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

Camille Maumet[1] and Thomas E. Nichols[1,2]

[1] Warwick Manufacturing Group, The University of Warwick, Coventry, UK.
[2] Statistics Department, The University of Warwick, Coventry, UK.

Abstract. Meta-analysis is a powerful statistical tool to combine results from a set of studies. When image data is available for each study, a number of approaches have been proposed to perform such meta-analysis 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, often only standardised estimates are shared, reducing the possible meta-analytic approaches. 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 eight meta-analytic approaches on simulated and real data. In one-sample tests, combination of contrast estimates into a random-effects General Linear Model or non-parametric statistics provide a good approximation of the reference approach. If only standardised statistical estimates are shared, permutations of z-score is the preferred approach.

1 Introduction

TODO: check number of meta-analytic approaches (8 or 9) and update everywhere

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 [11] 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 Intensity-Based Meta-Analysis (IBMA) that combines the effect estimates and their 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. Various efforts are currently underway to facilitate sharing of neuroimaging data but emphasis is usually on statistical maps (see, e.g. [2]). There are three evident approaches to sharing summary data from each study i:

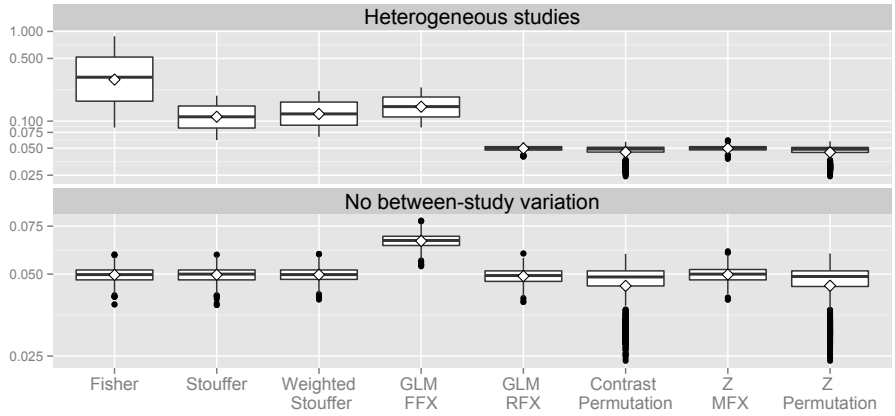

Fig. 1. False positive rates of the meta-analytic estimators under the null hypothesis for $p < 0.05$.

1. the contrast estimates $\hat{\beta}_i$ and contrast variance estimates $\hat{S}_i^2$.
2. the contrast estimates $\hat{\beta}_i$.
3. the standardized statistical 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. In fMRI, units will depends on data, model and contrast vector scaling and are typically different across neuroimaging software due to different data scaling approaches [10].

TODO: cite Gang's paper

Given the growing interest in data sharing in the neuroimaging community, and the relative easiness of sharing 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 $\hat{S}_i^2$'s, 2 $\hat{\beta}_i$'s only and 5 Z_i's. We compare the validity and the accuracy of the eight 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 4.

2 Methods

2.1 Theory

For study $i = 1, \ldots, k$ we have contrast estimate $\hat{\beta}_i$, its contrast variance estimate $\hat{S}_i^2$ (i.e. squared standard error), its 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 just a column of ones for a one-sample test) and $\epsilon \sim \mathcal{N}(0, W)$ is the residual error term. Eq. (1) can be solved by weighted least squares giving:

TODO: add contrast

$$\hat{\gamma} = (X^T W^{-1} X)^{-1} X^T W^{-1} \hat{\beta} \tag{2}$$

$$\mathrm{Var}(\hat{\gamma}) = (X^T W^{-1} X)^{-1} \tag{3}$$

In a meta-analysis random-effects (RFX) model, we have $W = \mathrm{diag}(\sigma_1^2 + \tau^2, \ldots, \sigma_k^2 + \tau^2)$ where τ^2 denotes the between-study variance. Approximating σ_i^2 by $\hat{S}_i^2$ and given $\hat{\tau}^2$ an estimate of τ^2 we obtain the statistics detailed in Table 1 for one-sample tests. 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), we have $W = \mathrm{diag}(\sigma_1^2 \ldots \sigma_k^2)$ where σ_i^2 denotes the contrast variance for study i.

