Standardized Effect Sizes and Image-Based Meta-Analytical Approaches for fMRI Data.

Han Bossier^{1*}, Thomas E. Nichols² & Beatrijs Moerkerke¹

¹Ghent University, Henri-Dunantlaan, 2, 9000, Ghent, Belgium

²Big Data Institute, Li Ka Shing Centre for Health Information and Discovery,
Nuffield Department of Population Health, University of Oxford, Oxford, OX3
7LF, United Kingdom

*Corresponding author: Han.Bossier@Ugent.be

Abstract: Scientific progress is based on the ability to compare opposing theories and thereby develop consensus among existing hypotheses or create new ones. We argue that data aggregation (i.e. combine data across studies or research groups) for neuroscience is an important tool in this process. An important prerequisite is the ability to directly compare fMRI results over studies. In this paper, we discuss how an observed effect size in an fMRI data-analysis can be transformed into a standardized effect size. We demonstrate how these enable direct comparison and data aggregation over studies. Furthermore, we also discuss the influence of key parameters in the design of an fMRI experiment (such as number of scans and the sample size) on (statistical) properties of standardized effect sizes. In the second part of the paper, we give

an overview of two approaches to aggregate fMRI results over studies. The first corresponds to extending the two-level general linear model approach as is typically used in individual fMRI studies with a third level. This requires the parameter estimates corresponding to the group models from each study together with estimated variances and meta-data. Unfortunately, there is a risk of running into unit mismatches when the primary studies use different scales to measure the BOLD response. To circumvent, it is possible to aggregate (unitless) standardized effect sizes which can be derived from summary statistics. We discuss a general model to aggregate these and different approaches to deal with between-study heterogeneity. Furthermore, we hope to further promote the usage of standardized effect sizes in fMRI research.

Keywords: replicability, standardized effect sizes, GLM, data aggregation, fMRI

Conflict of interest: The authors declare no conflict of interest

1 Introduction

The analysis of neuroimaging data such as those obtained by functional magnetic resonance imaging (fMRI) is a complex endeavour. There are several challenges that need to be addressed. First there is a high cost to scan subjects. Second the data are typical of a high dimensionality and are characterized by a low signal to noise ratio ([Button et al., 2013](#)). Moreover, due to its complexity there is a high flexibility in the analysis and reporting of results ([Carp, 2012](#)). Given these challenges, it has been pointed out in literature that statistical power to detect common effect sizes in the neuroscience literature is rather low ([Poldrack et al., 2017](#)), resulting in an increased prevalence of false positive published effects ([Ioannidis, 2005](#)) and results that cannot be replicated ([Patil et al., 2016](#)) using an independent sample of new data ([Turner et al., 2018](#)). Several initiatives to address these issues have been proposed in literature. The first is the development of open source tools and standardization protocols to facilitate good research practices ([Poldrack and Poline, 2015](#)). For instance NeuroPower from the NeuroPowerTools¹ can be used to calculate a minimal sample size to obtain sufficient statistical power when planning whole brain fMRI analyses ([Durnez et al., 2016](#)). Next the design of an fMRI experiment can be optimized (in terms of statistical efficiency) using NeuroDesign from the same software library. While these tools focus on planning an experiment, there is also an increased effort to standardize the organization and storage of

¹<http://neuropowertools.org>

obtained fMRI data. Examples are the Brain Imaging Data Structure ([Gorgolewski et al., 2016](#)) or the Neuroimaging Data Model ([Maumet et al., 2016](#)) which facilitates representation of meta-data. Combined, these protocols enable easy sharing of data between research groups.

A second solution directly addresses insufficient power by collecting data of thousands of individuals through large scale collaborations. Examples are the Human Connectome Project ([Van Essen et al., 2013](#)) or the UK Biobank ([Sudlow et al., 2015](#)). Finally, as an individual researcher it is possible to increase a sample size by aggregating data collected earlier (either published or not). Scientifically, this can be done using different meta-analytical techniques ([Wager et al., 2007](#); [Costafreda, 2009](#)). The general idea is to synthesize results while weighting the different studies by reliability (e.g. studies with a larger sample size typically have a higher weight in a meta-analysis). Note that such initiatives are only possible when researchers engage in a data sharing culture ([Poline et al., 2012](#)).

In common between these solutions is the need to compare results over fMRI studies. This is not only an obvious prerequisite in order to combine results in a meta-analysis ([Bowman, 2012](#); [Borenstein et al., 2009](#)) but also to gain insight into the magnitude of an effect. Most studies focus on null hypothesis significance testing, where the null hypothesis corresponds to the absence of an effect ([Friston et al., 1995](#)), and only report on the statistical significance. In case of fMRI data, an effect reflects the amplitude of the blood-oxygenated-level dependent (BOLD) signal obtained by contrasting

(several) experimental conditions. The strength of the evidence in favor of the alternative hypothesis depends on the sample size as precision increases monotonically with the latter. Mathematically, it is true that any real effect (small or large) can be declared statistically significant as long as the sample size is large enough (Friston, 2012; Bzdok and Yeo, 2017). For this reason, the appropriate question is not only whether there is a true effect (and the direction of the effect) but also what the magnitude of the underlying effect is. This is termed practical significance (Kirk, 1996) and can be quantified by a (voxelwise) point estimate for the effect as well as an estimate of its precision. If effects are measured on a meaningful scale, then there is no need to standardize these (Cummings, 2011) before comparing or when combining over studies. However, the BOLD signal has an arbitrary unit (Chen et al., 2017) which creates a need to standardize fMRI effect sizes.

To our knowledge, the description of standardized effect sizes and an overview of methods to combine these estimates have not yet been fully documented in the field of neuroimaging. The goals of this paper are (1) to describe the advantages of standardized effect sizes, (2) to give an overview of the calculations and statistical properties of standardized effect sizes specifically related to fMRI data and (3) to discuss several methods to pool data across studies while dealing with heterogeneity between studies.

Note that several papers in literature describe various methods and tools to aggregate neuroimaging data (see for an overview: Wager et al., 2007) or provide specific guidelines on how to perform a meta-analysis with neuroimaging

data (Müller et al., 2017). Most of these methods and diagnostic tools focus on coordinate-based meta-analyses that combine results over studies using only the 3D coordinates of peak activation from the individual studies (Acar et al., 2018; Albajes-Eizagirre and Radua, 2018; Eickhoff et al., 2009, 2012; Turkeltaub et al., 2002, 2012; Radua et al., 2012; Tench et al., 2017; Wager et al., 2009, 2004). In this paper, we focus on the context where full brain images are available and hence where information for each voxel is combined over studies. This is termed image-based meta-analysis. For a detailed comparison between both, see: Salimi-Khorshidi et al. (2009).

2 Effect sizes

We start this section with a more detailed argumentation to incorporate standardizing effect size estimates for fMRI data. We will contrast these estimates with the BOLD percent signal change. Thereafter, we will discuss the computations underlying standardized effect sizes and third, describe its relationship with fMRI data.

2.1 Why standardize?

The BOLD percent signal change is a frequently used measure of the magnitude of an effect (Chen et al., 2017). It is measured in response to an experimental condition and corresponds to the peak amplitude of the fMRI time series in relation to a baseline signal. The latter represents an average

signal within each voxel over all time points. However, an average fMRI signal can vary between subjects or scanning sessions due to various external factors (e.g. caffeine level of the subject). Therefore, a normalization step is executed where the entire signal in each voxel within each subject and session is scaled to a target value. Written otherwise, if we denote Y as the observed signal, then the normalization corresponds to: $\frac{Y}{\text{mean}(Y)} \times \text{target value}$. The target value ensures different sessions and/or participants are comparable *within* an analysis. The main issue is that there are incompatibilities between various software libraries for the analysis of fMRI data. More specifically, there are differences in 1) techniques to calculate an average signal (i.e. $\text{mean}(Y)$) and 2) the target values chosen to scale the signal. AFNI (Cox, 1996) for instance calculates the average signal within each voxel separately while SPM²(Statistical Parametric Mapping: Wellcome Department of Cognitive Neurology, London, UK) and FSL³ (Smith et al., 2004) perform a grand mean scaling where the average signal is calculated over all masked voxels (Chen et al., 2017).

