Published in final edited form as:

Neuroimage. 2020 August 01; 216: 116760. doi:10.1016/j.neuroimage.2020.116760.

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]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 represent hypotheses) are tested in the same general linear model. We argue that a correction for this multiplicity must be performed to avoid excess of false positives. Various methods for correction have been proposed in the literature, but few have been applied to 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 on each voxel or vertex of an image representation of the brain. However, an equally common situation in which such multiplicity occurs is when multiple contrasts of parameter estimates of the same general linear model (GLM) are considered, or even 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

*Corresponding author. biancaalberton@alunos.utfpr.edu.br (B.A.V. Alberton).

CRediT authorship contribution statement

Bianca A.V. Alberton: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing -review & editing, Visualization. **Thomas E. Nichols**: Writing - original draft, Conceptualization, Validation. **Humberto R. Gamba**: Resources, Funding acquisition, Project administration. **Anderson M. Winkler**: Conceptualization, Methodology, Software, Validation, Formal analysis, Resources, Writing - original draft, Writing - review & editing, Supervision.

Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2020.116760.

reused and reanalysed in different ways to consider different models and hypotheses. If left uncontrolled, such multiplicity can lead to an undesirably high number of false positives.

For example, in the Human Connectome Project (HCP), the task f_{MRI} 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 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:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon \tag{1}$$

where $\mathbf{y}$ is $N \times 1$ vector containing the image data for N subjects at a given voxel (or vertex), $\mathbf{X}$ is the $N \times R$ full column rank design matrix with the modelled R explanatory variables, $\boldsymbol{\beta}$ is the $R \times 1$ vector with the (to be estimated) parameters that linearly combine the variables in $\mathbf{X}$ to explain the variability observed in $\mathbf{y}$, and $\boldsymbol{\varepsilon}$ is the $N \times 1$ vector of random errors. The coefficients can be estimated via ordinary least squares as $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. The interest is to test K, $K \geq 1$, null hypotheses, each of them represented as $\mathcal{H}_k^0 : \mathbf{C}_k'\boldsymbol{\beta} = 0, k = \{1, \ldots, K\}$, $\mathbf{C}_k$ is a $R \times S_k$ full rank matrix that defines the contrasts of parameter estimates. In ANOVA designs, the interest is a global (*omnibus*) null hypothesis of no difference among all groups, which can be tested using the F-statistic:

$$F = \frac{\widehat{\boldsymbol{\beta}}'\mathbf{C}_k\left(\mathbf{C}'_k(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_k\right)\mathbf{C}_k'\widehat{\boldsymbol{\beta}}}{\nu_1} \Big/ \frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{\upsilon_2} \tag{2}$$

where $\nu_1 = \text{rank}(\mathbf{C}_k)$ and $\nu_2 = N - \text{rank}(\mathbf{X})$ are the degrees of freedom. Large values of F provide evidence against the null, and favours the alternative hypothesis $\mathcal{H}_k^1 : \mathbf{C}_k'\boldsymbol{\beta} \neq 0$. If $S_k = 1$, $\mathbf{C}_k'\widehat{\boldsymbol{\beta}}$ is scalar and a t-statistic can be obtained from F as $t = \text{sign}\left(\mathbf{C}_k'\widehat{\boldsymbol{\beta}}\right)\sqrt{F}$, or equivalently:

$$t = \frac{\mathbf{C}'_k \widehat{\boldsymbol{\beta}} \left(\mathbf{C}'_k (\mathbf{X}'\mathbf{X})^{-1} \mathbf{C}_k \right)^{-\frac{1}{2}}}{\sqrt{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}} / \nu_2}} \quad (3)$$

Large values of t provide evidence against the null, and favours the alternative hypothesis $\mathscr{H}_k^1 : \mathbf{C}_k' \boldsymbol{\beta} > 0$. In either case, the test is said to be *significant* at the level α, $0 < \alpha < 1$, if the probability of observing a random statistic T larger or equal than t (or F) is smaller or equal than α when the null hypothesis is true, i.e., if $\mathbb{P}(T \geqslant t) \leqslant \alpha$; in that case, the null hypothesis is rejected in favour of the alternative hypothesis. The probability $\mathbb{P}(T \geqslant t)$ if the null is true is the p-value; it can be computed based on distributional assumptions on t, or through a resampling procedure, such as a permutation test (Pesarin and Salmaso, 2010). For any test performed, if the probability of rejecting the null hypothesis when it is in fact true is exactly equal to α, the test is said to be *exact*; if the probability is smaller than α, the test is said to be *conservative*; and if the probability is larger than α, the error rate is not controlled and the test is *invalid*.

Permutation is a non-parametric method in which the distribution of the test statistic is ascertained by explicitly calculating all (or a large number of) the possible values that it could assume should the null hypothesis be true. The fundamental assumption that ensures validity of permutation methods is that of *exchangeability* under the null hypothesis, that is, the joint distribution of the data remains unchanged after permutations (see Helwig, 2019b, for a recent review). In other words, any permutation of the data is, under the null, as likely to have been observed as the original, unpermuted data. The distribution of the test statistic is then 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, $j = \{1, 2, \ldots, J\}$, for which larger values provide more evidence against the null hypothesis, 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 statistic 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).

2.2. Multiple testing

As more tests are conducted, the more likely it will be that at least one 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 (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 *familywise 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 case it is changed to accommodate the multiplicity of tests (the p-values remain unchanged), or the p-values can be *adjusted* (p_{adj}), in which case 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 as 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), to distinguish them from the usual multiplicity due to the many measurements taken across space (MTP-I)[1] (Winkler et al., 2016b). The multiplicity of contrasts of model parameter estimates belongs to that class. As with correction over voxels or vertices (MTP-I), which is performed across all image points of interest (e.g., the whole brain, or within a region of interest), it is desirable that the correction in the MTP-IIX considers only the hypotheses of interest. In other words, it is not always the case that all and every possible contrast of parameter estimates is relevant or meaningful to be tested; for example, when investigating G groups, one 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 pairwise group comparisons.

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 (see also Table 1). Although the methods below target control over the FWER, the conceptualisation of correction across contrasts, particularly in the case 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 $\alpha_{cor} = 1 - (1-\alpha)^{1/\tau}$, 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, $p_{adj} = 1 - (1 - p_{unc})^{\tau}$, 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.

[1]Both MTP-I and MTP-II refer to the excess of false positives that arises when multiple tests are performed. The MTP-I should not be confused with *type II* errors, which refer to the false negative test results.

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, a global 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.

Bonferroni.—Based on Boole's inequality, which states that, given a series of events, the probability that at least one happens is smaller or equal to the sum of the probabilities of each of the individual events, Bonferroni (1936) suggested an approximation for the corrected test level as $\alpha_{\text{cor}} = \alpha/\tau$. 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, Bonferroni's method is slightly conservative when compared to Dunn–Šidák (which is exact). If the tests are not independent, the correction becomes yet more conservative. For the relation between Bonferroni and Dunn–Šidák, see Appendix A.

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 studentised range distribution (SRD), that is, $p_{\text{adj}} = 1 - Q_{\text{cdf}}\left(t\sqrt{2}; G, \nu_2\right)$, where Q_{cdf} represents the cumulative distribution function (cdf) from the SRD, G is the number of groups and ν_2 is the number of degrees of freedom. Unlike Fisher's LSD, this method does not require an initial F-test. 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). 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 signals against a single reference regressor). 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.

Scheffé.—Under the null hypothesis that all group means are equal, the adjusted p-value can be obtained as $p_{\text{adj}} = 1 - F_{\text{cdf}}\left(\frac{\text{sign}(t)t^2}{G-1}; G-1, \nu_2\right)$, where F_{cdf} is the cumulative distribution function of the F distribution, with $\nu_1 = G - 1$ and ν_2 degrees of freedom, and $\text{sign}(t)$ is the sign of t-statistic.[2] This method was proposed by Scheffé (1953) as a way to correct the statistic across all possible contrasts, and not only pairwise group comparisons. Therefore, it

has the advantage of correcting across contrasts that investigate more complex comparisons among groups, which is not possible with Tukey or Fisher's LSD. However, when only pairwise comparisons are of interest, this method is much more stringent than Fisher's LSD or Tukey methods.

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 $p_{\text{adj}} = 1 - Q_{\text{cdf}}\left(t\sqrt{2}; G-1, \nu_2\right)$. 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 (Seaman et al., 1991).

Westfall–Young (permutation).—Using permutation tests to calculate the distribution of the maximum statistic across the set of tests that are being corrected, the adjusted p-value can be computed as $p_{\text{adj}} = 1 - t^{\max}_{\text{cdf}}(t) = \frac{1}{J}\sum_j^J I\left(t_j^{\max} \geqslant t\right)$, 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[3] (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 tests that includes it, such that this approach constitutes a *closed testing procedure* (Marcus et al., 1976). Such closed procedures are known to strongly control the FWER; for a recent discussion, see Proschan and Brittain (2020); for strong and weak control, see Hochberg and Tamhane (1987); Nichols and Hayasaka (2003). The use of permutation tests allows the application of Westfall–Young correction for data following any distribution, as long as they are exchangeable and that the calculated statistic is pivotal (more specifically, it is required that the distribution of the test statistic for each test does not depend on others being considered, a condition known as *subset pivotality*).

Wang–Cui.—It has been suggested that a lower critical level for the first stage (the F-test) of the Fisher–Hayter method can be found in a single, holistic procedure that integrates the results of all possible pairwise group comparisons from the second stage. To accomplish this, a Monte Carlo distribution of the F-statistic is computed under the assumption of complete normality but, crucially, for a given Monte Carlo realization m, whenever $\left|t_{ij}^{(m)}\right| < Q^{-1}_{\text{cdf}}(\alpha; G-1, \nu_2)$ for all $i \neq j$ groups, $F^{(m)}$ is defined as 0, as opposed to the actual calculated F-statistic for that realization. Wang and Cui (2017) argue that this procedure is more powerful than Fisher–Hayter, being exact for balanced designs, and conservative otherwise (yet more powerful than Fisher–Hayter).

