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). When image data is available for each study, 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

A growing literature is focusing on the lack of statistical power in neuroimaging studies (see, e.g. [2]), feeding the debate on the validity and reproducibility of published neuroimaging results. Meta-analysis, by providing inference based on the results of previously conducted studies, provides an essential method to increase power and hence confidence in neuroimaging.

A number of methods have been proposed for neuroimaging meta-analysis (see [14] for a review). As the results of neuroimaging studies are usually conveyed by providing a table of peak coordinate and statistics, most of these meta-analyses are restricted to combining coordinate-based information. Nevertheless the best practice method is an Image-Based Meta-Analysis (IBMA) that combines the effect estimates and standard errors from each study [1].

In order for IBMA to be possible in neuroimaging, tools for sharing 3D volumes obtained as a result of a statistical analysis are needed. NeuroVault [7] is an example of one such plateform which facilitates sharing of neuroimaging results data but emphasis is mainly on statistical maps. There are three evident approaches to sharing summary data from each study i:

1. the contrast estimates $\hat{\beta}_i$ and contrast variance estimates s_i^2.
2. the contrast estimates $\hat{\beta}_i$.
3. the standardized statistic maps Z_i.

Depending on how much data is shared, different strategies can be used to combine the available results into a meta-analysis. While the first option is the best practice, leading to statistically optimal estimates [5], it requires the contrasts to be expressed with in the same units and inference relies on asymptotic results (i.e under large sample sizes). In fMRI, units will depends on the field strength [4] as well as data, model and contrast vector scaling [12] and the number of samples included in a meta-analysis is usually small.

Given the growing interest in data sharing in the neuroimaging community, and the relative easiness of sharing and combining just (unitless) statistic maps, there is a need to identify what is the minimal data to be shared in order to allow for future IBMA.

Here we compare the use of IMBA using 9 meta-analytic approaches: 2 approaches use $\hat{\beta}_i$'s and s_i^2's, 2 $\hat{\beta}_i$'s only and 5 Z_i's. We compare the validity and the accuracy of the nine meta-analytic approaches on simulated and real data including 21 studies of pain in control subjects.

Section 2 describes the meta-analytic estimates along with the experiments undertaken on simulated and real data to assert their validity. The results are described in section 3. Finally, we conclude in section 5.

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

	$\hat{\gamma}$	$\mathrm{Var}(\hat{\gamma})$	Assumptions
MFX GLM	$\left(\sum \kappa_i \hat{\beta}_i\right) / \left(\sum \kappa_i\right)$ with $\kappa_i = 1/(\hat{\tau}^2 + s_i^2)$	$1/\sum \kappa_i$	IGE.
RFX GLM	$\sum \hat{\beta}_i / k$	σ_C^2/k	IGE; $\tau^2 + \sigma_i^2 = \sigma_C^2 \ \forall i$
FFX GLM	$\left(\sum \hat{\beta}_i \times n_i/\sigma_i^2\right) / \left(\sum n_i/\sigma_i^2\right)$	$1/(\sum n_i/\sigma_i^2)$	IGE; $\tau^2 = 0$.
Contrast Perm.	$\sum \hat{\beta}_i / k$	Empirical	ISE.
Z MFX	$\sum Z_i / k$	σ_C^2/k	IGE; $1 + \tau^2/\sigma_i^2$ cst.
Z Perm.	$\left(\sum_{i=1}^{k} Z_i\right) / \sqrt{k}$	Empirical	ISE.

Table 1: One-sample meta-analytic estimates, sampling variance and associated assumptions. Note: $P_i = \Phi(-Z_i)$

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 gold standard approach is to fit contrast estimates and their standard error with a hierarchical general linear model (GLM) [5], creating a third-level (level 1: subject; level 2: study; level 3: meta-analysis). The general formulation for the study-level data is:

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

where γ is the meta-analytic parameter to estimate, $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1 \ldots \hat{\beta}_k]^T$ is the vector of contrast estimates, X is the $k \times p$ study-level matrix (typically just a column of ones for a one-sample test) and $\epsilon \sim \mathcal{N}(0, W)$ is the residual error term.

In the most general case of a random-effects (RFX) meta-analysis, 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 contrast variance for study i. Eq. (1) can be solved by weighted least squares giving:

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

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:

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

Approximating σ_i^2/n_i by s_i^2 and given $\hat{\tau}^2$ an estimate of τ^2 we obtain the estimate detailed in Table 1 for a one-sample test. Asymptotic theory shows that inference can be carried out by comparing the statistic to a Student distribution with $k-1$ degrees of freedom $+ADD_REF$ as depicted in Table 2. This reference approach will be referred to as **Mixed-effects (MFX) GLM**.

In a **fixed-effects (FFX) GLM**, i.e. assuming no or negligible between-study variance, $W = \mathrm{diag}(\sigma_1^2/n_1, \ldots, \sigma_k^2/n_k)$. Approximating σ_i^2/n_i by s_i^2 we obtain the feasible generalied least squares estimate detailed in Table 1 for a one-sample test. Asymptotic theory shows that inference can be carried out by comparing the statistic to a Student distribution with $(\sum_i n_i * k) - 1$ degrees of freedom $+ADD_REF$ as depicted in Table 2.

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 $W = \mathrm{diag}(\sigma_C^2, \ldots, \sigma_C^2)$ where σ_C^2 is the combined within and between-subject variance, 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:

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

Table 2: 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.

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

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

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 **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 $+ADD_REF$.

