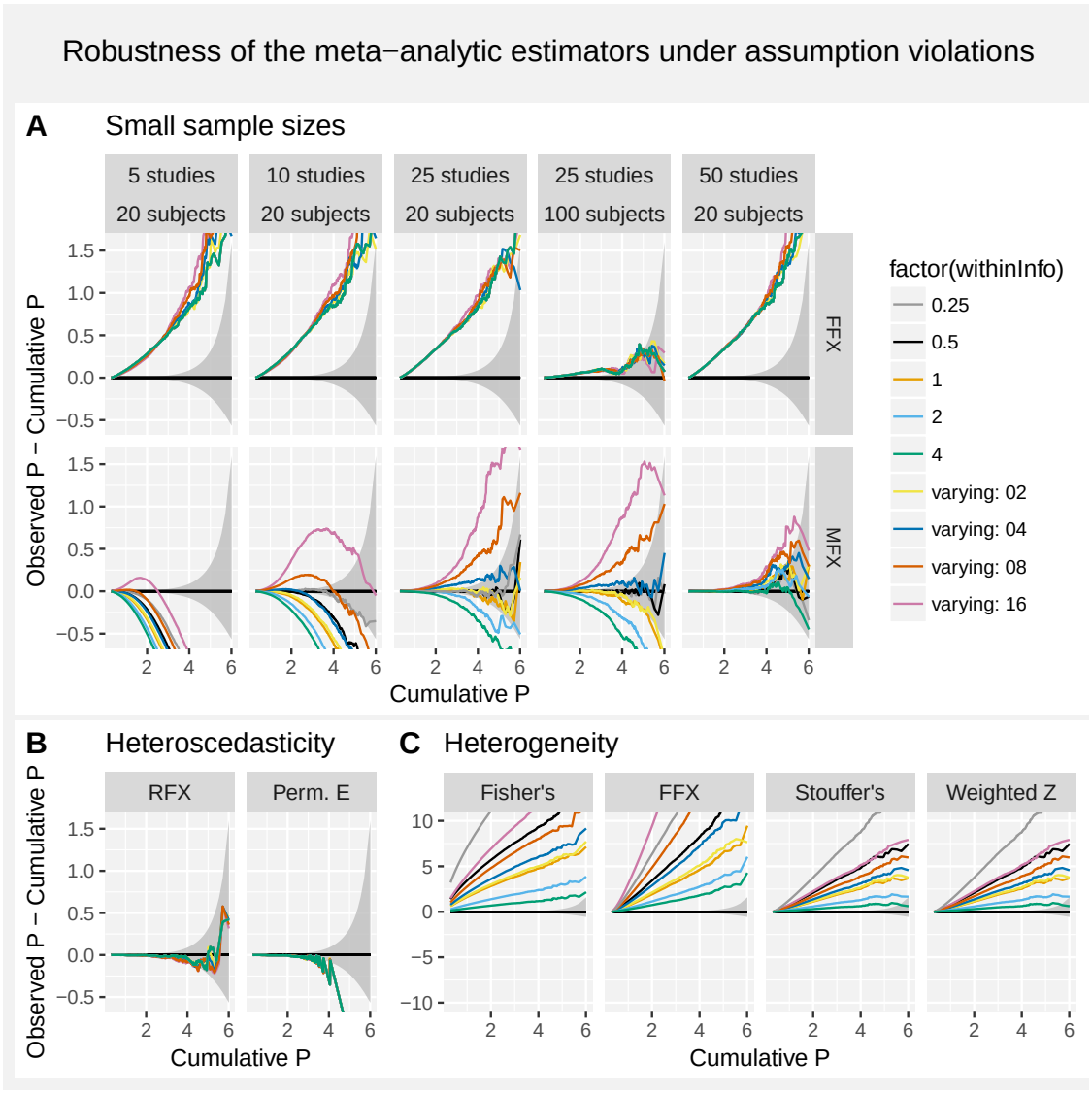

Figure 2: Deviation from theoretical P-values in one-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). P-values are displayed using a negative $\log_{10}$ scale.

3.3. Accuracy on real data

Fig. 4A. displays the ROC curves for all the meta-analytic estimators under varying levels of heterogeneity. As expected, the fixed-effects approaches are the most affected by heterogeneity. Hence, Fisher's method is both the most sensitive method under low heterogeneity and the less sensitive under large heterogeneity. Random-effects approaches are relatively insensitive to the level of heterogeneity. Amongst random-effects approaches the most optimal are Stouffers RFX and Z Permutation that display nearly identical ROC curves, followed by MFX GLM, Contrast Permutation and finally RFX GLM. Differences between MFX GLM and RFX GLM are likely to be non-significant as they both had p-values within the 95% confidence interval.

Fig. 4B. displays the ROC curves for all the meta-analytic estimators under varying levels of heteroscedasticity. Again, the fixed-effects approaches are the most sensitive to heteroscedasticity. This can be explained by the fact that under high heteroscedasticy, some studies will present a low (or high) within-study variance, relatively increasing the between-study variance by comparison to the within-study variance. Amongst the random-effects approaches the most optimal are again Stouffer MFX and Z permutation that display nearly identical ROC curves.

3.3.1. Presence of heterogeneity and heteroscedasticity in real data

Fig. 5 presents the results of Cochran's Q test of heterogeneity on real data. Overall, 88% of the voxels displayed significant heterogeneity at FDR-corrected $P < 0.05$. Fig. 6 displays the I^2 statistic. 27% of the voxels presented

