Highlights

Multiple testing correction over contrasts for brain imaging

Bianca A. V. Alberton, Thomas E. Nichols, Humberto R. Gamba, Anderson M. Winkler

- A correction for multiple testing across contrasts is necessary
- Permutation was the only method that performed well in all simulation scenarios
- ANOVA followed by pairwise (*post hoc*) comparisons does not control the error rate
- Some well-known methods are conservative when a subset of the contrasts is tested

Multiple testing correction over contrasts for brain imaging

Bianca A. V. Alberton[a,*], Thomas E. Nichols[b], Humberto R. Gamba[a], Anderson M. Winkler[a,c]

[a] *Graduate Program in Electrical and Computer Engineering, Universidade Tecnológica Federal do Paraná, Curitiba,* PR, *Brazil.*
[b] *Li-Ka Shing Big Data Institute, University of Oxford,* UK
[c] *Section on Development and Affective Neuroscience (*SDAN*), National Institute of Mental Health (*NIMH*), National Institutes of Health (*NIH*), Bethesda,* MD, USA

Abstract

The multiple testing problem arises not only when there are many voxels or vertices in an image representation of the brain, but also when multiple contrasts of parameter estimates (that is, hypotheses) are tested in the same general linear model. Here we argue that a correction for this multiplicity must be performed to avoid excess of false positives. Various methods have been proposed in the literature, but few have been applied in brain imaging. Here we discuss and compare different methods to make such correction in different scenarios, showing that one classical and well known method is invalid, and argue that permutation is the best option to perform such correction due to its exactness and flexibility to handle a variety of common imaging situations.

Keywords: multiple comparisons, multiple testing, brain imaging, permutation tests, contrast correction.

1. Introduction

A well known problem in brain imaging is the multiplicity of tests, which arises given the fact that a statistical test is performed in each voxel or vertex of an image representation of the brain. However, an equally common

*Corresponding author.
Email addresses: biancaalberton@alunos.utfpr.edu.br (Bianca A. V. Alberton), thomas.nichols@bdi.ox.ac.uk (Thomas E. Nichols), humberto@utfpr.edu.br (Humberto R. Gamba), anderson.winkler@nih.gov (Anderson M. Winkler)

situation in which such multiplicity occurs is when multiple contrasts of parameter estimates of the same general linear model (GLM) are considered, or even many multiple different such models. In effect, data acquisition is one of the most expensive and laborious stages of an experiment, such that often the same data are reused and reanalysed in different ways to consider different models and hypotheses. If left uncontrolled, such multiplicity can lead to an undesirably high number false positives.

For example, in the Human Connectome Project (HCP), the task fMRI N–back working memory experiment is conducted using different classes of stimuli (faces, places, tools and body parts) (Barch et al., 2013). Analyses involving these classes can be corrected using Bonferroni, but results would be overly conservative due to dependence among the tests. Comparisons among the classes using methods such as analysis of variance (ANOVA) followed by pairwise comparisons do not control the error rate, as we demonstrate later in this paper.

Even though a number of methods have been proposed for correction for similar problems in non-imaging fields (Hochberg and Tamhane, 1987; Hsu, 1996), most of these have seen little use in the brain imaging. In this short technical note, we assert that such correction is necessary, discuss and compare a few existing methods through which it can be implemented, and provide an approach based on permutation tests that provides an exact (as opposed to conservative) control over the error rate, even when the multiple hypotheses being tested are not independent.

2. Theory

2.1. Notation and general aspects

Consider the general linear model (GLM) expressed by:

$$\mathbf{y} = \mathbf{M}\boldsymbol{\psi} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{y}$ is $N \times 1$ vector containing the image data for N subjects at a given voxel (or vertex), $\mathbf{M}$ is the $N \times R$ design matrix with the modeled R explanatory variables, $\boldsymbol{\psi}$ is the $R \times 1$ vector with the (to be estimated) parameters that scale the variables in $\mathbf{M}$ so as to explain the variability observed in $\mathbf{y}$, and $\boldsymbol{\epsilon}$ is the $N \times 1$ vector of random errors. The coefficients can be estimated via ordinary least squares as $\hat{\boldsymbol{\psi}} = \mathbf{M}^{+}\mathbf{y}$, where the symbol $^{+}$ represents the Moore–Penrose pseudo-inverse. The interest is to test K, $K \geqslant 1$, null hypotheses, each of them represented as $\mathcal{H}_k^0 : \mathbf{C}_k'\boldsymbol{\psi} = 0$,

$k = \{1, \ldots, K\}$, where $\mathbf{C}_k$ is a $R \times S_k$ matrix that defines the contrasts of parameter estimates. When $\text{rank}\,(\mathbf{C}_k) = 1$, a suitable test statistic can be calculated as:

$$t = \frac{\mathbf{C}_k' \hat{\boldsymbol{\psi}} (\mathbf{C}_k' (\mathbf{M}'\mathbf{M})^{-1} \mathbf{C}_k)^{-\frac{1}{2}}}{\sqrt{\hat{\boldsymbol{\epsilon}}' \hat{\boldsymbol{\epsilon}} / \nu}} \tag{2}$$

where $\nu = N - \text{rank}(\mathbf{M})$ are the degrees of freedom. The test is said to be *significant* at the level α, $0 \leqslant \alpha \leqslant 1$, if the probability of observing a random variable T larger or equal than t is smaller or equal than α, that is, $\mathrm{P}(T \geqslant t) \leqslant \alpha$. This is the p-value, and can be computed based on distributional assumptions on t, or through a resampling procedure, such as a permutation test (Pesarin and Salmaso, 2010).

2.2. Permutation inference

Permutation is a non-parametric method in which the distribution of t is ascertained by explicitly calculating all (or a large number of) the possible values that it could assume should the null hypothesis be true. These values are obtained by randomly rearranging the data as if the null hypothesis were indeed true (Nichols and Holmes, 2002; Winkler et al., 2014). For a number J of permutations of the data, each one associated with a test statistic t_j, $j = \{1, 2, \ldots, J\}$, the p-value can be calculated as the number of occurrences of a random (after permutation) t_j that is larger or equal to the observed, original statistic t (obtained without any permutation) divided by the number of permutations performed. In other words, t_j takes the role of the random variable T described above. Note that t itself (computed from the model without any permutation) should also be counted among the set of J statistics computed after permutation, such that the smallest possible p-value in a permutation test cannot be smaller than $1/J$, and therefore cannot be zero (Phipson and Smyth, 2010).

2.3. Multiple testing

As more tests are conducted, the more likely it will be that any will be declared significant even if no actual effect exists. The problems associated with the multiplicity of tests across points (e.g., voxels or vertices) in an image representation of the brain are well known, and various strategies have been devised over the years (for reviews, see Nichols and Hayasaka, 2003; Farcomeni, 2008; Nichols, 2012). In general, approaches target the control of one of two different error quantities: the *family-wise error rate*

(FWER), which is defined as the chance of any false positive across all tests, and the *false discovery rate* (FDR), which is the expected proportion of false positives across all tests in which an effect has been found significant.