[2]For negative arguments, F_{cdf} evaluates as zero; for two-tailed t-tests, the term sign(t) can be omitted.

[3]If multiple image points are also considered, then $t_j^{\max}$ is the maximum value of t_j across both image points and contrasts.

2.3. Use in permutation tests

In principle, all methods discussed above and summarised in Table 1 could be used with permutation tests. Bonferroni and Dunn–Šidák, which use as starting point the uncorrected p-values, can be adjusted regardless of whether the p-values are obtained parametrically or non-parametrically; for these methods, all that is needed is independence among the tests and that, under the complete null hypotheses, the p-values are uniformly distributed; if tests are not independent, the resulting inferences are conservative, yet valid. Fisher's LSD, Tukey, Scheffé, Fisher–Hayter and Wang–Cui, as originally proposed, require that the test statistic at one or two stages are compared to some reference distribution, under the assumption of normality. However, nothing prevents that these distributions are also obtained non-parametrically, via permutations. For example, the *F*-distribution used in Fisher's LSD, Fisher–Hayter, and Scheffé can be obtained applying a permutation test using all possible unique (linearly independent) comparisons among the groups. For Tukey, Fisher–Hayter and Wang–Cui, the SRD *Q*-distribution can likewise be obtained using the permutation distribution of the largest difference among all comparisons, using only permutations that change subjects among those groups that are being tested in a given contrast (Petrondas and Gabriel, 1983). For each of these cases, a specific algorithm would be built, and permutations could make these tests robust to departures from normality.

The Westfall–Young method stands distinct from these other, potential methods in that it is intrinsically non-parametric, whereby the distribution of the maximum statistic is obtained via permutations. As with such other potential methods, normality is not assumed, nor independence, since the synchronized permutations over which the maximum statistic is obtained implicitly captures eventual dependencies among the tests. Furthermore, the procedure is optimal in that it targets the very definition of familywise error rate: if the maximum is significant, then surely at least one rejection of the null hypothesis has happened; if the null is true for all tests, then that is a familywise error. The distribution of the maximum statistic is, therefore, a direct way to control the FWER. Moreover, it is algorithmically simpler than permutation versions of the other tests, and requires the permutation of data only for the hypotheses of interest. The Westfall–Young is the *de facto* method for permutation-based FWER-corrected inference. Henceforth, and consistent with the literature, when referring to correction using permutations, we are referring to this method.

3. Evaluation methods

3.1. Synthetic data

To investigate the error rates and power of each method, we considered one-way ANOVA designs, with simulated data representing 2000 voxels (these could also be construed as vertices, or any other type of imaging element). The dataset consisted of random variables following a normal distribution with zero mean and unit variance. We considered 2500 realisations of 8 different scenarios where three simulation parameters were manipulated: 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 $\mathbf{X}\boldsymbol{\beta}$ with $\boldsymbol{\beta} = [\beta_1, \beta_2, 0, \ldots, 0]'$, where $\beta_1 = +s_t$, $\beta_2 = -s_t$ and $s_t = t_{\mathrm{cdf}}^{-1}(1-\alpha; \nu_2)\left(\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}\right)^{1/2}/2$. $\mathbf{C} = [+1, -1, 0, \ldots, 0]'$ 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, both with magnitudes determined by s_t, an approximate power of 50% irrespective to the sample size or number of groups can be expected in all simulations for the contrast comparing these two groups before any adjustment has been made for the multiplicity of tests (in this case, multiplicity of contrasts). Details on how to add signal that confers a specific power before the correction are described in Appendix B. Also, although stronger effects are expected to be detected in the 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 contrasts used to calculate power were excluded from the FWER calculation.

The simulations used sample sizes of 40, 50, 70 or 90 subjects. For the balanced scenarios, subjects were divided in 4, 5, 7 or 9 groups respectively, always with 10 subjects per group. For the unbalanced scenarios, subjects were likewise divided into 4, 5, 7 or 9 groups, for the same respective sample sizes, 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 were greater than every one of the other groups. The reason to investigate a correction over a subset of independent contrasts is that all possible pairwise group differences imply dependencies among them (for example, for three groups, the difference between groups 2 and 3 is fully determined once differences between groups 1 and 2, and 1 and 3 are known). Such dependencies do not exist for a subset of contrasts that are independent from each other, and by selecting all contrasts involving group 3, we obtain the minimum set of contrasts that can be tested while representing all possible groups. This type of model is very common, for example, when testing various patient groups against one control group, and is also a case that demonstrates that not all possible pairwise group comparisons might be of interest. For the analyses using this subset, power can be calculated as the proportion of detected true effects in contrasts where group 3 is greater than groups 1 and 2. The greatest proportion of signal is expected in the contrast testing that group 3 is greater than 2, which gives an approximate (expected) power of 20.6% before any correction is applied. The FWER is estimated as the proportion of tests in which the null hypothesis is rejected for contrasts that compare whether group 3 is greater than each of the groups from 4 to G.

All methods were evaluated using custom code written in GNU Octave (Eaton et al., 2014). Tukey, Fisher–Hayter and Wang–Cui correction, however, used critical values pre–calculated with the functions “ptukey” and “qtukey” from the R statistical software (R Core Team, 2018), whereas Westfall–Young permutation method used PALM – Permutation Analysis of Linear Models (Winkler et al., 2014) with 2500 permutations. Bonferroni and Dunn–Šidák were applied over parametric p-values computed using Student’s t distribution.

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 as 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,[4] 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 of each of these two variables, so that the grouping of subjects of BSMSS was different from the grouping of ECW. 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 West-fall-Young, Dunn–Šidák or Bonferroni), 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, 98 females) were included as nuisance variables.

Two different sets of contrasts were tested: one (set 1), with four contrasts, was used to investigate whether subjects from the first group (those with lowest BSMSS and ECW) would

[4]Currently there are no guidelines for the ideal threshold for the Euler number, which may be dataset- or population-specific (Rosen et al., 2018). The value −80 was chosen based on visual inspection of pial and inflated surfaces of both hemispheres, and on the signal-to-noise ratio and the percent artifact voxels reported by MRIQC, to exclude images with too many defects both on the voxels and on the surfaces; there is no intention that the threshold used here becomes a reference.

have smaller cortical volume than subjects from the other groups; another (set 2), with 20 contrasts, was used to investigate all possible pairwise group differences. The reason for using two sets is that methods that rely on the SRD, i.e., Tukey, Fisher–Hayter and Wang–Cui, assume that there is interest in investigating all pairwise group differences; other methods do not make this assumption. Investigating both ways provides a more comprehensive view of error rates and power for the two situations. Fig. 1 shows the contrasts tested.

Statistical analysis used PALM with 10,000 permutations and tail approximation (Winkler et al., 2016a) to calculate p-values for the *t*-statistic, both uncorrected (for contrasts) and corrected with West-fall-Young permutation method. All other methods were investigated using custom code written on GNU Octave. Bonferroni and Dunn–Šidák were applied over parametric p-values using Student's *t* distribution. As with the synthetic data, for Tukey, Fisher–Hayter and Wang–Cui, we invoked the functions "ptukey" and "qtukey" from R. As requested by an anonymous reviewer, a side analysis using the *G*-statistic (Winkler et al., 2014), which is robust to heteroscedasticity, was also performed. Not all methods can easily accommodate statistics that are robust to heteroscedasticity, such that only Westfall–Young is considered in this case; these additional results are shown in the Supplementary Material.

4. Results

4.1. Synthetic data

Fig. 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.

However, in the presence of signal for some of the contrasts, Fisher's LSD substantially exceeded the error rate of 5% for those contrasts that did not have signal (Fig. 2, B and E). Although with 4 groups the FWER was only slightly above the nominal level (5.51% and 5.62% for balanced and unbalanced groups, respectively), with 9 groups the FWER reached 14.30% for balanced and 15.07% for unbalanced models, which is around 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 had a similar FWER for all group configurations, as well as Fisher–Hayter and Wang–Cui when 4 and 5 groups were used. Bonferroni had a FWER slightly smaller than Dunn–Šidák, with a ratio around 0.9770 between them.[5] The most conservative method was Scheffé, with the highest observed FWER of 0.67%, when 5 groups were considered.

When all possible pairwise comparisons were analysed, and considering only the methods that control the FWER at the nominal level, Wang–Cui had the greatest mean power, with 23.58% and 24.61% for balanced and unbalanced models, respectively, when 4 groups were considered. Permutation had a slightly smaller power of 23.41% and 24.16% for balanced and unbalanced models, and was closely followed by Tukey, which had an observed power

[5]For 4, 5, 7 and 9 groups, there are 12, 20, 42 and 72 contrasts being tested. The expected ratio between Bonferroni and Dunn–Šidák is 0.9760 for 12 tests and 0.9751 for 72 tests (see Appendix A).

of 23.34% and 23.82% for balanced and unbalanced models. It should be noted, though, that Wang–Cui, Westfall–Young and Tukey did not differ significantly (see confidence intervals in Fig. 2, C and F). Fisher–Hayter started with a competitive power of 24.18% in unbalanced models with 4 groups, but as the number of groups increased, the power of Fisher–Hayter decreased at a faster rate than the other methods (Fig. 2, C and F). Dunn–Šidák, with a power of 21.04% when 4 unbalanced groups were considered, showed a slightly greater power than Bonferroni, with a power of 20.76% in the same configuration. Scheffé was the most conservative method, with a power of 16.56% in balanced and 16.91% unbalanced models with 4 groups.