With respect to scaling, SPM for instance sets the target value to approximately 100 while FSL (Smith et al., 2004) scales to ± 10.000 due to historical differences in the implementation of the software⁴. Given that the BOLD percent signal change is calculated with respect to the baseline, the value depends on the technique to calculate the latter (i.e. software dependent) prohibiting direct comparison across studies using different software. In other words, con-

²<http://www.fil.ion.ucl.ac.uk/spm>

³<https://fsl.fmrib.ox.ac.uk/fsl/fslwiki>

⁴<https://blog.nisox.org/2012/07/31/spm-plot-units/>

sider two related studies that report a 3% BOLD signal change within an area of activation. The researchers of the first study relied on AFNI while the others on FSL. Unfortunately, the reported effects will be slightly different due to differences in the calculation of the average signal. Therefore, the BOLD percent signal change should be regarded as a non-standardized effect size. One of the solutions is to report standardized effect sizes (that are comparable across studies) next to the BOLD percent signal change.

To continue, we discuss the calculation of these standardized effect sizes and its relation to fMRI in section [2.3](#).

2.2 Standardized effect sizes

While standardized effect size calculations for a mean difference between two populations are well described ([Borenstein et al., 2009](#); [Zakzanis, 2001](#)), this is not the case for a one sample average effect. In a group analysis of an fMRI study however, this is often of interest when pooling results over subjects. Therefore, we provide a general description for this specific effect size measure. Assume that the observed response variable Y for a given study i with $i = 1, \dots, k$ is distributed as $Y \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Here, μ_i is the true population effect of study i and σ_i^2 the corresponding within-study variance. The true standardized mean effect, δ_i , is defined as:

$$\delta_i = \frac{\mu_i}{\sigma_i}. \quad (1)$$

For ease of notation, we will drop the subscript i . Usually, δ is estimated using Glass’s estimator ([Glass, 1976](#)) or similarly for a one-sample case using Cohen’s d ([Cohen, 1988](#)):

$$d = \frac{\bar{Y}}{S}, \quad (2)$$

where S corresponds to the square root of the unbiased estimator for the sampling variability. Using equation (2) results in an expression of the observed effect ($\bar{Y}$) in units of S , which is a standardization procedure. However as shown in [Hedges \(1981\)](#) and demonstrated in section (5.1) of the Appendix, equation (2) is a biased estimator for δ . Relying on d results in an overestimation of δ , especially for small sample sizes. Using the derivations in [Hedges](#)

(1981), the exact unbiased estimator (g_e) is defined as:

$$g_e = \frac{\bar{Y}}{S} \times h, \quad (3)$$

where

$$h = \frac{\Gamma\left(\frac{N-1}{2}\right)}{\sqrt{\frac{(N-1)}{2}} \Gamma\left(\frac{N-2}{2}\right)}, \quad (4)$$

Γ represents the gamma function and N equals the sample size of the study.

Due to historical limitations in computing h , Hedges (1981) suggested an approximation (g_a):

$$g_a = \frac{\bar{Y}}{S} \times J, \quad (5)$$

where

$$J = \left(1 - \frac{3}{4(N-1) - 1}\right). \quad (6)$$

Note that for $N \geq 10$, we have $g_a - g_e < 0.001$. Even though the difference between both effect size calculations is small, computational resources nowadays enable calculating g_e instead of g_a . The induced overestimation using Cohen's d for small sample sizes and the difference between g_e and g_a is illustrated in Figure 1. On the y-axis we plot the ratio between the true effect size (i.e. δ) over the expected value of d (i.e. $E(d)$) hence demonstrating the overestimation.

Figure 1: Left panel: ratio of the true effect size (δ) versus the expected value of Cohen's d . Estimating an effect size with d thus results in an overestimation of δ for small sample sizes. The ratio equals the correction factor h used in the exact unbiased estimator (i.e. g_e). The right panel shows the difference between the exact expression and the approximation (i.e. g_a using correction factor J) when the $df \leq 7$. To summarize: $h = \frac{\delta}{E(d)}$ and $J \approx \frac{\delta}{E(d)}$.

Finally the expectation of g_e and its variance are given as:

$$E(g_e) = \frac{\mu}{\sigma} \quad (7)$$

$$= \delta \quad (8)$$

$$\text{Var}(g_e) = \text{Var}(d) \times h^2 \quad (9)$$

$$= \left[\frac{(N-1)(1+N\delta^2)}{N(N-3)} - \frac{\delta^2}{h^2} \right] \times h^2. \quad (10)$$

To estimate the within-study variability of the standardized effect, we plug in the estimate for δ using equation (2). We refer to the Appendix, section 5.1 for derivations.

For completeness, we provide the estimator for a standardized mean difference (g_Δ) between a control group with N_c subjects and an experimental group with N_e subjects together with its expectation and variance (Hedges, 1984; Borenstein et al., 2009).

$$g_\Delta = \frac{\bar{Y}_e - \bar{Y}_c}{S_{pooled}} \times h_\Delta, \quad (11)$$

where

$$S_{pooled}^2 = \frac{(N_e - 1)S_e^2 + (N_c - 1)S_c^2}{N_e + N_c - 2} \quad (12)$$

$$h_\Delta = \frac{\Gamma\left(\frac{N_e + N_c - 2}{2}\right)}{\sqrt{\frac{(N_e + N_c - 2)}{2}} \Gamma\left(\frac{N_e + N_c - 3}{2}\right)} \quad (13)$$

And then:

$$E(g_{\Delta}) = \frac{\mu_e - \mu_c}{\sigma} \quad (14)$$

$$\text{Var}(g_{\Delta}) = \frac{(N_e + N_c - 2)}{(N_e + N_c - 4) \frac{N_e N_c}{N_e + N_c}} \left[1 + \frac{N_e N_c}{N_e + N_c} \delta^2 \right] - \frac{\delta^2}{h_{\Delta}^2} \times h_{\Delta}^2 \quad (15)$$

2.3 Standardized effect sizes for fMRI data

To discuss the relation between g_e and fMRI data, we need to give a brief overview of the general model fitting procedure that is typically used in a mass univariate fMRI data analysis.

2.3.1 Statistical analysis of fMRI data

Due to computational limitations, the statistical analysis of fMRI data is split-up in two stages/levels. The first stage consists of fitting a general linear model (GLM) with the expected time series under the experimental condition and possible nuisance parameters to the observed BOLD signal. This is done for each voxel and subject separately. We will not use a subscript to denote the voxel for ease of notation. For a single subject s , we have (Friston et al., 1994):

$$\mathbf{Y}_s = \mathbf{X}_s \boldsymbol{\beta}_s + \boldsymbol{\varepsilon}_s, \quad s = 1, \dots, N, \quad (16)$$

where $\mathbf{Y}_s$ is a vector of length T containing the BOLD signal on T measurements, $\mathbf{X}_s$ is the design matrix containing a convolution of the stimulus onset

function with a hemodynamic response function (HRF) (Henson and Friston, 2007) as well as nuisance covariates, β_s is a vector with the parameters and ϵ_s is a vector of length T containing within-subject random error. Note that due to low-frequency drift, the within-subject error is not independently distributed but given by $\epsilon_s \sim \mathcal{N}(0, \sigma_s^2 \mathbf{V}_s)$, where $\mathbf{V}$ is the temporal autocorrelation (Mumford and Nichols, 2006). To obtain efficient estimators, we assume that the data, model and errors are decorrelated (whitened). This is usually done by premultiplying $\mathbf{Y}_s$ and $\mathbf{X}_s$ with a matrix $\mathbf{K}_s$ such that $\mathbf{K}_s \mathbf{V}_s \mathbf{K}_s^T = \mathbf{I}$ (Poline and Brett, 2012). By doing so, we have $\text{Var}(\epsilon_s) = \sigma_s^2 \mathbf{I}$ for each subject. Furthermore, note that for a fixed true β_s , we have

$$\text{Var}_\beta(\mathbf{c}\hat{\beta}_s) = \mathbf{c}(\mathbf{X}_s^T \mathbf{V}_s^{-1} \mathbf{X}_s)^{-1} \mathbf{c}^T \sigma_s^2, \quad (17)$$

with $\hat{\beta}_s$ the vector of parameter estimators and where $\mathbf{c}$ represents a contrast vector forming a linear combination of parameter estimates (Mumford and Nichols, 2009).

