



Puoliväli, T., Palva, S. and Palva, J. M. (2020) Influence of multiple hypothesis testing on reproducibility in neuroimaging research: a simulation study and Python-based software. *Journal of Neuroscience Methods*, 337, 108654.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

<http://eprints.gla.ac.uk/212261/>

Deposited on: 28 April 2020

Enlighten – Research publications by members of the University of Glasgow
<http://eprints.gla.ac.uk>

Influence of multiple hypothesis testing on reproducibility in neuroimaging research: A simulation study and Python-based software

Tuomas Puoliväli^{a,*}, Satu Palva^{a,b}, J. Matias Palva^{a,b,c}

^a*Helsinki Institute of Life Science (HiLIFE), Neuroscience Center, University of Helsinki, Finland*

^b*Centre for Cognitive Neuroimaging, Institute of Neuroscience and Psychology, University of Glasgow, UK*

^c*Department of Neuroscience and Biomedical Engineering, Aalto University, Finland*

Abstract

Background

Reproducibility of research findings has been recently questioned in many fields of science, including psychology and neurosciences. One factor influencing reproducibility is the simultaneous testing of multiple hypotheses, which entails false positive findings unless the analyzed p-values are carefully corrected. While this multiple testing problem is well known and studied, it continues to be both a theoretical and practical problem.

New Method

Here we assess reproducibility in simulated experiments in the context of multiple testing. We consider methods that control either the family-wise error rate (FWER) or false discovery rate (FDR), including techniques based on random field theory (RFT), cluster-mass based permutation testing, and adaptive FDR. Several classical methods are also considered. The performance of these methods is investigated under two different models.

Results

We found that permutation testing is the most powerful method among the considered approaches to multiple testing, and that grouping hypotheses based on prior knowledge can improve power. We also found that emphasizing primary and follow-up studies equally produced most reproducible outcomes.

Comparison with Existing Method(s)

We have extended the use of two-group and separate-classes models for analyzing reproducibility and provide a new open-source software “MultiPy” for multiple hypothesis testing.

Conclusions

Our simulations suggest that performing strict corrections for multiple testing is not sufficient to improve reproducibility of neuroimaging experiments. The methods are freely available as a Python toolkit “MultiPy” and we aim this study to help in improving statistical data analysis practices and to assist in conducting power and reproducibility analyses for new experiments.

*Correspondence should be addressed to Tuomas Puoliväli

Email addresses: tuomas.puolivali@ieee.org (Tuomas Puoliväli),
satu.palva@helsinki.fi (Satu Palva), matias.palva@helsinki.fi (J. Matias Palva)

Keywords: false discovery rate, family-wise error rate, multiple hypothesis testing, neurophysiological data, Python, reproducibility

1. Introduction

The reproducibility of published research has been recently called into question in many fields, including psychology and neurosciences (Button et al., 2013; Open Science Collaboration, 2015; Baker, 2016; Poldrack et al., 2017; Poldrack, 2019). Reproducibility is affected by many factors, one of which is the simultaneous testing of multiple hypotheses (Ioannidis, 2005). Testing multiple hypotheses entails false positive findings unless the analyzed p-values are appropriately corrected. Hence, there is demand for tools to evaluate multiple-testing data-analysis plans and to perform the required corrections using appropriate methods. Although numerous techniques exist for multiple hypothesis testing, it has remained incompletely understood how much the choice of method and relative emphasis on a primary study over its follow-up study, or vice versa, influences the reproducibility of the observations. Here we developed a Python-based software library that implements techniques for controlling the family-wise error rate (FWER) and the false discovery rate (FDR), and then developed a novel model-based approach to compare their relative performance in terms of their power, false positive rate, and reproducibility.

In this study, we assess differences between methods that control the FWER or FDR by comparing their performance using simulated data generated under two related spatial models. The first one is an extension of the classic two-group model that consists of distinct signal and noise regions (Bennett et al., 2009). The second one is an extension of Efron's (2008) separate-classes model; it combines two two-group models to enable representation of distributed effects. These models can be used to represent hypothetical effects in neurophysiological and neuroimaging data such as evoked or induced activity in an electroencephalography (EEG) or magnetoencephalography (MEG) time-frequency analysis, or relatively focal effects in functional or anatomical magnetic resonance imaging (MRI) data. In addition, these models allow performing numerical prospective power analyses to facilitate planning of new experiments, including the determination of sample and effect sizes that are required for observing true effects reproducibly. Since fundamental effects in systems-level neuroscience data are often spatially or temporally continuous (Penny & Friston, 2003; Heller et al., 2006; Chumbley et al., 2010), as well as distributed, these model are suitable for analyzing a wide range research questions. Moreover, only few previous studies to date have investigated properties of multiple testing procedures under models with two or more simultaneous distinct effects. Efron (2008) analyzed inference under a separate-classes model, which was motivated by diffusion tensor imaging data from dyslexic and control participants indicating distinct effects in anterior and posterior parts of the brain. Their results showed that since the underlying data had distinct effects and structures, performing two separate corrections was better than a single combined analysis (Efron, 2008).

In this study we advance a software called MultiPy, which is a Python-based open-source and freely-available toolkit for multiple hypothesis testing. Python has become in the recent

past the programming language of choice for many scientists across several disciplines. Accordingly, there already exists general-purpose packages for scientific computing (van der Walt et al, 2011), data visualization and manipulation (McKinney, 2010), and machine learning (Pedregosa et al, 2011) to name a few. In neuroimaging, specialized Python software has been developed for example for the pre-processing, analysis, and source reconstruction of EEG and MEG data (Gramfort et al, 2013; 2014). While these and other packages allow efficient analysis of single-subject neurophysiological and neuroimaging data, most provide only a limited number of options for correcting group-level results for multiple comparisons. Thus, our aim was to develop software that fills this gap.

Taken together, we here present a novel model-based approach for comparing multiple hypothesis testing methods, and quantify with simulated primary and follow-up experiments how large effect and sample sizes are needed to detect true effects reproducibly.

2. Methods

2.1. Definitions and notation

In statistical null hypothesis significance testing (NHST), there are four possible mutually exclusive outcomes as summarized in Table 1. Specifically, there are two desired outcomes, which are true positives and true negatives, and two undesired outcomes, which are false positives and false negatives. We will focus here on the number of incurred false positives \mathbf{V} while testing m hypotheses simultaneously under a specified critical level α , and how this count can be controlled using various procedures. We use the variables defined in Table 1 to denote the outcomes of NHST throughout the manuscript. Bold upper-case letters are used to denote random variables. Parts of the text also refer to variable π_0 , which is the proportion of true null hypotheses m_0 among m tests, and also to $\mathcal{N}(\mu, \sigma^2)$ which is the normal distribution with location and scale parameters μ and σ^2 and respectively. The variable \hat{x} denotes an estimate of the variable x . Since the simulated data in this study are always normally distributed, effect sizes are reported throughout the manuscript using Cohen’s d , which is the difference between means divided by standard deviation.

Table 1: Possible outcomes in NHST.

	Declared non-significant	Declared significant	Total
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	$m - m_0$
	$m - \mathbf{R}$	R	m

FWER is the probability of making one or more false positive conclusions while testing m hypotheses simultaneously. For independent tests, it can be described mathematically with the equation $\text{FWER} = \mathbb{P}(\mathbf{V} \geq 1) = 1 - (1 - \alpha)^m$, which is visualized in Figure 1A for

the three conventionally used critical levels 0.001, 0.01, and 0.05. In practice, it can often be advantageous to control the FDR instead of FWER (Benjamini & Hochberg, 1995), which is the expected proportion of discoveries that are false. Controlling the FDR allows exchanging a negligible number of false positives to increased power.