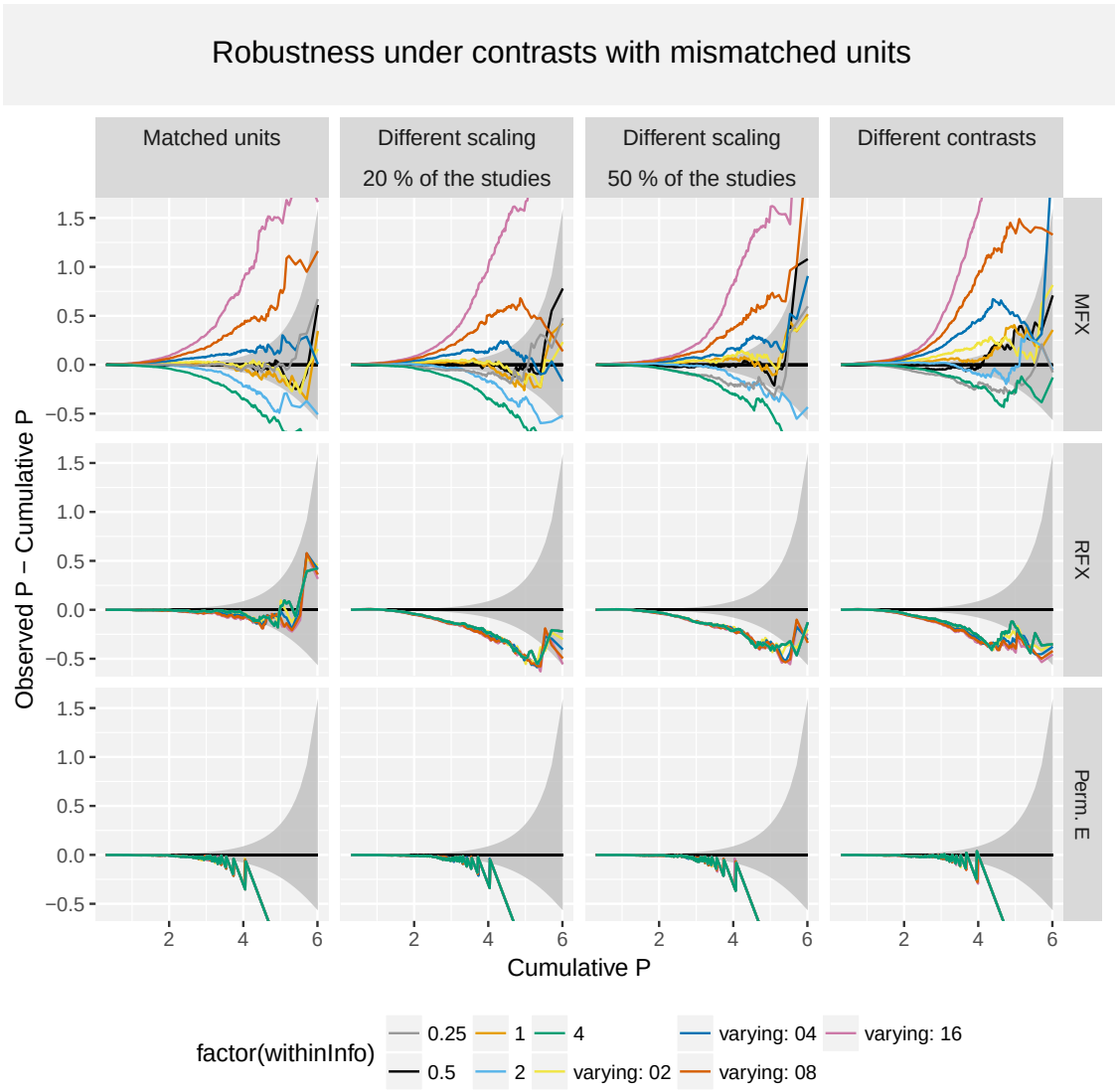

Figure 3: Deviation from theoretical P-values in one-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). Deviation from theoretical Z in two-sample tests with unit mismatch, under ideal circunstances for each statistical approach ($\tau^2 = 1$ and $k = 5, 25, 50$ with matched ("nominal") or mismatched ("different scaling target", "different scaling algorithm", "different contrast vector scaling") units.

a larger between-study variance. Both statistics therefore support the presence of study heterogeneity (non negligible between-study variance) in this collection of studies.

Fig. 7 displays the variability of the study parameter estimate across voxels. This plots supports the presence of heteroscedasticity in this dataset.

4. Discussions

With the growing availability of summary image data for published neuroimaging studies, image-based meta-analysis becomes feasible. Here, we investigated the validity and accuracy of nine meta-analytic estimators under conditions that are typically observed in fMRI and which might invalidate the underlying assumptions of each method.

As expected, fixed-effects methods were shown to be invalid in the presence of heterogeneity. On the other hand, homoscedastic methods were shown to be robust to heteroscedasticity which is in line with published literature on group fMRI statistics [24]. More surprisingly, MFX GLM was showed to be invalid in the presence of large heteroscedasticity due to its approximations not being accurate in small samples.