The second stage, denoted with subscript G , is used to estimate a population effect by combining all participants. The response variable is the vector of estimated first level contrasts $\mathbf{Y}_G = \{\mathbf{c}\hat{\beta}_1, \dots, \mathbf{c}\hat{\beta}_N\}^T$. We now get

$$\mathbf{Y}_G = \mathbf{X}_G \beta_G + \epsilon_G, \quad (18)$$

where $\mathbf{X}_G$ is a group design matrix and ϵ_G a group error vector. In the simplest case, $\mathbf{X}_G$ corresponds to the group average which is a vector of length N with ones. Importantly, ϵ_G is a mixed effects error component as it contains both variability of the imperfect intra-subject fit at the first stage (i.e. $\mathbf{c}\hat{\beta}_s$)

and between-subject variability, denoted as σ_G^2 (Mumford and Nichols, 2006, 2009). To simplify the notation here, we look at the unusual case where each participant has the same design matrix $\mathbf{X}$ and we assume there is no temporal autocorrelation in the time series of each subject (or we work with pre-whitened data). We thus have $(\mathbf{X}_j^T \mathbf{X}_j)^{-1} = (\mathbf{X}_l^T \mathbf{X}_l)^{-1}$ for all j, l after which we drop the subscript in $\mathbf{X}$. The variance of the error component in model (18) becomes:

$$\text{Var}_G(\boldsymbol{\varepsilon}_G) = \text{Var}_\beta(\mathbf{Y}_G) + \sigma_G^2 I_N. \quad (19)$$

with

$$\widehat{\text{Var}}_\beta(\mathbf{Y}_G) = \text{diag}\{\widehat{\sigma}_1^2 \mathbf{c}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{c}^T, \dots, \widehat{\sigma}_N^2 \mathbf{c}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{c}^T\} \quad (20)$$

where diag indicates the diagonal operator and β denoting this is an intra-subject variance.

Next we assume that the within-subject variance is homogeneous over all subjects (i.e. $\sigma_j^2 = \sigma_l^2$ for all j, l). This simplifies our discussion on standardized effect sizes in the following section. Furthermore, by doing so we are able to rely on the ordinary least squares (OLS) approach (Holmes and Friston, 1998) to estimate the parameters of equation (18). Finally, we get for a fixed true β_G :

$$\text{Var}_G(\mathbf{Y}_G) = [\mathbf{c}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{c}^T \sigma^2 + \sigma_G^2] I_N. \quad (21)$$

We only need to estimate a single (combined within- and between-subject) variance term. We shall denote S_G^2 as the sample variance of Y at the group level (i.e. $S_G^2 = \widehat{\text{Var}}_G(\mathbf{Y}_G)$). Inference for the simple group average case proceeds with a one sample t -test. The test statistic (i.e. a t -value) of a parameter of interest (i.e. a specific β_G) corresponds to $t = \frac{\hat{\beta}_G}{S_G/\sqrt{N}}$ and is compared to a t -distribution with $N - 1$ degrees of freedom. The result is a statistical parametric map (SPM) containing a test statistic for each voxel.

2.3.2 Model fitting procedure and effect size estimation

While it is possible to calculate a voxelwise standardized effect for each subject in an fMRI experiment, the general interest is to measure a standardized group effect. Specifically for fMRI, the one-sample standardized mean effect (i.e. $\mathbf{X}_G$ = a column vector of 1) is obtained by combining equations (3), (18) and (21) as:

$$g_e = \frac{\hat{\beta}_G}{S_G} \times h. \quad (22)$$

Note that it is also possible to obtain an effect size using the t -value in each voxel. More particularly, we have $g_e = \frac{t}{\sqrt{N}} \times h$.

Due to the two-stage procedure when fitting the GLM, there are some characteristics/effects on the standardized effect size worth noting. To start with, consider the expression of δ which contains both an estimate of the average effect as well as one of the sampling variability. Therefore, both estimators for μ and σ need to be unbiased. For instance, we make the assumption that

the individual time series are effectively decorrelated. Furthermore we need to assume the correct statistical model is used to estimate the parameters of the GLM (e.g. movement parameters are included if necessary). In other words, we assume researchers are measuring a relevant effect instead of an artefact. Second, the derivations for the distributional properties of the standardized effect sizes are obtained using the assumption of a normally distributed outcome variable. In the case of fMRI, we thus assume normally distributed β_G 's, which follows from the central limit theorem if N is sufficiently large. Next, there is an effect of both the length of the individual time series (T) and the number of participants (N) on the standardized effect size and its variance. First note that the sampling variability (S_G^2) will decrease as T increases. This is the case as the intra-subject variability due to estimating the single-subject parameters in the first stage (i.e. the first part in equation 21) decreases. In turn, δ will increase. Thus one obtains larger effect sizes using longer scanning sequences. However note that larger effect sizes are associated with larger variances of the effect size estimator as can be seen in equation (10). Secondly, while the effect size is independent of the sample size, the precision to estimate δ will increase with N . To demonstrate both the effect of T and N , we set up a small Monte-Carlo simulation study. The code for this simulation and to obtain the resulting figures can be found at: <https://github.com/NeuroStat/ESfMRI>. Using the *neuRosim* package (Welvaert et al., 2011) in R, we generate individual fMRI BOLD time series by convoluting a block design experiment with a canonical HRF and add white

random noise to the signal. Within-subject variability is homogeneous across all subjects and a small amount of between-subject variability is added. We then vary the length of the individual time series with $T \in [100, \dots, 500]$ and let $N = 20$ or $N = 100$. A standard two-stage GLM is fitted and we estimate the group level standardized effect size with its variance using the formula provided above. Results are shown in Figures 2 and 3. First we visualize standard properties of the OLS estimator for the parameters of interest in the second level GLM. The estimator for β_G is unbiased and its precision increases with both T and N (i.e. lower variability over the simulations). Second we observe higher values for g_e as T increases and a higher precision as N increases. Furthermore, the variance of g_e increases with T . This can be seen in panel **D** of Figure 5.3 where the red line indicates the expected value for the variance. Note that the increase is rather small. Indeed, the empirical variance (i.e. over Monte-Carlo simulations) of the estimator for g_e increases, albeit this is too small to be detected in panel **C** of Figure 5.3. A demonstration of the increase in the empirical variance is given in Appendix, section 5.2, where we calculate the variance over the simulations of the estimates for g_e marginally over the number of subjects. Note that even though the empirical distribution of $\text{Var}(g_e)$ is skewed (see panel **D** of Figure 5.3), the estimator is unbiased. A demonstration of this property is given in the Appendix, section 5.3. Furthermore, note how the variance for g_e decreases with the number of subjects. This can be seen in both panel **C** and **D** of Figure 5.3 and is calculated explicitly in the Appendix, section 5.2. This effect is more substantial than the increase

in the variance associated with the number of scans.

Finally, there will be an effect of spatial smoothing in the pre-processing on the estimated standardized effect size. As smoothing directly influences the signal and noise of within-subject time series, the smoothed standardized effect size will be different from the raw standardized effect.

An important remark is the distinction between effect size estimation and null hypothesis significance testing for fMRI. Both procedures are mainly used within a massive univariate context ([Lindquist, 2008](#)). However, inference procedures for the parameters of the GLM critically relies on correcting for multiple testing ([Nichols and Hayasaka, 2003](#); [Nichols, 2012](#)). While effect sizes can be calculated on thresholded maps, the resulting estimates will be over-optimistic as the amount of tests increases. This is true as only voxels with the smallest amount of noise will survive the threshold for significance, irrespective of the underlying true effect size ([Reddan et al., 2017](#)). Therefore, we suggest to complement significance testing with estimating the standardized effect sizes on non-thresholded images.

Figure 2: Empirical distributions over 1000 Monte-Carlo simulations of the estimated group level parameters in a typical two-stage fMRI general linear model (see equation (18)). The length of the individual time series and the total sample size increases to demonstrate its effect on the parameter estimates. Panel **A** depicts the estimated group level parameters, while panel **B** gives S_G^2 .

Figure 3: Empirical distributions over 1000 Monte-Carlo simulations of the estimated g_e (panel **C**) and its variance (panel **D**) using equation (10). The length of the individual time series and the total sample size increases to demonstrate its effect on the parameter estimates. The solid red line in panel **D** corresponds to the expected value.

3 Meta-analysis

There are several methods available to aggregate fMRI data. However, the amount of information available to the analyst will largely determine which method can be used. We discuss two strategies aggregating results when full brain images are available (i.e. image-based meta-analysis). In an fMRI meta-analysis, three levels can be distinguished: subjects, studies/groups and the meta-analytical level.

In the first method, full brain images from different studies are modeled by extending the two-stage hierarchical general linear model presented in section (2.3.1). To do so, the analyst needs the parameter estimates of the GLM of each study (equation 18) together with an image containing the estimate of its variance. Furthermore, depending on the software used within each study, meta-data will be required to scale data in order to avoid unit mismatches between studies. The second method relies on a transformation of the estimates from each study to standardized effect sizes (discussed in section 2.3.2). This approach can be used if only images containing the test-statistic of each study are available.

