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

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 [20] 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 [4], 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 [3] as well as data, model and contrast vector scaling [13] and the number of samples included in a meta-analysis is usually small.

Given the growing interest in data sharing in the neuroimaging community [18, 14], 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.

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

	$\hat{\gamma}$	$\text{Var}(\hat{\gamma})$	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: $P_i = \Phi(-Z_i)$

Section 2 provides theoretical background on each of the nine meta-analytic approaches. Experiments undertaken on simulated and real data are then described. The results are described in section 3. Discussions, inlcluding our recommendations are provide in 4 Finally, we conclude in section 5.

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) [4], 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 p=1 with 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, i.e. assuming non-zero between-study variance, we have $W = \text{diag}(\sigma_1^2/n_1 + \tau^2, \ldots, \sigma_k^2/n_k + \tau^2)$ 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:

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

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:

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

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, $W = \text{diag}(\sigma_1^2/n_1, \ldots, \sigma_k^2/n_k)$. 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 $(\sum_i n_i xk) - 1$ degrees of freedom $+ADD_REFandcheckdofvalueinFSL$ as depicted in Table 2.

	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.

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 ($\sigma_i^2/n_i = \sigma^2 \quad \forall i$) 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-study 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:

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

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

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, 15] 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 [5]. **Stouffer**'s approach combines directly the standardised statistic [21]. In [24] following [10], 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, 15] 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 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. Finally, all contrast methods (MFX GLM, RFX GLM, Contrast permutation and FFX GLM) require the contrasts to be expressed with in the same units.

2.2. Experiments

2.2.1. 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 [17]. A total of 144 parameter sets (9 σ_i^2 x 2 τ^2 x 4 k x 2 n_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 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). 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..

2.2.2. Accuracy on real data

We then compared the nine meta-analytic estimators on a dataset of 21 studies of pain analyzed with FSL that we made available at: http://neurovault.org/collections/XRJCFIYG/. 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.

We used Neurosynth's automated large scale coordinate-based meta-analysis of pain (available at http://neurosynth.org/analyses/ as ground truth of areas that should present an effect. 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 Receiver-Operating-Characteristics (ROC) curves. ROC curves are useful to provide a quantitative evaluation of the trade-off between sensitivity and specificity provided by a given classifier (here the meta-anlytic approaches).

The code used to compute the real data analysis is available at: `https://github.com/cmaumet/zmeta_rocs`.

2.2.3. Software and computing resources

Simulated and real data meta-analyses were computed with Matlab R2016b. The code used to compute the simulations is available at: `https://github.com/cmaumet/zmeta_buster` and used the IBMA toolbox [11], SPM (RRID:SCR_007037) [6], FSL (RRID:SCR_002823) [9] and SnPM (RRID:SCR_002092) [15]. Simulations were computed on a High Permformance Computing cluster provided by the Statistics Departement at the University of Warwick. All other analyses were computed on a Mac Book Pro laptop.

Figures displayed in this manuscript were computed with R [19] and used ggplot2 [22] and cowplot [23] packages. The code used to compute the figures displayed in this manuscript is available at: `https://github.com/cmaumet/zmeta` in the form of R jupyter notebooks [16].