When only one test is considered (that is, in the absence of multiple testing), all that is needed for a decision on rejecting or not the null hypothesis are the p-value p and a (usually pre-defined) test level α. When more than one test is performed, either the test level can be *corrected* (α_{cor}), in which it is changed so as to accommodate the multiplicity of tests (the p-values remain unchanged), or the p-values can be *adjusted* (p_{adj}), in which these are changed instead (the test level remains then unchanged). In either case, the modifications are such that the FWER or the FDR is controlled at the level α (though for FDR, test levels are often denoted q).

Multiple image points, such as voxels or vertices, are not the only way in which multiple testing can occur in brain imaging; various other sources of multiplicity that are not simply multiple spatial tests are possible; some of us have previously used the term *multiple testing problem type* II (MTP-II), so as to distinguish them from the usual multiplicity due to the many measurements taken across space (Winkler et al., 2016b). The multiplicity of contrasts of model parameter estimates belongs to this class. As with correction over voxels or vertices, which is performed across all image points of interest (MTP-I, e.g., the whole brain, or within a region of interest), the correction in the MTP-II case can consider only the hypotheses of interest. In other words, it is not always the case that all and every possible contrast of parameter estimates is of interest; for example, when testing G groups we may only be interested in each group individually (G tests), or the interest may lie on differences among specific groups (and not necessarily on all the $G(G-1)$ possible tests).

Methods to account for the multiplicity of contrasts include the same that can be used over any set of p-values, such as Bonferroni or FDR, or can be specific to the context of the general linear model. Below we briefly summarise some of these methods. Although here the focus is on FWER, the conceptualization of the correction across contrasts, as well as the application of permutation tests, remains similar for FDR.

Dunn–Šidák. The probability that two independent events can occur simultaneously is given by the product of their probabilities. Thus, when multiple independent statistical tests are conducted, the corrected test level can be computed as $\alpha_{\text{cor}} = 1-(1-\alpha)^{1/\tau}$, where τ is the number of tests (when the

correction is applied to multiple contrasts, and using our notation, $\tau = K$). This idea was considered by Tippett (1931), and later proved in a related context by Dunn (1958) and Šidák (1967). The Dunn–Šidák (DS) equation is often described as an adjustment of one or more p-values instead of correcting α, that is, $p_{\text{adj}} = 1 - (1 - p_{\text{unc}})^{\tau}$, where p_{unc} is the original (uncorrected) p-value obtained in a given test. In this case, the FWER is controlled for the adjusted p-values without the need to modify the test level α. The equality holds only if the tests are independent; the adjustment is conservative otherwise.

Bonferroni. Bonferroni (1936) suggested an approximation for the corrected test level as $\alpha_{\text{cor}} = \frac{\alpha}{\tau}$. The simplicity and intuitive appeal of this approximation were certainly determinants to its enormous popularity across all scientific domains over many decades. However, even for independent tests, this approximation is slightly conservative when compared to Dunn–Šidák (which is exact). If the tests are not independent, the correction becomes yet more conservative. As Bonferroni is an approximation to DS, and DS is known to be the more powerful of the two, only the latter is analyzed in this study.

Fisher's least significant difference. This is a two-step procedure suggested by Fisher (1935) as a way to identify which tests were responsible for driving the overall (*omnibus*) result of an ANOVA (it does not appear that Fisher was concerned with multiple testing when proposing this test). In this method, an *omnibus* F-test is performed to detect if there are *any* group differences; if and only if the F-test is significant at α, a second stage is performed, in which follow up (*post hoc*) t-tests between each pair of groups are evaluated at the same level α. However, it has been demonstrated (Hayter, 1986) that the maximum probability of finding at least one incorrect result can greatly exceed the test level, growing rapidly as the number of groups exceed 3, and is therefore not recommended (Hsu, 1996). The test remains valid for up to and including 3 groups, though.

Tukey. For one-way ANOVA layouts, Tukey (1953) proposed that p-values for the comparison between each pair of group means could be computed with reference to the studentized range distribution (SRD), that is, $p_{\text{adj}} = Q_{\text{cdf}}\left(t\sqrt{2}; G, \nu\right)$, where Q_{cdf} represents the cumulative distribution function (cdf) from the SRD, ν is the number of degrees of freedom, and G is the number of groups. Although this method has been applied mainly for ANOVA

designs, it could also be considered for scenarios in which multiple comparisons are performed among regression coefficients for continuous variables (e.g., comparing different continuous signal against a single reference regressor). This procedure assumes that all possible pairwise comparisons could be of interest, where the greatest difference between means is the most likely to be rejected. When the groups are unbalanced, the procedure is known to be conservative, and the degree of conservativeness varies with the number of groups and the severity of the unbalance (Hochberg and Tamhane, 1987). This method, and variants of it, have received a number of different names in different settings, including *T-Procedure*, *wholly significant difference* (WSD), *honestly significant difference* (HSD), and, when applied to unbalanced models, *Tukey–Kramer test*. For simplicity, here we call it simply "Tukey" for both balanced and unbalanced experiments.

Fisher–Hayter. Hayter (1986) proposed that the p-value could be adjusted by changing the second step of Fisher's LSD to accommodate the SRD as $p_{\text{adj}} = Q_{\text{cdf}}\left(t\sqrt{2}; G-1, \nu\right)$. In comparison with Tukey's method, this procedure has greater power due to the loosening of G to $G-1$, and still maintains control of the FWER due to the F-test applied in the first step of the procedure. Without this first step, the FWER would be slightly greater than the defined α level. Being a modification of Fisher's LSD, this method is sometimes named Fisher–Hayter (FH) procedure (Keppel and Wickens, 2004).

Permutations tests. Permutation, as described above, can also be used to correct for multiple testing. This is done by computing the adjusted p-value using as reference the distribution of the *maximum* statistic across the set of tests that are being corrected, that is, $p_{\text{adj}} = \frac{1}{J}\sum_j^J I(t_j^{\max} \geqslant t)$, where J is the total number of permutations, $I(\cdot)$ is the indicator function and $t_j^{\max}$ is the maximum value of t_j across the contrasts (Westfall and Young, 1993). The reason why this works is that the test whose statistic is maximum across all tests being considered is also the one that has the maximum statistic for any subset of these tests that includes it, such that this approach constitutes a *closed testing procedure* (Marcus et al., 1976), which controls the FWER.

3. Evaluation Methods

3.1. Synthetic data

To investigate the error rates and power of each method, we considered 5000 realisations of an one-way ANOVA design, with simulated data repre-

senting 2000 voxels (these could also be construed as vertices, or any other type of imaging element) following a normal distribution with zero mean and unit variance. We considered 8 different scenarios by manipulating 3 simulation parameters: presence or absence of signal, balanced or unbalanced designs, and the correction over all possible pairwise group comparisons or only the largest subset of linearly independent contrasts.