3.1 Extending the GLM

We start with the case where most information is available. It is possible to extend the two-stage hierarchical GLM presented in equation (18) with a third level. To do so, we need at least the following information from each

study i with $i = 1, \dots, k$: the estimated contrast parameters of interest from the GLM $\hat{\beta}_{G,i}$ together with their estimated variances, the sample size or preferably the effective degrees of freedom for each voxel and finally meta-data such as the mean intensity of each study image. If we use M to denote the third (meta-analytical) level and $\mathbf{c}_G$ as a contrast forming vector at group level (e.g. contrasting patients with controls) in each study, then we have $\mathbf{Y}_M = \{\mathbf{c}_G \hat{\beta}_{G,1}, \dots, \mathbf{c}_G \hat{\beta}_{G,k}\}^T$. The corresponding GLM can be written as:

$$\mathbf{Y}_M = \mathbf{X}_M \beta_M + \boldsymbol{\varepsilon}_M, \quad (23)$$

where $\mathbf{X}_M$ represents the meta-analytical design matrix used to aggregate data. Similar to the two-stage model presented earlier, the error term consists of two parts. In this case: within-study variability and between-study variability (σ_M^2). Note however that within-study variability on its own consists of the intra-subject and between-subject variability. Written formally, we thus have:

$$\text{Var}(\boldsymbol{\varepsilon}_M) = \text{Var}_G(\mathbf{Y}_M) + \sigma_M^2 \mathbf{I}_k, \quad (24)$$

with

$$\widehat{\text{Var}}_G(\mathbf{Y}_M) = \text{diag}\{\hat{\sigma}_1^2 \mathbf{c}_G (\mathbf{X}_{G,1}^T \mathbf{X}_{G,1})^{-1} \mathbf{c}_G^T, \dots, \hat{\sigma}_k^2 \mathbf{c}_G (\mathbf{X}_{G,k}^T \mathbf{X}_{G,k})^{-1} \mathbf{c}_G^T\}$$

and diag again indicating the diagonal operator and for each study i ($i = 1, \dots, k$), $\sigma_i^2 = \text{Var}_G(\boldsymbol{\varepsilon}_{G,i})$ as in expression 19. Note that differences in sample sizes may cause within-study variability being non-homogeneous. In turn,

this violates an assumption underlying the OLS approach to estimate the parameters of model 23. As explained in Mumford and Nichols (2009), it is possible to relax the assumption of homogeneous within-study variances by using a generalized least squares approach. Since this requires an estimation of multiple variance terms, we need to use iterative algorithms such as restricted maximum likelihood. For fMRI, it is possible to use a general linear mixed model implemented as *FLAME1* or *FLAME1+2* in FSL (Woolrich et al., 2004) or the approach implemented in fMRIstat (Worsley et al., 2002). Note that if the assumption of within-study homogeneity is violated, one should avoid the summary statistic approach in SPM12⁵. However, it is possible to use a full general linear mixed model implemented as *spm_mfx.m* in SPM12 and described in (Friston et al., 2005).

While modeling the subject- or study-level data using a hierarchical three-stage GLM seems straightforward, special attention needs to be given to the units of the data. Ideally, the effect of an experimental paradigm (i.e. the β parameters in the GLM) is expressed in the same unit for each study before going into model (23). However as mentioned in section 2.1, the BOLD signal has an arbitrary unit. This will not only depend on the scaling technique used when normalizing the time series (as discussed earlier), but also on the scale of the design matrix and the scaling of the contrast of interest at the first level of the GLM in each study (Maumet and Nichols, 2015). With respect to the design matrix, ideally predictors have a peak amplitude (height)

⁵<https://www.fil.ion.ucl.ac.uk/spm/software/spm12/>

of 1 ensuring the effect is expressed in the same unit as the data. This is usually the case for block event designs, but rarely for event related designs with short inter-trial intervals. With respect to the scaling of the contrast, ideally the positive part sums to 1 and the negative part sums to -1 (see also <https://blog.nisox.org/2012/07/31/spm-plot-units/>). In most cases, the analyst will not be able to change these parameters. Avoiding unit mismatches between studies will therefore become challenging.

To demonstrate this, assume that two individual fMRI studies rely on the same technique to calculate an average signal and both studies are interested in the same experimental contrast, measuring a relevant effect. As discussed above, a typical data analysis consists of fitting a general linear model (GLM) in each voxel. Even though both studies could measure the same BOLD percent signal change, the parameter estimates in the GLM will differ if the target value of the normalization step differs. To compare these parameter estimates over both studies (possibly with the aim to aggregate them in a meta-analysis), one then needs to use an extra parameter to scale the design matrix from one study to another. An illustration of this problem is given in Figure 4. Using R (R Core Team, 2015), we run a simulation⁶ where we generate 1000 time series for $N = 50$ subjects each. We design two scenarios. These correspond to the signal being normalized with a target value in the normalization step corresponding to either 10.000 (option A) or 100 (option B). We demonstrate how a difference in normalization results in parameter estimates that cannot

⁶The code can be found at: <https://github.com/NeuroStat/ESfMRI>

be combined unless the analyst knows how to scale data between studies (using a scaling parameter c on Figure 4).

If raw data (i.e. subjects) are available, one could re-run the entire analysis within one software library thus avoiding differences in scaling techniques between software libraries and with the ability to re-code contrasts vectors. Furthermore, one could obtain for each study the effective peak height of the predictors in the design matrix and use this value to scale the parameter estimates accordingly before fitting the second-level GLM.

If only the second-level group data are available (i.e. parameter estimates and their estimated variances), the analyst will need to rely on meta-data (e.g. software library used, type of experiment, the chosen contrast of interest,...) to scale the estimates accordingly before running a third-level. Standardization protocols such as the Neuroimaging Data Model (Maumet et al., 2016) attempt to provide machine-readable descriptions of studies together with the necessary meta-data to overcome these issues.

If no meta-data is available or only a limited amount of information is available (e.g. only the statistical parametric map), then the analyst has to rely on techniques that do not depend on the unit of the data. We will discuss these now.

Figure 4: Illustration of pooling two studies (A and B) where each have a different target value in the normalization step. This results in regression coefficients (i.e. β_{A1} and β_{B1}) that cannot be compared directly. Two solutions are possible: 1) one can scale the design matrix (i.e. X) of one study to the other (resulting in β_{C1}) or 2) one could standardize the observed effect size (i.e. $s\overline{ES}$). To generate the observed time series, we first create a true signal by convoluting a block design experiment (20 sec ON/OFF) with a canonical hemodynamic response function (HRF) (Henson and Friston, 2007). The true signal is scaled so that it corresponds to 3% BOLD signal change relative to the baseline. We then add white random noise to the true signal and a small amount of homogeneous between-subject variability. Note: c corresponds to the scaling parameter needed to combine study A and B .

3.2 Aggregating standardized effect sizes

An alternative method to aggregate fMRI studies relies on transforming the data beforehand. As discussed in section 2.3.2, it is possible to transform a test-statistic to standardized effect sizes. Furthermore, note that the non-systematic variability of standardized effect sizes is proportional to the inverse of the sample size ($\frac{1}{n}$) of the study (see equation 10). Moreover, the variance is completely determined by the estimate of the effect size and the sample size. This has three important implications. First, effect size estimates have a different error variance due to different sample sizes between studies (Hedges, 1984). These differences are accounted for when aggregating effect sizes by using a weighting approach (discussed below). Second, aggregating studies is more feasible as one only needs the estimate of the effect magnitude (obtained using a test statistic) and the corresponding study sample size. For fMRI it is thus possible to do a meta-analysis using the statistical parametric maps in conjunction with the sample sizes. Finally, it is possible to use all degrees of freedom between the different effect size estimates to estimate systematic differences between studies (Hedges, 1984). In other words: there is no need to iteratively estimate within-study and between-study variance components as could be the case in the three-stage GLM from equation (24).