Combining contrast estimates If the $\hat{S}_i^2$ are unavailable, the contrast estimates $\hat{\beta}_i$ can be combined by assuming that the within-study contrast variance σ_i^2 is roughly constant ($\sigma_i^2 \simeq \sigma^2$) or negligible in comparison to the between-study variance($\sigma_i^2 \ll \tau^2$). 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:

$$\hat{\gamma} = (X^T X)^{-1} X^T \hat{\beta} \tag{4}$$

$$\mathrm{Var}(\hat{\gamma}) = (X^T X)^{-1} \sigma_C^2 \tag{5}$$

Given $\hat{\sigma}_C^2$ an estimate of σ_C^2 we obtain the statistics presented in Table 1 for one sample tests. This approach will be referred to **RFX GLM** in the following.

	Meta-analysis statistic	Nominal H_0 distrib.	Inputs	Assumptions
FFX GLM	$\left(\sum_{i=1}^{k} \frac{\hat{\beta}_i}{\hat{S}_i^2}\right) / \sqrt{\sum_{i=1}^{k} 1/\hat{S}_i^2}$	$\mathcal{T}_{(\sum_{i=1}^{k} n_i - 1) - 1}$	$\hat{\beta}_i, \hat{S}_i^2$	IGE; σ_i^2 cst; $\tau^2 = 0$.
MFX GLM	$\left(\sum_{i=1}^{k} \frac{\hat{\beta}_i}{\hat{S}_i^2 + \hat{\tau}^2}\right) / \sqrt{\sum_{i=1}^{k} 1/(\hat{S}_i^2 + \hat{\tau}^2)}$	$\mathcal{T}_{k-1}$	$\hat{\beta}_i, \hat{S}_i^2$	IGE; $\tau^2 = \hat{\tau}^2$.
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; $\tau^2 + \sigma_i^2$ cst.
Ctrst Perm.	$\left(\sum_{i=1}^{k} \frac{\hat{\beta}_i}{\sqrt{k}}\right) / \widehat{\sigma}_C^2$	Empirical	$\hat{\beta}_i$	ISE.
Fisher's	$-2 \sum_{i=1}^{k} \ln(\Phi(-Z_i)))$	$\chi^2_{(2k)}$	Z_i	IGE; $\tau^2 = 0$.
Stouffer's	$\left(\sum_{i=1}^{k} Z_i\right) / \sqrt{k}$	$\mathcal{N}(0,1)$	Z_i	IGE; $\tau^2 = 0$.
Wght Stouff.	$\left(\sum_{i=1}^{k} \sqrt{n_i} Z_i\right) / \sqrt{\sum_{i=1}^{k} n_i}$	$\mathcal{N}(0,1)$	Z_i, n_i	IGE; $\tau^2 = 0$.
Z MFX	$\left(\sum_{i=1}^{k} Z_i\right) / \sqrt{k}\hat{\sigma}$	$\mathcal{T}_{k-1}$	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.

As an alternative to parametric approaches, non-parametric inference [6, 9] can be performed by comparing the RFX GLM T-statistic to the distribution obtained "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. This approach will be referred to as **Contrast permutation**.

TODO: We should be able to do something if we have the sample sizes by assuming constant within subject variance

Combining standardised statistics When only test statistic images are available there are a several alternate approaches available. **Fisher's** meta-analysis provide a statistic to combine the associated p-values [5]. **Stouffer's** approach combines directly the standardised statistic [14]. In [15] following [7], 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 suited only for one-sample tests. The corresponding statistics are presented in Table 1. 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 1

With contrasts, non-parametric inference [6, 9] can be obtained by sign flipping on the Z_i's. This approach will be referred to as **Z permutation**.

Approximations In practice, all of the methods based on contrast data have approximate parametric null distributions. The nominal distributions listed in Table 1 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 $\hat{S}_i^2$'s. Further, even under homoscedasticity, MFX GLM's null is approximate due to iterative estimation of $\hat{\tau}^2$.