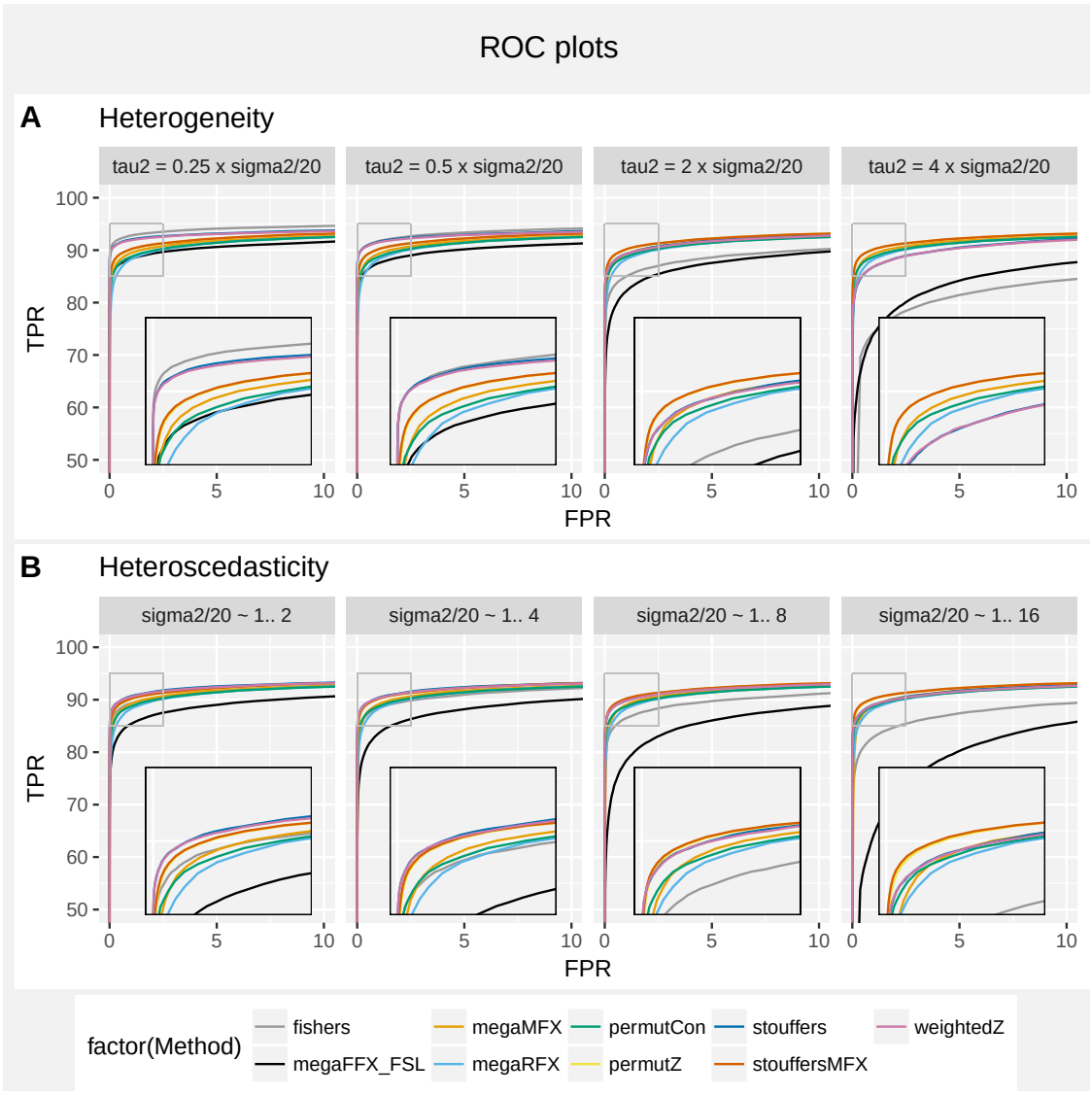

Figure 4: ROC curves of the meta-analytic estimators where true positive rate were computed using a real data meta-analysis of 21 studies of pain and false positive rates using simulated data under various levels of heterogeneity (A) and hetoroscedasticity (B).

In the presence of mismatched units, GLM RFX appeared conservative, while contrast permutation provided the best behaviour, closest to nominal. As confirmed by our real data analysis, we do expect heterogeneity to be present in meta-analytic studies due for instance to variations in the analytic procedure including varying experimental designs, analysis workflows or even due to different imaging instruments. The dataset we used in our real data analysis was created within a single lab and using the same neuroimaging analysis software (FSL), our estimates of heteroscedasticity and heterogeneity are likely to be lower than would be typically observed in a dataset pulling more heterogeneous studies. Similarly, we do expect heteroscedasticity to be present for the same reasons as well as due to varying sample sizes across studies. Finally meta-analysis sample sizes are still relatively small.

Given the still relatively small sample sizes that can be achieved in IBMA as of today, we recommend using RFX GLM, Contrast Permutation, Stouffers MFX or Z Permutation that do not rely on small sample approximations and are robust to both heterogeneity and heteroscedasticity. Unit mismatch across contrasts is likely to be an issue as, even when statistic maps are shared in support of a publication, it is very rare for it to be accompagnied by metadata describing full details on the analysis. Although z-based meta-analyses are suboptimal [6] until full reporting is routinely done, we suggest to use Z-based methods that are insensitive to units.

Because true areas of activations in real data are unknown, we relied on an external source to compute the ground truth activations: the Neurosynth platform which provides very large-scale automatic coordinate-based meta-analysis of the literature. Because this ground truth was determined using a coordinate-based (and not image-based) meta-analysis,

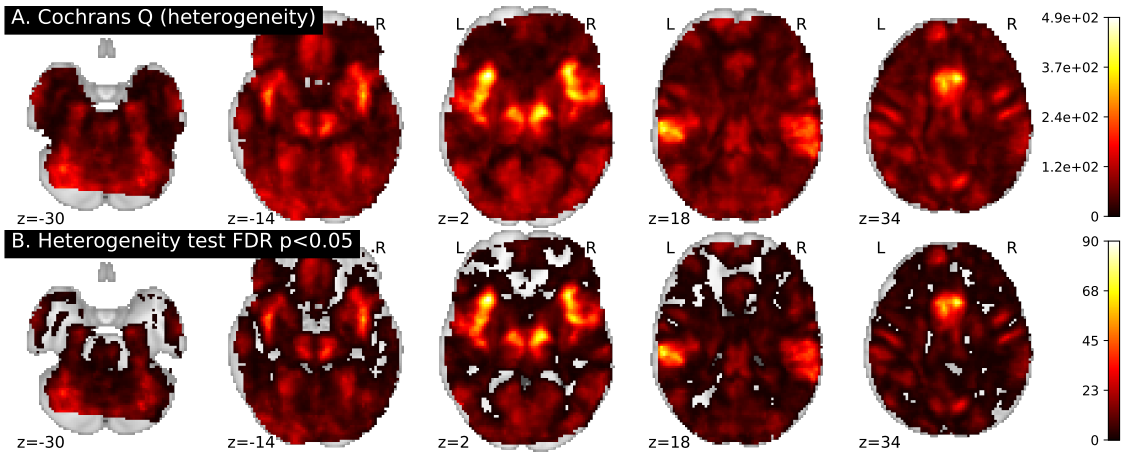

Figure 5: Test of heterogeneity using Cochran's Q test: Q statistic (A) and significant P-values in $-\log_{10}$ at FDR-corrected $P < 0.05$.

it is very likely to be missing some true activated areas with small effect sizes (that would effectively not pass the threshold at the level of a single study). Our estimated true positive rate are therefore likely to be overestimations.

Because true areas of no activations in real data are unknown, we relied on simulated data to compute the false positive rates. We only had a single real dataset for meta-analysis and therefore only computed true positive rates once and combined them with the different false positive rates computed by simulations. In practice the sensibility is very likely to vary under different level of heterogeneity, heteroscedasticity and one would need a larger real dataset to fully investigate sensitivity. The most faithful are the ones that correspond to the settings that are closer to the one observed in the real dataset, i.e. is sigma2/20 1...16.

Finally, in our simulations we investigated heteroscedasticity due to varying within-study variances but not due to varying sample sizes. We do expect the results to be consistent (ignoring the X^2 dof) but would have to be double checked.

5. Conclusion

We have compared nine meta-analytic approaches in the context of one-sample test. Through simulations, we found the expected invalidity of fixed-effects approaches in the presence of study heterogeneity, but also of FFX GLM even with no between-study variation. In a real dataset of 21 studies of pain, there was evidence for substantial between-study variation that supports the use of random-effects meta-analytic statistics. When only contrast estimates are available, RFX GLM was valid. When only standardised estimates are available, permutation is the preferred option as the one providing the most faithful results.