With respect to signal, when it was added, it was to all voxels of subjects in groups 1 and 2, and with size defined as $\mathbf{M}\boldsymbol{\psi}$ with $\boldsymbol{\psi} = [\psi_1\ \psi_2\ 0\ \ldots\ 0]'$, where $\psi_1 = s_t$, $\psi_2 = -s_t$, and $s_t = t^{-1}_{\text{cdf}}(1-\alpha;\nu)\left(\mathbf{C}'\left(\mathbf{M}'\mathbf{M}\right)^{+}\mathbf{C}\right)^{1/2}/2$. $\mathbf{C} = [1\ -1\ \ 0\ \ldots\ 0]'$ is the contrast vector and $\alpha = 0.05$ is the test level. By adding a positive signal to group 1 and a negative signal to group 2, with magnitudes determined by s_t, an approximate power of 50% can be expected for the contrast comparing these two groups before any adjustment has been made for the multiplicity of tests (in this case, multiplicity of contrasts). Also, although stronger effects are expected to be detected in comparisons between groups 1 and 2, smaller effects can also be detected when groups 1 or 2 are compared with any other group. Thus, power calculations below consider effects in any comparison involving groups 1 or 2.

The simulations used sample sizes of 50, 100 or 150 subjects. For the balanced scenarios, subjects were divided in 5, 10 or 15 groups respectively, always with 10 subjects per group. For the unbalanced scenarios, subjects were likewise divided into 5, 10 or 15 groups respectively, however with each group consisting of a random number of participants under the constraint that each group had at least 4 subjects.

The comparisons considered (1) all possible pairwise group differences and (2) a subset of linearly independent comparisons. In the latter case, we tested the hypotheses that group 3 was greater than every one of the other groups. The reason to investigate a correction over a subset of independent contrasts is because all possible pairwise group differences imply dependencies among them (for example, for three groups, the difference between the second and third is fully determined once differences between first and second and first and third are known). Such dependencies do not exist for a subset of contrasts that are independent from each other. For the analyses using this subset, power can be estimated as the proportion of detected true effects in contrasts where group 3 is greater than groups 1 and 2, and to estimate the FWER as the proportion of false detected effects in contrasts where group 3 is greater than groups 4 to G.

All methods were evaluated using custom code written in GNU Octave

(Eaton et al., 2014). Tukey and Fisher–Hayter correction, however, used the functions "ptukey" and "qtukey" from the R statistical software (R Core Team, 2018), whereas permutation used PALM – Permutation Analysis of Linear Models (Winkler et al., 2014) with 5000 permutations.

3.2. Real data

We used data from the Healthy Brain Network (HBN) (Alexander et al., 2017) made publicly available by the Child Mind Institute (CMI, New York, NY, USA) to evaluate how correction over the number of contrasts could affect a realistic data analysis. The data collection, as well their distribution in anonymized format, was approved by their institutional ethics review board. Structural, T_1-weighted magnetic resonance images were processed using FreeSurfer (Dale et al., 1999; Fischl et al., 1999). Quality control was performed based on visual inspection of the reconstructed pial and inflated surfaces, with particular emphasis on the subjects with extreme values for the metrics produced by the tool MRIQC (Esteban et al., 2017), and most extreme Euler numbers for either of the hemispheres (Rosen et al., 2018); the threshold for the Euler number was -80, and subjects in which either hemisphere had values more negative than this threshold were excluded. This allowed selection of 278 subjects that successfully finished the FreeSurfer processing, that passed quality control, and further, that had complete data for the variables that we selected to be used in the statistical analysis (described below).

As dependent variables we investigated the volume of the subcortical structures automatically segmented by FreeSurfer: thalamus, caudate, putamen, pallidum, hippocampus, amygdala, nucleus accumbens, and ventral diencephalon (VDC, a group of structures whose precise limits are not typically discernible with T_1-weighted scans, and that includes hypothalamus, mammillary bodies, lateral and medial geniculate nuclei, subthalamic nuclei, substantia nigra and nucleus ruber) (Fischl et al., 2002, 2004). We also investigated associations with the estimated total intracranial volume (eTIV) (Buckner et al., 2004).

As independent variables, we chose two disparate measures for investigation, one related to social factors (Barratt simplified measure of social status; BSMSS) (Barratt, 2006) and another related to physiological factors (extracellular water, ECW). Two analyses of covariance (ANCOVA) designs were considered, one to investigate BSMSS and another for ECW. The subjects were divided in 5 groups using the 20th, 40th, 60th and 80th percentiles

$$
\begin{array}{llcl}
 & & & \begin{array}{ccccccc} \text{G}_1 & \text{G}_2 & \text{G}_3 & \text{G}_4 & \text{G}_5 & \text{age} & \text{sex} \end{array} \\
\mathbf{C}_1 & (\text{G}_1 > \text{G}_2) & = & [\ +1 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0\ ]' \\
\mathbf{C}_2 & (\text{G}_1 > \text{G}_3) & = & [\ +1 \quad 0 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0\ ]' \\
\mathbf{C}_3 & (\text{G}_1 > \text{G}_4) & = & [\ +1 \quad 0 \quad 0 \quad -1 \quad 0 \quad 0 \quad 0\ ]' \\
\mathbf{C}_4 & (\text{G}_1 > \text{G}_5) & = & [\ +1 \quad 0 \quad 0 \quad 0 \quad -1 \quad 0 \quad 0\ ]' \\
\mathbf{C}_5 & (\text{G}_2 > \text{G}_1) & = & [\ -1 \quad +1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0\ ]' \\
\mathbf{C}_6 & (\text{G}_2 > \text{G}_3) & = & [\ 0 \quad +1 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0\ ]' \\
\mathbf{C}_7 & (\text{G}_2 > \text{G}_4) & = & [\ 0 \quad +1 \quad 0 \quad -1 \quad 0 \quad 0 \quad 0\ ]' \\
\mathbf{C}_8 & (\text{G}_2 > \text{G}_5) & = & [\ 0 \quad +1 \quad 0 \quad 0 \quad -1 \quad 0 \quad 0\ ]' \\
\mathbf{C}_9 & (\text{G}_3 > \text{G}_1) & = & [\ -1 \quad 0 \quad +1 \quad 0 \quad 0 \quad 0 \quad 0\ ]' \\
\mathbf{C}_{10} & (\text{G}_3 > \text{G}_2) & = & [\ 0 \quad -1 \quad +1 \quad 0 \quad 0 \quad 0 \quad 0\ ]' \\
\mathbf{C}_{11} & (\text{G}_3 > \text{G}_4) & = & [\ 0 \quad 0 \quad +1 \quad -1 \quad 0 \quad 0 \quad 0\ ]' \\
\mathbf{C}_{12} & (\text{G}_3 > \text{G}_5) & = & [\ 0 \quad 0 \quad +1 \quad 0 \quad -1 \quad 0 \quad 0\ ]' \\
\mathbf{C}_{13} & (\text{G}_4 > \text{G}_1) & = & [\ -1 \quad 0 \quad 0 \quad +1 \quad 0 \quad 0 \quad 0\ ]' \\
\mathbf{C}_{14} & (\text{G}_4 > \text{G}_2) & = & [\ 0 \quad -1 \quad 0 \quad +1 \quad 0 \quad 0 \quad 0\ ]' \\
\mathbf{C}_{15} & (\text{G}_4 > \text{G}_3) & = & [\ 0 \quad 0 \quad -1 \quad +1 \quad 0 \quad 0 \quad 0\ ]' \\
\mathbf{C}_{16} & (\text{G}_4 > \text{G}_5) & = & [\ 0 \quad 0 \quad 0 \quad +1 \quad -1 \quad 0 \quad 0\ ]' \\
\mathbf{C}_{17} & (\text{G}_5 > \text{G}_1) & = & [\ -1 \quad 0 \quad 0 \quad 0 \quad +1 \quad 0 \quad 0\ ]' \\
\mathbf{C}_{18} & (\text{G}_5 > \text{G}_2) & = & [\ 0 \quad -1 \quad 0 \quad 0 \quad +1 \quad 0 \quad 0\ ]' \\
\mathbf{C}_{19} & (\text{G}_5 > \text{G}_3) & = & [\ 0 \quad 0 \quad -1 \quad 0 \quad +1 \quad 0 \quad 0\ ]' \\
\mathbf{C}_{20} & (\text{G}_5 > \text{G}_4) & = & [\ 0 \quad 0 \quad 0 \quad -1 \quad +1 \quad 0 \quad 0\ ]'
\end{array}
$$