As an alternative to parametric approaches, non-parametric inference [8, 11] 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 [6]. **Stouffer**'s approach combines directly the standardised statistic [17]. In [18] following [9], 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 Stouffer's**. All these meta-analytic statistics assumes no or negligible between-study variance and are only suited for one-sample tests. The corresponding statistics are presented in Table 2. As suggested in [1], to get a kind of MFX with Stouffer's approach, the standardised statistical estimates Z_i can be combined in an OLS analysis. The corresponding estimate, referred as **Z MFX** is also provided in 2

Non-parametric inference [8, 11] can also be obtained by sign flipping on the Z_i's. This approach will be referred to as **Z permutation**.

Approximations. In practice, methods based on FGLS (MFX and FFX) have approximate parametric null distributions. The nominal distributions of RFX 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. Finally, all contrast methods (MFX, RFX, Contrast permutation and FFX) require the contrasts to be expressed with in the same units.

2.2. Experiments

2.2.1. Simulations

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 $\sigma_i^2 = \sigma^2 \alpha_i$ with $\sigma^2 \in n \times [0.25, 0.5, 1, 2, 4]$ and α_i is either equal to 1 for all studies or is taken from $\alpha_i \in [1, 2, 4, 8, 16]$ to simulate varying within-study variances, $\tau^2 \in [0, 1]$ is the between-study variance. Four different number of studies per meta-analysis were used: $k \in [5, 10, 25, 50]$. We set the number of subjects per studies $n = 20$ which is a common sample size in existing neuroimaging studies [13]. A total of 72 parameter sets (9 σ_i^2 x 2 τ^2 x 4 k) 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 on the mean effect in 1 group of k), a two-sample meta-analysis (testing significance in mean differences between two groups of k each) and an unbalanced two-sample meta-analysis (testing significance in mean differences between two groups of $2*k/5$ and $2*k*4/5$ respectively).

We conducted simulations to evaluate the validity of each estimator in small samples and under violations of their assumptions, namely inhomogeneity of contrast variances s_i^2, 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: $\hat{\beta}_i i^* = \hat{\beta}_i a_i$ and $s_i^{2*} = s_i^2 a_i^2$. 2 types of unit mismatched were investigated:

- Mismatch in scaling of the contrast vectors: a_i linearly sampled between 0.4 and 1.6 (mean is 1).
- Mismatch in data scaling from different sofware: for each group a_i=1 for $i <= n_{soft1}$ a_i=1 for $i > n_{soft1}$ with $n_{soft1} \in \{1/5, 1/2\} * n_g$ where n_g is the number of studies in the group.

The full set of simulations is summarised is table **??**. Code is available at: `https://github.com/cmaumet/zmeta_buster`.

2.2.2. Real data

We then compared the nine meta-analytic estimators to the reference approach, MFX GLM, on a dataset of 21 studies of pain. Comparability of contrast estimates depends on equivalent scaling of the data, models, and contrast vectors. Data scaling was consistently performed by FSL, setting median brain intensity to 10,000; model were all created by FSL's Feat tool; and contrasts were constructed to preserve units, with sum of positive elements equal to 1, sum of negative elements equal to -1.

To investigate the presence of between-study variation, we computed the ratio of the between-study variance (estimated using FSL's FLAME [16]) to the total variance (sum of between- and within-study variances), as suggested in [3]. Here we use the average (across study) within-study variance as an estimate of within-study variance in the denominator: $\hat{\tau}^2/(\hat{\tau}^2 + \sum_{i=1}^k s_i^2)$. 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 RFX models.

Then for each estimator we compared the standardised meta-analytic statistic to the z-statistic obtained with the reference approach. Overestimation of z-statistic leads to overly optimistic detections while underestimation outline a reduced sensitivity of the approach.

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 and MFX, under ideal conditions otherwise (i.e. τ^2=0 for FFX and τ^2=1 for MFX). When the number of subjects is small, FFX 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. Surprinsingly, while