6. Acknowledgements

We gratefully acknowledge the use of the pain dataset from the Tracey pain group, FMRIB, Oxford. The majority of this work was conducted while TEN and CM were at the University of Warwick and used the High Performance Computing cluster of the Departement of Statistics, University of Warwick.

7. References

[1] Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. *NeuroImage*, 45(3):810–23, 2009.

[2] Alawi A. Alsheikh-Ali, Waqas Qureshi, Mouaz H. Al-Mallah, and John P. A. Ioannidis. Public Availability of Published Research Data in High-Impact Journals. *PLOS ONE*, 6(9):e24357, September 2011.

[3] Matthew Brett, Michael Hanke, Ben Cipollini, Marc-Alexandre Côté, Chris Markiewicz, Stephan Gerhard, Eric Larson, Gregory R. Lee, Yaroslav Halchenko, Erik Kastman, and et al. nibabel: 2.1.0. Aug 2016.

[4] Katherine S Button, John P a Ioannidis, Claire Mokrysz, Brian a Nosek, Jonathan Flint, Emma S J Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews. Neuroscience*, 14(5):365–76, 2013.

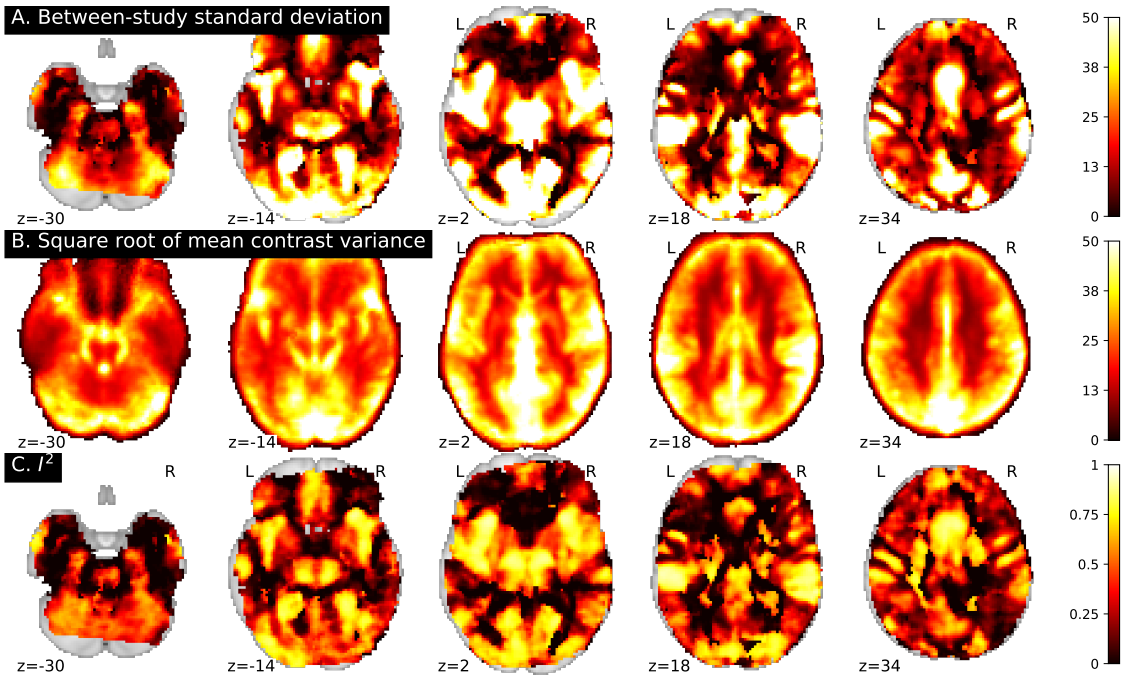

Figure 6: Estimated between-study variance (A), estimated average within-study contrast variance (B) and ratio of the estimated between study-variance onto total variance (i.e. I^2) (C).

[5] Gang Chen, Paul A Taylor, and Robert W Cox. Is the Statistic Value All We Should Care about in Neuroimaging? *bioRxiv*, (September):064212, 2016.

[6] P. Cummings. Meta-analysis based on standardized effects is unreliable. *Archives of pediatrics & adolescent medicine*, 158(6):595–7, 2004.

[7] Lloyd T. Elliott, Kevin Sharp, Fidel Alfaro-Almagro, Sinan Shi, Karla L. Miller, Gwenaëlle Douaud, Jonathan Marchini, and Stephen M. Smith. Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature*, 562(7726):210, October 2018.

[8] R.A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1932.

[9] Peter T. Fox and Jack L. Lancaster. Mapping context and content: the BrainMap model. *Nature Reviews Neuroscience*, 3(4):319–321, April 2002.

[10] K.J. Friston, J. Ashburner, S.J. Kiebel, T.E. Nichols, and W.D. Penny, editors. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press, 2007.

[11] 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, Tal Yarkoni, and Daniel S. Margulies. NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in Neuroinformatics*, 9(8), April 2015.

[12] WH William H . Greene. *Econometric analysis*, volume 97. 2012.

[13] Annamaria Guolo and Cristiano Varin. Random-effects meta-analysis: the number of studies matters. *Statistical Methods in Medical Research*, 26(3):1500–1518, June 2017.

[14] Julian Higgins and Simon G Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, 21(11):1539–1558, 2002.

[15] A. P. Holmes, R. C. Blair, G. J.D. Watson, and I. Ford. Nonparametric Analysis of Statistic Images from Functional Mapping Experiments. *Journal of Cerebral Blood Flow and Metabolism*, 16:7–22, 1996.