Figure 1: List of contrasts testing all possible pairwise comparisons among the 5 groups ($\text{G}_1, \ldots, \text{G}_5$) as encoded by the design matrix (not shown). The last two regression coefficients modelled age and sex, in an ANCOVA design.

of each of these two variables. By dividing the continuous variable into discrete units, we can more easily consider an AN(C)OVA scenario for which some of the correction methods were originally devised (even though no such limitation exists for Dunn–Šidák or permutation tests), and further, we can accommodate the possibility of certain non-linear effects. It should be emphasised, however, that this division is completely arbitrary, and is done here solely for convenience. Age (5 – 15 years, mean = 9.85, standard deviation = 2.65) and sex (180 males and 98 females) were included as nuisance variables. All pairwise differences were tested. Contrasts 1 through 4 tested whether group 1 would be larger than the others, contrasts 5 through 8 tested whether group 2 would be larger than the others, and so forth, for a total of 20 contrasts covering all possible pairwise group comparisons in both directions (Figure 1).

Statistical analysis used PALM with 10000 permutations and tail approx-

imation (Winkler et al., 2016a) to calculate the uncorrected (for contrasts) and corrected p-values. Dunn–Šidák, Fisher's LSD, Tukey, and Fisher–Hayter were applied to the data using custom code written on GNU Octave. As with the synthetic data, for Tukey and Fisher–Hayter we invoked the functions "ptukey" and "qtukey" from R.

4. Results

4.1. Synthetic data

Figure 2 shows the error rates and power of each simulation testing all pairwise comparisons. All evaluated procedures controlled the FWER in the absence of signal, that is, when the null hypothesis was true for all $K = G(G-1)$ contrasts, for both balanced and unbalanced models.

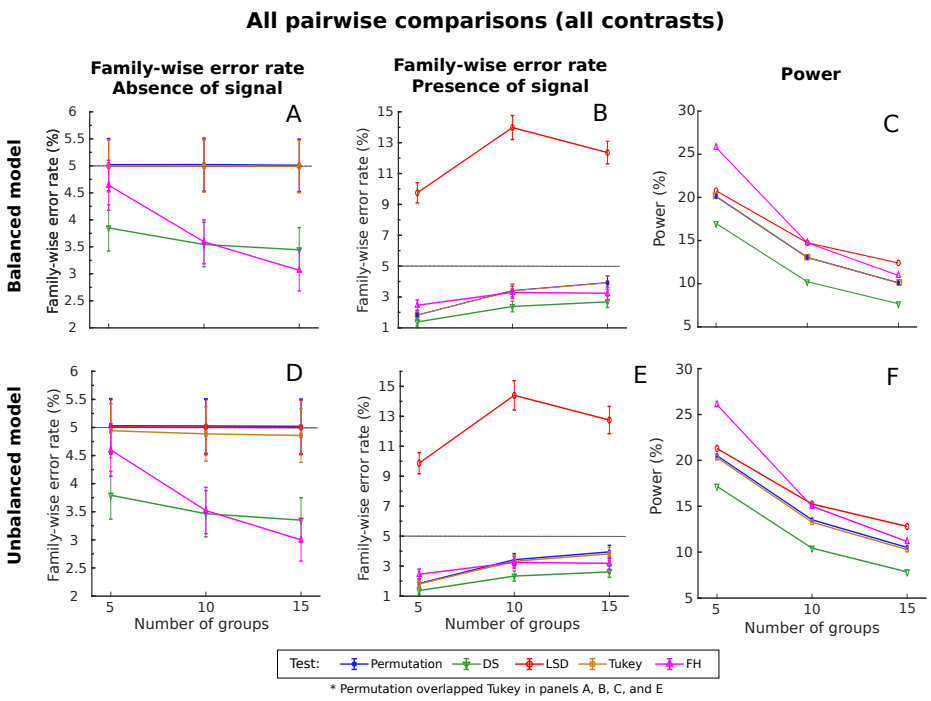

Figure 2: Mean family-wise error rate, its standard deviation and power after correcting across contrasts using permutation, Dunn–Šidák (DS), Fisher LSD, Fisher–Hayter (FH) and Tukey when testing all pairwise comparisons. Starting with balanced models, (A) shows the FWER results in the absence of signal when all contrasts are considered, (B) shows the FWER in the presence of signal, but considering the contrasts that have no signal, and (C) the respective power, i.e., the ability to detect signal for the contrasts that had signal; panels (D), (E), and (F) show, respectively, the same, for unbalanced models.

However, in the presence of signal for some of the contrasts, LSD substantially exceeded the error rate of 5% for those contrasts that did not have signal (Figure 2, B and E). With 10 groups, for example, the FWER was 13.98% for balanced and 14.39% for unbalanced models, which is almost three times higher than the expected nominal level of 5%. All other methods maintained the control over the FWER in the presence of signal for some of the contrasts. Permutation and Tukey procedure had a similar FWER of around 3.4%, while Dunn–Šidák was the most conservative, with an observed FWER of 1.38% for balanced and 1.37% for unbalanced models.