3.2.1 Models for aggregating effect sizes

In the past, it was custom to ignore any possible between-study variability (i.e. fixed-effects models) when aggregating effect sizes. However there is a growing

consensus that more complex models are preferred over these simple fixed-effects models (Hunter and Schmidt, 2000; Viechtbauer, 2005). Similar to the three-stage hierarchical model, variability of effect sizes are then decomposed into two parts. The first is the sampling variance (i.e. within-study variability) determined by the sample size and the second equals heterogeneity between studies due to observed effect sizes being drawn from a population of effect sizes (i.e. between-study variability). Typically, models incorporating the latter are denoted as a random effects model. However we will discuss a more general case which allows for moderator variables to be added. We introduce new notation to differentiate models for standardized effects from those in BOLD percent signal change discussed earlier (section 3.1). We can write a meta-analysis for all $i = 1, \dots, k$ studies as a linear mixed-effects model (e.g. Raudenbush, 2009; Viechtbauer et al., 2015):

$$\mathbf{Y}^* = \mathbf{X}^* \boldsymbol{\beta}^* + \mathbf{u} + \boldsymbol{\varepsilon}^*, \quad (25)$$

where $\mathbf{Y}^*$ denotes the $(k \times 1)$ vector of observed (standardized) effect sizes and $\mathbf{X}^*$ is the $k \times (p + 1)$ meta-analysis design matrix with p possible moderator variables. Next, $\boldsymbol{\beta}^*$ is the column vector of length $(p + 1)$ with parameters, $\mathbf{u}$ is the vector of random effects distributed as $\mathcal{N}(0, \tau^2)$. Then $\boldsymbol{\varepsilon}^*$ is the within-study error with the individual terms distributed as $\varepsilon_i^* \sim \mathcal{N}(0, \sigma_i^{*2})$. The mentioned random effects model without moderator variables corresponds to the case where $\mathbf{X}^*$ is a column vector of 1's. The intercept in $\boldsymbol{\beta}^*$ then corresponds to the population average. Furthermore τ^2 reflects all between-study

heterogeneity not accounted for by the moderators. A weighted least squares approach is used to estimate the parameters in β^* . The usual estimator for β^* is denoted by B^* and is expressed as:

$$B^* = (X^{*T} \widehat{W} X^*)^{-1} X^{*T} \widehat{W} Y^*. \quad (26)$$

In this, $\widehat{W}$ denotes a diagonal weight matrix with elements $w_i = 1/(\sigma_i^{*2} + \widehat{\tau}^2)$. We thus give less weight to studies with a lower sample size (i.e. lower precision).

A great deal of attention in the literature deals with methods to estimate τ^2 , the between-study heterogeneity (Higgins and Thompson, 2002; Veroniki et al., 2016; Viechtbauer, 2005). Choosing an appropriate method to estimate τ^2 is important for two reasons. First, a biased estimator leads to an over- or underestimation of β^* . Second, the variance-covariance matrix of B^* used in a standard Wald-type test can be expressed as $(X^{*T} \widehat{W} X^*)^{-1}$ (Viechtbauer et al., 2015) and hence, the precision to estimate the parameters is also a function of τ^2 .

In the following section, we will discuss three methods that are worth mentioning to estimate τ^2 for fMRI data. These are the popular DerSimonian and Laird estimator (DerSimonian and Laird, 1986), the Hedges estimator and the restricted maximum likelihood estimator. Moreover, we will discuss these using an empirical assessment of the amount of between-study heterogeneity that one can expect in an fMRI meta-analysis. Obtaining such an estimate has three benefits. First it can be used in subsequent Monte-Carlo simula-

tion studies by providing sensible values for simulation parameters. Second, results of these assessments challenge the analyst to think ahead about potential between-study heterogeneity when planning a meta-analysis. Finally, the expression for τ^2 can attain negative values at which they are usually truncated at zero. This leads to biased estimates (Bohning, 2002). Hence it is interesting to know whether one can encounter this situation in an fMRI meta-analysis. We collect a real data set of 33 studies to discuss the estimators for τ^2 and the observed distribution of between-study heterogeneity within a region of interest.

3.2.2 Between-study heterogeneity in fMRI studies.

We collect a database of whole-brain fMRI studies involving a general experience of *pain* versus *no pain*. We obtain 33 studies using NeuroVault where we transform the SPMs to standardized effect sizes. We then estimate between-study heterogeneity using 3 possible estimators for τ^2 (discussed below). Furthermore we add two measurements that quantify the degree of between-study heterogeneity. We plot all the observed values over all voxels within a region of interest in Figure 5. The different steps are detailed below and the code is available at our GitHub repository.

Step 1: collecting studies

All studies are collected using the search term *pain* in NeuroVault (last search dates from January 2017). The contrast of interest is the effect of experiencing

pain versus a baseline or versus experiencing no pain. We apply a lenient definition of the experience of pain to obtain upper bounds on the observed between-study heterogeneity. The database contains multiple stimuli, ranging from auditory, thermal to mechanical stimuli. Results are manually checked for content and/or errors. If needed, images are re-sampled to MNI space with a voxel size of $2 \times 2 \times 2$ mm. Note that we do not explicitly follow guidelines (such as the PRISMA statement⁷) on how to collect/report studies for a meta-analysis/systematic review. Nor do we aim to assess whether publication bias known as the ‘file drawer problem’ (Rosenthal, 1979) is present. Hence the research question is not necessarily on the mapping of the experience of pain in the human brain. The database contains 33 studies from which 21 correspond to the work of one research group. Furthermore, 6 studies provide individual subject data containing only the beta estimates of the fitted GLM. We re-create the studies for these by fitting group level models using a standard ordinary least squares approach (Holmes and Friston, 1998), assuming homogeneous within-subject variability. The total sample size equals 659 subjects.

Step 2: defining a region of interest

As is done in Poldrack et al. (2017), we restrict our analysis to regions that are functionally relevant to the experience of pain. To obtain a region of interest (ROI), we run an automated meta-analysis using NeuroSynth⁸ (Yarkoni et al., 2011). This tool automatically extracts the coordinates of foci of activation

⁷<http://www.prisma-statement.org/>

⁸<http://neurosynth.org/>

from published papers. Out of the entire database of published fMRI studies, 420 studies reported activation associated with *pain*. For each voxel, a cross table *mentioned pain in a study (YES/NO) × reported activation (YES/NO)* can then be constructed. Next the probability that activation in a given brain voxel is associated with the term *pain* is calculated (i.e. $P(\text{activation}|\text{pain})$). The Pearson χ^2 test is used to test the null hypothesis of independence between each brain voxel and *pain*. A binary mask (i.e. functionally relevant: 1/0) is then created by controlling the false discovery rate at level 0.01.

Step 3: estimating between-study heterogeneity

Using the *NeuRRoStat* package⁹ in R, we first transform the *t*-values within the ROI to standardized effect sizes (i.e. g_e) based on equation (3). Within-study variability is calculated using equation (10). Next we use the *metafor* package (Viechtbauer, 2010) in R to fit a mixed-effects model as given in equation (25) where a fixed moderator variable is added to control for the 21 studies corresponding to the same research group. The weighted least squares approach requires an estimation of τ^2 - the between-study variability. To condense the amount of formula in the text, we refer to the appendix section 5.4 for the mathematical description of the three estimators for τ^2 .

The first is the popular DerSimonian and Laird (DL) estimator (DerSimonian and Laird, 1986). This method of moments estimator has the advantage of being very quick to calculate. However, as discussed in Viechtbauer (2005),

⁹<https://github.com/NeuroStat/NeuRRoStat>

the estimator is only unbiased assuming that within-study variances (σ^{*2}) are known (i.e. not estimated). Obviously this assumption does not hold and becomes especially problematic if the sample sizes of individual studies are small. Furthermore, even though the estimator does not require an assumption of normality for the random effects, it has been shown in a simulation study how other methods are preferred when the amount of studies in the meta-analysis is small (Jackson et al., 2010). A small adaptation of the DL estimator is given by the Hedges estimator (HE) (Viechtbauer, 2005). This estimator is proven to be unbiased when assuming the within-study variances are estimated rather than being known and has the same computational advantage as the DL estimator. However as shown in Viechtbauer (2005), the HE estimator is the least efficient, meaning the variance of the estimates is the highest among the three estimators. A final alternative is an iterative approach using the restricted maximum likelihood (REML). This estimator benefits from a good compromise between efficiency and being unbiased. The disadvantage is the increased computational time and possible convergence problems. We estimate τ^2 using all three approaches presented above.