[16] John D Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

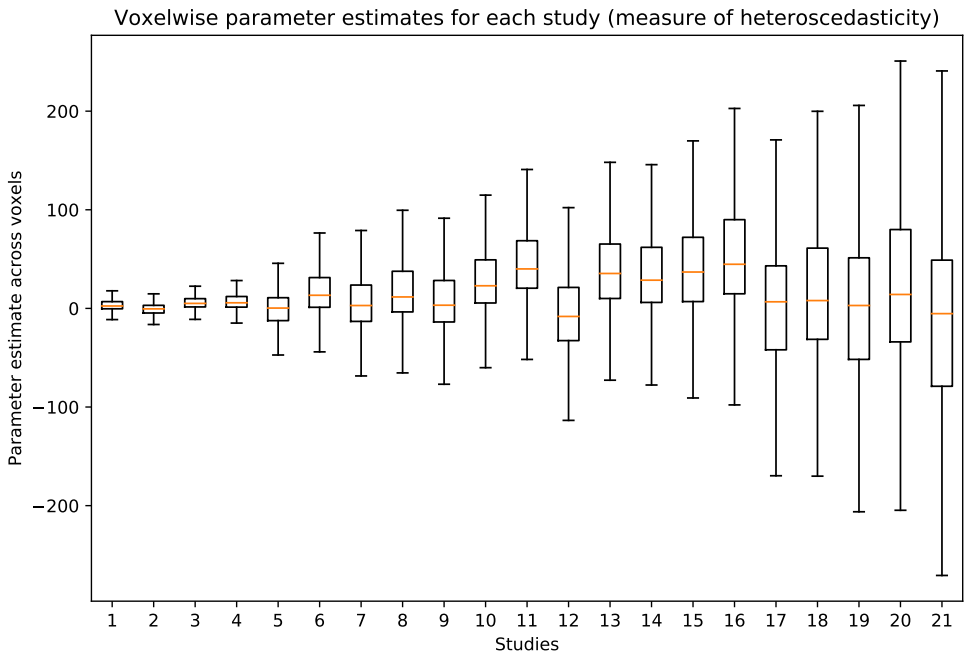

Figure 7: Boxplot of the parameter estimates across voxels by studies. For each box, the orange line corresponds to the median and the top and bottom lines of the black square are the upper and lower quartiles of the distribution. The whiskers cover the data points that are located up to 1.5 times the inter-quartile distance. Points falling out of the whiskers are not displayed. Studies are sorted from smaller to larger interquartile ranges.

[17] Mark Jenkinson, Christian F. Beckmann, Timothy E.J. Behrens, Mark W. Woolrich, and Stephen M. Smith. Fsl. *NeuroImage*, 62(2):782 – 790, 2012. 20 YEARS OF fMRI.

[18] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed ¡today¿].

[19] I. Kosmidis, A. Guolo, and C. Varin. Improving the accuracy of likelihood-based inference in meta-analysis and meta-regression. *Biometrika*, 104(2):489–496, June 2015.

[20] T. Liptak. On the combination of independent tests. *Magyar Tud. Akad. Mat. Kutato Int. Kozl.*, 3:171–197, 1958.

[21] Camille Maumet and Thomas E. Nichols. IBMA: An SPM toolbox for NeuroImaging Image-Based Meta-Analysis. In *7th INCF Congress of Neuroinformatics*, volume 8, Leiden, Netherlands, August 2014.

[22] Maarten Mennes, Bharat B. Biswal, F. Xavier Castellanos, and Michael P. Milham. Making data sharing work: The FCP/INDI experience. *NeuroImage*, 82:683 – 691, 2013.

[23] Karla L. Miller, Fidel Alfaro-Almagro, Neal K. Bangerter, David L. Thomas, Essa Yacoub, Junqian Xu, Andreas J. Bartsch, Saad Jbabdi, Stamatios N. Sotiropoulos, Jesper L. R. Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Thomas W. Okell, Peter Weale, Iulius Dragonu, Steve Garratt, Sarah Hudson, Rory Collins, Mark Jenkinson, Paul M. Matthews, and Stephen M. Smith. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, 19(11):1523–1536, November 2016.

[24] J. A. Mumford and T. E. Nichols. Simple group fMRI modeling and inference. *NeuroImage*, 47(4):1469–75, 2009.

[25] Thomas E. Nichols. Spm plot units, 2012.

[26] Thomas E Nichols, Samir Das, Simon B Eickhoff, Alan C Evans, Tristan Glatard, Michael Hanke, Nikolaus Kriegeskorte, Michael P Milham, Russell A Poldrack, Jean-baptiste Poline, Erika Proal, Bertrand Thirion, David C Van Essen, Tonya White, and B T Thomas Yeo. Best practices in data analysis and sharing in neuroimaging using MRI. *Nature Neuroscience*, 20(3):299–303, feb 2017.

[27] Thomas E. Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25, jan 2002.

[28] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[29] Fernando Pérez and Brian E. Granger. IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29, May 2007.

[30] 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. *Nature Publishing Group*, 2017.

[31] Russell A. Poldrack and Krzysztof J Gorgolewski. Making big data open : data sharing in neuroimaging. *Nature neuroscience*, 17(11), 2014.

[32] Jean-Baptiste Poline, Janis L. Breeze, Satrajit Ghosh, Krzysztof Gorgolewski, Yaroslav O. Halchenko, Michael Hanke, Christian Haselgrove, Karl G. Helmer, David B. Keator, Daniel S. Marcus, Russell A. Poldrack, Yannick Schwartz, John Ashburner, and David N. Kennedy. Data sharing in neuroimaging research. *Frontiers in Neuroinformatics*, 6, April 2012.

[33] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[34] MATLAB Release 2016b. The mathworks. *Inc., Natick, Massachusetts, United States*, 488, 2016.

[35] S. Smith, P. R. Bannister, C. Beckmann, M. Brady, S. Glare, D. Flitney, P. Hansen, M. Jenkinson, D. Leibovici, B. Ripley, M. Woolrich, and Y. Zhang. FSL : New Tools for Functional and Structural Brain Image Analysis. (6):2001, 2001.

[36] S. Stouffer, L. DeVinney, and E. Suchmen. *The American Soldier: Adjustment During Army Life*, volume 1. Princeton University Press, Princeton, NJ, 1949.

[37] unknown. Feat/userguide - fslwiki, 2017.

[38] Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[39] Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

[40] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

[41] Claus O. Wilke. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*, 2016. R package version 0.7.0.

[42] Mark W. Woolrich, Timothy E J Behrens, Christian F. Beckmann, Mark Jenkinson, and Stephen M Smith. Multilevel linear modelling for FMRI group analysis using Bayesian inference. *NeuroImage*, 21(4):1732–47, apr 2004.

[43] Tal Yarkoni, Russell A. Poldrack, Thomas E. Nichols, David C. Van Essen, and Tor D. Wager. Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665–670, August 2011.

[44] D V Zaykin. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of evolutionary biology*, 24(8):1836–41, 2011.