When all possible pairwise comparisons were analysed, and considering only the methods that control the FWER at the nominal level, Fisher–Hayter had the greatest power, with of 25.76% and 26.05% for balanced and unbalanced models, respectively, when 5 groups were considered. In this configuration, permutation and Tukey had similar power of approximately 20%, and Dunn–Šidák was the most conservative procedure with a power of 16.97% in balanced and 17.17% in unbalanced models. As the number of groups increased, the power of Fisher–Hayter decreased at a faster rate than of the other methods (Figure 2, C and F).

In general, the FWER and the power as shown in Figure 2 had similar values for experiments with both balanced and unbalanced designs. The exception was in the simulation in which signal was absent (Figure 2, D), in which Tukey exhibited a slight reduction in the observed FWER that was proportional to the increase of number of groups in the unbalanced model. However, the same trend did not appear in the presence of signal.

The greatest difference in the performance of the various methods appeared when only a subset of linearly independent contrasts was used (Figure 3). In this case, Tukey and Fisher–Hayter were very conservative and had an almost null power in the simulation with 15 groups. In the subset of linearly independent contrasts, the total signal power is around 10.8%, since the only contrasts with signal are the ones that test whether group 3 is greater than 1 or 2. In this context, permutation had the greatest power among the valid methods, 9.17% and 9.44% in the experiment with 5 groups, for balanced and unbalanced respectively.

4.2. Real data

Division of the subjects into 5 groups using the 20th, 40th, 60th and 80th percentiles resulted in unbalanced groups with 56 subjects on average. Table 1 shows the range from BSMSS and ECW, as well as the values used to

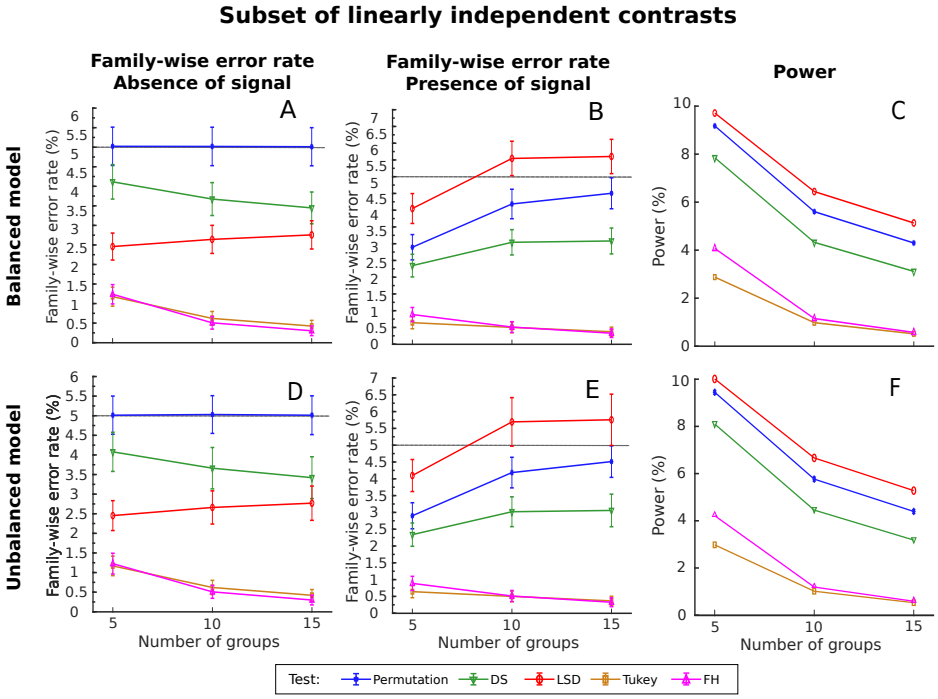

Figure 3: Mean family-wise error rate, its standard deviation and power after correcting across contrasts using permutation, Dunn–Šidák (DS), Fisher LSD, Fisher–Hayter (FH) and Tukey when testing only a subset of linearly independent contrasts. Starting with balanced models, (A) shows the FWER results in the absence of signal when all contrasts are considered, (B) shows the FWER in the presence of signal, but considering the contrasts that have no signal, and (C) the respective power, i.e., the ability to detect signal for the contrasts that had signal; panels (D), (E), and (F) show, respectively, the same, for unbalanced models.

divide the subjects into these discrete groups.

Table 1: Measures and its percentiles used to divided the data into groups.

Measure	Range	Percentiles			
		20th	40th	60th	80th
BSMSS	9 – 66	35.05	48	54.5	61
ECW (liters)	4.23 – 49.59	11.28	15.39	19.24	26.20

Without correction for the multiplicity of contrasts, some effects were detected among BSMSS groups (Figure 4): subjects in group 2 had greater mean cortical volume in the right nucleus accumbens than the subjects from groups 3 and 4 (contrasts 6 and 7); group 4 showed greater mean volume in the left pallidum than group 1 (contrast 14); and subjects in group 5