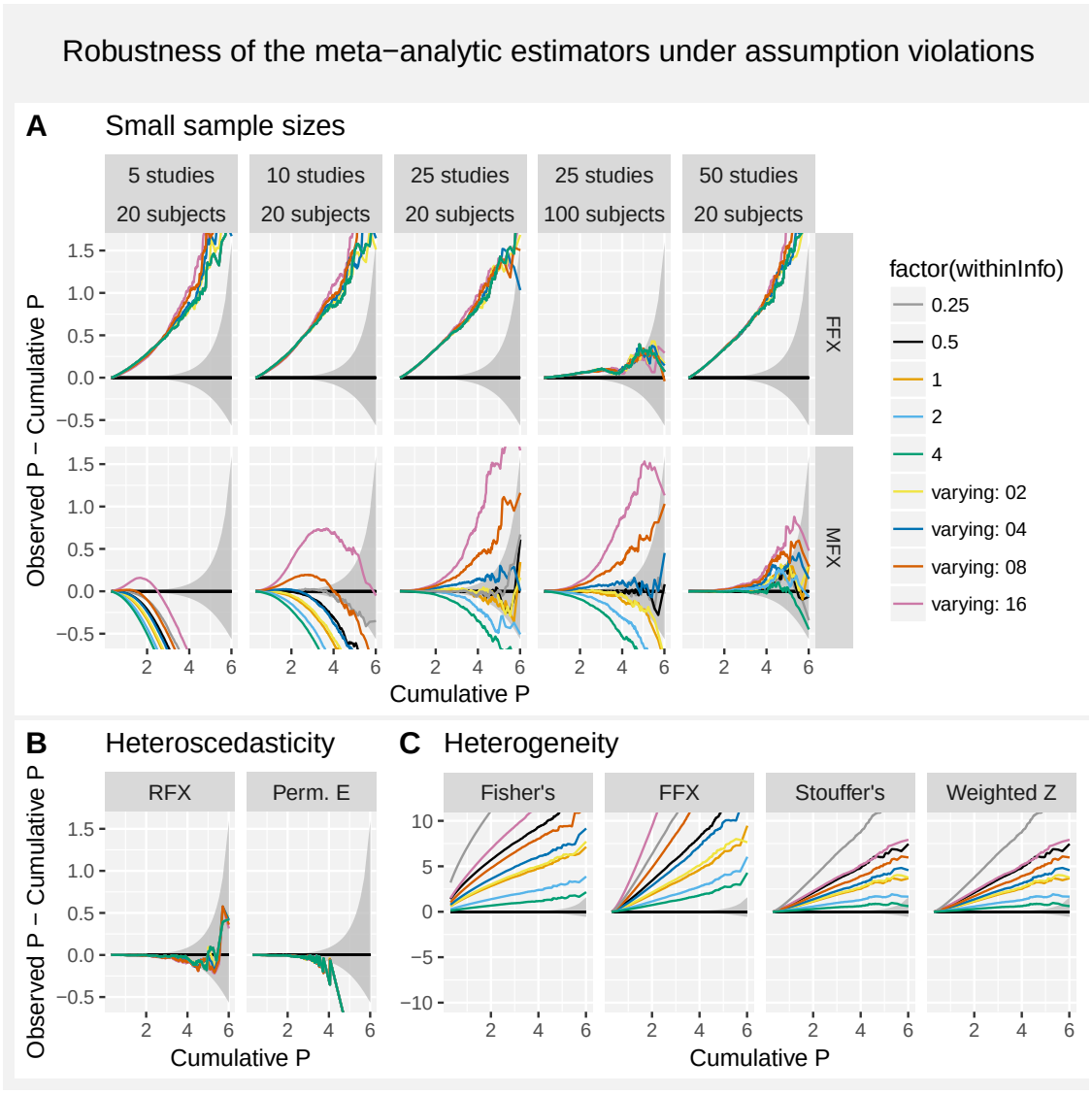

Figure 2: Deviation from theoretical P-values in one-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). P-values are displayed using a negative $\log_{10}$ scale.

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.

In the nominal case, i.e. when the units are matched across studies and contrasts, RFX GLM and Contrast Permutation are valid, as expected. For small P-values, Contrast Permutation is conservative as expected due to the discrete nature of its distribution. In the presence of a high within-study variance, MFX GLM also appears to be conservative. RFX GLM displays the best behaviour with a pattern that is within the 95% confidence interval of the theoretical Z for all within-study variance studied and only slightly conservative when the within-study variances are varying across studies.

For small number of studies, permutation methods (including Contrast Permutation and Z permutation) are conservative as expected due to the discrete nature of their distribution (cf. supplementary figure TODO).

Other approaches (TODO) have a nominal behaviour under small sample sizes as expected according to theory. Only Stouffers MFX presents some invalidity, which can be explained by the fact that this is an had hoc not recommended statistic (cf. supplementary figure TODO).

Fig. 2B presents the one-sample simulation results under heteroscedasticity (σ_i^2 non constant over studies). We focus here on methods fow which valitidy is only guaranteed under homoscedasticity: RFX and Contrast permutation, with a sample of 25 studies under ideal conditions otherwise (i.e. τ^2=1). RFX and Perm E. are robust to heteroscedasticity for all settings studied. RFX 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 Fixed-effects GLM, with a sample of 25 studies. 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