2.2. *Models for evaluating multiple hypothesis testing methods and performing numerical power and reproducibility analyses*

2.2.1. *Spatial two-group model for comparing FWER and FDR controlling methods*

Fundamental effects in systems-level neuroscience data are often spatially or temporally continuous (Penny & Friston, 2003; Heller et al., 2006; Chumbley et al., 2010). To compare different methods for controlling the FWER and FDR under these circumstances, we first used the spatial two-group model suggested by Bennett and co-authors (2009). This model consists of a $n_v \times n_v$ variable two-dimensional grid with a $n_s \times n_s$ variable signal region in the middle. For each location on the grid, there are N samples in each of two groups denoted by A and B . The samples of the groups A and B are distributed as $\mathcal{N}(0, 1)$ and $\mathcal{N}(\Delta, 1)$ respectively in the signal region; the parameter controls the effect size (Cohen’s d). In all other locations, the samples of both groups are distributed as $\mathcal{N}(0, 1)$ to model random noise. In other words, there is a true effect at every location within the signal region, and no true effects elsewhere. This model was used for the numerical comparison of the different multiple testing procedures, as well as for performing prospective power and reproducibility analyses.

2.2.2. *Spatial separate-classes model for comparing FWER and FDR controlling methods when the true effects are distributed*

In addition to being continuous in space, time, or frequency, fundamental effects in neuroscience data are also often distributed across temporal, spatial, or spectral scales (Heller et al., 2006; Chumbley et al., 2010). To compare the relative performance of multiple testing methods numerically when the true effects are distributed, we developed a separate-classes model which is a combination of two separate two-group models similar to Efron’s earlier study (2008) but with the spatial structure suggested by Bennett and co-authors (2009). In the model, there are now groups A and B in the first signal region and groups C and D in the second region. Similar to the two-group model, all samples outside of the signal regions are distributed as $\mathcal{N}(0, 1)$ to model random noise. Within the signal regions, samples in groups A and C are distributed as $\mathcal{N}(0, 1)$, and samples in groups B and D as $\mathcal{N}(\Delta_1, 1)$ and $\mathcal{N}(\Delta_2, 1)$ respectively. This model was used for testing the difference between performing two separate analyses and a single analysis, as well as for testing how differences in the two effect sizes Δ_1 and Δ_2 influence the multiple testing results. The model is visualized schematically in Figure 1C.

2.3. *Multiple hypothesis testing methods*

2.3.1. *Classic methods for controlling FWER*

Classic methods for controlling the FWER include the Bonferroni correction, Šidák’s correction (Šidák, 1967), Holm-Bonferroni procedure (Holm, 1979), and Hochberg’s procedure

(Hochberg, 1988). The first two of these methods are similar with respect to computing new critical levels $\alpha_{\text{bonferroni}} = \alpha/m$ and $\alpha_{\text{sidak}} = 1 - (1 - \alpha)^{\frac{1}{m}}$ by a direct adjustment of the original level with the number of performed comparisons. In turn, the Holm-Bonferroni and Hochberg procedures apply the exact same threshold $\alpha/(m - k + 1)$ while correcting the k th ascendingly sorted p-value. The distinction between these two procedures is that they process the p-values in opposite orders and are hence categorized as step-down and step-up procedures.

2.3.2. Benjamini-Hochberg FDR method

FDR controls the expected proportion of discoveries that are false, that is $\mathbb{P}(\mathbf{R} > 0) \cdot \mathbb{E}(\mathbf{V}/\mathbf{R} | \mathbf{R} > 0)$ (Benjamini & Hochberg, 1995). Controlling the FDR implies willingness to accept a small fraction of false positives among the tests that are declared significant in exchange for improved power. The Benjamini-Hochberg FDR procedure is one of the most widely used FDR procedures, and can be understood graphically: p-values are sorted into ascending order and the hypotheses corresponding to the p-values that are below the line $y = \alpha/m + 0$ are rejected until the first p-value crossing the line is seen (see Figure 1B).

2.3.3. Adaptive FDR methods

The Benjamini-Hochberg FDR method assumes that $\pi_0 = 1$ which can make it too conservative when the number of tested hypotheses is large (Storey & Tibshirani, 2003). Therefore, modern adaptive methods such as the Storey-Tibshirani q-value method (Storey & Tibshirani, 2003) and the two-stage procedure by Benjamini and colleagues (2006) attempt to estimate π_0 as part of the correction procedure. While a successful estimation of π_0 can allow greater power than the Benjamini-Hochberg method, there is a possible caveat: under some circumstances, applying the adaptive procedures may result in *more* significant discoveries than there were in the original uncorrected data (Reiss et al, 2012). In the study by Reiss and colleagues (2012), this kind of paradoxical results were found to occur in an MRI study with spatially widespread effects. The q-value method was found to be more vulnerable than the two-stage procedure (Reiss et al., 2012).

2.3.4. Permutation testing

In comparison to the other approaches that control the FWER, permutation tests provide a non-parametric but more computationally expensive alternative. Instead of manipulating a set of p-values, they process the data under analysis directly. Permutation testing proceeds in two stages. In the first stage, the null hypothesis, exchangeability of observations under the null hypothesis, and the test statistic are specified (Nichols & Holmes, 2001). In the second stage, the permutation distribution is built by repeatedly relabeling the observations and computing the corresponding test statistics, which then allows calculating the significance of the correct labeling (Nichols & Holmes, 2001). Here, we performed the permutation tests using the procedure described by Maris and Oostenveld (2007) with a cluster-mass test statistic (Bullmore et al., 1999). However, in contrast to the original algorithm, we transformed the final p-values into their upper bounds using the method suggested by Phipson and Smyth (2010) to avoid problems with zero p-values. In the original procedure described by Maris

and Oostenveld (2007), obtaining zero p-values is possible when all permutation test statistics are less extreme than the observed one. The reason is that the p-value was defined as $p = b/m$ where b is the number of times the permutation test statistic is equal or more extreme than the observed test statistic and m is the number of permutations. Therefore, the p-value could be underestimated by approximately $1/m$ when b is zero (Phipson & Smyth, 2010). However, this issue is avoided by defining the p-value instead as the upper bound $p_u = (b + 1)/(m + 1)$ which is always strictly positive (Phipson & Smyth, 2010).

2.3.5. *Random field theory*

Random field theory (RFT) is an approach for controlling the FWER that has been widely used in the analysis of functional and structural MRI and positron emission tomography (PET) data (Worsley et al, 1992; Friston et al, 1994, Worsley et al, 1996), especially as part of the statistical parametric mapping (SPM) software (Frackowiak, 1997; Ashburner, 2012). Previous studies have also explored its applicability for the analysis of source-reconstructed EEG and MEG signals (Carbonell et al, 2004; Pantazis et al, 2005), and its strengths and weaknesses over permutation testing have been recently explored and discussed in depth (Eklund et al, 2016). Application of RFT to neuroimaging data proceeds typically in three subsequent steps. First, the smoothness of the analyzed statistical map is estimated, which gives its resolution element or resel count. Here, we approximated the number of resels with the analyzed area divided by the squared full-width-at-half-maximum (FWHM) of the applied Gaussian smoothing kernel. Second, the expected Euler characteristic is computed for a range of thresholds, which can be informally defined as the number of distinct continuous regions that survive the thresholding. For two-dimensional data, the expected Euler characteristic is given by the equation $\mathbb{E}[EC] = R(4 \log_e 2)(2\pi)^{-3/2} \exp(-\frac{1}{2}Z_t^2)$ where Z_t is the z-score threshold (Worsley et al., 1992). Finally, the empirical data is compared to the theoretical expectation for declaring statistically significant differences.