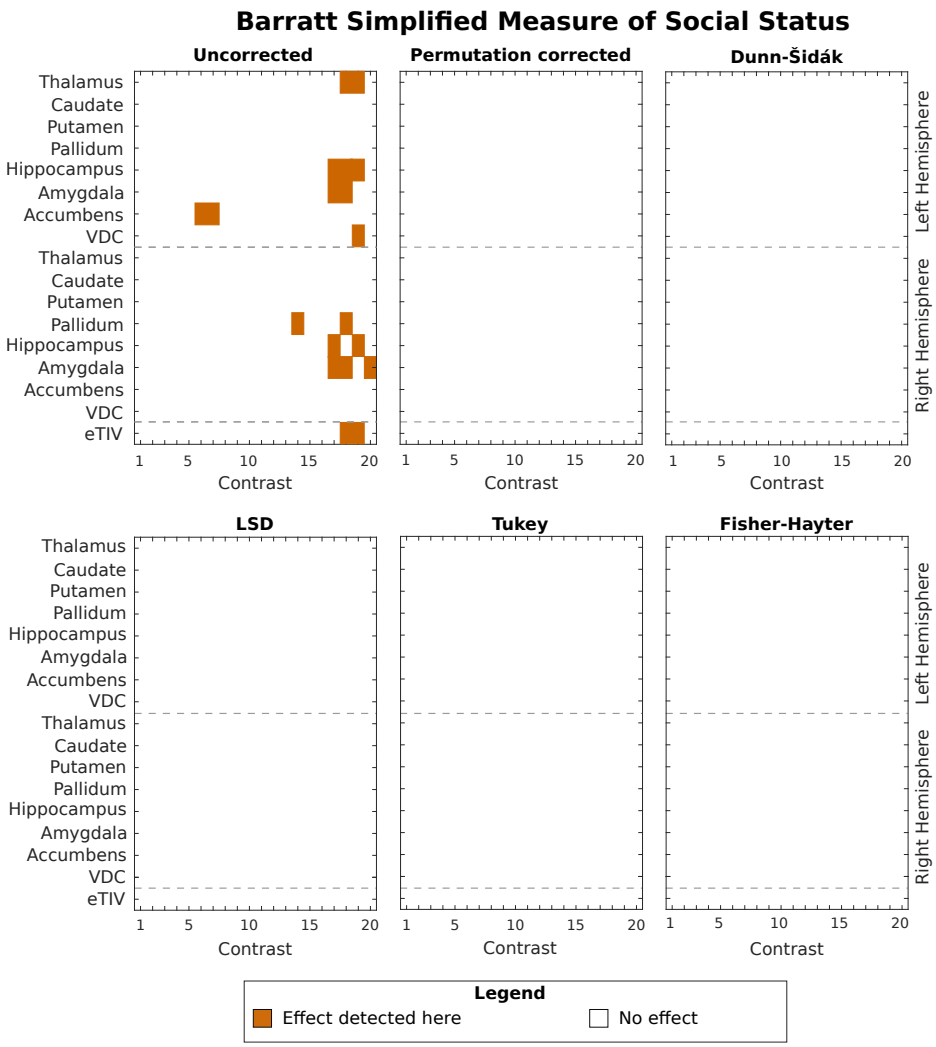

Figure 4: Effect in both hemispheres per contrast detected when dividing the subjects into five groups using the percentiles from BSMSS.

had greater estimated total intracranial volume, as well as larger thalamus, hippocampus, amygdala, pallidum and ventral diencephalon volumes than some of the other groups (contrast 16 to 20). After correction, no such effects were observed.

For ECW, a number of regions were found to have significantly larger volumes for all ECW groups in relation to group 1 (contrasts 5, 9, 13 and 17). From these, eTIV, as well as right pallidum, amygdala and accumbens did not survive the correction across contrasts. These results are summarised

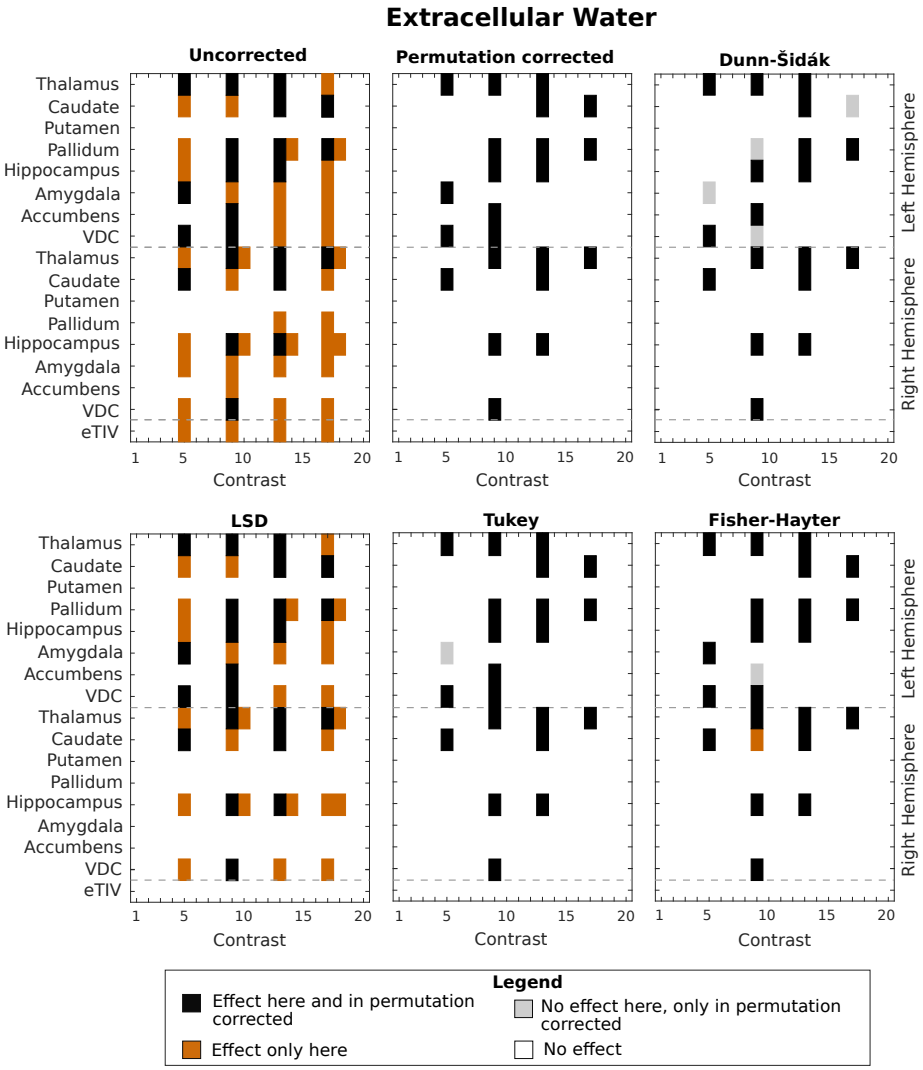

Figure 5: Effect in both hemispheres per contrast detected when dividing the subjects into five groups using the percentiles from extracellular water.

in Figure 5. Among the correction methods, permutation and Tukey had similar performance, differing only in the detection of greater left amygdala volume for group 2 in relation to group 1 (contrast 5). Permutation results differed a bit more when comparing with those obtained with Fisher–Hayter, in which a difference in the volume of the left accumbens that was significant with permutation was no longer so after Fisher–Hayter (contrast 9, $G_3 > G_1$), and another effect was significant with Fisher–Hayter in the right caudate, but not with permutation (also contrast 9). A much sparser set of results was observed with the Dunn–Šidák approach, which did not identify the same

effects as permutation in the left hemisphere in the amygdala, hippocampus, ventral diencephalon and caudate (contrasts 5, 9 and 17), nor the same effects that remained after the correction with Fisher–Hayter in the right caudate. While a substantial number of comparisons remained significant after Fisher's LSD correction, the simulations had already demonstrated that these results are invalid; nonetheless, these are also shown in Figure 5.

5. Discussion

5.1. Error rates and power

Permutation was the only method that performed well in all simulation scenarios: it always controlled the FWER at the test level (was an exact procedure). While methods such as Fisher–Hayter and Tukey had a good performance in some scenarios, they have limitations: they do not extend trivially to designs that do not follow an AN(C)OVA design, they concern all possible pairwise comparisons and do not perform as well when only a subset of them are used.

When some of the pairwise comparisons contained signal, LSD vastly exceeded the 5% error rate for the comparisons that did not contain signal, confirming the results from Hayter (1986), and showing that LSD offers only weak control of the FWER, i.e., it controls the FWER only when no true effect is present in any of the hypotheses being tested, and becoming invalid (that is, extrapolating the test level) when signal was present in some of the contrasts. The other methods we assessed in this technical note, in turn, offered strong control, that is, they ensured an FWER equal to or smaller than the test level, both in the absence and presence of true effects (for the definition of strong and weak control, see Hochberg and Tamhane, 1987; Nichols and Hayasaka, 2003).