In general, the FWER and the power had similar values for experiments with both balanced and unbalanced designs. The exception was in the simulation of unbalanced groups in complete absence of signal (Fig. 2, D), in which Tukey exhibited a small reduction in the observed FWER and power in relation to the balanced case. This slight reduction was proportional to the number of groups in the unbalanced model. However, the same trend did not appear in the FWER 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 (Fig. 3). In this case, Tukey, Fisher–Hayter and Wang–Cui were very conservative and had a power smaller than 1.4% in the simulation with 9 groups. In the subset of linearly independent contrasts, West-fall-Young had the greatest power among the valid methods, 10.95% and 11.35% in the experiment with 4 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 2 shows the range from BSMSS and ECW, as well as the values used to divide the subjects into these discrete groups.

Without correction for the multiplicity of contrasts, four effects were detected among BSMSS groups in the contrast set 1, which consisted of the first group compared to all others: group 1 had smaller volume in the hippocampus and amygdala from both hemispheres when compared to group 5 (contrast 4). These results are shown in Fig. 4. After correction using Westfall–Young, only the effect in the left hippocampus and right amygdala remained significant; using the *G*-statistic did not affect these results, as shown in the Supplementary Material. After correcting with Dunn–Šidák or Bonferroni, only the effect in the left hippocampus remained. No effect survived the correction using Fisher's LSD, Tukey, Scheffé, Fisher–Hayter or Wang–Cui.

In the contrast set 2, that consisted of all possible pairwise group comparisons, some other effects were detected between BSMSS groups before correction across contrasts, as shown in Fig. 5: subjects in group 2 had smaller mean cortical volume in the right pallidum than group 4 (contrast 7), smaller volume in the left thalamus, left hippocampus, both amygdalas, right pallidum and eTIV than group 5 (contrast 8), and greater volume in the left nucleus accumbens than subjects from groups 3 and 4 (contrasts 10 and 14); subjects from group 3 had smaller cortical volume in the hippocampus, left thalamus, left VDC and eTIV (contrast

12); and subjects from group 4 showed smaller right amygdala volume than group 5 (contrast 16). However, after correcting with any of the methods, none of these effects remained significant.

For ECW, contrast set 1, a total of 54 effects were detected without correction, pointing to a number of regions with smaller volumes for group 1 in relation to all other groups, as shown in Fig. 6: thalamus, caudate, hippocampus, amygdala, and ventral diencephalon from both hemispheres, as well as the eTIV and the left pallidum (contrasts 1 to 4); the accumbens and the right pallidum from group 1 were found smaller than some of the other groups (contrasts 2 to 4). After correction across contrasts, 40 effects survived with Westfall–Young; among all methods, this was the one that detected the largest number of effects. Dunn–Šidák and Bonferroni identified the same effects as each other, and when compared to Westfall–Young, they did not identify a significant difference between group 1 and groups 3 and 4 in the right amygdala (contrasts 2 and 3). Wang–Cui, Fisher–Hayter and Tukey detected almost the same effects as each other, except by an effect in the right caudate, showing smaller volume in group 1 than in group 3 (contrast 2) that was detected with Wang–Cui and Fisher–Hayter, but not with Tukey; the latter detected 21 effects in total. Scheffé was the most conservative method and detected only 6 effects (contrasts 2 and 3).

When all pairwise comparisons (contrast set 2) were tested for ECW, a number of regions were found to have smaller volumes for group 1 than all other ECW groups (contrasts 1 to 4), as well as smaller left pallidum, right thalamus and right hippocampus in the group 2 than some of the groups 3 to 5 (contrasts 6 to 8) using uncorrected p-values; these results are summarised in Fig. 7. From these, effects in eTIV, as well as right pallidum, amygdala and accumbens did not survive correction across contrasts using any method. Among the correction methods, Fisher–Hayter and Wang–Cui led to the same results. The results with Westfall–Young and Tukey were similar to each other, differing only in the detection of smaller left amygdala volume in group 1 in relation to group 2 (contrast 1); the amount of differences was more pronounced when compared with those obtained with Fisher–Hayter or Wang–Cui, in which a difference in the volume of the left accumbens that was significant with Westfall–Young was no longer so after Fisher–Hayter (contrast 2, $G_1 < G_3$), and another effect was significant with Fisher–Hayter and Wang–Cui in the right caudate, but not with permutation (also contrast 2). A much sparser set of results was observed with the Dunn–Šidák, Bonferroni and Scheffé approaches. Dunn–Šidák did not identify the same effects as permutation in the left hemisphere in the caudate, pallidum, amygdala, and ventral diencephalon (contrasts 1, 2 and 4), nor the same effects that remained after the correction with Fisher–Hayter or Wang–Cui in the right caudate. Furthermore, the effect in the left pallidum (contrast 4) was not detected with Bonferroni, whereas the smallest amount of effects, only in the thalamus, left pallidum, and right hippocampus (contrasts 2 and 3), were detected with Scheffé. 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 Fig. 7.

5. Discussion

5.1. Error rates and power

The Westfall–Young permutation method was the only procedure that performed well in all simulation scenarios and always controlled the FWER at the test level. While methods such as Wang–Cui, Fisher–Hayter and Tukey had good performance in some cases, they have limitations: they do not extend trivially to studies that do not follow an AN(C)OVA design, they take into account all possible pairwise comparisons, and do not perform as well when only a subset of them are of interest. Moreover, Fisher–Hayter and Wang–Cui require, along with LSD, an initial *F*-test, whereas for the others this step is bypassed altogether, with the correction applicable directly to the *t*-tests.

When some of the pairwise comparisons contained signal, Fisher's LSD vastly exceeded the 5% error rate for the comparisons that did not contain signal, confirming the results obtained by Hayter (1986), and showing that Fisher's LSD offers only weak control of the FWER, i.e., it controls the FWER only when no true effects are present in any of the hypotheses being tested, and becoming invalid (that is, extrapolating the test level) when signal is present in some of the contrasts. The other methods, in turn, offered strong control when their assumptions were met, that is, they ensured an FWER equal to or smaller than the test level, both in the absence and in the presence of true effects. It should be noted that since Westfall–Young is a closed testing procedure, it controls the FWER in the strong sense for any number of contrasts, independently of their dependency structure, as long as the exchangeability assumption holds and that the sampling distribution of the test statistic does not depend on unkwown parameters.

Among the methods tested, Scheffé led to the lowest observed FWER among all methods in all scenarios. When all pairwise group comparisons are considered, that is, the case in which there are dependencies among the tests, Dunn–Šidák and Bonferroni were also very conservative, substantially below the nominal test level of 5%, resulting in low power. With linearly independent contrasts, however, Dunn–Šidák and Bonferroni became less conservative, and their difference in power when compared to Westfall–Young was always below 1.5%. Between Dunn–Šidák and Bonferroni, the former is more powerful, and it is exact when the tests performed are independent (whereas even in these cases, Bonferroni is slightly conservative).

In the complete absence of signal, as the number of groups increased, the FWER for most methods either remained stable, or became smaller, towards conservativeness, except for Fisher's LSD. With signal in groups 1 and 2, the observed FWER tended to approach the test level as the number of groups increased, presumably due to the smaller proportion of contrasts that contain signal. This is not surprising, although the rate with which these methods approached the test level differed substantially.

From the simulation results presented in Figs. 2 and 3, it is clear that, although there are performance differences among the methods, as more tests are conducted, the more strict the correction becomes. Although here the focus is on controlling the FWER across contrasts, ensuring a high power, i.e., lower false negative rate, is equally important, and thus, a careful

definition of the research hypotheses must precede statistical analysis. This, again, favours permutation methods, and in this case also Dunn–Šidák and Bonferroni. The reason is that all other methods perform a broad correction that implicitly considers all possible pairwise comparisons (or even any possible contrasts, as in Scheffé), many of which might not be of any interest, a feature that unduly penalises power (Fig. 3, panels C and F).

5.2. Correction over contrasts and brain imaging

Accommodating correction over contrasts in 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 based on 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: the correction can use the distribution of the maximum statistic across imaging units (voxels, vertices, regions) *and* also across contrasts. This is simple enough to be included in any permutation testing algorithm.

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 (RFT; Worsley et al., 1996) 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 across 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 used for one kind of multiple testing, correction for the other can be included in the same algorithm, with negligible further computational overhead.

Furthermore, methods relying on the SRD have fewer applications since this distribution assumes that differences between quantities (e.g., between groups) are being tested. Although we focused on pairwise comparisons, AN(C)OVA designs may include more complex contrasts involving multiple groups, as for example, averaging two groups and comparing them against a third group. Contrasts may also test hypotheses that involve arbitrary combination of continuous and discrete regressors, or some contrasts comparing groups while others test continuous regressors (for example, in a design where categorical groups, age and sex are modelled, there may be interest in testing group differences, as well as linear and/or quadratic effects of age, or sex differences). Such heterogeneity, that goes beyond mere group comparisons, is not easily accommodated by methods based on the SRD, such as Tukey, Fisher–Hayter or Wang–Cui, but are handled easily by the West-fall-Young permutation method, which is very general. More complex designs as these are very common in neuroimaging; Table 3 shows the applicability of the different methods for various common designs.

Moreover, if one-sided *t*-tests are used to investigate both positive and (separately) negative effects, further correction is necessary for the fact that two tests are being considered, a point brought forward recently by Chen et al. (2019), and that can be solved trivially with the West-fall-Young permutation method, as not only it allows correction for the multiplicity of tests that occur when testing any type of regressors, but it also allows inclusion of both positive and negative tests along with any other combination of contrasts of interest, possibly involving other regressors. The method provides valid inferences while retaining the directional information for each individual test, holds even for cases in which positive and negative tests are not independent, and further, can be considered for tests of regression coefficients estimated for different designs; all that is necessary for Westfall–Young is that the assumption of exchangeability holds, as well as the condition of subset pivotality. Additionally, to the extent that the permutation tests can accommodate statistics that are robust to heteroscedasticity (Guillaume et al., 2014; Winkler et al., 2014; DiCiccio and Romano, 2017; Helwig, 2019a), the Westfall–Young method remains valid also in these cases; the same cannot be said for the other, mostly parametric methods.