3. Results

3.1. Robustness to violation of model assumptions

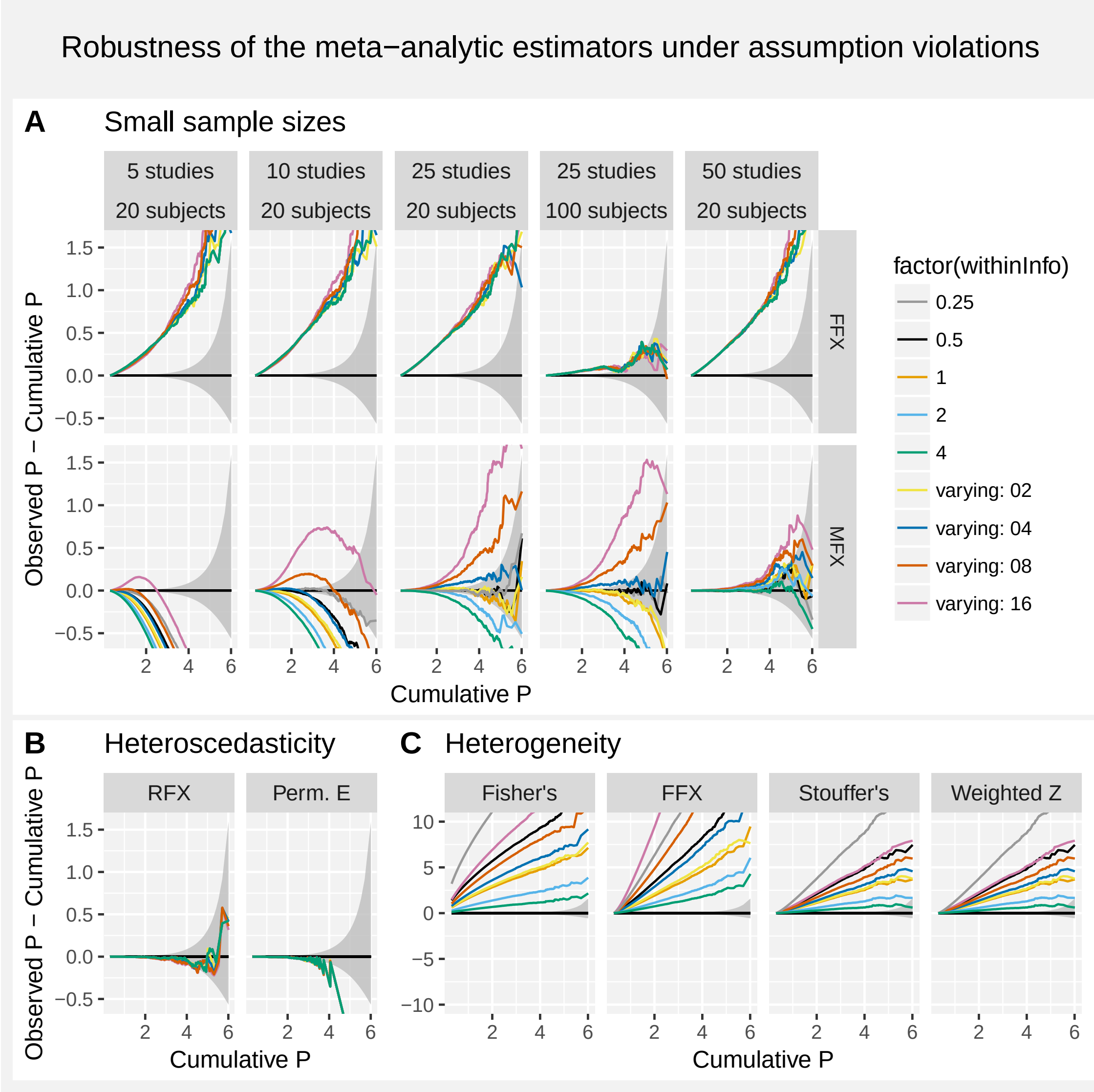

Figure 2: Deviation from theoretical P-values in one-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). P-values are displayed using a negative $\log_{10}$ scale.

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 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 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