Step 4: visualizing between-study heterogeneity

To complement the analysis, we add two commonly used estimates of between-study heterogeneity. The first one being H^2 which is defined as $\frac{Q}{k-1}$, where Q is the amount of observed dispersion between studies. Note that Q on itself depends on the estimator for τ^2 and H^2 should therefore not be considered as

an independent check of heterogeneity. The expected value of dispersion under the assumption of homogeneous effects (i.e. there is no between-study heterogeneity) equals $df = k - 1$ (i.e. equal to the degrees of freedom). Therefore, H^2 is a measure of the ratio of observed dispersion over the expected amount of dispersion. We also add I^2 which is defined as $\frac{Q-df}{Q} \times 100\%$ and thus the ratio between excess dispersion over the total (observed) dispersion. We visualize the distributions for these various parameters within the ROI in Figure 5. We observe values for H^2 and I^2 that indicate substantial between-study heterogeneity. For instance, the maximum values for H^2 and I^2 using the Hedges estimators equal 8.52 and 88.3% respectively. There is no rule of thumb to describe the observed heterogeneity as either low, mild or severe (Higgins and Thompson, 2002). However, we believe the observed dispersion can be considered substantial. With respect to the different estimators for τ^2 (i.e. DL, HE and REML), we note the following observations. First, the HE estimator is associated with more outlying estimates of τ^2 , followed by the REML estimator. However, this does not seem to affect the estimation of β^* as these distributions are nearly identical. The range of B^* over the three estimators is between -0.573 and 1.24 (panel C of Figure 5). Note that τ^2 is estimated on the same scale. Furthermore, we can construct a 95% credibility interval of the estimated B^* 's (for each estimator of τ^2) using the assumption of normally distributed effects around B^* (i.e. indicating the amount of variability one can expect). For instance, using the HE estimator, the interval associated with the voxel with the largest estimated B^* (i.e. 1.24) in the ROI is equal

to $[0.385; 2.09]$. The interval for the lowest observed B^* (i.e. -0.573) equals $[-1.36; 0.218]$. As a reference, the interval for the voxel with the highest value for $\hat{\tau}^2$ (i.e. 0.48) using HE equals $[-0.891; 1.82]$. Written otherwise, there is at least some between-study heterogeneity present that cannot be explained by the moderator variables included in the analysis. Moreover, we observe only a few number of zero-valued estimates. This implies we do not need to worry about bias induced to truncating the variance estimates to zero. Finally we plot the observed effects together with the estimates for τ_{DL}^2 on an $2 \times 2 \times 2$ MNI template in Figure (6). We can see how larger values for B^* are associated with larger values for τ_{DL}^2 .

To conclude, we show how a typical fMRI meta-analysis will need to deal with a substantial amount of between-study heterogeneity.

Figure 5: Empirical distributions for various parameters obtained after fitting a moderator meta-analysis for the contrast PAIN > NO PAIN within a ROI. Panel **A** and **B** visualize two measurements for the amount of between-study heterogeneity based on Hedges estimator. Panel **C** shows the estimated effect in standardized units. Panel **D** depicts the estimated between-study variability. To estimate the latter, we use the DerSimonian and Laird (LD) estimator, the Hedges (HE) estimator and an approach using the Restricted Maximum Likelihood (REML).

Figure 6: Estimated weighted effect (B^*) for the PAIN > NO PAIN meta-analysis on the left and corresponding between-study variability (τ^2) using the DerSimonian and Laird estimator on the right for voxels within the ROI. The MNI coordinates of the slices correspond to $x = 0$, $y = 0$ and $z = 0$. Larger estimates for B^* are associated with higher between-study heterogeneity.

3.2.3 Alternative methods

While transforming a test statistic to standardized effect sizes before aggregating is one way to avoid differences in the unit between studies, alternative methods are possible. These focus on combining test statistics as these have no unit either. For instance, Fisher’s combined probability method ([Fisher, 1925](#)) converts Z -values from independent tests (i.e. the different studies) to P -values and then combines these into a χ^2 test-statistic. Another approach is Stouffer’s method and can be used to average Z -statistics while also weight studies by their sample size ([Liptak, 1958](#)). Furthermore, it is also possible to avoid distributional assumptions when performing statistical inference using non-parametric methods such as permutation testing. An overview of these methods together with simulation results tailored to fMRI data are given in ([Maumet and Nichols, 2016](#)).

4 Conclusion

In this paper, we first demonstrate the main benefit of standardizing effect size estimates when reporting single fMRI studies. As standardized effect magnitudes have no unit, it becomes possible to compare the reported estimates over studies. In contrast, the BOLD percent signal change depends on the technique to calculate an average (i.e. baseline) signal which varies with software packages. We then discuss two main techniques to aggregate fMRI data over studies (i.e. a meta-analysis). The first one is the extension of the two-stage

GLM approach (typically used in the analysis of single fMRI studies) with a third level. Ideally, the data is modelled using a mixed-effects approach (incorporating both within- and between-study variability). One benefit here is the usage of the original data. No transformation of the original data is needed before the meta-analysis. However, the analyst needs to be careful for unit mismatches in the model between studies in the database. This might occur when pre-processing of the individual studies is done using different software packages. An alternative technique avoids this problem by first transforming the reported study-level effect sizes to standardized effect sizes. These are then modelled again using a mixed-effects approach, incorporating both within- and between-study variability. Studies with small sample sizes (i.e. higher within-study variability) are down-weighted when estimating the population effect. Finally, we show that considerable between-study heterogeneity (which is not explained by moderator variables) can be expected in whole-brain fMRI meta-analyses.

We hope to further encourage data sharing practices so that complete fMRI images can be aggregated using appropriate methods.

5 Appendix

5.1 Distribution of a standardized mean effect

In this section, we provide the expectation and variance of a standardized mean effect. The derivations are obtained using Hedges (1981) which provides the expectation and variance of a standardized mean difference between an experimental and control group. The following can be used for estimating a standardized mean effect in one sample.

First consider a response variable Y of study i with $i = 1, \dots, k$ and sample size n_i . We assume $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Furthermore, we are interested in estimating the mean effect $\delta_i = \frac{\mu_i}{\sigma_i}$ using $d_i = \bar{Y}_i/S_i$, where S_i is the square root of the unbiased sample variance of study i . Note the following property of the sample variance, S^2 :

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2, \quad (27)$$

where χ^2 is a Chi-squared random variable with $n-1$ degrees of freedom (df).

Before we derive the expectation and variance of d , consider the expression of a non-central t -distributed random variable:

$$T = \frac{\mathcal{Z} + \theta}{\sqrt{\frac{V}{\nu}}}, \quad (28)$$

where $\mathcal{Z} \sim \mathcal{N}(0, 1)$, V is a Chi-squared distributed random variable with $n-1$ degrees of freedom and θ is the non-centrality parameter.

Now we re-write the expression for d where we drop the subscript i for ease of notation.

$$\begin{aligned} d &= \frac{\bar{Y}}{S} = \frac{(\sigma/\sqrt{n})^{-1}\bar{Y}}{(\sigma/\sqrt{n})^{-1}S} \\ &= \frac{(\sigma/\sqrt{n})^{-1}\bar{Y}}{\frac{\sqrt{n-1}}{\sqrt{n-1}}(\sigma/\sqrt{n})^{-1}S} \\ &= \frac{\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} + \frac{\mu}{\sigma/\sqrt{n}}}{\sqrt{n}\sqrt{\frac{S^2(n-1)}{\sigma^2(n-1)}}}. \end{aligned}$$

Denote $\mathcal{Z}$ again as a standard normal distributed random variable. Furthermore using property (27), we write $V = \frac{(n-1)S^2}{\sigma^2}$ and $\nu = n-1$. Hence we get:

$$\begin{aligned} d &= \frac{\mathcal{Z} + \sqrt{n}\delta}{\sqrt{n}\sqrt{\frac{V}{\nu}}} \\ &= \frac{1}{\sqrt{n}} \left[\frac{\mathcal{Z} + \sqrt{n}\delta}{\sqrt{\frac{V}{\nu}}} \right]. \end{aligned} \tag{29}$$

Using equation (28), it follows that d equals $\frac{1}{\sqrt{n}}$ times a non-central t -distributed random variable with $n-1$ degrees of freedom and $\sqrt{n}\delta$ as the non-centrality parameter.

Using expression (29) and the first moment of the non-central t -distribution

(Johnson and Welch, 1940), we have for $\nu > 1$:

$$E(d) = \delta \times \left[\frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\frac{(n-1)}{2}} \Gamma\left(\frac{n-2}{2}\right)} \right]^{-1} \quad (30)$$

$$\approx \delta \times \left(1 - \frac{3}{4(n-1)-1} \right)^{-1}, \quad (31)$$

where Γ is the gamma function: $\Gamma(x) = (x-1)!$

Furthermore, we define:

$$J = \left(1 - \frac{3}{4(n-1)-1} \right), \quad (32)$$

which is an approximation for $\frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\frac{(n-1)}{2}} \Gamma\left(\frac{n-2}{2}\right)}$ in equation (30) (Hedges, 1981).

When $n \geq 10$, then $\frac{\delta}{J} - E(d) < 0.001$.