Among the methods tested, Dunn–Šidák held the lowest observed FWER among all methods when all pairwise group comparisons are considered, that is, the case in which there are dependencies among the tests, being substantially below the nominal test level of 5%, and therefore being generally the most conservative (although for 15 groups, Fisher–Hayter became even more conservative). This method had also the lowest power. When testing linearly independent contrasts, however, Dunn–Šidák became less conservative, and its difference from permutation, which held the greatest power among the valid methods, was always less than 1.5%.

In the complete absence of signal, the FWER from most methods either remained stable, or tended to become more conservative, except LSD. With

signal in groups 1 and 2, the observed FWER tended to approach the test level as the number of groups increased, which is due to the smaller proportion of contrasts that contain signal when the number of groups is larger. While this is not surprising, the rate with which these methods approached the test level differed substantially, with permutation generally being the closest to the test level among the methods that are valid (thus, excluding Fisher's LSD).

It should be noted, in addition to these results, that permutation can be applied not only to contrasts involving all pairwise comparisons or a linearly independent subset of contrasts in an ANC(C)OVA design, but also to hypotheses in the general linear model that involve any arbitrary combination of continuous and discrete regressors. Moreover, permutation tests are particularly interesting in experiments in which there are only a few subjects, or if the assumptions that underlie parametric tests cannot be confirmed (Ludbrook and Dudley, 1998).

From the simulation results present in Figures 2 and Figures 3, it is clear that, although there are differences among the performance of the different methods, the more tests are conducted, the more strict the correction applied to the contrasts becomes. To reduce power losses, a careful definition of the research hypotheses must be done beforehand, which once again favours permutation tests. The reason is that the correction using the permutation distribution of the maximum can consider only the specific hypotheses that the researcher is interested in, as opposed to blanket corrections that implicitly consider all possible comparisons, many of which might not be of any interest.

5.2. Contrast correction and brain imaging

Accommodating correction over contrasts with brain imaging requires that both types of multiple testing, that is, across space (MTP-I), and across contrasts (MTP-II) (Winkler et al., 2016b) are considered together. For the former, permutation tests offer a solution that is valid, powerful, and with minimal assumptions (Westfall and Young, 1993; Nichols and Holmes, 2002; Winkler et al., 2014), and that extends to the latter in a quite simple manner that can be included in any permutation testing algorithm, that is, the correction can use the distribution of the maxima across imaging units (voxels, vertices, regions) *and* also across contrasts.

The same cannot be said for the other methods discussed: corrections that would bypass the need for permutations for both contrasts and imaging

units would need to rely not only on non-permutation methods for correction across contrasts for AN(C)OVA designs such as those presented here, but also on the many assumptions associated with methods as the random field theory (Worsley et al., 1996, RFT) for correction across image points. Furthermore, there are no known RFT results for fields following the SRD. The converse, that is, correcting first for the number of image points, and then the correction over contrasts, would find other difficulties since, likewise, there are no known results for the Euler characteristic across multiple, possibly non-independent, search volumes in the context of the RFT. All these would impose substantial challenges to guarantee control over the FWER. The most direct way to solve either of these is to use a permutation test, and once that is is used for one kind of multiple testing, correction for the other can be included in the same algorithm, with negligible further computational overhead.

5.3. Real data

We have shown an example of how contrast correction can be applied to an ANCOVA, here generating discrete groups by dividing subjects into groups using the percentiles of two continuous variables. Although such correction should be done even when testing continuous regressors in the GLM, Fisher's LSD, Tukey and Fisher–Hayter methods can only be used to test differences between groups, and do not extend trivially to studies that do not follow an AN(C)OVA design.

After correction for the multiplicity of contrasts, no significant differences were observed between the BSMSS groups. The BSMSS score ranges between 8 and 66 and can be used as a proxy for the social status by assessing, for a child, the occupation of their parents and level of schooling (Barratt, 2006). It does not measure the social class directly, neither the economic status. Although some studies have found brain regions, such as the hippocampus, that appear to be correlated with socioeconomic status (SES), income, and/or stress related to SES (Hanson et al., 2011; Hair et al., 2015; Hanson et al., 2015; Jednoróg et al., 2012; Luby et al., 2013; Yu et al., 2018; Dufford et al., 2018; McDermott et al., 2019), none of them investigated the relation between the brain morphology and only the social status. Even though the score of social status provided by BSMSS and the socioeconomic status are related, they are not equivalent (Barratt, 2006). Besides, other studies classified the subjects as "in" or "out" of the poverty class, while in this ANCOVA example we are dividing the subjects into 5 groups

using the data as available publicly. Although the first percentile represents the subjects in the lower social classes, this does not implicate that they are also in the poverty class. Therefore, the findings without correction shown in Figure 4 might be indeed false positives.

For ECW, more than half of the effects found without correction did not survive after correction using the valid methods (thus, excluding Fisher's LSD). However, those that did survive, did so consistently across most of the methods, including Dunn–Šidák. As the latter is the most conservative, the fact that the results are generally similar across the methods, with only small differences compared to permutation (Figure 5) and the other valid methods, suggests that it is unlikely that these results are mere false positives. To the best of our knowledge, there are no studies investigating the correlation between the body ECW and the cortical volume of any areas. The possibility that these ECW effects are true positives would be strengthened after correcting for the number of regions considered, that is, the 8 regions from each hemisphere, plus the eTIV (these here take the role of image points). However, doing so in this analysis would unduly punish all methods except permutation and DS, since only the latter two can accommodate directly the MTP-I.

6. Conclusions

We compared different methods for multiple testing correction across contrasts in the context of the GLM, for both synthetic and real data, and argued that such correction is necessary to avoid excess of false positives. Among those methods, permutation offers a set of key advantages, some of which were demonstrated. It controls the error rate close to the nominal level, and it is also the most flexible method, as it can be used with arbitrary GLM designs, corrects over specific hypotheses of interest, and allows correction to various sources of multiplicity, all of which can be implemented in the same algorithm with minimal cost.

Acknowledgements

B.A.V.A. was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). A.M.W. was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; 211534/2013-7). T.E.N. was supported by the Wellcome Trust, 100309/Z/12/Z. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

References

Alexander, L.M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., Vega-Potler, N., Langer, N., et al., 2017. Data Descriptor: An open resource for transdiagnostic research in pediatric mental health and learning disorders. Scientific Data 4, 1–26. doi:10.1038/sdata.2017.181.

Barch, D.M., Burgess, G.C., Harms, M.P., Petersen, S.E., Schlaggar, B.L., Corbetta, M., Glasser, M.F., Curtiss, S., Dixit, S., Feldt, C., et al., 2013. Function in the human connectome: task-fmri and individual differences in behavior. Neuroimage 80, 169–189.