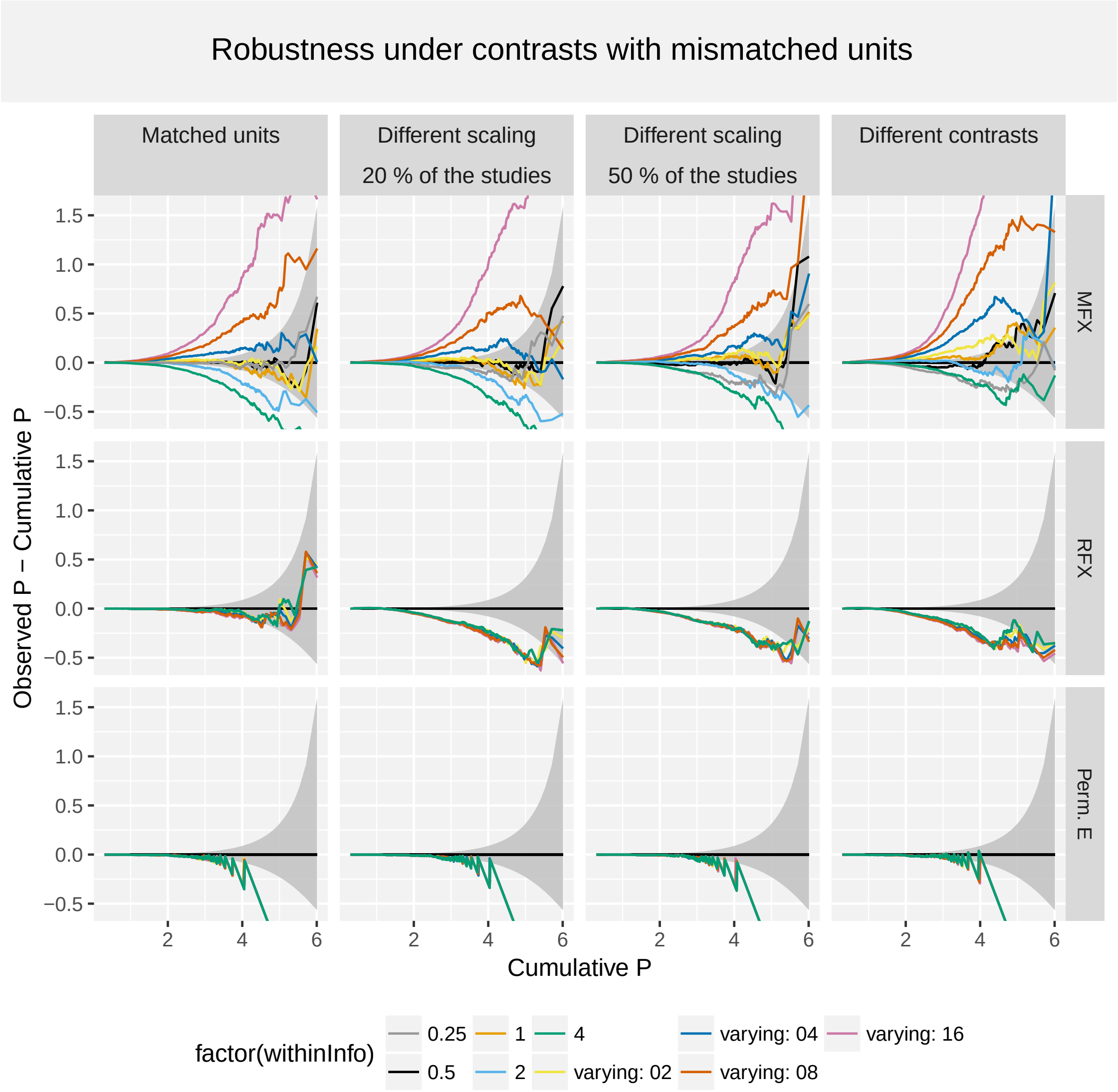

Figure 3: Deviation from theoretical P-values in one-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). Deviation from theoretical Z in two-sample tests with unit mismatch, under ideal circumstances for each statistical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