TODO: Clarify what is approximate in each

2.2 Experiments

Simulations Due to the approximate nature of the sampling distributions, we conduct simulations to evaluate the validity of each estimator under inhomogeneity of contrast variances $\hat{S}_i^2$ and under the presence of non-negligible between-study variance.

To verify the validity of each estimator under the null hypothesis we estimated the false positive rate at $p < 0.05$ uncorrected. For each meta-analysis, we simulated $\hat{\beta}_i$ and $\hat{S}_i^2$ such as:

$$\hat{\beta}_i \sim \mathcal{N}(0, \frac{\sigma_i^2}{n_i} + \tau^2) \tag{6}$$

$$\hat{S}_i^2 \sim \frac{\sigma_i^2}{n_i - 1}\chi^2_{(n_i-1)} \tag{7}$$

where $\sigma_i^2 \in [1/2, 1, 2, 4]$ is the within-study variance, $\tau^2 \in [0, 1/20]$ is the between-study variance (fixed-effects models are strictly only appropriate for $\tau^2 = 0$). For different number of studies per meta-analysis we used: $k \in [5, 10, 25, 50]$, and set the number of subjects per studies n_i to vary across the common range of sample sizes in neuroimaging studies. In each simulated meta-analysis we simulated one study with exactly 20, 25, 10 and 50 subjects. For the remaining studies 1/4 of the n_i's were drawn from $\mathcal{U}(11, 20)$, 1/4 from $\mathcal{U}(26, 50)$ and the remaining from $\mathcal{U}(21, 25)$, where $\mathcal{U}(a, b)$ is the discrete uniform distribution on the integers a to b inclusive. A total of 32 parameter sets (4 σ_i^2 x 2 τ^2 x 4 k) was therefore tested and a total of 71^3 realisations were created.

Real data We then compared the 8 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 [13]) 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: $\hat{\tau}^2/(\hat{\tau}^2 + \sum_{i=1}^{k} \hat{S}_i^2)$. 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.

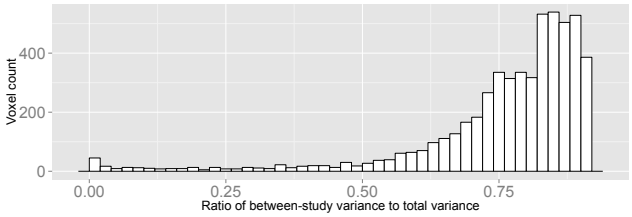

Fig. 2. Histogram of the between-study variance to the sum of the between-subject variance and the mean within-study variance.

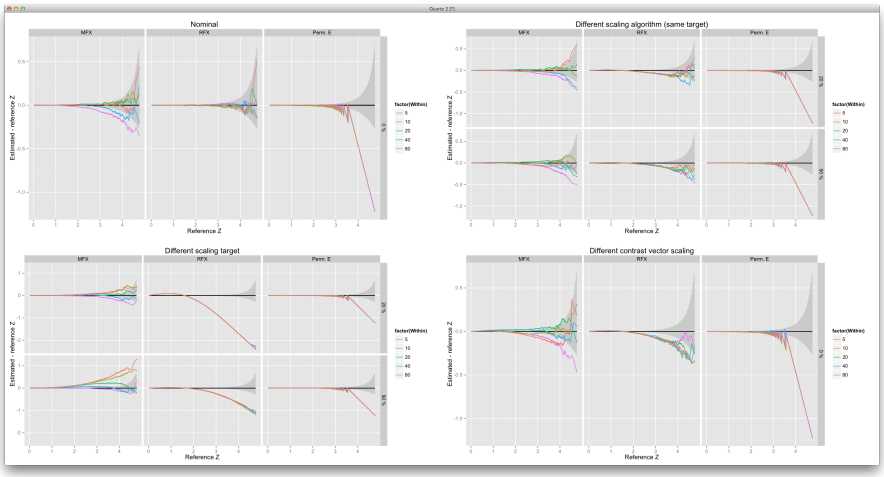

Fig. 3. Deviation from theoretical Z in one-sample tests with $\tau^2 = 1$ and $k = 25$ with matched ("nominal") or mismatched ("different scaling target", "different scaling algorithm", "different contrast vector scaling") units.

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 How bad is the units issue?