2.4. *Data and data analyses*

2.4.1. *Comparison of multiple hypothesis testing methods using the spatial two-group and separate-classes models*

The spatial two-group and separate-classes models were used to generate simulated data for numerically comparing the different FWER and FDR controlling procedures. The simulation using the two-group model was constructed as follows: the size of the signal region was 30×30 variables centered in the middle of a 90×90 variable grid containing no other true effects (i.e. the proportion of true effects was approximately 11%), and there were 25 samples in each of the two groups. The sample size was selected based on meta-analyses that suggest the median sample size in psychological and neuroscience research to be in the range 22–28 (Button et al, 2013; Szucs et al, 2017; Turner et al, 2018). The simulation was repeated for each effect size Δ in the range 0.5–1.5 with 0.05 increments. In the separate-classes model both signal regions were 15×15 variables and the overall grid size was 45×90 variables (i.e. the proportion of true effects was again approximately 11%), and the sample size was set at 25 in each group. The simulation was repeated with the effect size combinations $\{(\Delta_1, \Delta_2) | \Delta_1, \Delta_2 \in [0.5 - 1.5]\}$ with 0.05 increments. The critical level was set at

$\alpha = 0.05$ in both types of simulations. Also, the data was smoothed using a Gaussian kernel with FWHM set to match the size of the signal region when the RFT based method was applied. Further, all permutation tests were performed with 100 randomly drawn permutations with the t-threshold set at $t = 1$. For the other methods, p-values were obtained using Student’s t-tests. Furthermore, if the spline-based null density estimator in the q-value method did not converge to the interval $[0, 1]$, the conservative choice of setting it as $\hat{\pi}_0 = 1$ was made. Finally, the performance of each method was quantified as the number of rejected null hypotheses for which the alternative hypothesis was true (i.e. power). The main steps of generating data using the two-group model and evaluating the performance of a given multiple testing method are also visualized as flowchart in the supplementary figure S1.

2.4.2. *Reproducibility in the two-group model*

To estimate effect sizes needed to observe true effects reproducibly in the two-group model, we performed simulated primary and follow-up experiments. The simulations were constructed as follows: the critical level was set at $\alpha = 0.05$, the sample size at $N = 25$ in both of the two groups, the signal region was 30×30 variables in the middle of a 90×90 variable grid containing no other effects, the results were averaged over 20 realizations, and the tested effect sizes ranged from 0.5 to 1.5 with 0.05 increments. Tests that were declared significant in the primary experiment were selected for further testing in the follow-up experiment. To decide which hypotheses were reproducible across the simulated experiments, we used the FWER replicability method (Benjamini & Heller, 2008; Benjamini et al., 2009; Bogomolov & Heller, 2013). Briefly, this method defines weights c_p and c_f for the primary and follow-up studies respectively, with the constraint $c_p + c_f = 1$, and performs the multiple testing corrections at the critical levels $c_p\alpha$ and $c_f\alpha$. Those tests that are declared significant in both the primary and the follow-up experiment are considered to be replicable and the rest are not. The simulations were performed by emphasizing the primary study at values 0.02, 0.5, and 0.98, which correspond to the critical levels 0.001, 0.025, and 0.049. Here, the selection of the lowest tested emphasis value was further motivated by a close correspondence to a recent suggestion of starting to use 0.005 as the new standard critical level in primary neuroscience experiments (Benjamin et al., 2018). The simulated data were analyzed using the correction methods introduced previously. The performance of each method was quantified with a “reproducibility rate” that is defined as $r = m_r/m_1$ where m_r is the number of true effects declared reproducible and $m_1 = m - m_0$ is the number of all true effects. See supplementary Figure S1 for an illustration of how to evaluate reproducibility using a given multiple testing method.

2.5. *Comparison of computational complexities*

In addition to comparing the performance of the different multiple testing methods in terms of produced true and false positives and negatives, we evaluated their computational complexities using the well-known big O notation. This method evaluates each algorithm’s asymptotic performance with respect to the sizes of the inputs, which are here the number of tests and possible other parameters, such as the size of the smoothing kernel for the RFT-based approach.

2.5.1. Comparison of controlling FDR and FWER in analyzing a large-sample aging dataset

To illustrate differences in controlling the FWER and FDR in empirical data, we performed mass-univariate analyses on magnetic resonance images extracted from the open-access series of imaging studies dataset (OASIS) (Marcus et al, 2007). These data contain T1-weighted images from demented and non-demented participants aged 18 to 96 years, which allows investigating how age and age-related diseases influence cortical and sub-cortical structures and morphology. The dataset has been analyzed in several previous studies, which have found for example that total brain volume decreases with age and socioeconomic status (Fotenos et al, 2008), and that the shape of corpus callosum predicts progression of dementia (Ardekani et al, 2014). The dataset has also been used in methodological studies for benchmarking different data analysis approaches and testing new classification algorithms (Khajehnejad et al, 2017; Alam et al, 2017).

We performed the data analysis in three main steps. First, we selected to our analyses the healthy individuals to keep the subsequent statistical testing as simple as possible. Second, we extracted vertex-level cortical thickness estimates, computed using the FreeSurfer image analysis suite (Fischl & Dale, 2000). The data were spatially smoothed using a Gaussian 10 mm full-width at half-maximum kernel. Third, a linear regression was performed at each individual vertex between cortical thickness and age, yielding a large univariate statistical map. Correction for multiple comparisons was performed using Sidak's procedure, which controls the FWER, and the Benjamini-Hochberg procedure, which controls the FDR.

2.6. Software implementation

MultiPy has been written in the programming language Python from the Python Software Foundation (<http://www.python.org>) (version 2.7.14). The implementation partly builds on previously published Python software packages, including Scikit-image (van der Walt et al., 2014) (version 0.13.0), SciPy (Oliphant, 2007) (0.17.0), NumPy (van der Walt et al., 2011) (1.10.2), Seaborn (0.8.0), and Matplotlib (Hunter, 2007) (2.1.0). The software is documented and available in <http://github.com/puolival/multiply>, and published under a permissive open-source license to facilitate further development and integration to other software and data analysis pipelines.

3. Results

3.1. Permutation testing is the most powerful method for analyzing data generated using the two-group and separate-classes models

The two-group and separate-classes models were used to simulate data at different sample and effect sizes to compare the relative performance of different multiple testing methods. The simulations were repeated for several times to evaluate the average power of each compared method at each sample and effect size. Permutation testing was the most powerful method for analyzing data generated using both the two-group model and the separate-classes model. Methods based on controlling the FDR produced intermediate results, and the least numbers of true positive effects were detected using the other techniques that control the FWER. The differences in power between the different methods varied as a function of effect size. These results are visualized in Figures 2 and 3.

3.2. *Placing an identical emphasis on the primary and follow-up studies produces optimal results*

To evaluate what is the optimal multiple testing strategy between a primary study and a follow-up study, we simulated data using the two-group model at different sample and effect sizes and varied the relative emphasis placed on the primary study over its follow-up study as well as the used correction method. This process was repeated for several times to evaluate the rate of reproducibility of true effects. The performance difference between the partial conjunction and FWER replicability methods was dependent on the emphasis placed on the primary study in comparison to the follow-up study, assuming a conservative correction for multiple testing was carried out. When the primary study received larger emphasis than the follow-up study, the partial conjunction method outperformed the FWER replicability method. In contrast, when the follow-up study received higher emphasis than the primary study, the situation was reversed so that the FWER replicability method produced the better results. These differences varied in magnitude as a function of the effect size (Figure 4A). When the FDR was controlled, optimal results were obtained when the primary and follow-up studies were emphasized equally. This result was obtained repeatedly even if the parameters of the simulations were changed; the same was true for a range of null densities, as well as for different sample sizes. However, the magnitude of the difference was again dependent on the effect size. To further probe the consistency of the results, the analysis was repeated with different correction methods, yielding consistent outcomes. These results are visualized in Figures 4B–C and 5. Another result obtained using the same set of simulations is that sample sizes larger than those typically used in neuroscience studies were needed to observe true effects reproducibly. A total of 40–50 samples per group or condition (i.e. a total of 80–100 samples) was needed to obtain a rate of reproducibility higher than 80% when the primary and follow-up studies were emphasized equally or the primary study was emphasized more than the follow-up study.

3.3. *Computational costs increase with the performance of the multiple testing method*

To assess whether improved statistical results are related to increased computational requirements, we assessed the correction methods' running times using the big O notation.