5.3. Real data

Using real data, two examples where contrast correction can be applied to an ANCOVA were considered, one in which only some comparisons between groups were of interest (contrast set 1) and other where all pairwise comparisons were of interest (contrast set 2). Both used discrete groups by arbitrarily dividing subjects into subsets using the percentiles of two continuous variables. Although the correction for the multiplicity of contrasts should be done even when testing continuous regressors in the GLM, Fisher's LSD, Tukey, Fisher–Hayter and Wang–Cui methods are designed for use in tests of differences between groups, and do not extend trivially to studies that do not follow AN(C)OVA design, as discussed above.

In the contrast set 1, after correction with Westfall–Young, Dunn–Šidák or Bonferroni, subjects with BSMSS score lesser than or equal to 35 (group 1) were found to have smaller hippocampus volume in the left hemisphere than subjects with BSMSS score greater than or equal to 61 (group 5). Additionally, the volume of the right amygdala was also found smaller in group 1 than in group 5. However, in contrast set 2, in which all pairwise comparisons are analysed, and thus the correction for multiplicity of tests is more stringent due to a greater number of tests being performed, these effects were not found. 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). Although it does not measure the social class directly, neither the economic status, some studies have found correlations between socioeconomic status (SES), income, and/or stress related to SES and measurements derived from brain regions such as the hippocampus and the amygdala (Hanson et al., 2011, 2015; Hair et al., 2015; Jednoróg et al., 2012; Luby et al., 2013; Yu et al., 2018; Dufford et al., 2018; McDermott et al., 2019). Still, none of them investigated the relation between the brain morphology and only the social status. Moreover, other studies classified the subjects as "in" or "out" of the poverty class, while in the ANCOVA example subjects were divided into five groups using the data as available publicly. Thus, these results must be interpreted cautiously, and may represent false positives.

For ECW, in both contrast sets analysed, many subcortical structures were found to have smaller volume in group 1 than in other groups. The methods relying on the SRD assume that all pairwise comparisons among groups are of interest, and thus, they lead to the same results in both sets 1 and 2 for the contrasts that are present in both; Scheffé, which assumes that all comparisons among any groups are of interest, hence leading to fewer results compared to the other methods. Westfall–Young, Dunn–Šidák and Bonferroni consider only the contrasts that are truly of interest, and thus, when fewer contrasts are tested (as in set 1), they can potentially find more effects than when many contrasts are tested (as in set 2). In effect, in contrast set 1, Westfall–Young detected almost twice the number of effects found with methods that rely on the SRD. In set 2, the results are generally similar across the methods, with only small differences compared to Westfall–Young (Fig. 7) and the other valid methods (except for Scheffé, which is the most conservative). This 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 Westfall–Young, Bonferroni and Dunn–Šidák, since only these can accommodate directly the MTP-I.

It should be noted that any conclusive interpretation of these results, as to whether they represent true positives or not, need to take power into consideration: any approach to correct for multiple testing has the drawback of a drop in statistical power, as evidenced in the real data example, where groups were artificially divided using BSMSS or ECW scores. While one traditionally worries about type I error, type II error rate is also a consequence of low accuracy of parameter estimates in low powered studies. The noted discrepancies in the literature may be a consequence of random sampling error, since reductions in power are necessary to control the amount of error type I in the multiple testing context. Moreover, presumably much of the published literature on the topic suffers from low statistical power, regardless (for further discussion, see Mumford, 2012; Button et al., 2013; Cremers et al., 2017; Szucs and Ioannidis, 2017).

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 the methods, Westfall–Young 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, as it can be used with arbitrary GLM designs, allows correction over specific hypotheses of interest, and allows correction of various sources of multiplicity, all of which can be implemented with minimal computational cost.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank the anonymous reviewers for their thoughtful comments. We also thank the Child Mind Institute (CMI) for providing data. B.A.V.A. was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). A.M.W. received support from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; 211534/2013-7) and from the National Institutes of Health (ZIA-MH002781 and ZIA-MH002782). T.E.N. was supported by the Wellcome Trust, 100309/Z/12/Z. This work utilized computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

Appendix A.: Relation between Dunn–Šidák and Bonferroni

Dunn–Šidák can be seen as a special case of Kimball's inequality, which states that the probability of all events in a set is equal to or greater than the product of their individual probabilities, and where the equality holds only if these events are independent. Dunn–Šidák's inequality was proved valid for data following a multivariate normal distributions with positive definite correlation matrix (Hochberg and Tamhane, 1987), and therefore, can accommodate some negative correlations between tests. Bonferroni is an approximation of a similar inequality devised by Boole, which establishes the probability of any of the events in a set. Although always valid (independently of the correlation structure of the tests), it is an approximation and, therefore, is never exact. In general, the relation between Dunn–Šidák and Bonferroni is expressed as (Abdi, 2007):

$$\alpha_{\text{cor}} \geqslant \alpha_{\text{DS}} \geqslant \alpha_{\text{B}}$$

where α_{DS} and α_{B} are, respectively the Dunn–Šidák and Bonferroni's corrected test level. The equality between Dunn–Šidák and Bonferroni holds only for $\tau = 1$. As more tests are performed, α_{B} is always smaller than α_{DS}, stabilizing at a ratio of approximately 0.975 for an uncorrected test level $\alpha_{\text{unc}} = 0.05$, as shown in Figure A.1. The smaller the p-values, smaller is this ratio. If the tests are independent, the correction using Dunn–Šidák is exact, that is, $\alpha_{\text{cor}} = \alpha_{\text{DS}}$.

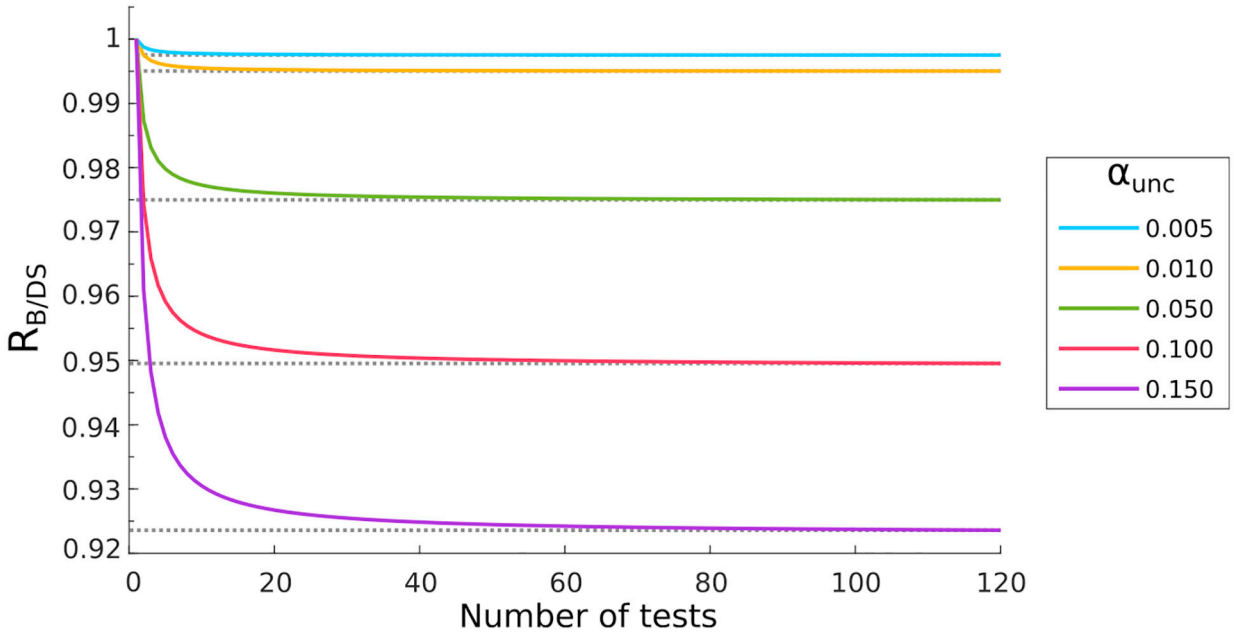

Fig. A.1.
Comparison between the correction performed with Bonferroni and Dunn–Šidák, where $R_{B/DS} = \alpha_B/\alpha_{DS}$. Note that the ratio increases with the number of tests and decreases with the uncorrected level α_{unc} used.

Appendix B.: Simulated signals with the t-statistic

In general, simulated datasets are obtained by sampling random variables that follow a specific distribution (e.g. normal) consistent with the null hypothesis. If the interest is in the alternative hypothesis, a signal with a specific size can be added as a function of the desired power *P*, the test statistic, test level, and sample size. In the context of the GLM, the sample size is represented in the design efficiency, expressed as $\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}^{1/2}$, where **X** and **C** are known, and the effect size is given by $s_t = t_s(\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}^{1/2}$, where $t_s = t_{cdf}^{-1}(1-\alpha;\nu_2) + t_{cdf}^{-1}(P;\nu_2)$ is the location of the peak of signal distribution. Figure B.1 shows two examples of signal added to the normal distribution, one with 50% and the other with 80% power. Note that with 50% power ($P = 0{:}50$), $t_s = t_{cdf}^{-1}(1-\alpha;\nu_2)$ because the peak distribution of the alternative hypothesis is placed over the critical α level. In ANOVA designs, where the group means are compared, this signal can be further divided among groups (e.g. half of the signal for each group, as described in Section 3.1).