	$\hat{\gamma_1}$	$\text{Var}(\hat{\gamma_1})$
Random effects	$\left(\sum \eta_i \hat{\beta}_i\right) / \left(\sum \eta_i\right)$ with $\eta_i = 1/(\tau^2 + \sigma_i^2/n_i)$	$1/\sum \eta_i$
Fixed effects	$\left(\sum \hat{\beta}_i \times n_i/\sigma_i^2\right) / \left(\sum n_i/\sigma_i^2\right)$	$1/(\sum n_i/\sigma_i^2)$

Table S1: One-sample weighted least squares (WLS) estimates and their sampling distributions for random-effects and fixed-effects meta-analyses. The FGLS estimates and assumed sampling distributions are obtained by substituing: $\tau^2 \leftarrow \hat{\tau}^2$ and $\sigma_i^2/n_i \leftarrow s_i^2$.

	$\hat{\gamma_1}$ - $\hat{\gamma_2}$	$\text{Var}(\hat{\gamma_1}$ - $\hat{\gamma_2})$
Random effects	$\left(\sum_{i \in G_1} \eta_i \hat{\beta}_i\right) / \left(\sum_{i \in G_1} \eta_i\right) - \left(\sum_{i \in G_2} \eta_i \hat{\beta}_i\right) / \left(\sum_{i \in G_2} \eta_i\right)$ with $\eta_i = 1/\left(\tau^2 + \sigma_i^2/n_i\right)$	$1/\sum_{i \in G_1} \eta_i + 1/\sum_{i \in G_2} \eta_i$
Fixed effects	$\left(\sum_{i \in G_1} \hat{\beta}_i \times n_i/\sigma_i^2\right) / \left(\sum_{i \in G_1} n_i/\sigma_i^2\right) - \left(\sum_{i \in G_2} \hat{\beta}_i \times n_i/\sigma_i^2\right) / \left(\sum_{i \in G_2} n_i/\sigma_i^2\right)$	$1/\left(\sum_{i \in G_1} n_i/\sigma_i^2\right) + 1/\left(\sum_{i \in G_2} n_i/\sigma_i^2\right.$

Table S2: Two-sample weighted least squares (WLS) estimates and their sampling distributions for random-effects and fixed-effects meta-analyses. The FGLS estimates and assumed sampling distributions are obtained by substituing: $\tau^2 \leftarrow \hat{\tau}^2$ and $\sigma_i^2/n_i \leftarrow s_i^2$.

	$\hat{\gamma_1}$	$\text{Var}(\hat{\gamma_1})$
RFX GLM	$\sum \hat{\beta}_i/k$	σ_C^2/k
Z MFX	$\sum Z_i/k$	σ_C^2/k

Table S3: For the meta-analytic approaches based on Ordinary Least Squares (OLS), one-sample meta-analytic estimates and sampling variances. Note that σ_C^2 will be different for each row of this table.

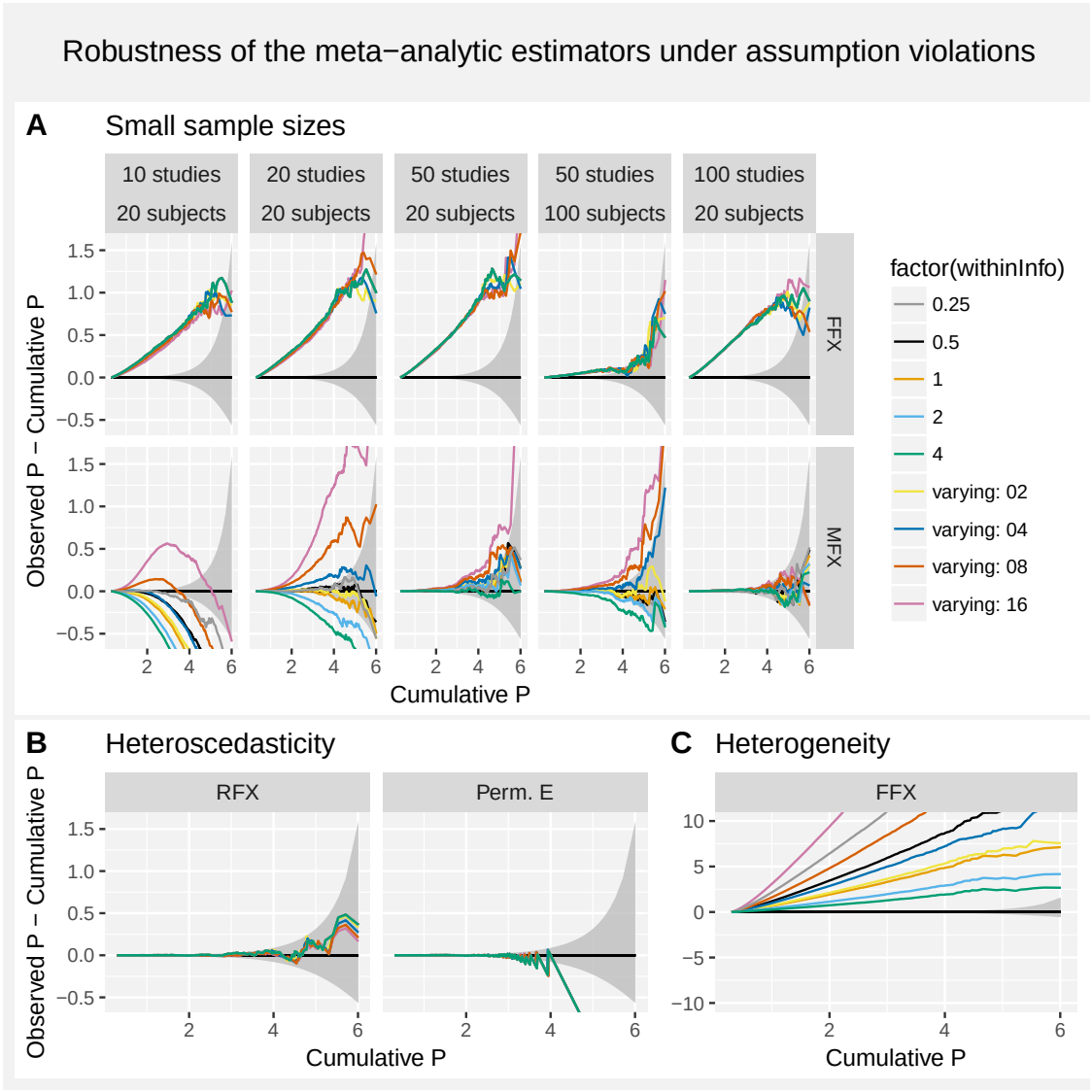

Figure S1: Deviation from theoretical P-values in two-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). P-values are displayed using a negative $\log_{10}$ scale.