Bonferroni's and Sidak's methods compute new critical levels based on the original critical level and the number of performed tests. The methods then proceed to compare the p-value of each test to the new critical level, and therefore have linear running times with respect to the number of performed tests. In contrast, the Hochberg, Holm-Bonferroni, and Benjamini-Hochberg methods all have running times $O(n \log n)$. These three methods sort the tests based on their p-values into ascending order, and then compare each p-value to a threshold based on its position in the ordered list. Since sorting can be accomplished in $O(n \log n)$ time, using for example the merge sort algorithm, and the other operations are less demanding, the overall running time is $O(n \log n)$. The two-stage procedure also has the running time $O(n \log n)$ but has a more complicated structure: in the first step, a routine that is algorithmically similar to the Benjamini-Hochberg method is performed, followed by a computation of the number of significant tests, which is a linear operation, and finally the last step is again algorithmically similar to the Benjamini-Hochberg method, so the overall

complexity is again governed by the sorting operation. In turn, the q-value method is more computationally expensive than the previous methods. It can be implemented in four main stages. First, the p-values are sorted into ascending order. Second, the number of p-values larger than k is computed for a range of values (e.g. $k = 0, 0.01, 0.02, \dots, 0.95$), making the complexity of this step $O(nm)$ where m is the number of thresholds, assuming a simple implementation based on nested repetition structure. Third, a spline function is fit to the data obtained in the second step, which can be accomplished in linear time. The final step is to decide which p-values are significant, which is also a linear-time operation. The overall complexity is thus $O(n(m + \log n))$, although it does seem that the $O(nm)$ part could be eliminated via dynamic programming techniques. The running time of the RFT-based approach depends mostly on the smoothing step, which is $O(nk^2)$ where k is the radius of the smoothing kernel. Permutation testing is even slower, since for p number of permutations, n number of tests, and m number of clusters must be handled.

In summary, and comparing these results to those reported in section 3.1, it appears that the statistical performance of a multiple testing method is indeed related to its computational complexity. The trend that can be observed is that statistically superior results are obtained using more computationally expensive methods. In particular, permutation testing was the most computationally expensive method, but also produced most powerful results.

3.4. Comparison of controlling FDR and FWER in analyzing a large-sample aging dataset

Using the OASIS data, we estimated the correlation of age with cortical thickness and investigated how multiple testing based on controlling the FWER and FDR affected the results. A total of 55% of vertices showed a statistically significant effect of aging on cortical thickness when the p-values were corrected for multiple testing using Sidak’s correction. In contrast, when the correction was based on the Benjamini-Hochberg FDR procedure, the number of significant vertices was 82%, suggesting a more widespread effect. The uncorrected critical level was set at $\alpha = 0.05$ in both analyses. These results are illustrated in supplementary Figure S2. Qualitatively, when the FWER is controlled using Sidak’s correction, it would appear that aging selectively influences parts of the cortex. In contrast, the FDR based result suggests that almost the whole cortex is affected. This result is in line with previous studies that compared FWER and FDR based approaches (Durnez et al., 2014) and reproduces the well-known observation that strict correction methods tend to miss likely true effects due to diminished power on even large-sample datasets. This result can be reproduced using MultiPy with example code available in <https://github.com/puolival/multiply/tree/master/multiply/example>.

4. Discussion

In this study, we quantified the performance of several multiple hypothesis testing methods in the context of the reproducibility. To this end, we developed an open-source Python software, which implements classic and advanced techniques for controlling the FWER and FDR, simulating data under two different models, and performing reproducibility and power analyses. The software includes both parametric and non-parametric correction techniques,

which were compared numerically with simulations performed using the two models. These models, namely the two-group model and the separate-classes model, allow capturing features typically observed in neurophysiological and neuroimaging data, but which are nevertheless as simple as possible. We obtained results that advance three points. First, the simulations showed that permutation testing is the most powerful approach for analyzing data generated under these models among the included methods, with the caveat of having a high computational cost. Second, we found that incorporating prior knowledge to the testing process in the form of grouping hypotheses (i.e. performing separate analyses for distinct structures) can improve power. Third, we found that the combination of low power and testing of multiple hypotheses leads to poor reproducibility. This implies that increasing sample sizes in neuroscience experiments from the current levels is essential for reproducible discoveries of true effects. Moreover, these findings show that the recent suggestion of using 0.005 as the new threshold of statistical significance, instead of 0.05, in primary inferential statistics would not be optimal for reproducibility.

4.1. Permutation testing is the most powerful method for analyzing data generated using the spatial two-group model

For the data simulated using the two-group model, the classic methods that control the FWER (i.e. the Bonferroni, Šidák, Holm-Bonferroni, and Hochberg’s methods) produced either identical or very similar results. In contrast, the Benjamini-Hochberg FDR procedure and the adaptive FDR methods detected expectedly more true positives while incurring a small fraction of false positives. Since the adaptive FDR procedures yielded similar results, it makes the two-stage procedure favorable over the q-value method due to its more stable null density estimator (Reiss et al, 2012). The permutation test outperformed all other approaches by a large margin, and the performance of the RFT based method was between the classic methods that control the FWER and the methods that control the FDR. Notably, while permutation testing was the most powerful approach here, it produced false positives that concentrated near the signal region boundaries. Indeed, one should be aware of its possible limitations in accurately establishing effect locations or latencies when applied to neuroimaging data (Sassenhagen & Draschkow, 2018). The power curves for representative methods for each class of correction methods are visualized in Figure 2H.

4.2. Incorporating prior knowledge to the multiple hypothesis testing process by grouping hypotheses of distinct structures increases power

The separate-classes model allows two signal regions with distinct effect sizes, which makes it possible to test whether prior information about the underlying data structure can be used to improve results obtained from the multiple testing process. Indeed, in the separate-classes model, detecting a second effect is more difficult if a single combined analysis is performed, in comparison to performing two separate analyses motivated by the available prior information. This result suggests that researchers should place more focus on considering how to combine neuroscience data for statistical analyses. Many studies have indeed performed statistical analyses separately for each frequency and signal domain (Hirvonen et al., 2016; Kulashekhar et al., 2016; Siebenhühner et al., 2016; Lobier et al., 2018; Hirvonen

et al., 2017; Zhigalov et al., 2015; Hirvonen et al., 2018) and found valuable information from the spectro-temporal patterns of evoked and event-related neuronal dynamics. Leveraging available information, of a particular task for example, could further enhance sensitivity of these types of analyses. Similar arguments can be put forward for functional and structural MRI data: consider for example the hemispheric lateralization of language, speech, and auditory functions.

4.3. Plan experiments for reproducibility

The recommended practice for sample size selection is to perform a priori power calculations. However, multiple testing has remained a controversial topic with no trivial solutions. In addition, reproducibility is typically not quantified in power calculations. Hence, it has remained largely unknown how well-powered studies are needed to observe true effects reproducibly in the context of multiple testing. The issue is important to consider, since multiple testing occurs in most neurophysiological and neuroimaging research, and since the average study has an estimated power of only 8–31% (Button et al., 2013). Indeed, some scientific journals have already started to urge researchers to plan for reproducibility before conducting their experiments (Editorial, *Nature Biomedical Engineering*, 2018; Editorial, *Nature Communications*, 2018). To draw attention to this issue, we performed simulated primary and follow-up experiments using the spatial two-group model to highlight what the reproducibility rates might be at effect sizes typically observed in neuroscience research. Briefly, the results indicate that in this model the choice of multiple testing method and emphasis on the primary study have large influence on observing true effects reproducibly (Figure 4), and therefore we suggest similar analyses to be carried out while planning new primary and replication experiments. Overall, observing true effects reproducibly was difficult at small and moderate effect sizes with sample sizes that are typically used in psychological and neuroscience research. Our results also suggest that performing a too strict correction in the primary study is not optimal due to a substantially increased number of false negatives, and that performing a too loose correction in the primary study is not optimal either, due to an increased number of false positive outcomes that must be subsequently verified at the replication stage. Instead, the optimal result was obtained by giving an equal emphasis for the primary and follow-up studies (Figure 4B); the same conclusion was reached using both the FWER replicability and the FDR r -value methods. Importantly, this finding is partly at odds with the recent suggestion of starting to use 0.005 as the new standard threshold of statistical significance in primary neuroimaging and neuroscience experiments as the solution to reproducibility problems (Benjamin et al., 2018). Indeed, it could be advantageous to consider how to combine data from the primary and follow-up experiments as optimally as possible. Ideally, researchers could first choose their desired critical level, and then perform prospective reproducibility analyses to plan their experiments. In practice, this would mean simulating primary and follow-up data for many times, using a model and parameters that are as similar as possible to the expected empirical effect, computing reproducibility for different sample sizes, and making the final selection with respect to desired level of confidence. Attaining this ideal, of course, can be a difficult task, especially when there is little prior information available.