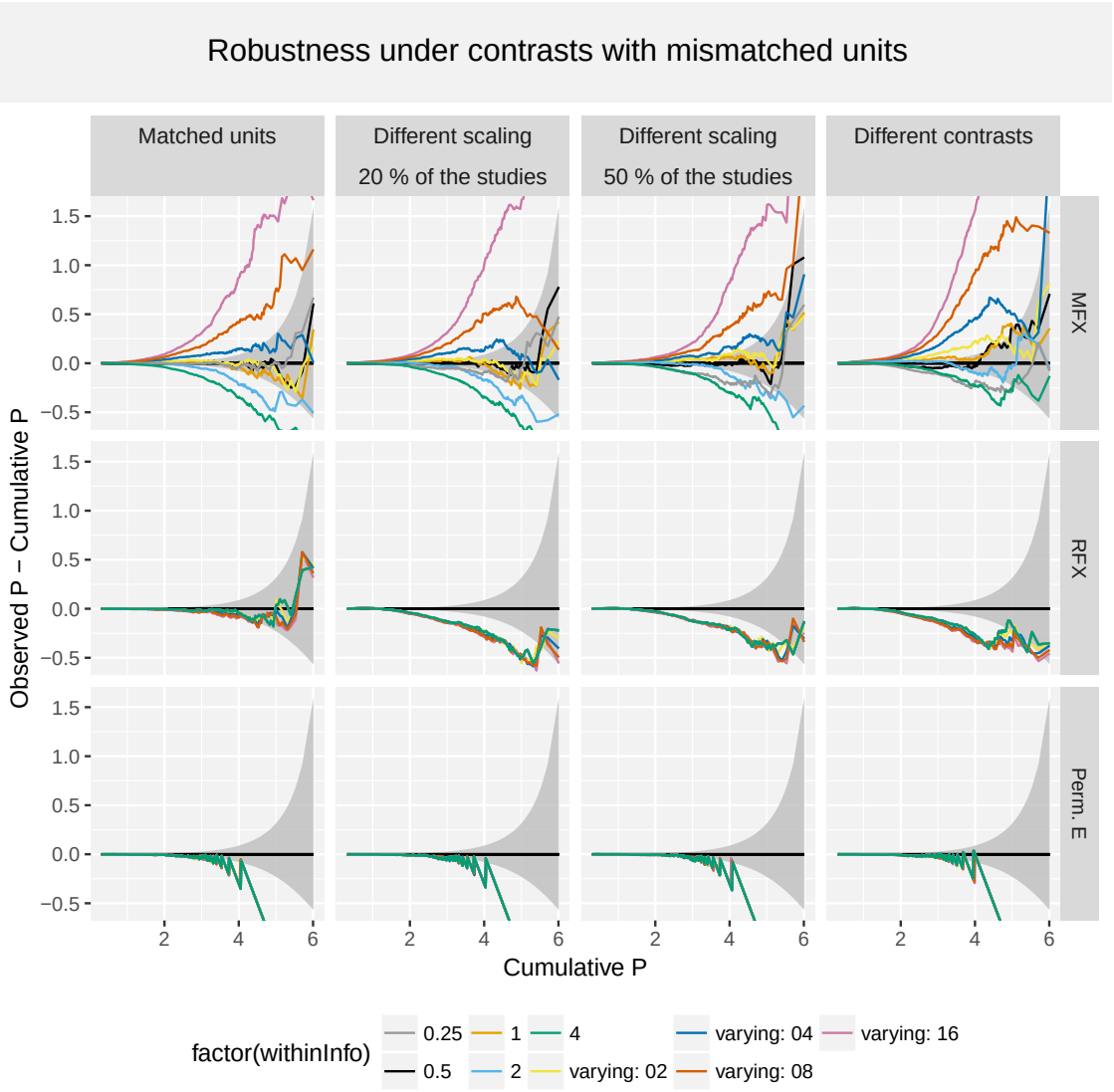

Figure 3: Deviation from theoretical P-values in one-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). Deviation from theoretical Z in two-sample tests with unit mismatch, under ideal circunstances for each statistical 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.

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 conservative. MFX GLM is robust to the presence of robust mimatches when the studies are homoscedastic. In the presence of strong heteroscedasticity, MFX GLM remains invalid as when the units are matched due to small sample size (cf. previous paragraph). In the presence of slight heteroscedasticity, unit mismatchs can cause invalidity.

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).

3.2.1. Group meta-analysis

Fig. **??** presents the simulation results for a one-sample test with $\tau^2 = 1$ and a sample size $k = 5, 25, 50$. For the nominal case, i.e. when the units are matched across studies and contrasts, MFX GLM, RFX GLM and Contrast Permutation are all valid, as expected. For small sample sizes ($k = 5$), MFX GLM and contrast permutation are both very conservative. For large values of Z, Contrast Estimation is conservative as expected due to the discrete nature of its distribution. More suprinsing, in the presence of a high within-study variance, MFX GLM also appears to be conservative. RFX GLM displays the best behaviour with a pattern that is always within the 95% confidence interval of the theoretical Z.

3.2.2. Balanced between-group meta-analysis

Fig. TODO presents the simulation results for a two-sample meta-analyses with $\tau^2 = 1$ and a sample size $k = 25$. For the nominal case, GLM RFX, GLM RFX and contrast estimation provide valid estimates. Contrast Permutation is conservative for large Z values. Both RFX GLM and MFX GLM display the best behaviour with a pattern that is within the 95% confidence interval of the theoretical Z.

In the extreme case of different scaling target, contrast permutation is always valid with a pattern very similar than its nominal behaviour. GLM RFX is valid for Z values greater than 1.5, which is the area of interest in detections, but display a strong conservativness, more pronounced than the Contarst Permutation. GLM MFX is slighlty invalid for all within-study variances except the largest one when 20% of the studies come from the second software.

3.2.3. Unbalanced between-group meta-analysis

Fig. TODO presents the simulation results for unbalanced two-sample meta-analyses with $\tau^2 = 1$ and a sample size $k = 25$. For the nominal case, MFX GLM, GLM RFX and contrast permutation provide valide estimate. As expected due to the discrete nature of its ampling distribution, contrast permutation is conservative for large Z value. GLM RFX is conservative. RFX GLM is closest to the theoretical behaviour with Z-values that are always within the 95% confidence interval.

In the extreme case of different scaling target, MFX GLM is always valid but slightly conservative. RFX GLM is valid for Z values greater than 1.5 (area of interest in detections) but conservative. Similarly contrast permutation is invalid for Z smaller than 1.5 and conservative otherwise. This can be explained by the violation of the exchangeability condition.

When different scaling algorithm are used, (same paragraph as for one-sample test)

When the contrast are scaled differently, we observe a very similar pattern than for different scaling algorithm with higher varaince of the estimates.

3.3. Simulations

Fig. 1 displays the false positive rate at $p < 0.05$ obtained for the nine estimators over all set of parameters in the absence and presence of between-study variation. As expected, the fixed-effects meta-analytic summary statistics, i.e. Fisher's, Stouffer's and weighted Stouffer's estimates, are liberal in the presence of study heterogeneity. The original Fisher's approach is the most invalid. More surprising, FFX GLM is also invalid with homogeneous studies. The explanation is over-estimation of degrees-of-freedom (DF); while DF is computed as $(\sum n - 1) - 1$, under heteroscasdicity (from σ_i or n_i) it will be much lower [15]. Z MFX and GLM RFX provide valid estimates, and the permutation estimates are valid but tend to be conservative with greater variation in false positive rates.

The impact of the number of studies involved in the meta-analysis and of the size of the within-study variance are investigated in Fig. TODO. Permutation inference is valid but conservative when 5 studies are used; this is because there are only $2^5 = 32$ possible permutations and thus $1/32 = 0.03125$ is largest attainable valid P-value. All approaches perform equally as soon as 10 or more studies are included in the meta-analysis.