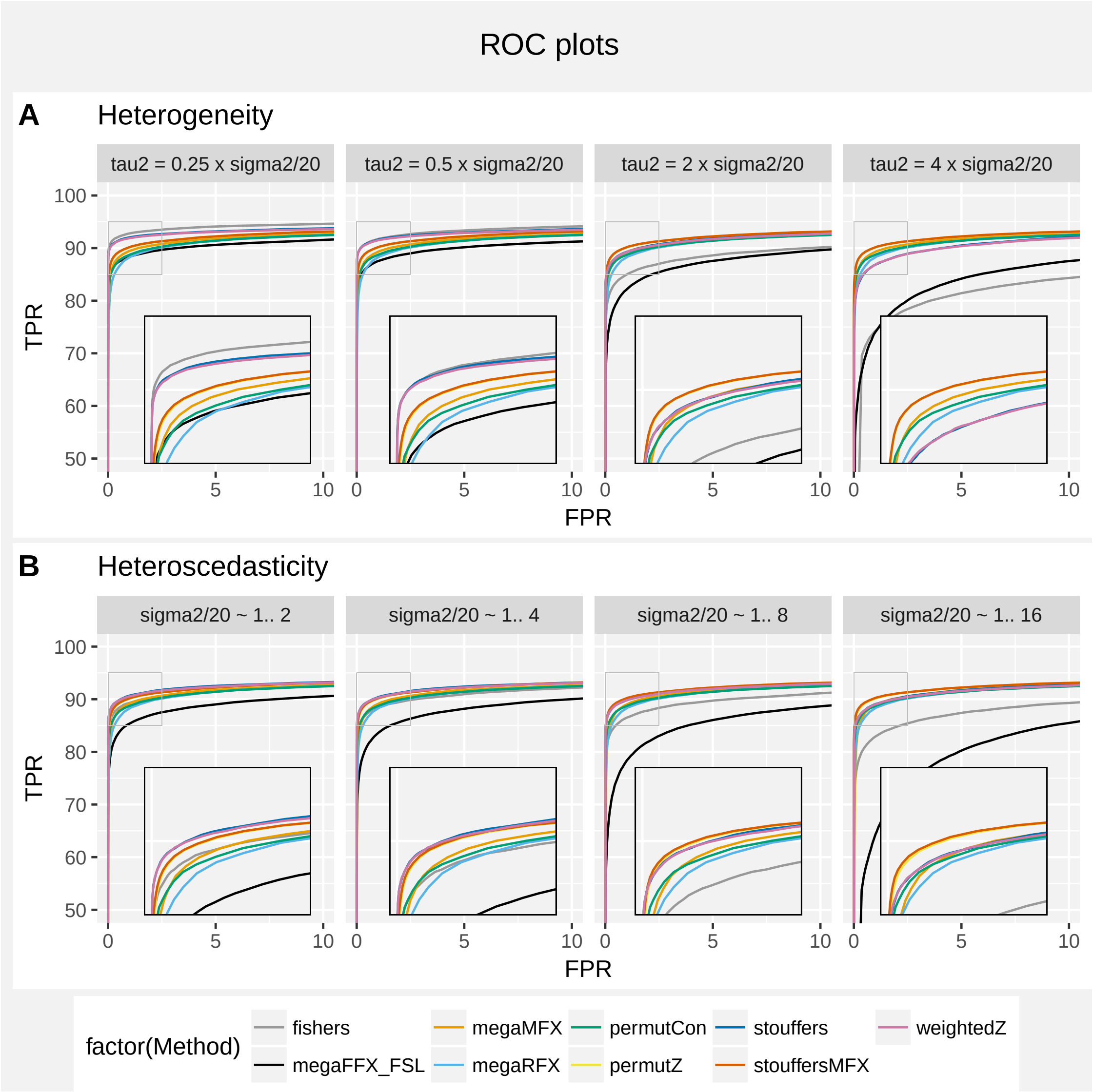

Figure 4: ROC curves of the meta-analytic estimators where true positive rate were computed using a real data meta-analysis of 21 studies of pain and false positive rates using simulated data under various levels of heterogeneity (A) and hetoroscedasticity (B).

3.3. 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 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 heteroscedasticity, 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.

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 valid under heteroscedasticity which is in line with published literature on group fMRI statistics [12]. More surprisingly, MFX GLM was showed to be invalid in the presence of large heteroscedasticity due to its approximations being not accurate in small samples.

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 design, analysis workflows or even imaging instruments. In our real data we are dealing with same lab + same soft, so even more heterogeneity/heteroscedasticity is expected in real meta-analyses. Similarly, we do expect heteroscedasticity to be present for the 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 or Stouffers Z MFX 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 very rarely full details about the analysis are provided. Although z-based meta-analyses are suboptimal [4] and might be affected by strong sample size differences which could happen in future with increased, until this is routinely done, we suggest to use Z-based methods that are insensitive to units.

Our simulated data were based on uncorrected p-values. In neuroimaging meta-analyses, results are more likely to be expressed using correction for multiple comparisons (FWE using RFT or FDR). Accuracy of uncorrected p-values is important for FDR (and FWE??) and for clusterwise thresholds. 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, 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). This means that our TPR are likely to be overestimated.