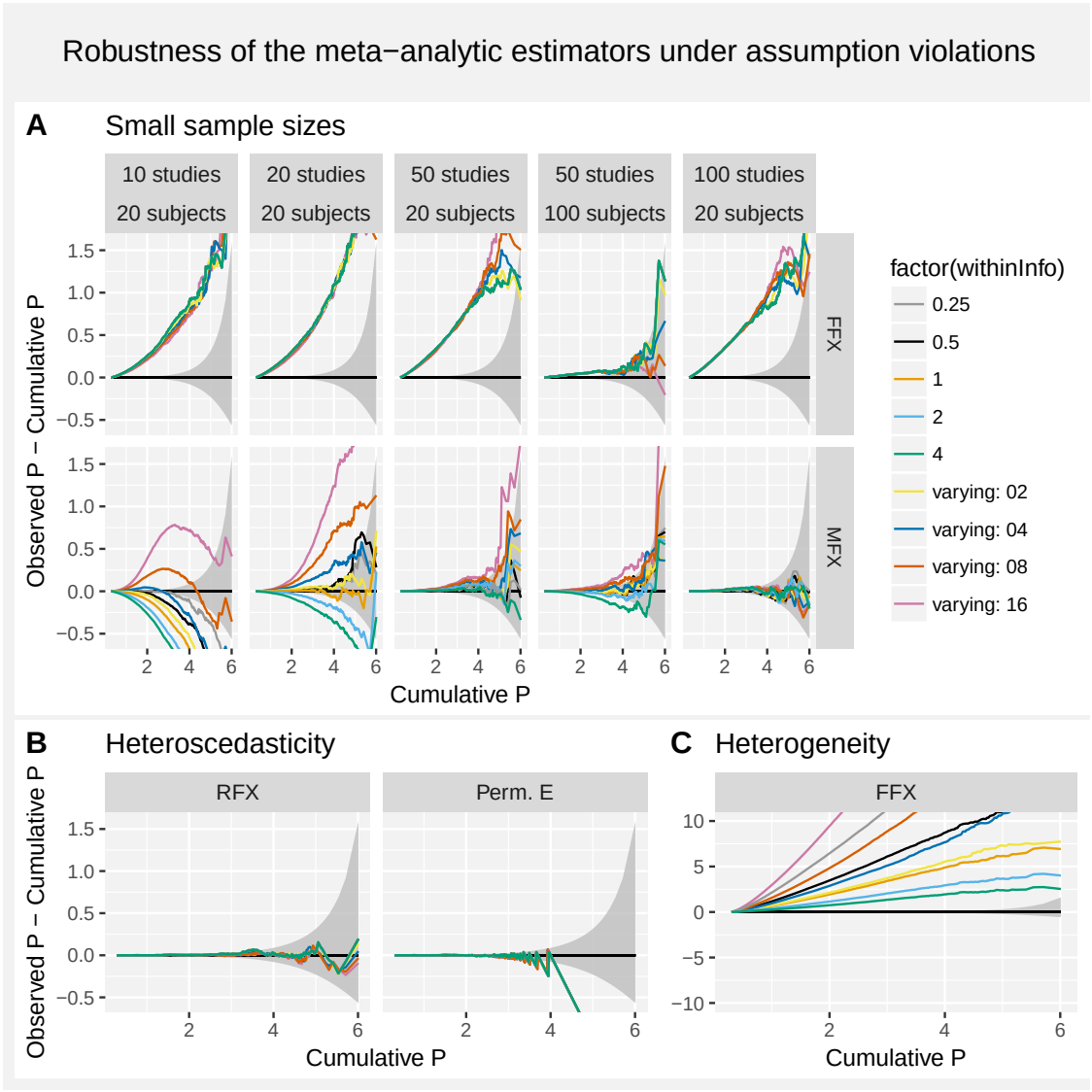

Figure S2: Deviation from theoretical P-values in unbalanced two-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). P-values are displayed using a negative $\log_{10}$ scale.