4.4. MultiPy enables the evaluation of data analysis pipelines and assessment of reproducibility of planned experiments

Here we implemented an array of parametric and non-parametric multiple testing methods as well as two models for their numerical evaluation in a single open-source toolkit. Therefore, it is possible to use the provided software as a platform for developing new multiple testing methods, since their performance can be evaluated directly against other existing state-of-the-art techniques. Indeed, the two-group models have been extensively used for this purpose in the existing literature; see for example Heller & Rosset (2019) for recent work on optimal control of the FDR in such a model. The software can be also used to validate existing custom neuroimaging data analysis pipelines, which are presently abundant among the different laboratories (Carp, 2012), by simulating data at chosen effect and sample sizes and testing whether the FWER or FDR is controlled. Moreover, the software enables its users to perform prospective numerical power analyses, and importantly, evaluate the reproducibility of their planned experiments when the two-group or separate-classes models are good approximations to the effects seen in the empirical data. In addition, the developed software can be used in future studies as a platform for developing new multiple testing techniques, since it allows evaluating their performance directly against other existing state-of-the-art techniques. It also allows researchers to validate their custom data analysis pipelines using the two-group and separate-classes models; data can be simulated at chosen effect and sample sizes and then analyzed similar to real empirical data to test whether the FWER or FDR is controlled.

4.5. How to choose between controlling the FWER and FDR

How should one then decide whether to control the FWER or the FDR? Generally, a strict control of false positives is often gained at the expense of false negatives, which may lead to important discoveries remaining unnoticed, especially when the number of tested hypotheses is large (see e.g. section 3.4). Therefore, such tight control is best justified when any false positive discoveries are expensive due to ethical, financial, time, or other constraints. For example, one might wish to conduct a detailed confirmatory follow-up study corresponding to each discovery in the primary study; the use of animals or high research costs could make false leads too expensive. On the opposite side of the spectrum, a looser control is also a sound choice for many types of data analyses, since often there are no theories or models concerning each individual test. Here, consider for example a typical EEG or MEG induced-response study with two conditions and a comparison of the responses in the time-frequency domain as the main data analysis. The statistical testing would likely involve some hundreds or thousands of tests depending on the length of the analysis time-window and frequency resolution (or spacing of frequencies in the case of wavelet-based methods), and there might not be a priori expectations for all of the time-frequency tuples that are tested. The objective here would be to look for consistent activity patterns over several frequencies and time points so that a small number of false positives would not compromise the results' interpretation. Overall, the optimal choice between controlling the FWER or the FDR seems to mostly depend on the cost of false positives in the considered study. Since the FWER can

be controlled using permutation while simultaneously preserving high power, we suggest its use when the associated computational burden is not a major limitation.

4.6. Future directions

In the future, the approach advanced here can be expanded into several new directions. For example, topics that have been omitted here include multiple testing of multimodal (Winkler et al., 2016) and network data. Further, there are extensions to some of the implemented methods, such as FDR based RFT approaches (Chumbley et al., 2010) and covariate-adjusted FDR techniques (Genovese et al., 2006; Ignatiadis et al., 2016; Basu et al., 2018). Furthermore, there are many more specialized techniques such as those developed for multiple testing of neuronal synchrony estimates (Maris et al., 2007; Singh & Phillips, 2010; Singh et al., 2011; Scott et al., 2015) that could be implemented as part of the software. In addition to these and other frequentist approaches, the software could be also developed to allow the use of Bayesian techniques, which are currently not widely supported by major neuroscience software packages. The software can be also extended to include more data-generating models, which would help researchers to perform more diverse simulation-based power and reproducibility analyses. Finally, the software could be made directly compatible with data structures of existing Python-based neuroimaging analysis tools, which would facilitate its integration into existing data analysis pipelines.

4.7. Limitations

In this study, we compared the performance of different multiple testing methods in the context of reproducibility, assuming that the two-group and separate-classes models are good proxies to typical effects observed in neuroscience and neuroimaging experiments. While these models appear to be suitable for representing simple effects for example in EEG or MEG time-frequency analyses that compare two different groups, conditions, repetitions, interventions, or similar experimental manipulations, as well as differences in two-dimensional functional or structural imaging data, we acknowledge that they are not representative of all types of neuroscience data. Indeed, complex experiments displaying widely distributed simultaneous effects would already require using different kinds of models. In addition, we have limited the simulations to cases where the p-values are “well-formed”, meaning that the assumptions of the performed statistical tests were not violated. The simulated data are also free of outliers and other artefacts, which are often a major concern in analyzing noisy real datasets. Therefore, the results reported in this manuscript should be considered optimistic, and future studies should explore how and to what degree these and other similar issues affect reproducibility in different kinds of multiple testing scenarios.

Acknowledgements

Authors thank Academy of Finland (grants 266745 and 281414) and Instrumentarium Science Foundation for funding, and Muriel Lobier, Sami Karadeniz, and Hamed Haque for helpful comments on an earlier version of the manuscript.

Conflicts of interest

The authors declare no competing financial or other interests.

Data availability statement

The data that support the findings of this study can be reproduced using the developed open-source software “MultiPy”, which is available online in <https://github.com/puolival/multiply>.

References

Alam S, Kwon GR, Kim JI, Park CS (2017): Twin SVM-based classification of Alzheimer’s disease using complex dual-tree wavelet principal coefficients and LDA. *Journal of Healthcare Engineering* 2017:8750506.

Ardekani BA, Bachman AH, Figarsky K, Sidtis JJ (2014): Corpus callosum shape changes in early Alzheimer’s disease: an MRI study using the OASIS brain database. *Brain Structure and Function* 219(1):343–352.

Ashburner J (2012): SPM: A history. *NeuroImage* 62–248(2):791–800. DOI: 10.1016/j.neuroimage.2011.10.025

Baker M (2016): Is there a reproducibility crisis? *Nature* 533:452–454. DOI: 10.1038/533452a

Basu P, Cai TT, Das K, Sun W (2018): Weighted false discovery rate control in large-scale multiple testing. *Journal of the American Statistical Association* 113(523):1172–1183.

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. (2018): Redefine statistical significance. *Nature Human Behavior* 2:6–10. DOI: 10.1038/s41562-017-0189-z

Benjamini Y, Heller R (2008): Screening for partial conjunction hypotheses. *Biometrics* 64:1215–1222. DOI: 10.1111/j.1541-0420.2007.00984.x

Benjamini Y, Heller R, Yekutieli D (2009): Selective inference in complex research. *Philosophical Transactions of the Royal Society A* 367:4255–4271. DOI: 10.1098/rsta.2009.0127

Benjamini Y, Hochberg Y (1995): Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1):289–300.