Because simulating realistic activations can be particularly challenging Because true areas of no activations in real data are unknown, we relied on simulated data to compute the false positive rates. This has the disadvantage of synthetic data listed above. We only had a single real dataset for meta-analysis and therefore only computed TPR once and then combined with the different FPR computed by simulations. In practice TPR is very likely to be affected by the different settings. The most faithful to reality is sigma2/20   1...16 which is consistent with our general recommendations stated above. 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 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 [12]. 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.

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

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

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

[6] 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.

[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] 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.

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

[11] 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.

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

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

[14] 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.

[15] 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.

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

[17] 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.

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

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

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

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

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

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

[24] 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.

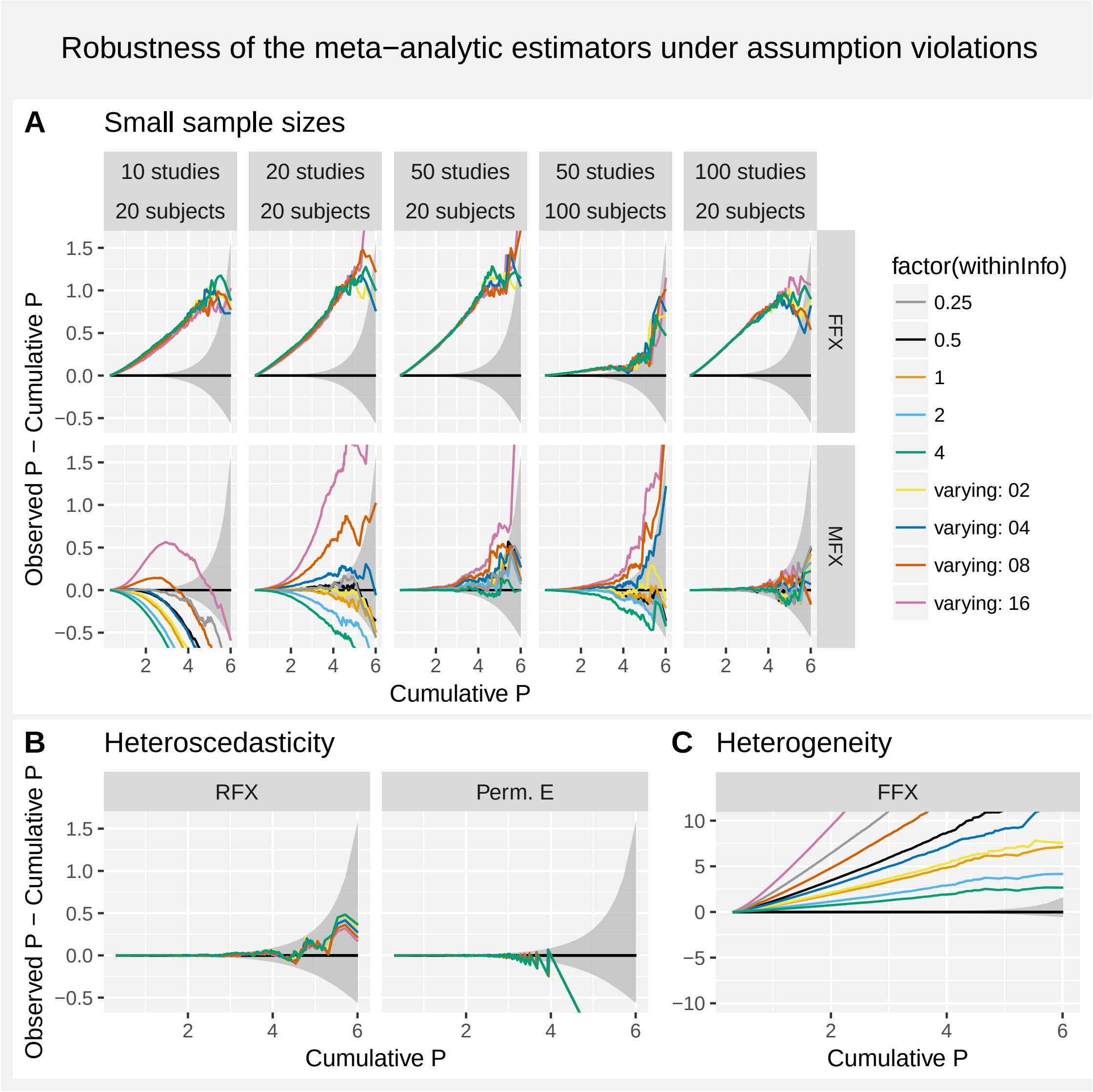

Figure S1: Deviation from theoretical P-values in two-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). P-values are displayed using a negative $\log_{10}$ scale.