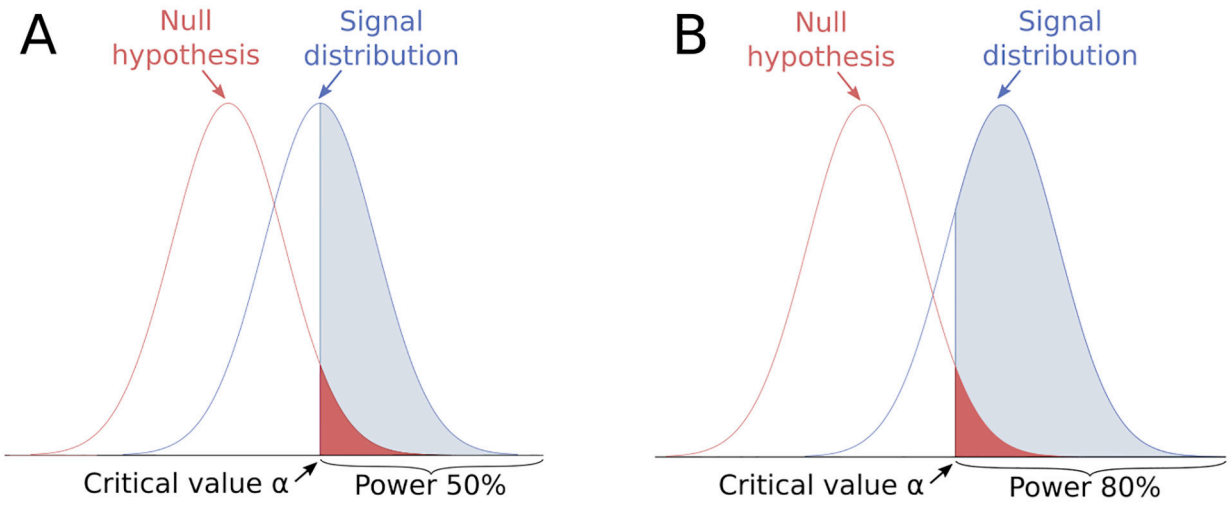

Fig. B.1.

Signal distribution and its relation with the null distribution. Two signals with different powers are exhibited, one with 50% (A) and the other with 80% (B) power.

References

Abdi H, 2007. The bonferonni and Šidák corrections for multiple comparisons. In: Encyclopedia of Measurement and Statistics. SAGE, pp. 1–9.

Alexander LM, 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.

Barch DM, Burgess GC, Harms MP, Petersen SE, Schlaggar BL, Corbetta M, Glasser MF, 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. [PubMed: 23684877]

Barratt W, 2006. The Barratt Simplified Measure of Social Status (BSMSS): Measuring SES. Indiana State University, Unpublished manuscript.

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.

Buckner RL, Head D, Parker J, Fotenos AF, Marcus D, Morris JC, Snyder AZ, 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, 724–738. [PubMed: 15488422]

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR, 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci 14, 365. [PubMed: 23571845]

Chen G, Cox RW, Glen DR, Rajendra JK, Reynolds RC, Taylor PA, 2019. A tail of two sides: artificially doubled false positive rates in neuroimaging due to the sidedness choice with t-tests. Hum. Brain Mapp 40, 1037–1043. [PubMed: 30265768]

Cremers HR, Wager TD, Yarkoni T, 2017. The relation between statistical power and inference in fMRI. PloS One 12, e0184923. [PubMed: 29155843]

Dale AM, Fischl B, Sereno MI, 1999. Cortical surface-based analysis: I. segmentation and surface reconstruction. Neuroimage 9, 179–194. [PubMed: 9931268]

DiCiccio CJ, Romano JP, 2017. Robust permutation tests for correlation and regression coefficients. J. Am. Stat. Assoc 112, 1211–1220.

Dufford AJ, Bianco H, Kim P, 2018. Socioeconomic disadvantage, brain morphometry, and attentional bias to threat in middle childhood. Cognit. Affect Behav. Neurosci 1–18. [PubMed: 29076064]

Dunn OJ, 1958. Estimation of the means of dependent variables. Ann. Math. Stat 1095–1111.

Eaton JW, 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, ISBN 1441413006. URL: https://www.gnu.org/software/octave/.

Esteban O, Birman D, Schaer M, Koyejo OO, Poldrack RA, Gorgolewski KJ, 2017. MRIQC: advancing the automatic prediction of image quality in MRI from unseen sites. PloS One 12, e0184661. [PubMed: 28945803]

Farcomeni A, 2008. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. Stat. Methods Med. Res 17, 347–388. [PubMed: 17698936]

Fischl B, Salat DH, 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. [PubMed: 11832223]

Fischl B, Salat DH, Van Der Kouwe AJ, Makris N, Ségonne F, Quinn BT, Dale AM, 2004. Sequence-independent segmentation of magnetic resonance images. Neuroimage 23, S69–S84. [PubMed: 15501102]

Fischl B, Sereno MI, Dale AM, 1999. Cortical surface-based analysis: ii: inflation, flattening, and a surface-based coordinate system. Neuroimage 9, 195–207. [PubMed: 9931269]

Fisher RA, 1935. The Design of Experiments, first ed. Oliver & Boyd.

Guillaume B, Hua X, Thompson PM, Waldorp L, Nichols TE, 2014. Fast and accurate modelling of longitudinal and repeated measures neuroimaging data. Neuroimage 94, 287–302. [PubMed: 24650594]

Hair NL, Hanson JL, Wolfe BL, Pollak SD, 2015. Association of child poverty, brain development, and academic achievement. JAMA pediatrics 169, 822–829. [PubMed: 26192216]

Hanson JL, Chandra A, Wolfe BL, Pollak SD, 2011. Association between income and the hippocampus. PloS One 6, e18712. [PubMed: 21573231]

Hanson JL, Nacewicz BM, Sutterer MJ, Cayo AA, Schaefer SM, Rudolph KD, Shirtcliff EA, Pollak SD, Davidson RJ, 2015. Behavioral problems after early life stress: contributions of the hippocampus and amygdala. Biol. Psychiatr 77, 314–323.

Hayter AJ, 1986. The maximum familywise error rate of Fisher's least significant difference test. J. Am. Stat. Assoc 81, 1000–1004.

Helwig NE, 2019a. Robust nonparametric tests of general linear model coefficients: a comparison of permutation methods and test statistics. Neuroimage 201, 116030. [PubMed: 31330243]

Helwig NE, 2019b. Statistical nonparametric mapping: multivariate permutation tests for location, correlation, and regression problems in neuroimaging. Wiley Interdisciplinary Reviews: Comput. Stat 11, e1457.

Hochberg Y, Tamhane AC, 1987. Multiple Comparison Procedures, first ed. John Wiley & Sons, Inc.

Hsu JC, 1996. Multiple Comparison Procedures: Theory and Methods, first 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. [PubMed: 22880000]

Luby J, Belden A, Botteron K, Marrus N, Harms MP, 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. [PubMed: 24165922]

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

McDermott CL, Seidlitz J, Nadig A, Liu S, Clasen LS, Blumenthal JD, Reardon PK, Lalonde F, Greenstein D, Patel R, et al., 2019. Longitudinally mapping childhood socioeconomic status associations with cortical and subcortical morphology. J. Neurosci 39, 1365–1373. [PubMed: 30587541]

Mumford JA, 2012. A power calculation guide for fMRI studies. Soc. Cognit. Affect Neurosci 7, 738–742. [PubMed: 22641837]

Nichols T, Hayasaka S, 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. Stat. Methods Med. Res 12, 419–446. [PubMed: 14599004]

Nichols TE, 2012. Multiple testing corrections, nonparametric methods, and random field theory. Neuroimage 62, 811–815. [PubMed: 22521256]

Nichols TE, Holmes AP, 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp 15, 1–25. [PubMed: 11747097]

Pesarin F, Salmaso L, 2010. Permutation Tests for Complex Data: Theory, Applications and Software. John Wiley & Sons.

Petrondas DA, Gabriel KR, 1983. Multiple comparisons by rerandomization tests. J. Am. Stat. Assoc 78, 949–957.

Phipson B, Smyth GK, 2010. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. Stat. Appl. Genet. Mol. Biol 9.

Proschan MA, Brittain EH, 2020. A primer on strong vs weak control of familywise error rate. Stat. Med 1–7. [PubMed: 31663647]

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 AF, Roalf DR, Ruparel K, Blake J, Seelaus K, Villa LP, Ciric R, Cook PA, et al., 2018. Quantitative assessment of structural image quality. Neuroimage 169, 407–418. [PubMed: 29278774]

Scheffé H, 1953. A method for judging all contrasts in the analysis of variance. Biometrika 40, 87–114.

Seaman MA, Levin JR, Serlin RC, 1991. New developments in pairwise multiple comparisons : somme powerful and practicable procedures. Psychol. Bull 110, 577–586.

Šidák Z, 1967. Rectangular confidence regions for the means of multivariate normal distributions. J. Am. Stat. Assoc 62, 626–633.

Szucs D, Ioannidis JPA, 2017. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLoS Biol. 15, e2000797. [PubMed: 28253258]

Tippett LH, 1931. Methods of Statistics. Williams Norgate, London.

Tukey JW, 1953. The Problem of Multiple Comparisons, pp. 1–301. Unpublished manuscript. See ?

Wang B, Cui X, 2017. An improved uniformly more powerful exact Fisher–hayter pairwise comparisons procedure. Biom. J 59, 767–775. [PubMed: 28436123]

Westfall PH, Young SS, 1993. Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment. John Wiley and Sons, New York.