Benjamini Y, Krieger AM, Yekutieli D (2006): Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* 93(3):491–507. DOI: 10.1093/biomet/93.3.491

Bennett CM, Wolford GL, Miller MB (2009): The principled control of false positives in neuroimaging studies. *NeuroImage* 63(4):1833–1840. DOI: 10.1093/scan/nsp053

Bogomolov M, Heller R (2013): Discovering findings that replicate from a primary study of high dimension to a follow-up study. *Journal of the American Statistical Association* 108(504):1480–1492. DOI: 10.1080/01621459.2013.829002

Bullmore E, Suckling J, Overmeyer S, Rabe-Hesketh S, Taylor E, Brammer M (1999): Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Transactions on Medical Imaging* 18:32–42.

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR (2013): Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14:365–376. DOI: 10.1038/nrn3475

Carp J (2012): The secret lives of experiments: methods reporting in the fMRI literature. *NeuroImage* 63(1):289–300. DOI: 10.1016/j.neuroimage.2012.07.004

Carbonell F, Galán L, Valdés S, Worsley K, Biscay RJ, Díaz-Comas L, Bobes MA, Parra M (2004): Random field-union intersection tests for EEG/MEG imaging. *NeuroImage* 22(1):268–276. DOI: 10.1016/j.neuroimage.2004.01.020

Chumbley J, Worsley K, Flandin G, Friston K (2010): Topological FDR for neuroimaging. *NeuroImage* 49(4):3057–3064. DOI: 10.1016/j.neuroimage.2009.10.090

Design preclinical studies for reproducibility (2018). *Nature Biomedical Engineering* 2:789–790. DOI: 10.1038/s41551-018-0322-y

Durnez J, Roels SP, Moerkerke B (2014): Multiple testing in fMRI: an empirical case study on the balance between sensitivity, specificity, and stability. *Biometrical Journal* 56(4):649–661.

Efron B (2008): Simultaneous inference: When should hypothesis testing problems be combined? *The Annals of Applied Statistics* 2(1):197–223. DOI: 10.1214/07-AOAS141

Eklund A, Nichols TE, Knutsson H (2016): Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *The Proceedings of the National Academy of Sciences of the United States of America* 113(33):7900–7905. DOI: 10.1073/pnas.1602413113

- Fischl B, Dale AM (2000): Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America* 97:11050–11055.
- Fotinos AF, Mintun MA, Snyder AZ, Morris JC, Buckner RL (2008): Brain volume decline in aging. *Archives of Neurology* 65(1):113–120.
- Frackowiak RSJ, Friston KJ, Frith CD, Dolan RJ, Mazziotta JC (1997): Human Brain Function. Academic Press, USA.
- Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, Evans AC (1994): Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping* 1(3):210–220. DOI: 10.1002/hbm.460010306
- Genovese CR, Roeder K, Wasserman L (2006): False discovery control with p-value weighting. *Biometrika* 93(3):509–524.
- Gramfort A, Luessi M, Larson E, Engemann D, Strohmeier D, Brodbeck C, Goj R, Jas M, Brooks T, Parkkonen L, Hämäläinen M (2013): MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience* 7:267. DOI: 10.3389/fnins.2013.00267
- Gramfort A, Luessi M, Larson E, Engemann D, Strohmeier D, Brodbeck C, Parkkonen L, Hämäläinen M (2014): MNE software for processing MEG and EEG data. *NeuroImage* 86:446–460. DOI: 10.1016/j.neuroimage.2013.10.027
- Gross J, Baillet S, Barnes GR, Henson RH, Hillebrand A, Jensen O, Jerbi K, Litvak V, Maess B, Oostenveld R, Parkkonen L, Taylor JR, van Wassenhove V, Wibral M, Schoffelen J-M (2013): Good practice for conducting and reporting MEG research. *NeuroImage* 65(100):349–363. DOI: 10.1016/j.neuroimage.2012.10.001
- Hayasaka S, Nichols TE (2003): Validating cluster size inference: Random field and permutation methods. *NeuroImage* 20(4):2343–2356. DOI: 10.1016/j.neuroimage.2003.08.003
- Heller R, Bogomolov M, Benjamini Y (2014): Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *The Proceedings of the National Academy of Sciences of the United States of America* 111(46):16262–16267. DOI: <https://doi.org/10.1073/pnas.1314814111>
- Heller R, Rosset S (2019): Optimal FDR control in the two-group model. *ArXiv* 1902.00892v1.
- Hirvonen J, Monto S, Wang SH, Palva JM, Palva S (2018): Dynamic large-scale network synchronization from perception to action. *Network Neuroscience* 2(4):442–463. DOI: 10.

1162/netn_a_00039

Hirvonen J, Palva S (2016): Cortical localization of phase and amplitude dynamics predicting access to somatosensory awareness. *Human Brain Mapping* 37(1):311–326. DOI: 10.1002/hbm.23033

Hirvonen J, Wibral M, Palva JM, Singer W, Uhlhaas P, Palva S (2017): Whole-brain source-reconstructed MEG-data reveal reduced long-range synchronization in chronic schizophrenia. *eNeuro* 17(4):5. DOI: 10.1523/eneuro.0338-17.2017

Hochberg Y (1988): A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75(4):800–802. DOI: 10.1093/biomet/75.4.800

Holm S (1979): A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2):65–70.

Hunter JD (2007): Matplotlib: A 2D graphics environment. *Computing in Science and Engineering* 9(3):90–95. DOI: 10.1109/MCSE.2007.55

Ignatiadis N, Klaus B, Zaugg JB, Huber W (2016): Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods* 13(7):577–580. DOI: 10.1038/nmeth.3885

Ioannidis JP (2005): Why most published research findings are false. *PLoS Medicine* 2(8):e124. DOI: 10.1371/journal.pmed.0020124

Khajehnejad M, Saatlou FH, Mohammadzade H (2017): Alzheimer’s disease early diagnosis using manifold-based semi-supervised learning. *Brain Sciences* 7(8):109.

Kulashekhar S, Pekkola J, Palva JM, Palva S (2016): The role of cortical beta oscillations in time estimation. *Human Brain Mapping* 37(9):3262–3281. DOI: 10.1002/hbm.23239

Lobier M, Palva JM, Palva S (2018): High-alpha band synchronization across frontal, parietal and visual cortex mediates behavioral and neuronal effects of visuospatial attention. *NeuroImage* 165:222–237. DOI: 10.1016/j.neuroimage.2017.10.044

Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL (2007): Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience* 19(9):1498–1507

Maris E, Oostenveld R (2007): Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods* 164(1):177–190. DOI: 10.1016/j.jneumeth.2007.03.024

Maris E, Schoffelen J-M, Fries P (2007): Nonparametric statistical testing of coherence differences. *Journal of Neuroscience Methods* 163(1):161–175.

McKinney W (2010): Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*. p. 51–56.

Nichols TE, Holmes AP (2001): Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping* 15:1–25. DOI: 10.1016/j.jneumeth.2007.02.011

Oliphant TE (2007): Python for scientific computing. *Computing in Science & Engineering* 9:10–20. DOI: 10.1109/MCSE.2007.58

Open Science Collaboration (2015): Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716. DOI: 10.1126/science.aac4716

Palva JM, Monto S, Kulashekhar S, Palva S (2010): Neuronal synchrony reveals working memory networks and predicts individual memory capacity. *The Proceedings of the National Academy of the United States of America* 107(16):7850–7855.

Pantazis D, Nichols TE, Baillet S, Leahy RM (2005): A comparison of random field theory and permutation methods for the statistical analysis of MEG data. *NeuroImage* 25:383–394. DOI: 10.1016/j.neuroimage.2004.09.040

Penny W, Friston K (2003): Mixtures of general linear models for functional neuroimaging. *IEEE Transactions on Medical Imaging* 22(4):504–514. DOI: 10.1109/TMI.2003.809140

Phipson B, Smyth GK (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:article39. DOI: 10.2202/1544-6115.1585

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011): Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, Nichols TE, Poline J-B, Vul E, Yarkoni T (2017): Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience* 18:115–126. DOI: 10.1038/nrn.2016.167

Poldrack RA (2019): The costs of reproducibility. *Neuron* 101:11–14. DOI: <https://doi.org/10.1016/j.neuron.2018.11.030>.