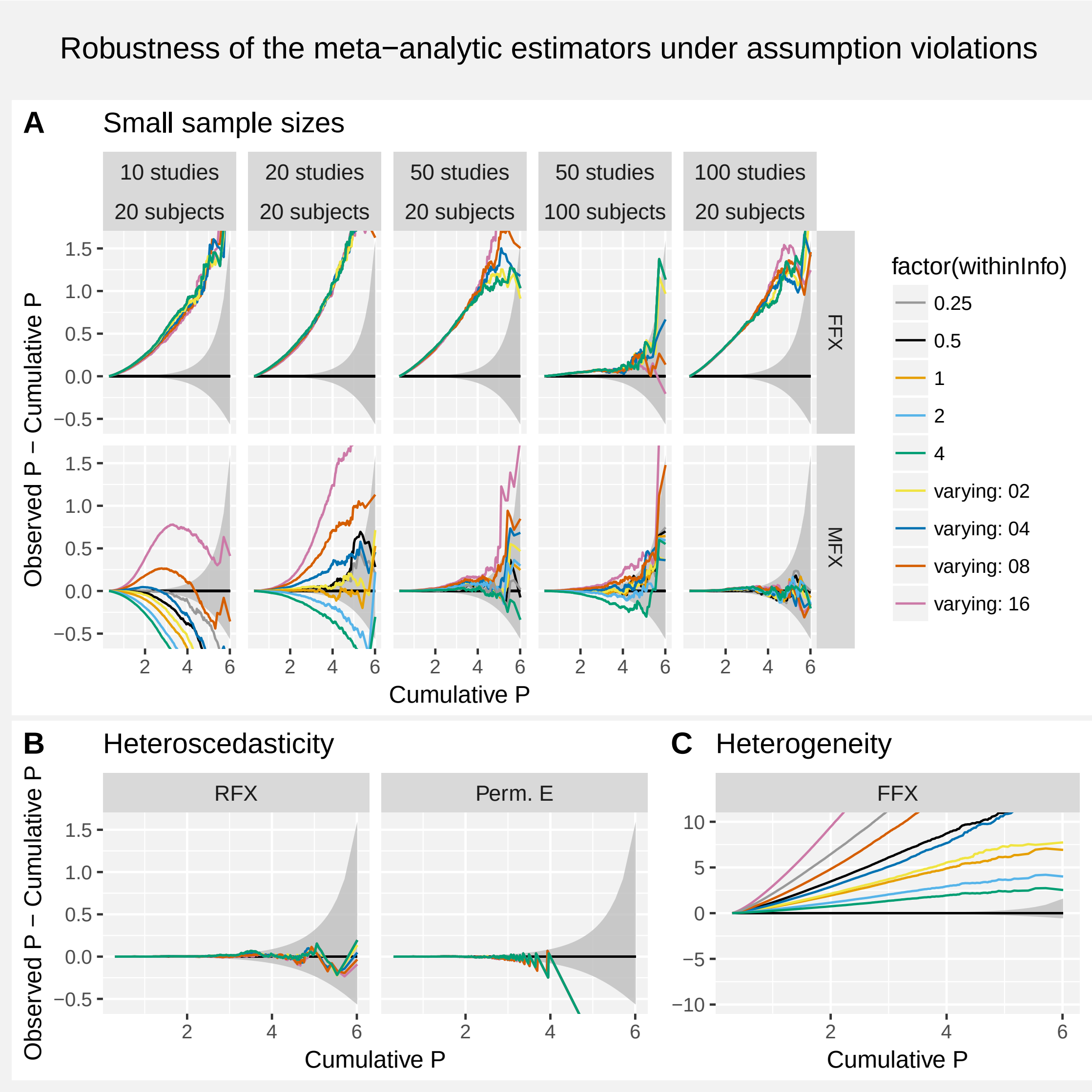

Figure S2: Deviation from theoretical P-values in unbalanced two-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). P-values are displayed using a negative $\log_{10}$ scale.