3.4. 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 sensitive to heterogeneity. The exterme being Fisher's method that is the most sensitive 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 MFX and Z permutation that display almost identical ROC curves, followed by MFX GLM, contrast permutation and finally RFX GLM. TODO: Why RFX is last??

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

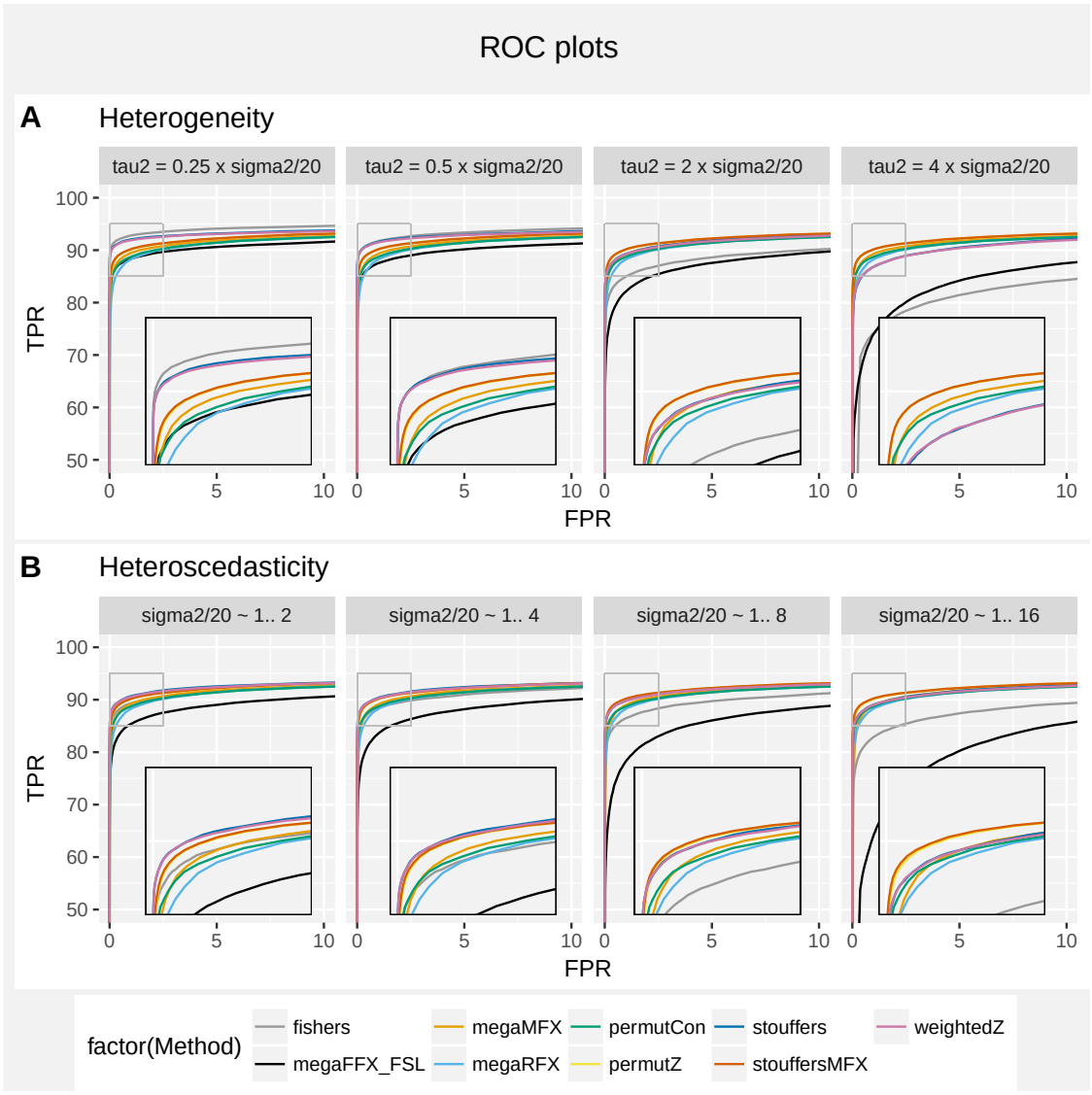

Figure 4: ROC curves of the meta-analytic estimators where true positive rate were computed using a real data meta-analysis of 21 studies of pain and false positive rates using simulated data under various levels of heterogeneity (A) and hetoroscedasticity (B).

variance by comparison to the within-study variance. Amongst the random-effects approaches the most optimal are again stouffers MFX and Z permutation that display almost identical ROC curves.

Fig. 4 plots the difference between the z-score estimated by each meta-analytic approach against the reference z-score computed with MFX GLM. All FFX statistics provide overly optimistic z-estimate suggesting, again, that study heterogeneity is present in the studied dataset. Among the RFX meta-analytic approaches, GLM RFX and contrast permutations provide z-scores estimate that are equal or smaller than the reference. Z permutation provides slightly larger z-scores between 1 and 3 (reference p-values between 0.16 and 0.0013) but is mostly in agreement with the reference z-scores. On the other hand, Z MFX is more liberal than the reference for z-score ranging from 3 to 5 (reference p-values between 0.0013 and 2.9e-07) and more stringent for z-scores smaller than 5.

4. Discussions

TODO: different scanners effect on units, event-related designs, analytical variability (maybe intro??) TODO: discuss cpmbination of real data TPR with FPR. Ideally we should have real data for all the settings.