For $\nu > 2$, the second moment of the non-central t -distribution, gives:

$$\begin{aligned} \text{Var}(d) &= \frac{1}{n} \left[\frac{n^2 \delta^2 - n \delta^2 + n - 1}{n - 3} \right] - \frac{\delta^2}{h^2} \\ &= \frac{(N-1)(1 + N \delta^2)}{N(N-3)} - \frac{\delta^2}{h^2}, \end{aligned} \quad (33)$$

where

$$h = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\frac{(n-1)}{2}} \Gamma\left(\frac{n-2}{2}\right)} \quad (34)$$

Note that it is again possible to replace h with J . Finally we plug in the estimate for δ in both equations (31) and (33).

As is clear from expression (31), $d = \frac{\bar{Y}}{S}$ is a biased estimator for δ . For this reason, Hedges (1981) suggested to use:

$$g = \frac{\bar{Y}}{S} \times J \quad (35)$$

as an unbiased estimator. The variance then simply becomes:

$$\text{Var}(g) = \text{Var}(d) \times J^2. \quad (36)$$

An alternative is the exact expression:

$$g = \frac{\bar{Y}}{S} \times h \quad (37)$$

$$\text{Var}(g) = \text{Var}(d) \times h^2 \quad (38)$$

5.2 Empirical variance of estimator for g_e

Figure 7: Empirical variances of the estimator for g_e . Variance is calculated on the estimates of g_e obtained in the 1000 Monte-Carlo simulation runs. In panel **A**, we calculate the variance marginally over the number of subjects. In panel **B**, we calculate the variance marginally over the number of scans.

5.3 Unbiased standardized effect size estimator

Figure 8: Empirical assessment of bias when estimating Hedges' g_e and its variance. Results are obtained using 1000 Monte-Carlo simulations. Individual fMRI time series are generated with Gaussian noise and homogeneous within-subject variability. Group level model parameters are estimated using the OLS estimators. Both the average over all simulations and the expected values are plotted for different time series lengths and sample sizes.

5.4 Estimators for τ^2

To begin with, we have Y_i^* , the (standardized) effect size from study i with $i = 1, \dots, k$. Next, σ_i^{*2} is the corresponding within-study sampling variance and a weight (w_i^*) for each study is defined as $w_i^* = 1/\sigma_i^{*2}$. The weighted average, $\bar{Y}^*$, is defined as

$$\bar{Y}^* = \frac{\sum_{i=1}^k w_i^* Y_i^*}{\sum_{i=1}^k w_i^*}. \quad (39)$$

The first estimator for τ^2 is the DerSimonian and Laird estimator ([DerSimonian and Laird, 1986](#)):

$$\hat{\tau}_{DL}^2 = \frac{\sum_{i=1}^k (Y_i^* - \bar{Y}^*)^2 - (k-1)}{c} \quad (40)$$

with

$$c = \sum_{i=1}^k w_i^* - \frac{\sum_{i=1}^k w_i^{*2}}{\sum_{i=1}^k w_i^*}.$$

The second estimator for τ^2 is the Hedges estimator.

$$\hat{\tau}_{HE}^2 = \frac{\sum_{i=1}^k (Y_i^* - \bar{Y}^*)^2}{k-1} - \frac{1}{k} \sum_{i=1}^k \sigma_i^{*2}. \quad (41)$$

It has been shown that the Hedges estimator in equation (41) is unbiased, even when substituting unbiased estimates for the σ_i^{*2} values ([Viechtbauer, 2005](#)).

The final estimator is based on the Restricted Maximum Likelihood approach (REML). In the approach presented here, the weights w_i^* are first defined as $w_i^* = 1/\sigma_i^{*2}$, but after estimating τ_{REML}^2 updated to $w_i^* = 1/(\sigma_i^{*2} +$

τ_{REML}^2). The estimator is given by

$$\hat{\tau}_{REML}^2 = \frac{\sum_{i=1}^k w_i^2 \left[(Y_i^* - \hat{\delta}_{ML}^*) - \sigma_i^{*2} \right]}{\sum_{i=1}^k w_i^{*2}} + \frac{1}{\sum_{i=1}^k w_i^*}, \quad (42)$$

where δ_{ML}^* is the maximum likelihood estimator equal to expression (39). The

estimator for τ_{REML}^2 is obtained by iterating between $\hat{\delta}_{ML}^*$ and

$\hat{\tau}_{REML}^2$ until convergence.

References

Acar, F., R. Seurinck, S. B. Eickhoff, and B. Moerkerke

2018. Assessing robustness against potential publication bias in activation likelihood estimation (ale) meta-analyses for fmri. *PLOS ONE*, 13(11).

Albajes-Eizaguirre, A. and J. Radua

2018. What do results from coordinate-based meta-analyses tell us? *NeuroImage*, 176(April):550–553.

Bohning, D.

2002. Some general points in estimating heterogeneity variance with the DerSimonian-Laird estimator. *Biostatistics*, 3(4):445–457.

Borenstein, M., L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein

2009. *Introduction to Meta Analysis*, 1 edition. Wiley-Blackwell.

Bowman, N. A.

2012. Effect Sizes and Statistical Methods for Meta-Analysis in Higher Education. *Research in Higher Education*, 53(3):375–382.

Button, K. S., J. P. a. Ioannidis, C. Mokrysz, B. a. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò

2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews. Neuroscience*, 14(5):365–76.

Bzdok, D. and B. T. Yeo

2017. Inference in the age of big data: Future perspectives on neuroscience. *NeuroImage*, 155(August 2016):549–564.
- Carp, J.
2012. The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, 63(1):289–300.
- Chen, G., P. A. Taylor, and R. W. Cox
2017. Is the statistic value all we should care about in neuroimaging? *NeuroImage*, 147(October 2016):952–959.
- Cohen, J.
1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.
- Costafreda, S. G.
2009. Pooling FMRI data: meta-analysis, mega-analysis and multi-center studies. *Frontiers in neuroinformatics*, 3(September):33.
- Cox, R. W.
1996. Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and biomedical research, an international journal*, 29 3:162–73.
- Cummings, P.
2011. Arguments for and against standardized mean differences (effect sizes). *Archives of Pediatrics and Adolescent Medicine*, 165(7):592–596.

DerSimonian, R. and N. Laird

1986. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–88.

Durnez, J., J. Degryse, B. Moerkerke, R. Seurinck, V. Sochat, R. Poldrack, and T. Nichols

2016. Power and sample size calculations for fMRI studies based on the prevalence of active peaks. *bioRxiv*, P. 049429.

Eickhoff, S. B., D. Bzdok, A. R. Laird, F. Kurth, and P. T. Fox

2012. Activation likelihood estimation meta-analysis revisited. *NeuroImage*, 59(3):2349–61.

Eickhoff, S. B., A. R. Laird, C. Grefkes, L. E. Wang, K. Zilles, and P. T. Fox

2009. Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. *Human brain mapping*, 30(9):2907–26.

Fisher, R.

1925. *Statistical methods for research workers*. Edinburgh Oliver & Boyd.

Friston, K.

2012. Ten ironic rules for non-statistical reviewers. *NeuroImage*, 61(4):1300–1310.

Friston, K. J., A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and

R. S. J. Frackowiak

1995. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4):189–210.

Friston, K. J., P. Jezzard, and R. Turner

1994. Analysis of functional MRI time-series. *Human Brain Mapping*, 1(2):153–171.

Friston, K. J., K. E. Stephan, T. E. Lund, a. Morcom, and S. Kiebel

2005. Mixed-effects and fMRI studies. *NeuroImage*, 24(1):244–52.

Glass, G. V.

1976. Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5(10):3–8.

Gorgolewski, K. J., T. Auer, V. D. Calhoun, R. C. Craddock, S. Das, E. P.

Duff, G. Flandin, S. S. Ghosh, T. Glatard, Y. O. Halchenko, D. A. Handwerker, M. Hanke, D. Keator, X. Li, Z. Michael, C. Maumet, B. N. Nichols, T. E. Nichols, J. Pellman, J.-B. Poline, A. Rokem, G. Schaefer, V. Sochat, W. Triplett, J. A. Turner, G. Varoquaux, and R. A. Poldrack

2016. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, 3:160044.

Hedges, L. V.

1981. Distribution Theory for Glass’s Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics*, 6(2):107–128.

Hedges, L. V.

1984. Advances in statistical methods for meta-analysis. *New Directions for Program Evaluation*, 1984(24):25–42.

Henson, R. and K. Friston

2007. Convolution Models for FMRI. In *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, F. K, A. J, K. S, N. T, and W. Penny, eds., Pp. 193–210. Academic Press.