Reiss PT, Schwartzman A, Lu F, Huang L, Proal E (2012): Paradoxical results of adaptive false discovery rate procedures in neuroimaging studies. *NeuroImage* 63(4):1833–1840. DOI: [10.1016/j.neuroimage.2012.07.040](https://doi.org/10.1016/j.neuroimage.2012.07.040)

Reproducibility: let's get it right from the start (2018). *Nature Communications* 9:3716. DOI: [10.1038/s41467-018-06012-8](https://doi.org/10.1038/s41467-018-06012-8)

Sassenhagen J, Draschkow D (2019): Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology* e13335. DOI: [10.1111/psyp.13335](https://doi.org/10.1111/psyp.13335)

Scott JG, Kelly RC, Smith MA, Zhou P, Kass RE (2015): False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association* 110(510):459–471

Šidák Z (1967): Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* 62(318):626–633. DOI: [10.2307/2283989](https://doi.org/10.2307/2283989)

Siebenhühner F, Wang SH, Palva JM, Palva S (2016): Cross-frequency synchronization connects networks of fast and slow oscillations during visual working memory maintenance. *Elife* e13451. DOI: [10.7554/elife.13451](https://doi.org/10.7554/elife.13451)

Singh AK, Asoh H, Phillips S (2011): Optimal detection of functional connectivity from high-dimensional EEG synchrony data. *NeuroImage* 58(1):148–156. DOI: [10.1016/j.neuroimage.2011.05.082](https://doi.org/10.1016/j.neuroimage.2011.05.082)

Singh AK, Phillips S (2010): Hierarchical control of false discovery rate for phase locking measures of EEG synchrony. *NeuroImage* 50(1):40–47. DOI: [10.1016/j.neuroimage.2009.12.030](https://doi.org/10.1016/j.neuroimage.2009.12.030)

Stodden V (2015): Reproducing statistical results. *Annual Review of Statistics and Its Applications* 2:1–19. DOI: [10.1146/annurev-statistics-010814-020127](https://doi.org/10.1146/annurev-statistics-010814-020127)

Storey JD, Tibshirani R (2003): Statistical significance for genomewide studies. *Proceedings of the National Academy of the United States of America* 100(16):9440–9445. DOI: [10.1073/pnas.1530509100](https://doi.org/10.1073/pnas.1530509100)

Szucs D, Ioannidis JPA (2017): Empirical assessment of published effect sizes and power in recent cognitive neuroscience and psychology literature. *PLoS Biology* 15(3):e2000797.

DOI: 10.1371/journal.pbio.2000797

Turner BO, Paul EJ, Miller MB, Barbey AK (2018): Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology* 1:62. DOI: 10.1038/s42003-018-0073-z

van der Walt S, Colbert SC, Varoquaux (2011): The NumPy array: A structure for efficient numerical computation. *Computing in Science Engineering* 13:22–30. DOI: 10.1109/MCSE.2011.37

van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, Gouillart E, Yu T, et al (2014): Scikit-image: Image processing in Python. *PeerJ* 2:e453. DOI: 10.7717/peerj.453

Winkler AM, Webster MA, Brooks JC, Tracey I, Smith SM, Nichols TE (2016): Non-parametric combination and related permutation tests for neuroimaging. *Human Brain Mapping* 37:1486–1511. DOI: 10.1002/hbm.23115

Worsley KJ, Evans AC, Marrett S, Neelin P (1992): A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow and Metabolism* 12:900–918. DOI: 10.1038/jcbfm.1992.127

Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ (1996): A unified statistical approach for determining significant voxels in images of cerebral activation. *Human Brain Mapping* 4:58–73. DOI: 10.1002/(SICI)1097-0193(1996)4:1<58::AID-HBM4>3.0.CO;2-O

Zhigalov A, Arnulfo G, Nobili L, Palva S, Palva JM (2015): Relationship of fast- and slow-timescale neuronal dynamics in human MEG and SEEG. *Journal of Neuroscience* 35(13):5385–5396. DOI: 10.1523/jneurosci.3880-14.2015

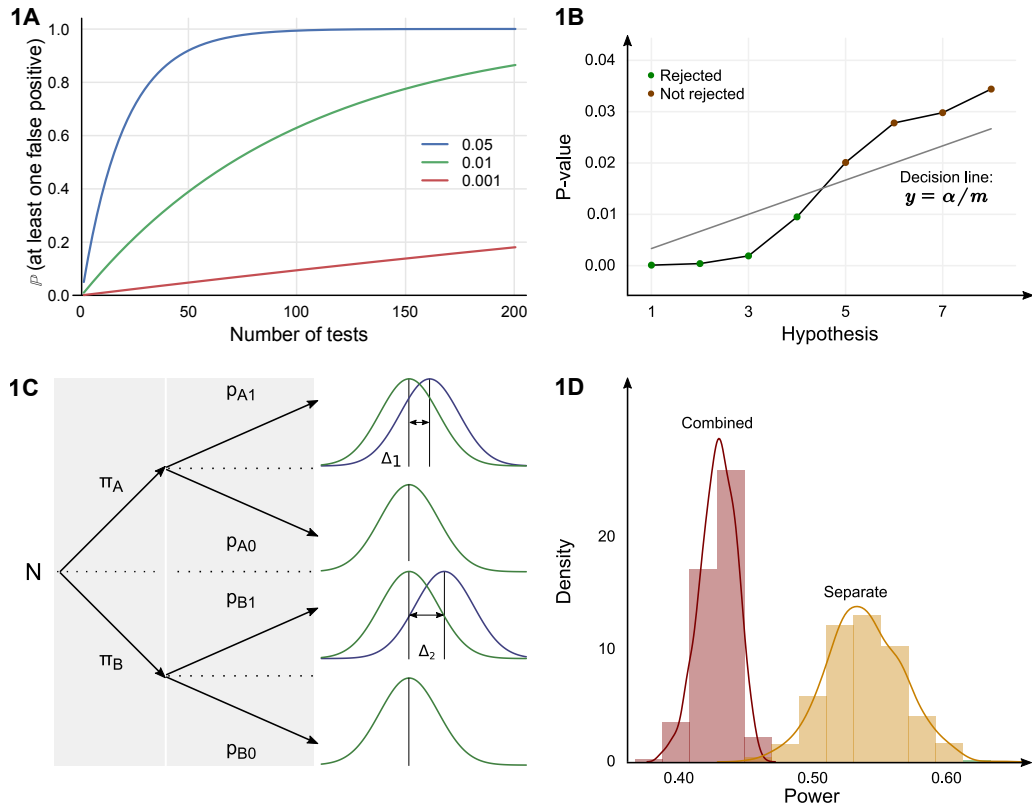


Figure 1: (A) Probability of observing at least one false positive outcome increases rapidly when more than some tens of independent tests are performed simultaneously and no adjustment for multiple testing is made. The blue, green, and red lines correspond to the three conventional critical levels 0.05, 0.01, and 0.001 respectively. (B) The Benjamini-Hochberg FDR procedure can be interpreted graphically as rejecting sorted p-values that fall below the line $y = \alpha/m + 0$ until the first non-significant p-value is seen. (C) Schematic of the structure of the separate-classes model. The arrows indicate how the sample is first divided into two separate classes A and B. In each class, only a proportion of the variables are associated with true underlying effects. The figure is partially adapted from Efron (2008). (D) The histograms and density estimates were constructed by simulating data under the separate-classes model for 1000 iterations with the effect sizes $\Delta_1 = 1.2$ and $\Delta_2 = 0.7$. Performing two separate analyses was more powerful than performing a single combined analysis when the two classes were known a priori.