5. Conclusion

We have compared nine meta-analytic approaches in the context of one-sample test. Through simulations, we found the expected invalidity of standard FFX 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 RFX meta-analytic statistics. When only contrast estimates are available, RFX GLM was valid. This is in line with previous results on within-group one-sample t-tests studies [10]. When only standardised estimates are available, permutation is the preferred option as the one providing the most faithful results. Further investigations are needed in order to assess the behaviour of these estimators in other configurations, including meta-analyses focusing on between-study differences.

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.

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] 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.

[3] G. Chen, Z. S. Saad, A. R. Nath, M. S. Beauchamp, and R. W. Cox. FMRI group analysis combining effect estimates and their variances. *NeuroImage*, 60(1):747–65, March 2012.

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

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

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

[7] 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.

[8] 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.

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

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

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

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

[13] 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.

[14] J. Radua and D. Mataix-Cols. Meta-analytic methods for neuroimaging data explained. *Biology of mood & anxiety disorders*, 2(1):6, 2012.

[15] F. E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2:110–114, 1946.

[16] 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.

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

[18] 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.

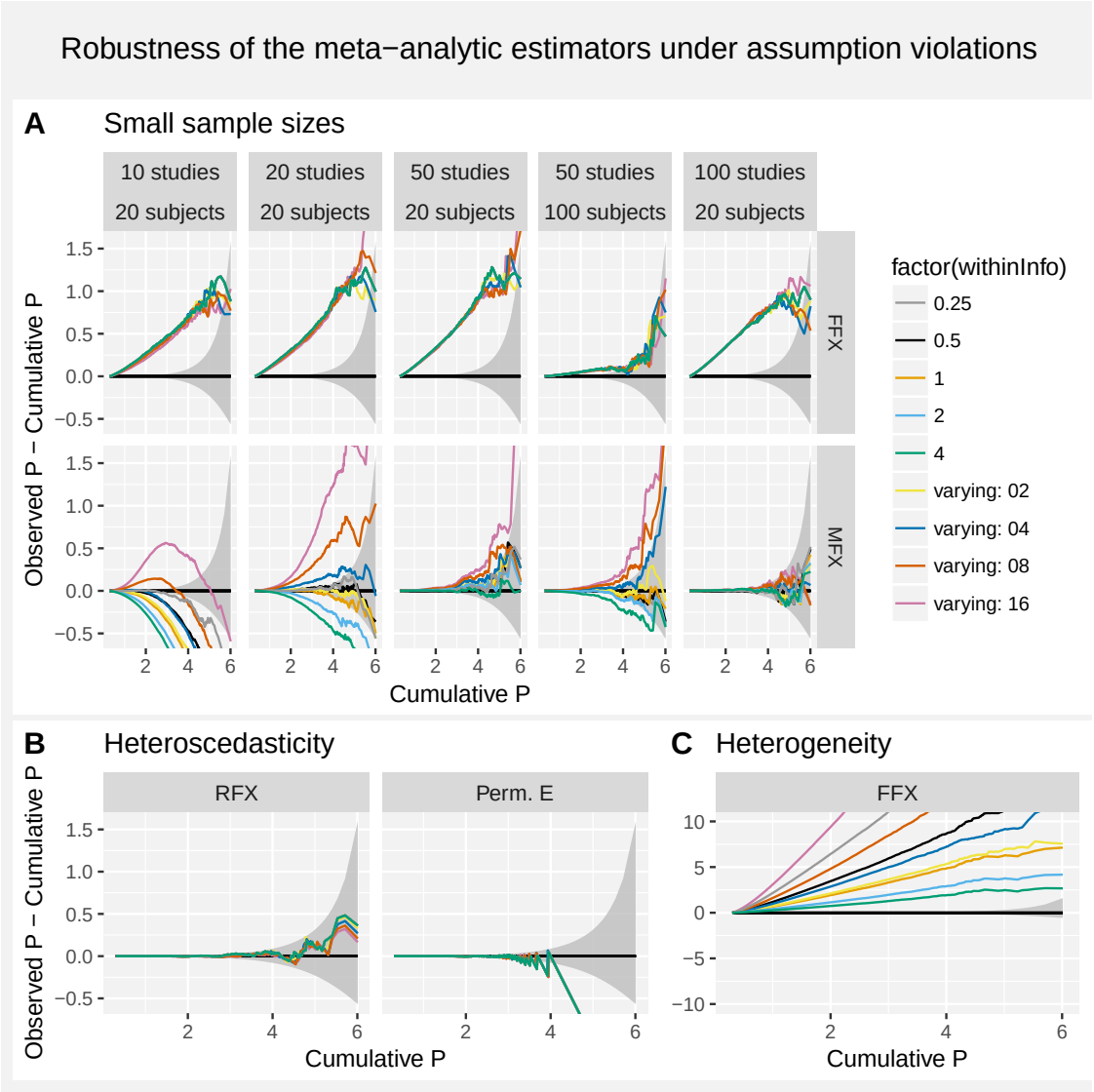

Figure S1: Deviation from theoretical P-values in two-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). P-values are displayed using a negative $\log_{10}$ scale.