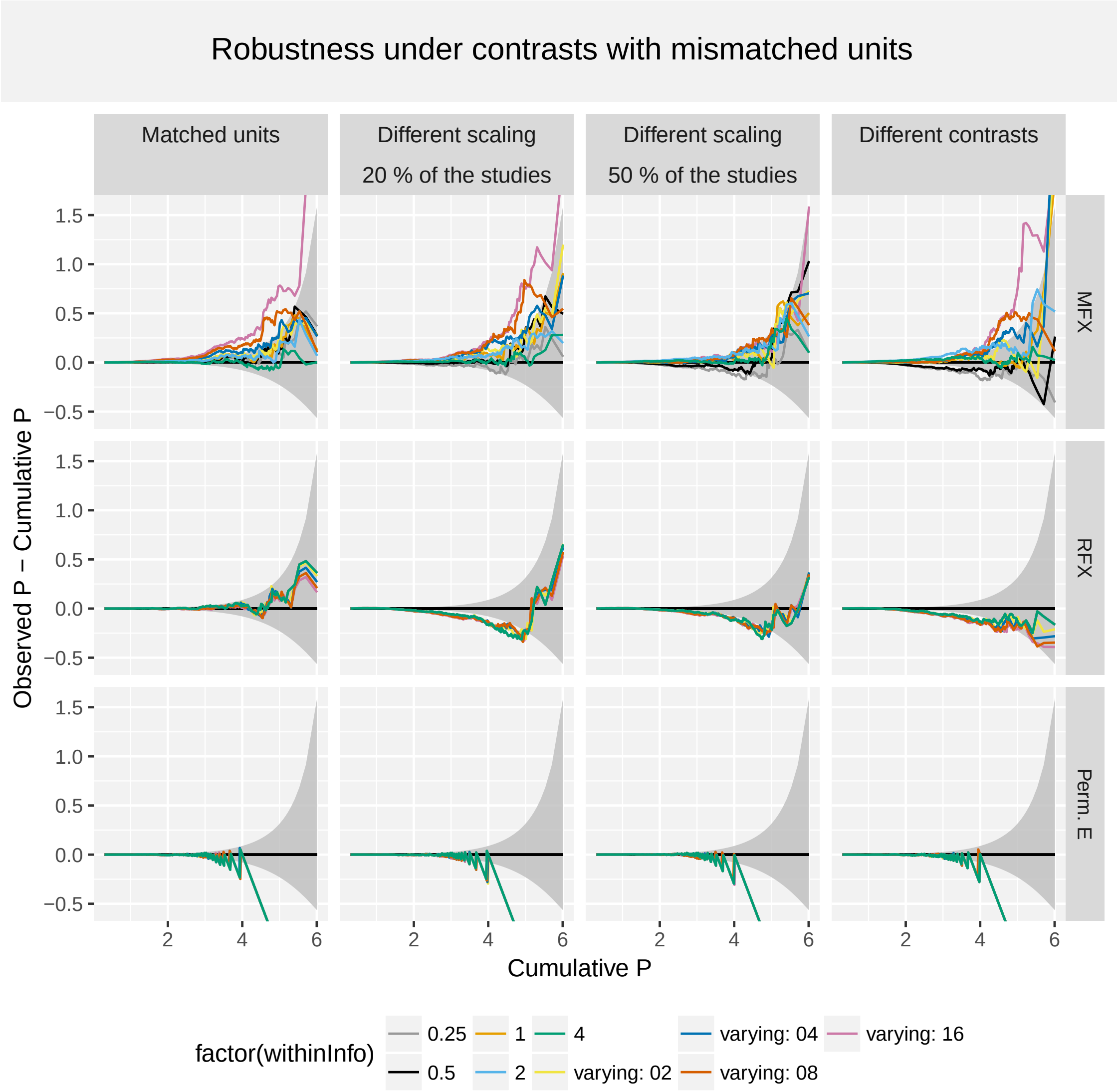

Figure S3: Deviation from theoretical P-values in one-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). Deviation from theoretical Z in two-sample tests with unit mismatch, under ideal circumstances for each statistical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