Barratt, W., 2006. The Barratt simplified measure of social status (BSMSS): Measuring SES. Unpublished manuscript, Indiana State University .

Bonferroni, C., 1936. Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze 8, 3–62.

Braun, H.I., Tukey, J.W., 1994. The Collected Works of John W. Tukey. Volume VIII: Multiple Comparisons (1948-1983). Chapman & Hall, London.

Buckner, R.L., Head, D., Parker, J., Fotenos, A.F., Marcus, D., Morris, J.C., Snyder, A.Z., 2004. A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. NeuroImage 23, 724–38.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis: I. segmentation and surface reconstruction. Neuroimage 9, 179–194.

Dufford, A.J., Bianco, H., Kim, P., 2018. Socioeconomic disadvantage, brain morphometry, and attentional bias to threat in middle childhood. Cognitive, Affective, & Behavioral Neuroscience , 1–18.

Dunn, O.J., 1958. Estimation of the means of dependent variables. The Annals of Mathematical Statistics , 1095–1111.

Eaton, J.W., Bateman, D., Hauberg, S., Wehbring, R., 2014. GNU Octave version 3.8.1 manual: a high-level interactive language for numerical computations. CreateSpace Independent Publishing Platform. URL: https://www.gnu.org/software/octave/. ISBN 1441413006.

Esteban, O., Birman, D., Schaer, M., Koyejo, O.O., Poldrack, R.A., Gorgolewski, K.J., 2017. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. PloS one 12, e0184661.

Farcomeni, A., 2008. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. Statistical methods in medical research 17, 347–388.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.

Fischl, B., Salat, D.H., Van Der Kouwe, A.J., Makris, N., Ségonne, F., Quinn, B.T., Dale, A.M., 2004. Sequence-independent segmentation of magnetic resonance images. Neuroimage 23, S69–S84.

Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical surface-based analysis: Ii: inflation, flattening, and a surface-based coordinate system. Neuroimage 9, 195–207.

Fisher, R.A., 1935. The design of experiments. 1st ed., Oliver & Boyd.

Hair, N.L., Hanson, J.L., Wolfe, B.L., Pollak, S.D., 2015. Association of child poverty, brain development, and academic achievement. JAMA pediatrics 169, 822–829.

Hanson, J.L., Chandra, A., Wolfe, B.L., Pollak, S.D., 2011. Association between income and the hippocampus. PloS one 6, e18712.

Hanson, J.L., Nacewicz, B.M., Sutterer, M.J., Cayo, A.A., Schaefer, S.M., Rudolph, K.D., Shirtcliff, E.A., Pollak, S.D., Davidson, R.J., 2015. Behavioral problems after early life stress: contributions of the hippocampus and amygdala. Biological psychiatry 77, 314–323.

Hayter, A.J., 1986. The Maximum Familywise Error Rate of Fisher's Least Significant Difference Test. Journal of the American Statistical Association 81, 1000–1004.

Hochberg, Y., Tamhane, A.C., 1987. Multiple Comparison Procedures. 1st ed., John Wiley & Sons, Inc.

Hsu, J.C., 1996. Multiple Comparison Procedures: Theory and Methods. 1st ed., Chapman & Hall/CRC.

Jednoróg, K., Altarelli, I., Monzalvo, K., Fluss, J., Dubois, J., Billard, C., Dehaene-Lambertz, G., Ramus, F., 2012. The influence of socioeconomic status on children's brain structure. PloS one 7, e42486.

Keppel, G., Wickens, T.D., 2004. Design and analysis: A researcher's handbook. Pearson Prentice-Hall.

Luby, J., Belden, A., Botteron, K., Marrus, N., Harms, M.P., Babb, C., Nishino, T., Barch, D., 2013. The effects of poverty on childhood brain development: the mediating effect of caregiving and stressful life events. JAMA pediatrics 167, 1135–1142.

Ludbrook, J., Dudley, H., 1998. Why permutation tests are superior to t and F tests in biomedical research. The American Statistician 52, 127–132.

Marcus, R., Peritz, E., Gabriel, K.R., 1976. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63, 655.

McDermott, C.L., Seidlitz, J., Nadig, A., Liu, S., Clasen, L.S., Blumenthal, J.D., Reardon, P.K., Lalonde, F., Greenstein, D., Patel, R., et al., 2019. Longitudinally mapping childhood socioeconomic status associations with cortical and subcortical morphology. Journal of Neuroscience 39, 1365–1373. doi:10.1523/JNEUROSCI.1808-18.2018.

Nichols, T., Hayasaka, S., 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. Statistical methods in medical research 12, 419–446.

Nichols, T.E., 2012. Multiple testing corrections, nonparametric methods, and random field theory. Neuroimage 62, 811–815.

Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Human Brain Mapping 15, 1–25.

Pesarin, F., Salmaso, L., 2010. Permutation tests for complex data: theory, applications and software. John Wiley & Sons.

Phipson, B., Smyth, G.K., 2010. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. Statistical applications in genetics and molecular biology 9.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.r-project.org/.

Rosen, A.F., Roalf, D.R., Ruparel, K., Blake, J., Seelaus, K., Villa, L.P., Ciric, R., Cook, P.A., et al., 2018. Quantitative assessment of structural image quality. NeuroImage 169, 407–418. doi:10.1016/j.neuroimage.2017.12.059.

Šidák, Z., 1967. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 62, 626–633.

Tippett, L.H., 1931. Methods of statistics. Williams Norgate: London.

Tukey, J.W., 1953. The problem of multiple comparisons. Unpublished manuscript. See Braun and Tukey (1994), pp. 1-301 .

Westfall, P.H., Young, S.S., 1993. Resampling-Based Multiple Testing: Examples And Methods for p-Value Adjustment. John Wiley and Sons, New York.

Winkler, A.M., Ridgway, G.R., Douaud, G., Nichols, T.E., Smith, S.M., 2016a. Faster permutation inference in brain imaging. NeuroImage 141, 502–516.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. NeuroImage 92, 381–397. doi:10.1016/j.neuroimage.2014.01.060.

Winkler, A.M., Webster, M.A., Brooks, J.C., Tracey, I., Smith, S.M., Nichols, T.E., 2016b. Non-parametric combination and related permutation tests for neuroimaging. Human brain mapping 37, 1486–1511.

Worsley, K.J., Marrett, S., Neelin, P., Vandal, A.C., Friston, K.J., Evans, A.C., 1996. A unified statistical approach for determining significant signals in images of cerebral activation. Human Brain Mapping 4, 58–73.

Yu, Q., Daugherty, A.M., Anderson, D.M., Nishimura, M., Brush, D., Hardwick, A., Lacey, W., Raz, S., Ofen, N., 2018. Socioeconomic status and hippocampal volume in children and young adults. Developmental science 21, e12561.