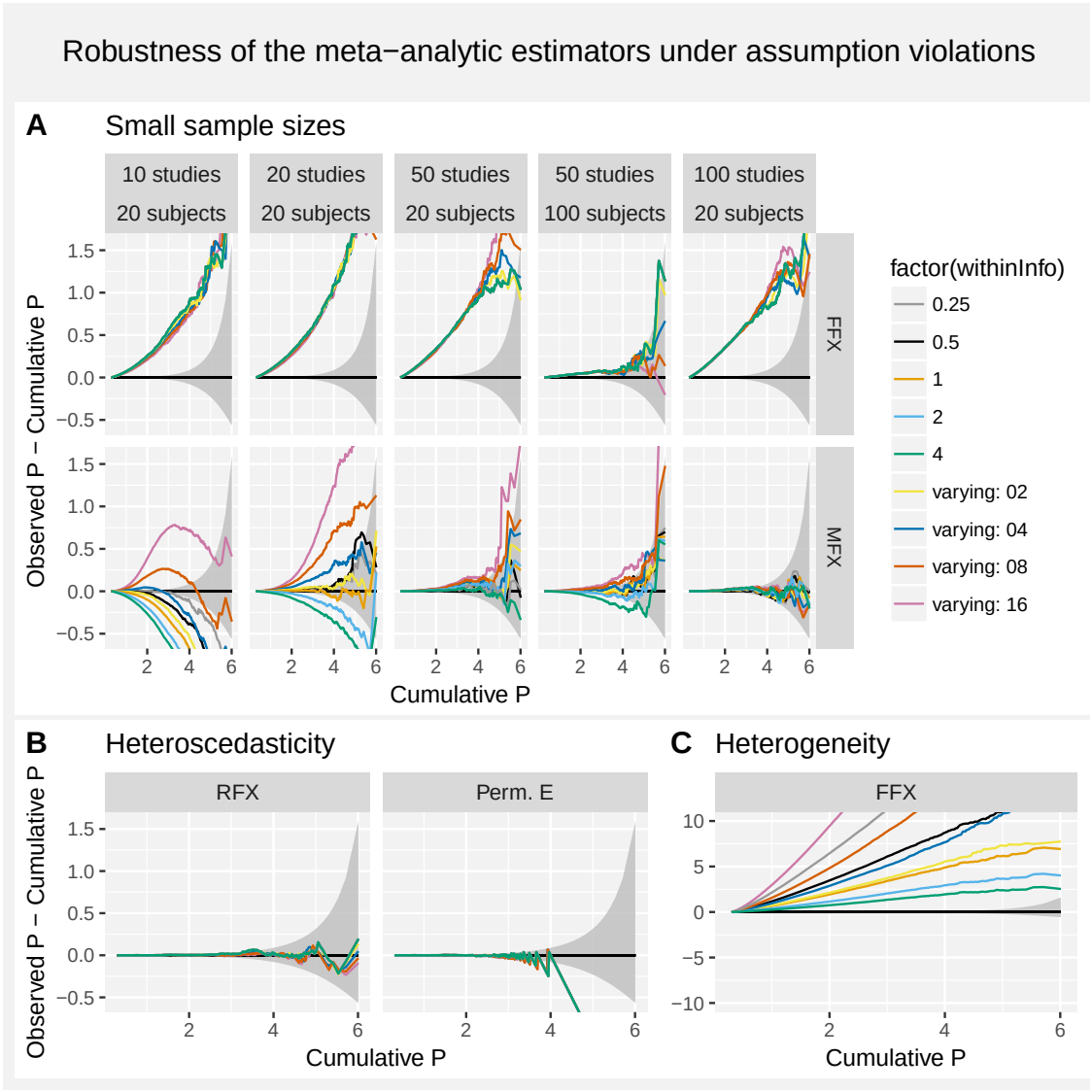

Figure S2: Deviation from theoretical P-values in unbalanced two-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). P-values are displayed using a negative $\log_{10}$ scale.