Winkler AM, Ridgway GR, Douaud G, Nichols TE, Smith SM, 2016a. Faster permutation inference in brain imaging. Neuroimage 141, 502–516. [PubMed: 27288322]

Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE, 2014. Permutation inference for the general linear model. Neuroimage 92, 381–397. [PubMed: 24530839]

Winkler AM, Webster MA, Brooks JC, Tracey I, Smith SM, Nichols TE, 2016b. Non-parametric combination and related permutation tests for neuroimaging. Hum. Brain Mapp 37, 1486–1511. [PubMed: 26848101]

Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC, 1996. A unified statistical approach for determining significant signals in images of cerebral activation. Hum. Brain Mapp 4, 58–73. [PubMed: 20408186]

Yu Q, Daugherty AM, Anderson DM, Nishimura M, Brush D, Hardwick A, Lacey W, Raz S, Ofen N, 2018. Socioeconomic status and hippocampal volume in children and young adults. Dev. Sci 21, e12561. [PubMed: 28464381]

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

Fig. 1.

List of contrasts used to compare groups. Set 1 used $\mathbf{C}_1$ through $\mathbf{C}_4$, that is, testing whether G_1 would have smaller gray matter volume than each one of the other four groups, whereas set 2 used $\mathbf{C}_1$ through $\mathbf{C}_{20}$, that is, testing all possible pairwise comparisons among the five groups (G_1, …, G_5) as encoded by the design matrix (not shown). The last two regression coefficients modelled age and sex, in an ANCOVA design.

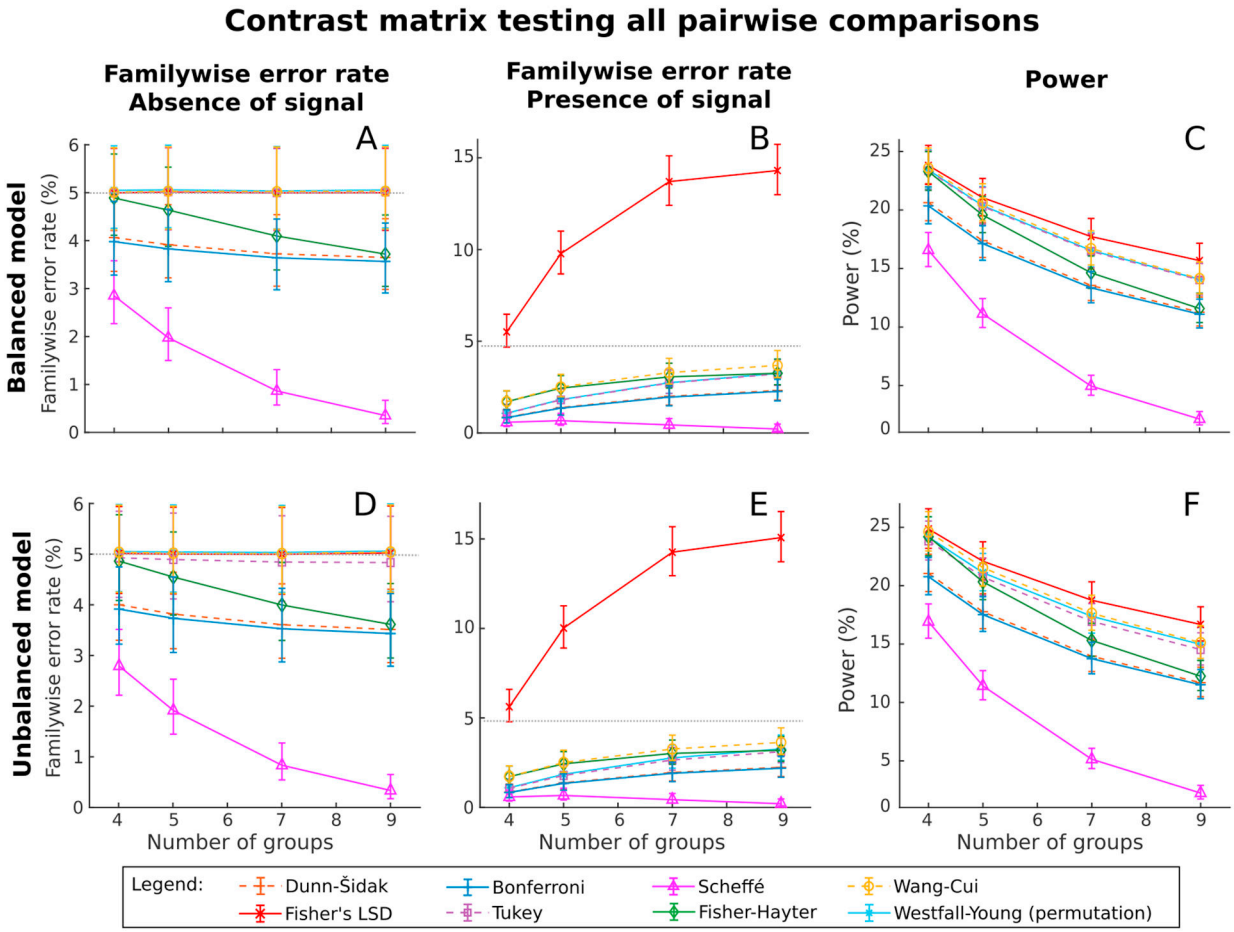

Fig. 2.

Mean familywise error rate and power after correcting across contrasts using Dunn–Šidák, Fisher LSD, Bonferroni, Tukey, Scheffé, Fisher–Hayter, West-fall-Young permutation method and Wang–Cui when testing all pairwise comparisons. Vertical bars represent the 95% confidence interval (tables with FWER and power, along with respective confidence intervals, are available in the Supplementary Material). 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. A power of 50% was expected before any correction was performed.

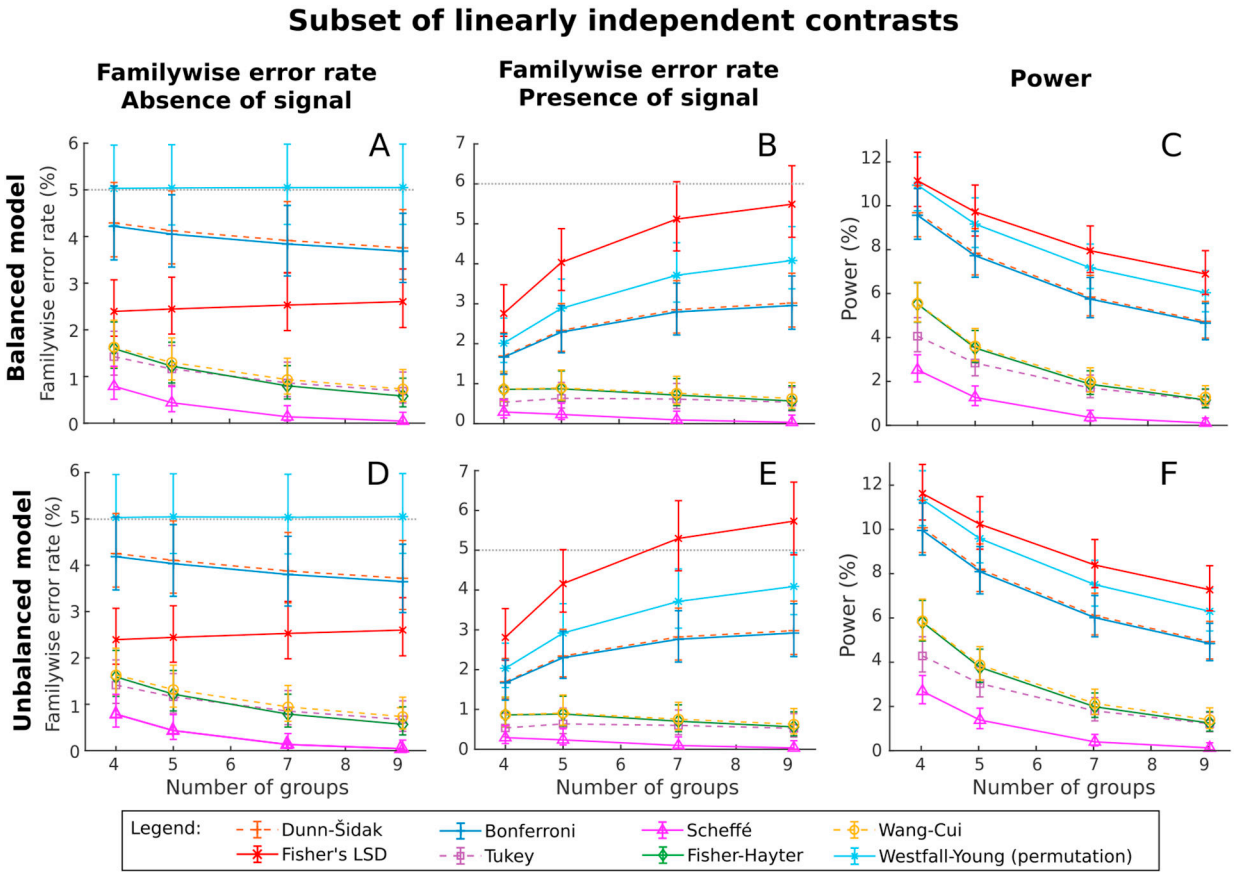

Fig. 3.

Mean familywise error rate and power after correcting across contrasts using Dunn–Šidák, Fisher LSD, Bonferroni, Tukey, Scheffé, Fisher–Hayter, West-fall-Young permutation method and Wang–Cui when testing only a subset of linearly independent contrasts. Vertical bars represent the 95% confidence interval (tables with FWER and power, along with the respective confidence intervals, are available in the Supplementary Material). 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. A power of 20.6% was expected before any correction was performed.