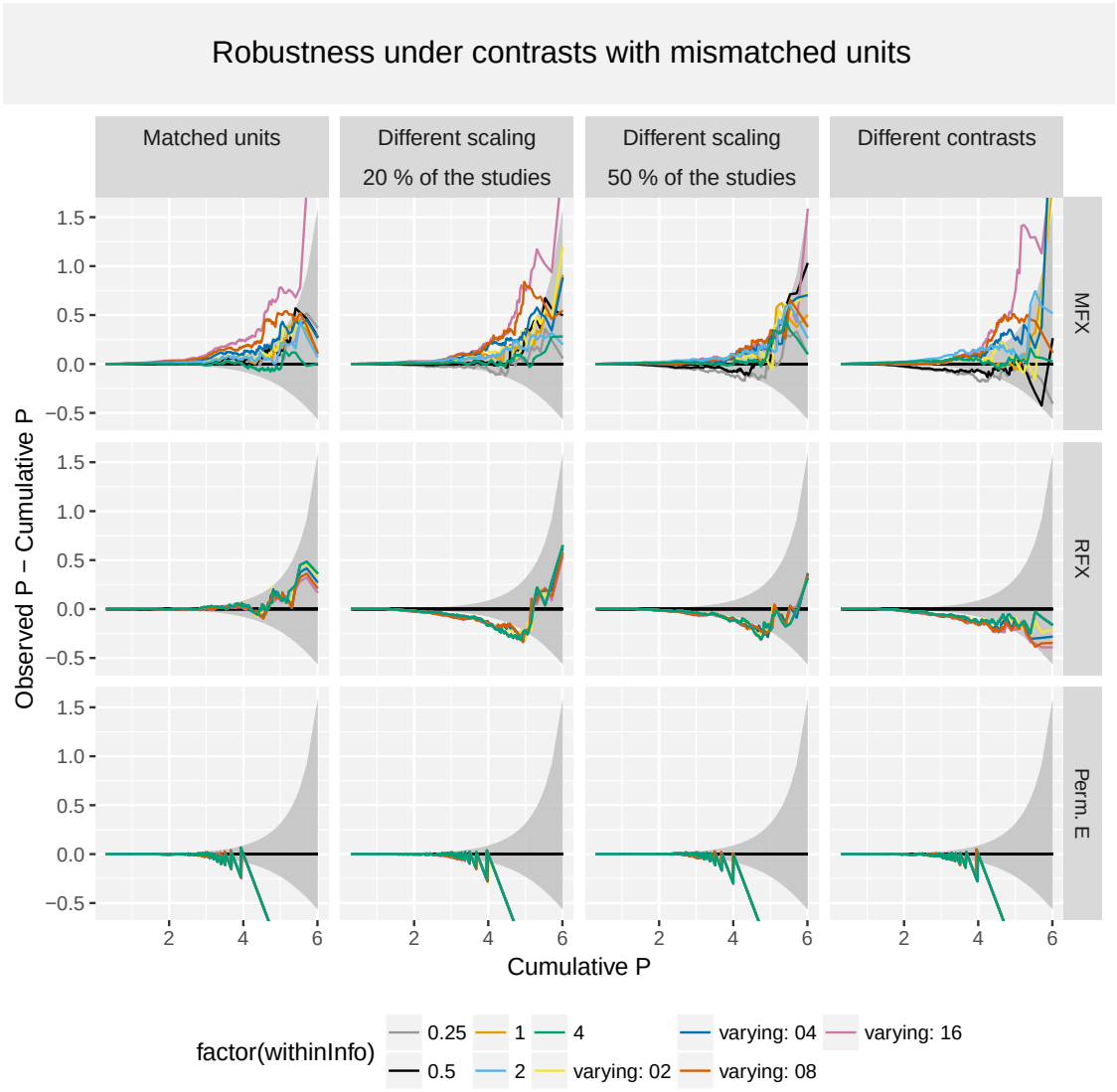

Figure S3: Deviation from theoretical P-values in one-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). Deviation from theoretical Z in two-sample tests with unit mismatch, under ideal circunstances for each statistical approach ($\tau^2 = 1$ and $k = 5, 25, 50$ with matched ("nominal") or mismatched ("different scaling target", "different scaling algorithm", "different contrast vector scaling") units.

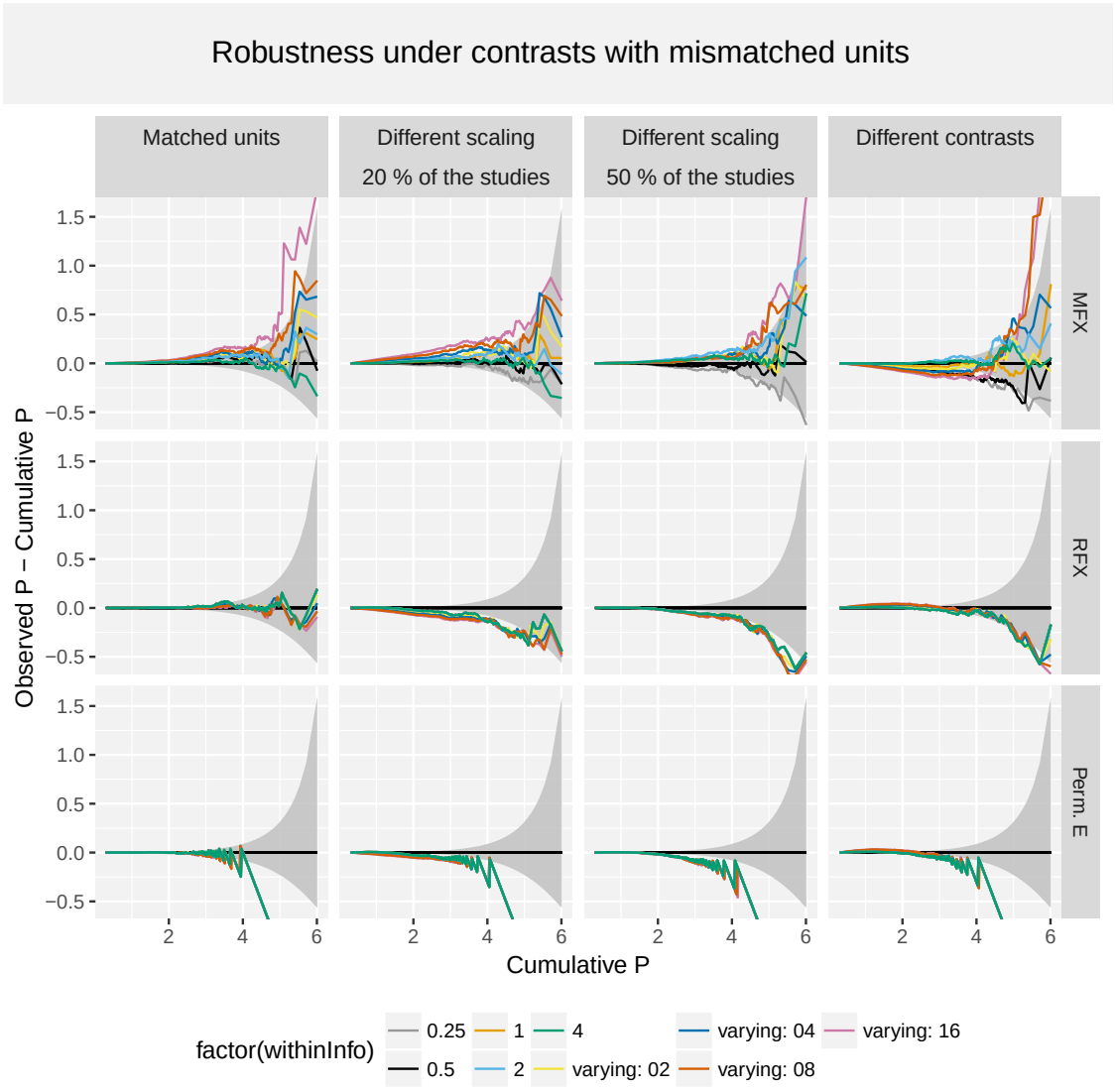

Figure S4: Deviation from theoretical P-values in one-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). Deviation from theoretical Z in two-sample tests with unit mismatch, under ideal circunstances for each statistical approach ($\tau^2 = 1$ and $k = 5, 25, 50$ with matched ("nominal") or mismatched ("different scaling target", "different scaling algorithm", "different contrast vector scaling") units.