Higgins, J. P. T. and S. G. Thompson

2002. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, 21(11):1539–58.

Holmes, A. P. and K. J. Friston

1998. Generalisability, Random Effects and Population Inference. *NeuroImage*, 7:S754.

Hunter, J. E. and F. L. Schmidt

2000. Fixed Effects vs. Random Effects Meta-Analysis Models: Implications for Cumulative Research Knowledge. *International Journal of Selection and Assessment*, 8(4):275–292.

Ioannidis, J. P. A.

2005. Why most published research findings are false. *PLoS Medicine*, 2(8):0696–0701.

Jackson, D., J. Bowden, and R. Baker

2010. How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? *Journal of Statistical Planning and Inference*, 140(4):961–970.
- Johnson, N. L. and B. L. Welch
1940. Applications of the Non-Central t-Distribution. *Biometrika*, 31(3/4):362–389.
- Kirk, R.
1996. Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56:746–759.
- Lindquist, M. a.
2008. The Statistical Analysis of fMRI Data. *Statistical Science*, 23(4):439–464.
- Liptak, T.
1958. On the combination of independent tests. *Magyar Tudományos Akademia Matematikai Kutató Intézetének Közleményei*.
- Maumet, C., T. Auer, A. Bowring, G. Chen, S. Das, G. Flandin, S. Ghosh, T. Glatard, K. J. Gorgolewski, K. G. Helmer, M. Jenkinson, D. B. Keator, B. N. Nichols, J. B. Poline, R. Reynolds, V. Sochat, J. Turner, and T. E. Nichols
2016. Sharing brain mapping statistical results with the neuroimaging data model. *Scientific Data*, 3:1–15.

Maumet, C. and T. E. Nichols

2015. Do the units matter? validity of intensity based meta-analysis in the presence of unit mismatches. 21st Annual Meeting of the Organization for Human Brain Mapping (OHBM), Honolulu, United States.

Maumet, C. and T. E. Nichols

2016. Minimal data needed for valid & accurate image-based fmri meta-analysis. *bioRxiv*.

Müller, V. I., E. C. Cieslik, A. R. Laird, P. T. Fox, J. Radua, D. Mataix-Cols, C. R. Tench, T. Yarkoni, T. E. Nichols, P. E. Turkeltaub, T. D. Wager, and S. B. Eickhoff

2017. Ten simple rules for neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews*.

Mumford, J. A. and T. Nichols

2006. Modeling and inference of multisubject fMRI data. *IEEE Engineering in Medicine and Biology Magazine*, 25(2):42–51.

Mumford, J. A. and T. Nichols

2009. Simple group fMRI modeling and inference. *NeuroImage*, 47(4):1469–1475.

Nichols, T. and S. Hayasaka

2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, 12(5):419–446.

Nichols, T. E.

2012. Multiple testing corrections, nonparametric methods, and random field theory. *NeuroImage*, 62(2):811–5.

Patil, P., R. D. Peng, and J. T. Leek

2016. A statistical definition for reproducibility and replicability. *bioRxiv*, P. 066803.

Poldrack, R. A., C. I. Baker, J. Durnez, K. J. Gorgolewski, P. M. Matthews,

M. R. Munafò, T. E. Nichols, J.-B. Poline, E. Vul, and T. Yarkoni

2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2):115–126.

Poldrack, R. A. and J. B. Poline

2015. The publication and reproducibility challenges of shared data. *Trends in Cognitive Sciences*, 19(2):59–61.

Poline, J.-B., J. L. Breeze, S. Ghosh, K. Gorgolewski, Y. O. Halchenko,

M. Hanke, C. Haselgrove, K. G. Helmer, D. B. Keator, D. S. Marcus, R. A.

Poldrack, Y. Schwartz, J. Ashburner, and D. N. Kennedy

2012. Data sharing in neuroimaging research. *Frontiers in Neuroinformatics*, 6(April):9.

Poline, J. B. and M. Brett

2012. The general linear model and fMRI: Does love last forever? *NeuroImage*, 62(2):871–880.

R Core Team

2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radua, J., D. Mataix-Cols, M. L. Phillips, W. El-Hage, D. M. Kronhaus, N. Cardoner, and S. Surguladze

2012. A new meta-analytic method for neuroimaging studies that combines reported peak coordinates and statistical parametric maps. *European psychiatry : the journal of the Association of European Psychiatrists*, 27(8):605–11.

Raudenbush, S. W.

2009. Analyzing effect sizes: Random-effects models. In *The handbook of research synthesis and meta-analysis*, H. Cooper, L. Hedges, and J. Valentine, eds., Pp. 295–315. New York, NY: Russell Sage Foundation.

Reddan, M. C., M. A. Lindquist, T. D. Wager, G. RJ, B. KS, L. MA, C. LJ, and H. T

2017. Effect Size Estimation in Neuroimaging. *JAMA Psychiatry*, 4(5):863.

Rosenthal, R.

1979. The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638–641.

Salimi-Khorshidi, G., S. M. Smith, J. R. Keltner, T. D. Wager, and T. E.

- Nichols
2009. Meta-analysis of neuroimaging data: A comparison of image-based and coordinate-based pooling of studies. *NeuroImage*, 45(3):810–823.
- Smith, S. M., M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. J. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney, R. K. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. De Stefano, J. M. Brady, and P. M. Matthews
2004. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23 Suppl 1:S208–19.
- Sudlow, C., J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins
2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3):1–10.
- Tench, C. R., R. Tanasescu, C. S. Constantinescu, D. P. Auer, and W. J. Cottam
2017. Coordinate based random effect size meta-analysis of neuroimaging studies. *NeuroImage*, 153(December 2016):293–306.
- Turkeltaub, P. E., G. F. Eden, K. M. Jones, and T. A. Zeffiro
2002. Meta-Analysis of the Functional Neuroanatomy of Single-Word Reading: Method and Validation. *NeuroImage*, 16(3):765–780.

Turkeltaub, P. E., S. B. Eickhoff, A. R. Laird, M. Fox, M. Wiener, and P. Fox
2012. Minimizing within-experiment and within-group effects in Activation Likelihood Estimation meta-analyses. *Human brain mapping*, 33(1):1–13.

Turner, B. O., E. J. Paul, M. B. Miller, and A. K. Barbey
2018. Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, 1(1):62.

Van Essen, D. C., S. M. Smith, D. M. Barch, T. E. J. Behrens, E. Yacoub, and K. Ugurbil
2013. The WU-Minn Human Connectome Project: an overview. *NeuroImage*, 80:62–79.

Veroniki, A. A., D. Jackson, W. Viechtbauer, R. Bender, J. Bowden, G. Knapp, O. Kuss, J. P. Higgins, D. Langan, and G. Salanti
2016. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, 7(1):55–79.

Viechtbauer, W.
2005. Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. *Journal of Educational and Behavioral Statistics*, 30(3):261–293.

Viechtbauer, W.
2010. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48.

Viechtbauer, W., J. A. López-López, J. Sánchez-Meca, and F. Marín-Martínez
2015. A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods*, 20(3):360–374.

Wager, T. D., J. Jonides, and S. Reading
2004. Neuroimaging studies of shifting attention: A meta-analysis. *NeuroImage*, 22(4):1679–1693.

Wager, T. D., M. Lindquist, and L. Kaplan
2007. Meta-analysis of functional neuroimaging data: current and future directions. *Social cognitive and affective neuroscience*, 2(2):150–8.

Wager, T. D., M. a. Lindquist, T. E. Nichols, H. Kober, and J. X. Van Snellenberg
2009. Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *NeuroImage*, 45(1 Suppl):S210–21.

Welvaert, M., Y. Rosseel, J. Durnez, B. Moerkerke, and G. Verdoolaege
2011. neuRosim: An R Package for Generating fMRI Data. *Journal of Statistical Software*, 44(10).

Woolrich, M. W., T. E. Behrens, C. F. Beckmann, M. Jenkinson, and S. M. Smith
2004. Multilevel linear modelling for fmri group analysis using bayesian inference. *NeuroImage*, 21(4):1732 – 1747.

Worsley, K. J., C. Liao, J. Aston, V. Petre, G. Duncan, F. Morales, and

A. Evans

2002. A general statistical analysis for fmri data. *Neuroimage*, 15(1):1–15.

Yarkoni, T., R. A. Poldrack, T. E. Nichols, D. C. Van Essen, and T. D. Wager

2011. Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665–670.

Zakzanis, K. K.

2001. Statistics to tell the truth, the whole truth, and nothing but the truth. Formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers. *Archives of Clinical Neuropsychology*, 16(7):653–667.