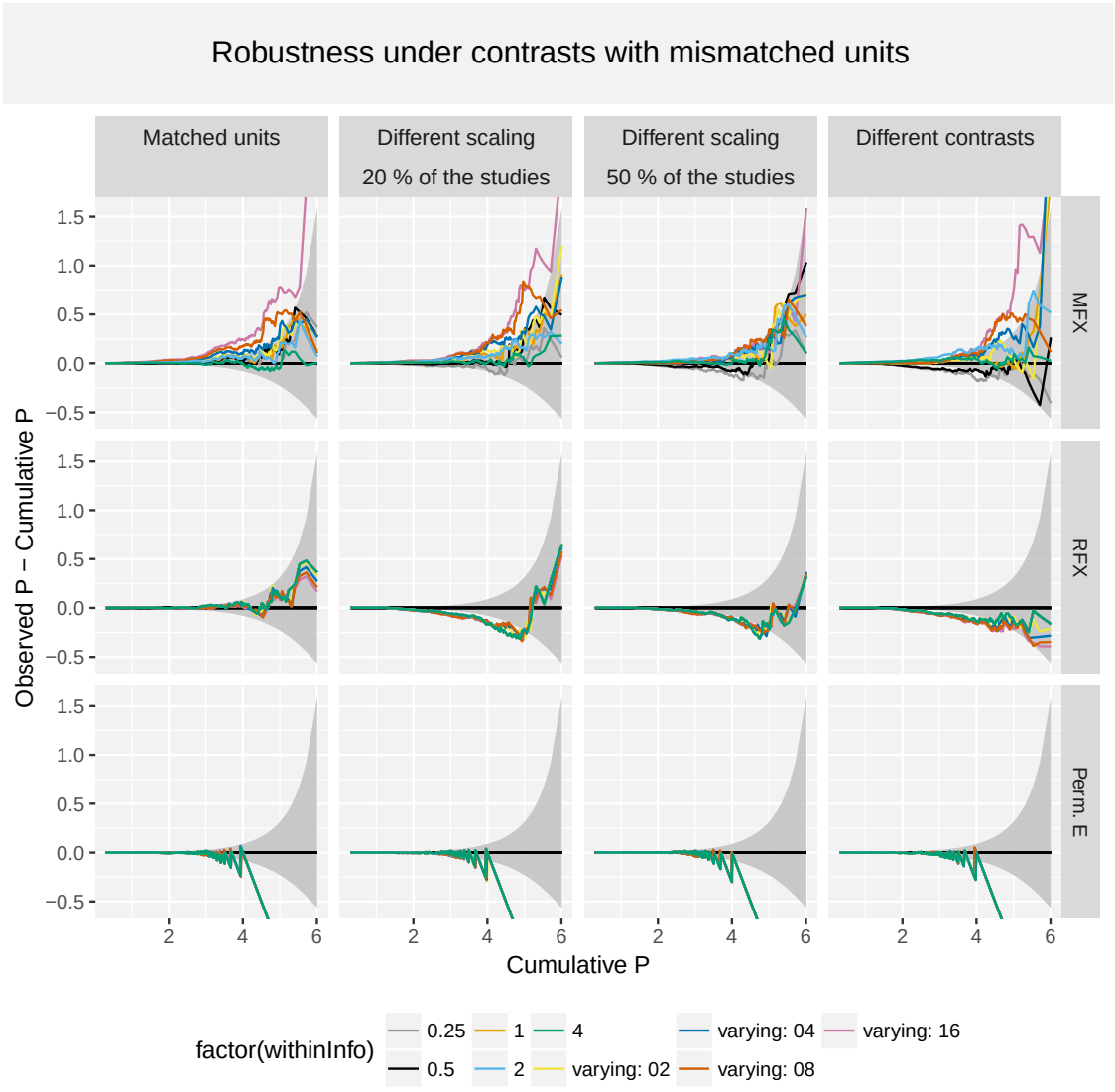

Figure S3: Deviation from theoretical P-values in one-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). Deviation from theoretical Z in two-sample tests with unit mismatch, under ideal circunstances for each statistical 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.

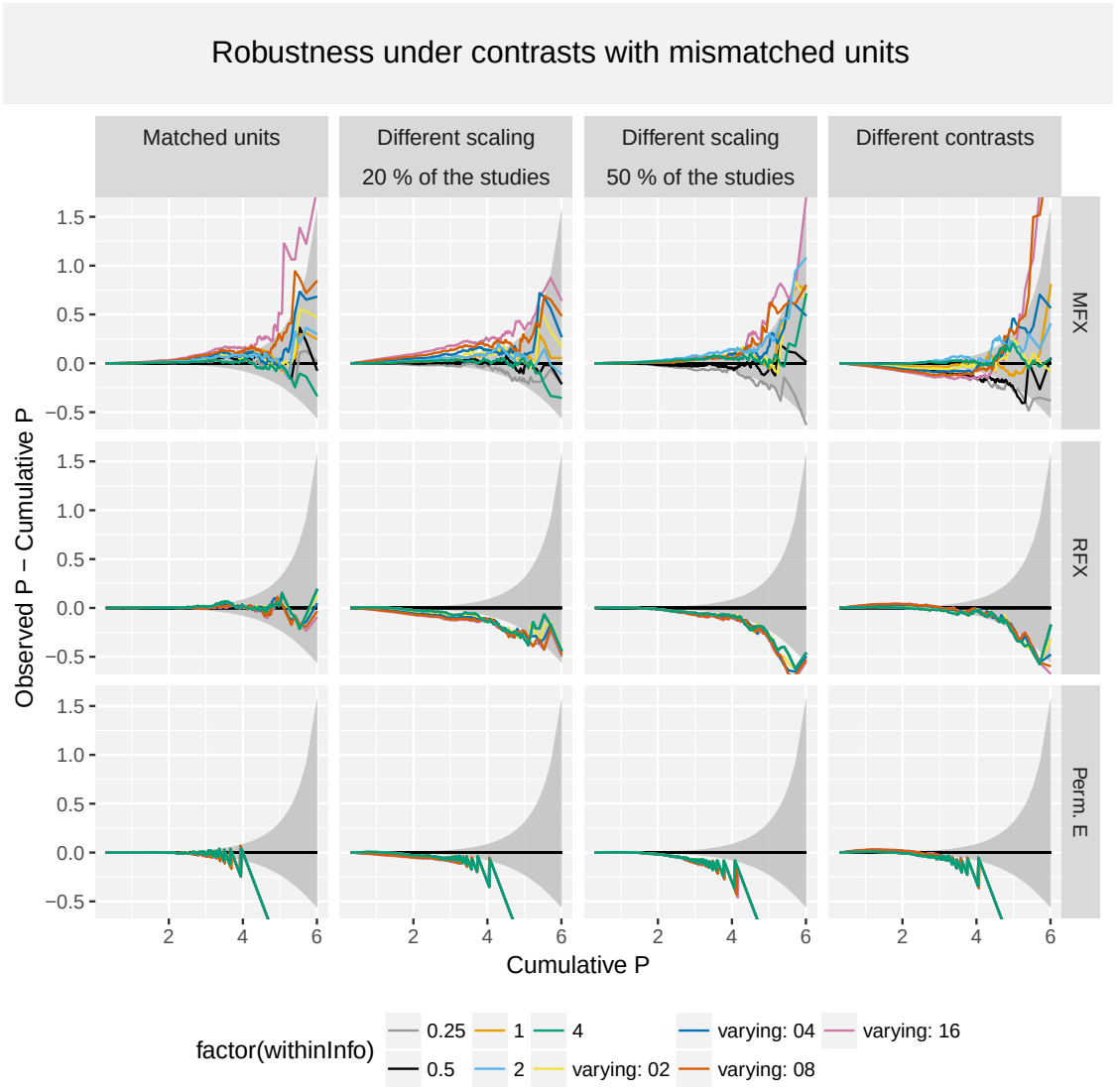

Figure S4: Deviation from theoretical P-values in one-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). Deviation from theoretical Z in two-sample tests with unit mismatch, under ideal circunstances for each statistical 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.