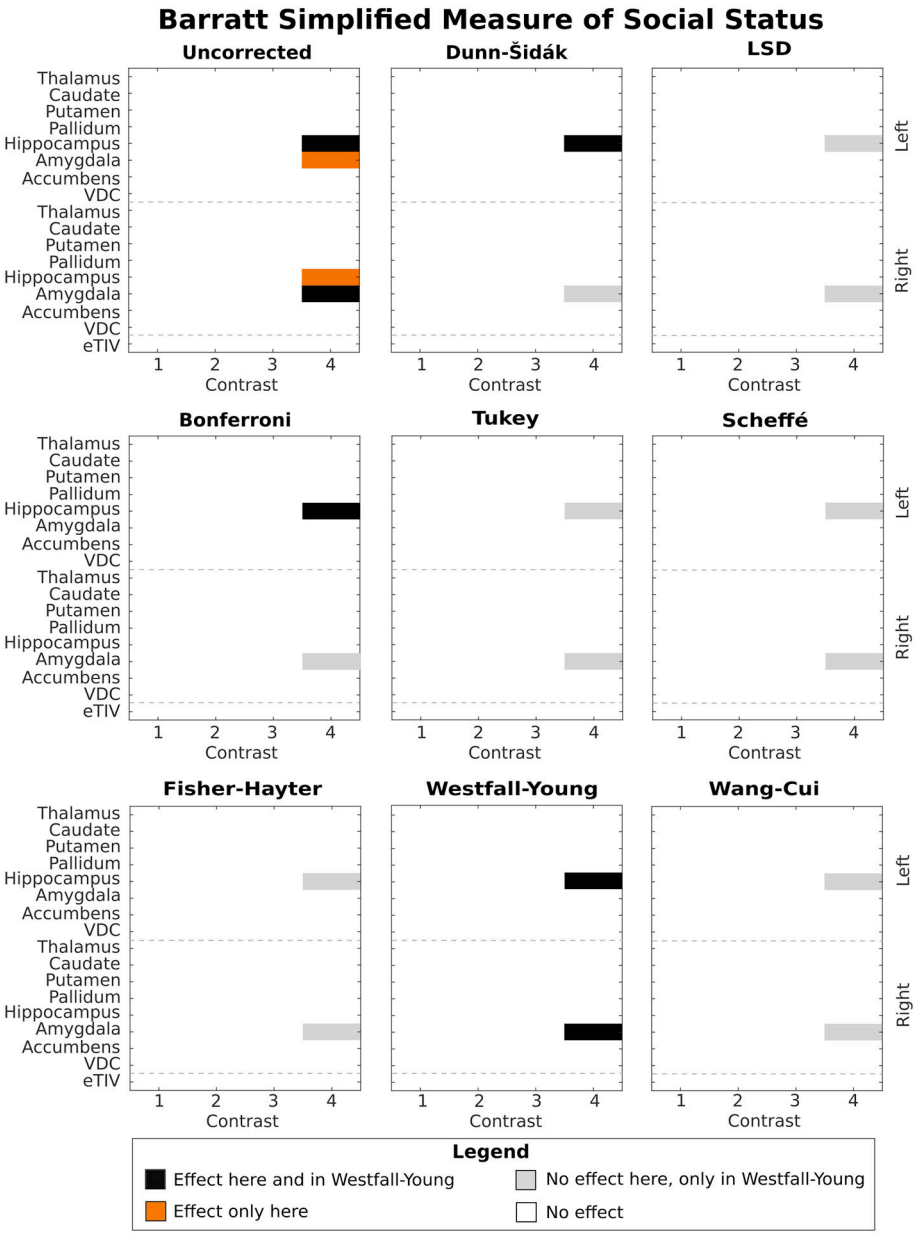

Fig. 4.

Effects in both hemispheres per contrast detected when dividing the subjects into five groups using the percentiles from BSMSS and testing contrast set 1, that is, when testing whether subjects with lower scores of BSMSS (group 1) have smaller cortical volume than subjects from the other four groups (with higher scores of BSMSS). Contrast 1 tests if group 1 has smaller volume than group 2, contrast 2 tests if group 1 has smaller volume than group 3, and so forth (see Fig. 1 for the list of contrasts tested).

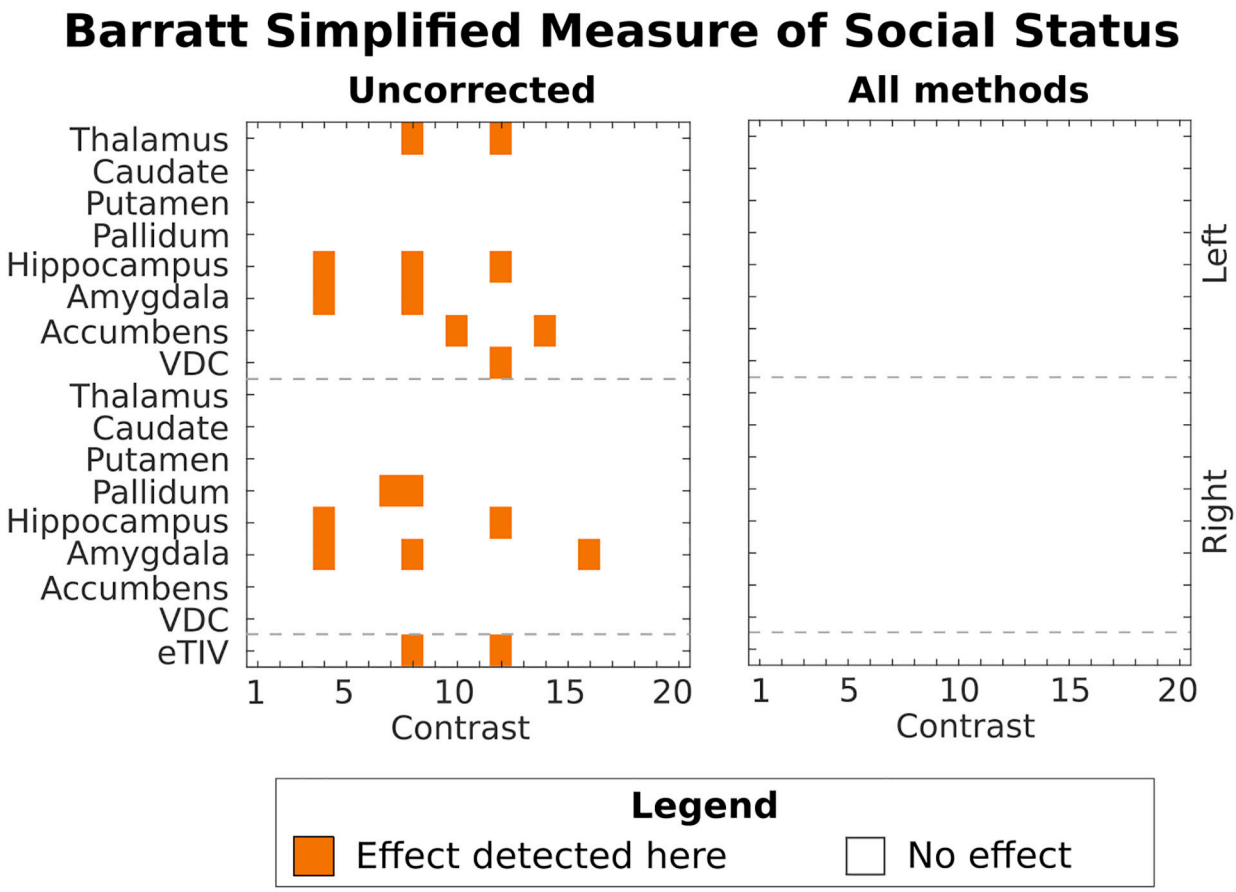

Fig. 5.

Effects in both hemispheres per contrast detected when dividing the subjects into five groups using the percentiles from BSMSS and all 20 pairwise comparisons were considered (contrast set 2). Before the correction across contrasts, some effects were detected (left panel), but no effect survived the correction performed with any of evaluated methods.