Group meta-analysis Fig. 3 presents the simulation results for a one-sample with $\tau^2 = 1$ and a sample size $k = 25$. 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 large values of Z, Contrast Estimation

is conservative as expected due to the discrete nature of its TODOxxTODO disribution. More suprinsingly, in the presence of a high within-subject variance, the 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.

In the extreme case of different scaling target, i.e. when data where scaled to a different mean (100 versus 10 000), Contrast Permutation displays a pattern that is very close to its nominal behaviour, namely it is valid but conservative for large Z. GLM RFX is valid for Z values that are greater than 1.5 (i.e. the area we are interested in in terms of detections) but very conservative, especially when the number of samples from each scaling factor are not equal. This behaviour is expected as the estimated between-study variance is inflated by the difference in scaling target. MFX GLM is invalid for small within-subject variances and conservative for large within-subject variances.

When different scaling algorithm are used, i.e. as in different neuroimaging software packages. Contrast Permutation still has a behaviour that is very similar to nominal. RFX GLM is valid with some conservativness for large Z but still less conservative than Contrast Permutation. MFX GLM is overall valid (TODO what about within=5 /20%) and conservative for large or small within-subject variance.

When the contrast are scaled differently, we observe a very similar pattern than for different scaling algorithm.

Balanced between-group meta-analysis Fig. 5 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-subject variances except the largest one when 20% of the studies come from the second software.

Unbalanced between-group meta-analysis Fig. 5 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

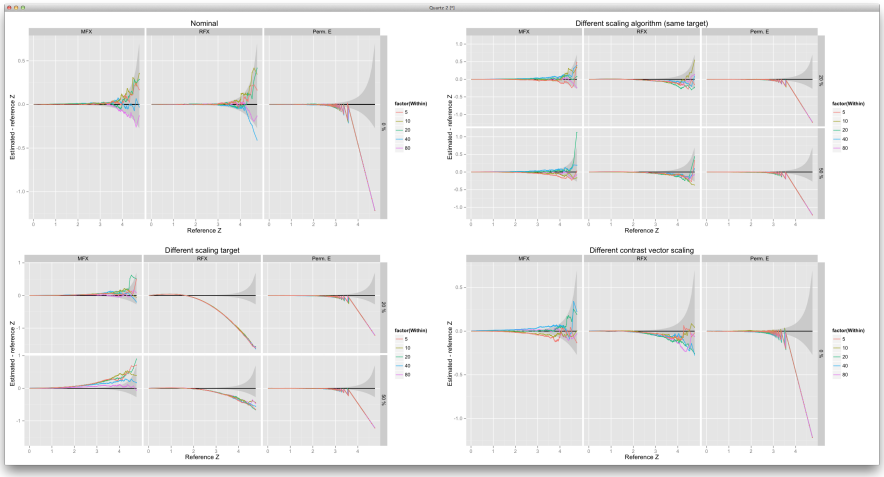

Fig. 4. Deviation from theoretical Z in balanced two-sample tests with $\tau^2 = 1$ and $k = 25$ with matched ("nominal") or mismatched ("different scaling target", "different scaling algorithm", "different contrast vector scaling") units.

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.

3.2 Simulations

Fig. 1 displays the false positive rate at $p < 0.05$ obtained for the eight 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 [12]. 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. **??**. 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