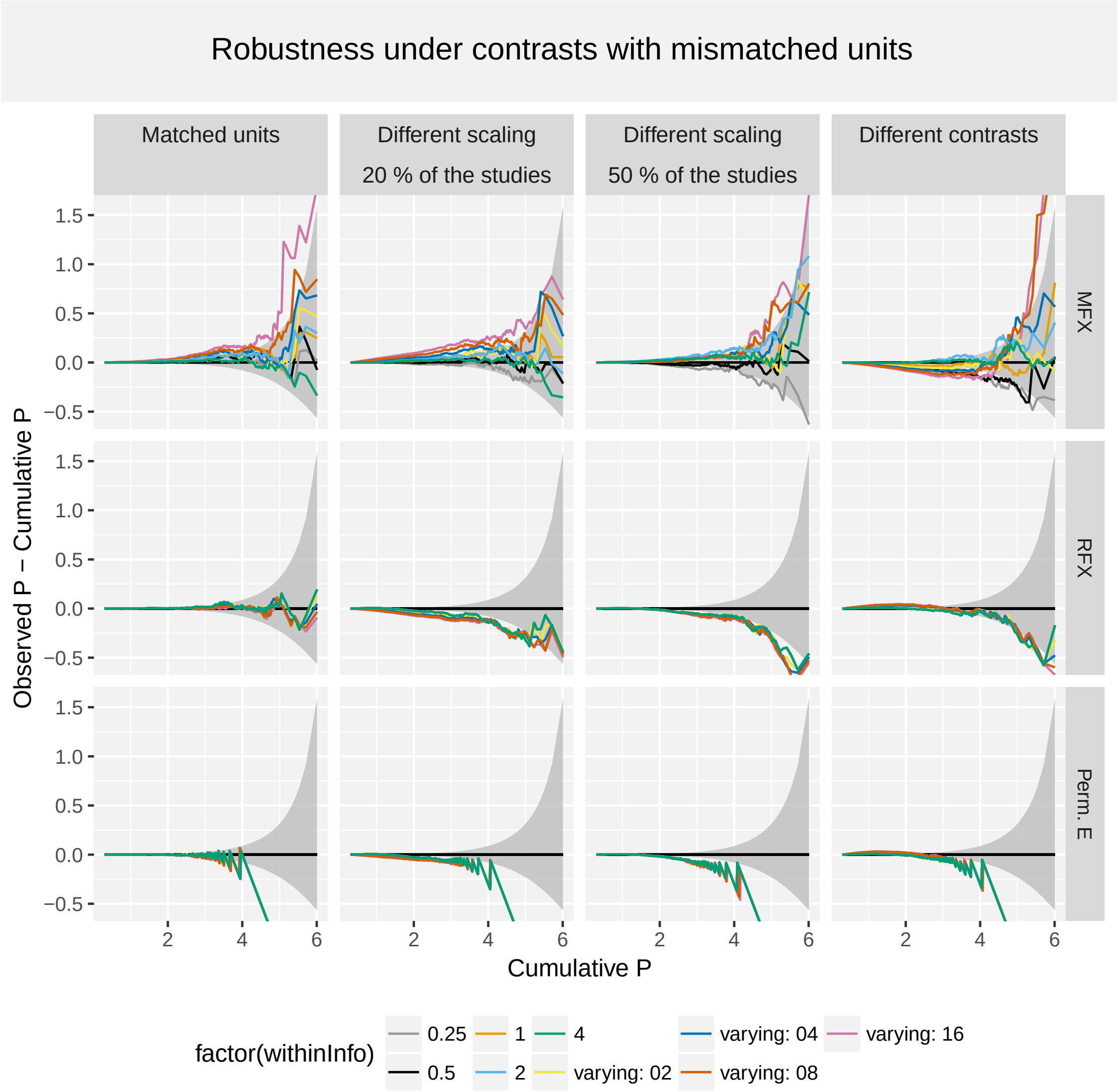

Figure S4: Deviation from theoretical P-values in one-sample tests under violations of the underlying model assumptions: small sample sizes (A), heteroscedasticity (B) and heterogeneity (C). Deviation from theoretical Z in two-sample tests with unit mismatch, under ideal circumstances for each statistical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