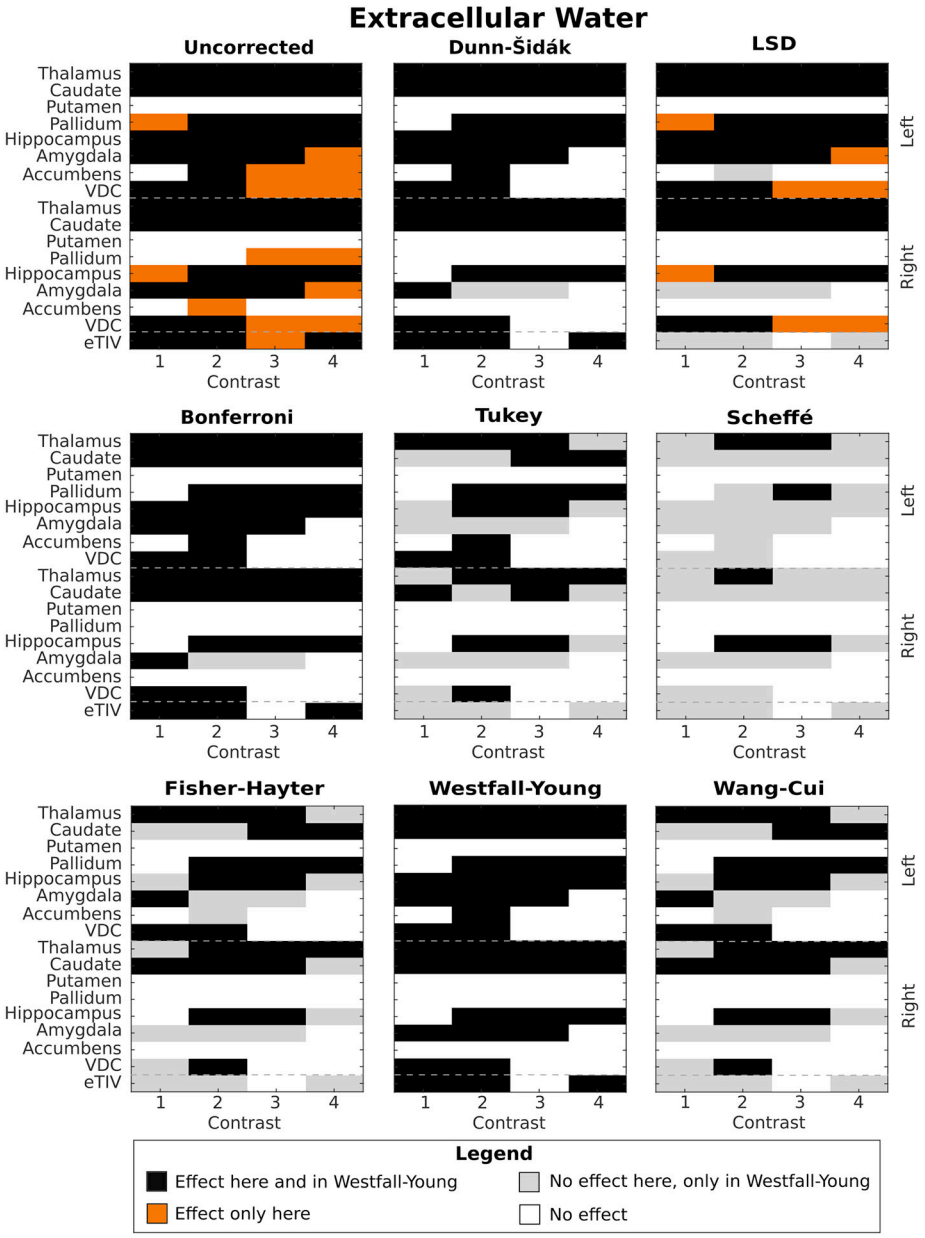

Fig. 6.

Effects in both hemispheres per contrast detected when dividing the subjects into five groups using the percentiles from extracellular water and testing contrast set 1, that is, testing whether subjects with lower extracellular water volume (group 1) have smaller cortical volume than subjects from the other 4 groups (with higher scores of extracellular water volume). Contrast 1 tests if group 1 has smaller volume than group 2, contrast 2 tests if group 1 has smaller volume than group 3, and so forth (see Fig. 1 for the list of contrasts tested).

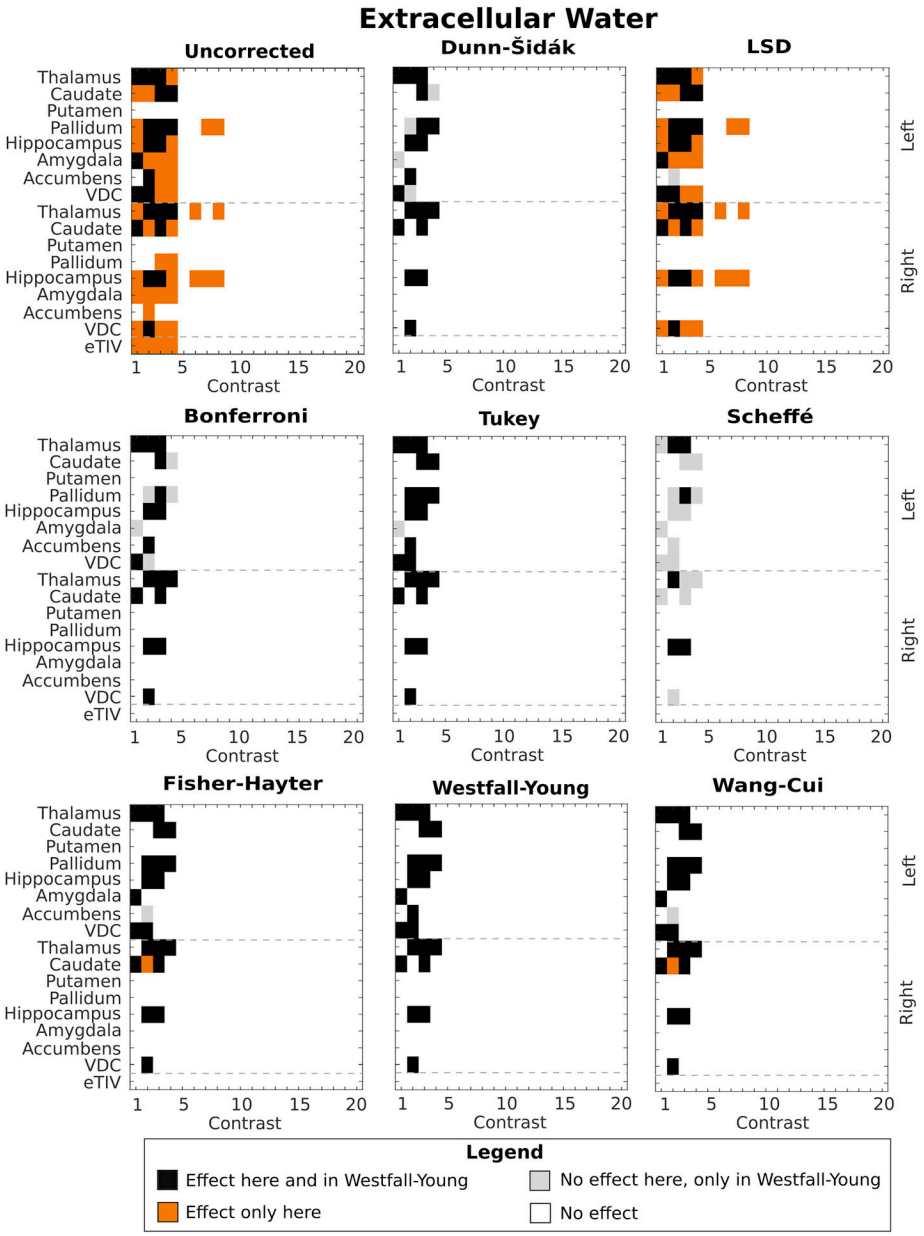

Fig. 7.

Effects in both hemispheres per contrast detected when dividing the subjects into five groups using the percentiles from extracellular water and all 20 pairwise comparisons were considered (contrast set 2).

Table 1

Summary comparison of methods for correction across contrasts.

Method	Global test p-value	Local tests' adjusted p-values
Dunn-Šidák	-	$1 - (1 - p_{\text{unc}})^{\tau}$
Bonferroni	-	$\tau \cdot p_{\text{unc}}$
Fisher's LSD	$1 - F_{\text{cdf}}(F; G - 1, \nu_2)$	$1 - t_{\text{cdf}}(t; \nu_2)$
Tukey	-	$1 - Q_{\text{cdf}}\left(t\sqrt{2}; G, \nu_2\right)$
Scheffé	-	$1 - F_{\text{cdf}}\left(\frac{\text{sign}(t)t^2}{G-1}; G - 1, \nu_2\right)$
Fisher-Hayter	$1 - F_{\text{cdf}}(F; G - 1, \nu_2)$	$1 - Q_{\text{cdf}}\left(t\sqrt{2}; G - 1, \nu_2\right)$
Westfall-Young	-	$1 - t^{\max}_{\text{cdf}}(t)$
Wang-Cui	$1 - F^{\text{wc}}_{\text{cdf}}(F; G - 1, \mathbf{n})$	$1 - Q_{\text{cdf}}\left(t\sqrt{2}; G - 1, \nu_2\right)$

Most of these methods are performed in a single, local step that produces the final, adjusted p-value; others require an initial global (*omnibus*) test, which, if significant, is followed by local tests (also known, in this context, as *post hoc* tests). If the initial global test is not significant, then the procedure stops, and $p_{\text{adj}} = 1$, that is, not significant. Fisher's LSD and Fisher–Hayter require an initial F-test with $(G - 1)$ and ν_2 degrees of freedom. Scheffé's adjusted p-values include the sign of t statistic, so that when a negative statistic is used, F_{cdf} evaluates as zero, and, thus, $p_{\text{adj}} = 1$ (for two tailed tests, the term sign(t) can be omitted). The global test in Wang–Cui use the F-statistic compared against a null distribution computed through Monte Carlo methods, and which depends on the number of groups and also on a vector **n** that contains the sizes of each group (thus explicitly accommodating unbalancedness). Note that the adjustment in Bonferroni can lead to p-values larger than 1, which reflects the fact that this method is an approximation (to Dunn–Šidák).

Table 2

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

Table 3

Applicability of methods to correct the multiplicity of tests across contrasts in common neuroimaging designs. Methods relying on the SRD, such as Tukey, Fisher–Hayter and Wang–Cui, can only be applied to tests of pairwise differences, whereas methods that do not assume that the data follow any distribution (i.e., Westfall–Young, Dunn–Šidák and Bonferroni) have many more applications. When there are dependencies among tests, the correction performed with Dunn–Šidák is conservative, but is exact if the tests are independent. Note that Fisher's LSD is only valid for experiments with 2 or 3 groups.

	Dunn–Šidák	Bonferroni	Fisher's LSD	Tukey	Scheffé	Fisher–Hayter	Westfall–Young	Wang–Cui
All pairwise comparisons between groups	•	•	✓	✓	•	✓	✓	✓
Pairwise comparisons among specific groups	•	•	✓	•	•	•	✓	•
All possible comparisons among any groups	•	•	✓	✗	✓	✗	✓	✗
Any comparisons among specific groups	•	•	✓	✗	•	✗	✓	✗
Simple regression (e.g. correlation)	•	•	✗	✗	✗	✗	✓	✗
Multiple regression (e.g., partial correlation)	•	•	✗	✗	✗	✗	✓	✗
Tests involving multiple models (designs)	•	•	✗	✗	✗	✗	✓	✗

✓ Applicable. • Applicable, possibly conservative. ✗ Not applicable.