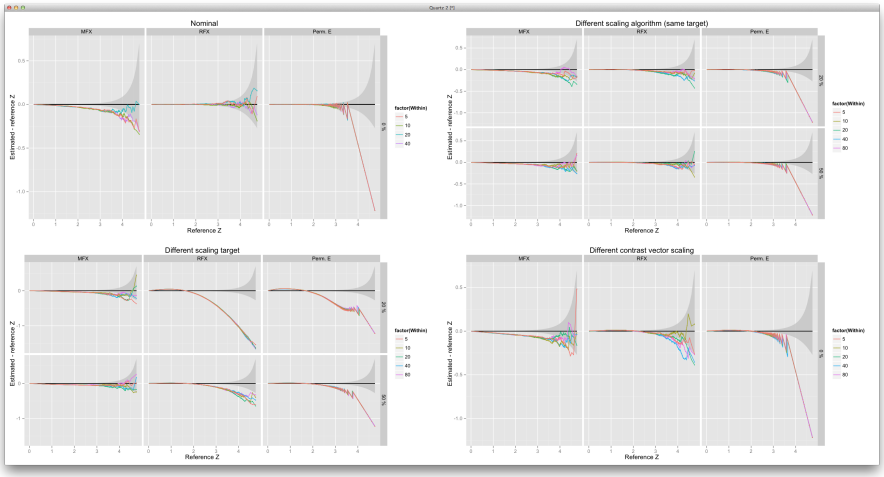

Fig. 5. Deviation from theoretical Z in unbalanced two-sample tests with $\tau^2 = 1$ and $k = 25$ with matched ("nominal") or mismatched ("different scaling target", "different scaling algorithm", "different contrast vector scaling") units.

attainable valid P-value. All approaches perform equally as soon as 10 or more studies are included in the meta-analysis.

3.3 Real data

The histogram of the ratio of between-subject variance to total variance is displayed in Fig. 2. From this graph it is clear that for most of the voxels the estimated between-study variance is greater than the within-study variance. We can therefore suppose the presence of study heterogeneity (non negligible between-study variance) in this collection of studies.

Fig. 6 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 Conclusion

We have compared eight meta-analytic approaches in the context of one-sample test. Through simulations, we found the expected invalidity of standard FFX

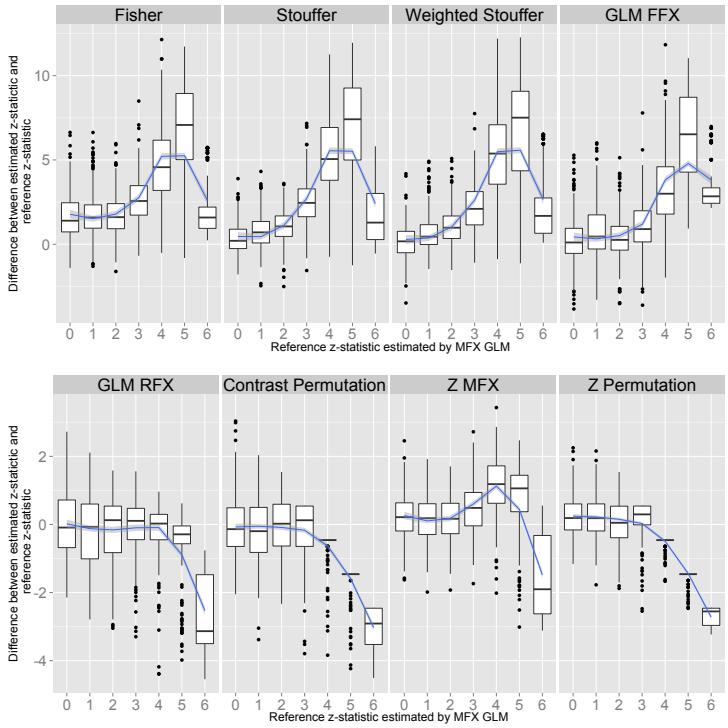

Fig. 6. Difference between the z-score estimated from each meta-analytic approach and the reference z-score from MFX GLM as a function of reference z-score.

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 [8]. 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.

5 Acknowledgements

We gratefully acknowledge the use of this data from the Tracey pain group, FMRIB, Oxford.

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. 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. 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.
7. T. Liptak. On the combination of independent tests. *Magyar Tud. Akad. Mat. Kutato Int. Kozl.*, 3:171–197, 1958.
8. J. A. Mumford and T. E. Nichols. Simple group fMRI modeling and inference. *NeuroImage*, 47(4):1469–75, 2009.
9. 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.
10. Thomas E. Nichols. Spm plot units, 2012.
11. J. Radua and D. Mataix-Cols. Meta-analytic methods for neuroimaging data explained. *Biology of mood & anxiety disorders*, 2(1):6, 2012.
12. F. E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2:110–114, 1946.
13. 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.
14. S. Stouffer, L. DeVinney, and E. Suchmen. *The American Soldier: Adjustment During Army Life*, volume 1. Princeton University Press, Princeton, NJ, 1949.
15. 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.