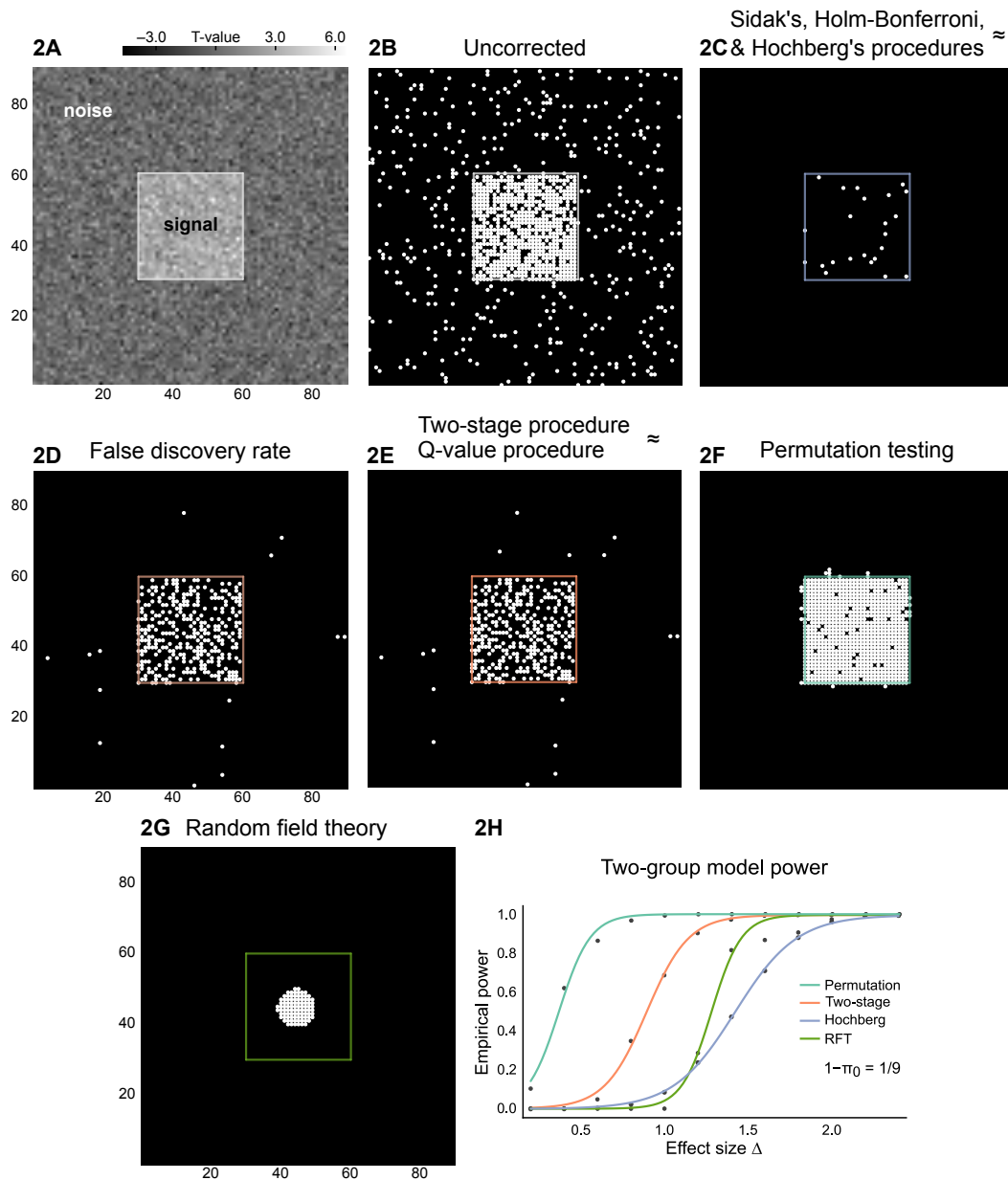


Figure 2: A single realization of the simulation performed using the two-group model. In each panel, there is a true effect at every location within the boxed region and no true effects elsewhere. Repeating the simulation for a large number of times using various parameter selections allows estimating the relative merits of the different testing procedures, as well as performing power and reproducibility analyses. (A) The t-statistic of each variable obtained using Student's t-test. (B) The uncorrected p-values corresponding to the t-values thresholded at the critical level $\alpha = 0.05$. (C–G) The significant p-values after the correction for multiple testing has been performed using each of the compared methods. In this particular instance, permutation testing detects the largest number of true effects while incurring a small number of false positives. (H) The average empirical power for a representative selection of the different methods at various effect sizes.

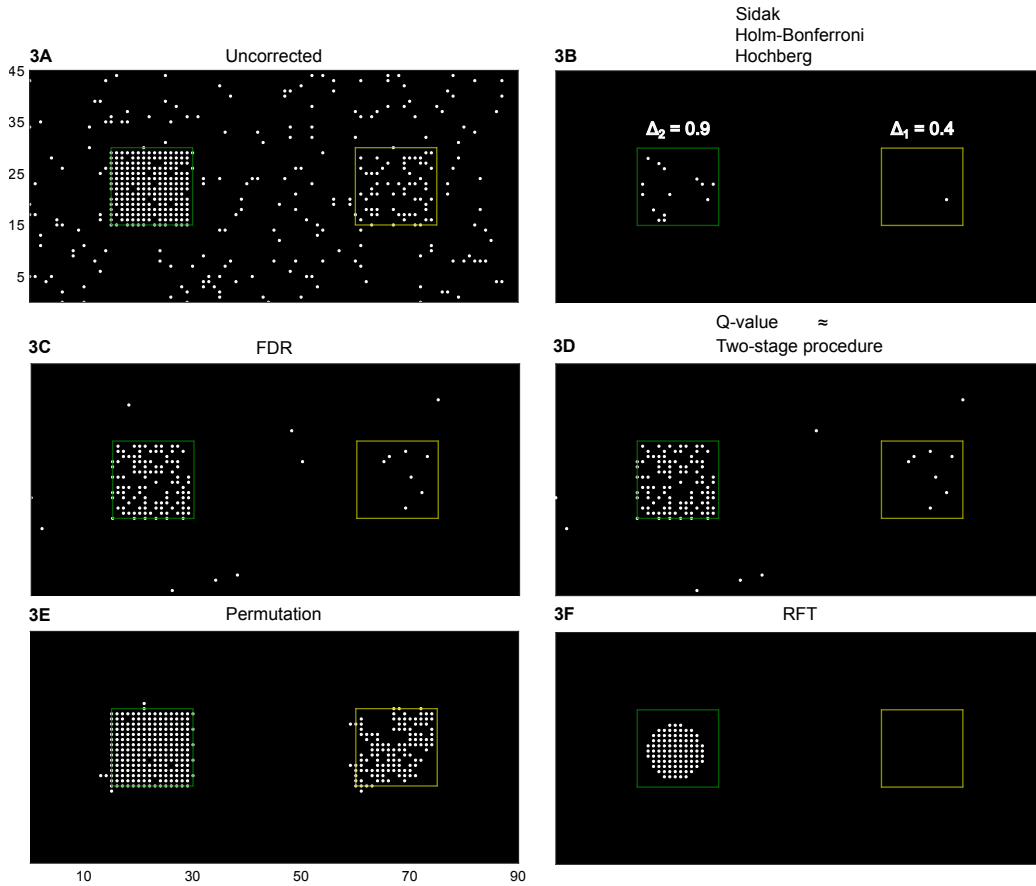


Figure 3: A single realization of a simulation performed using the spatial separate-classes model. There is a true effect at every location within the boxed regions and no true effects in the surrounding space. (A) The uncorrected p-values thresholded at $\alpha = 0.05$. (B) The classic methods that control the FWER produced identical results, and therefore they were collapsed into one panel. (C) P-values that were significant when FDR was controlled using the classic Benjamini-Hochberg procedure. (D) Similar to (B), the adaptive FDR methods gave almost identical results, and therefore the results were also collapsed into a single panel. (E–F) Permutation testing detects both true effects almost entirely whereas RFT detects only most of the larger effect. Generally, repeating the simulation allows estimating the relative merits of the different techniques in the presence of two distinct effects. In addition, it is possible to test which of the two options is better: to perform two separate or a single combined analysis.

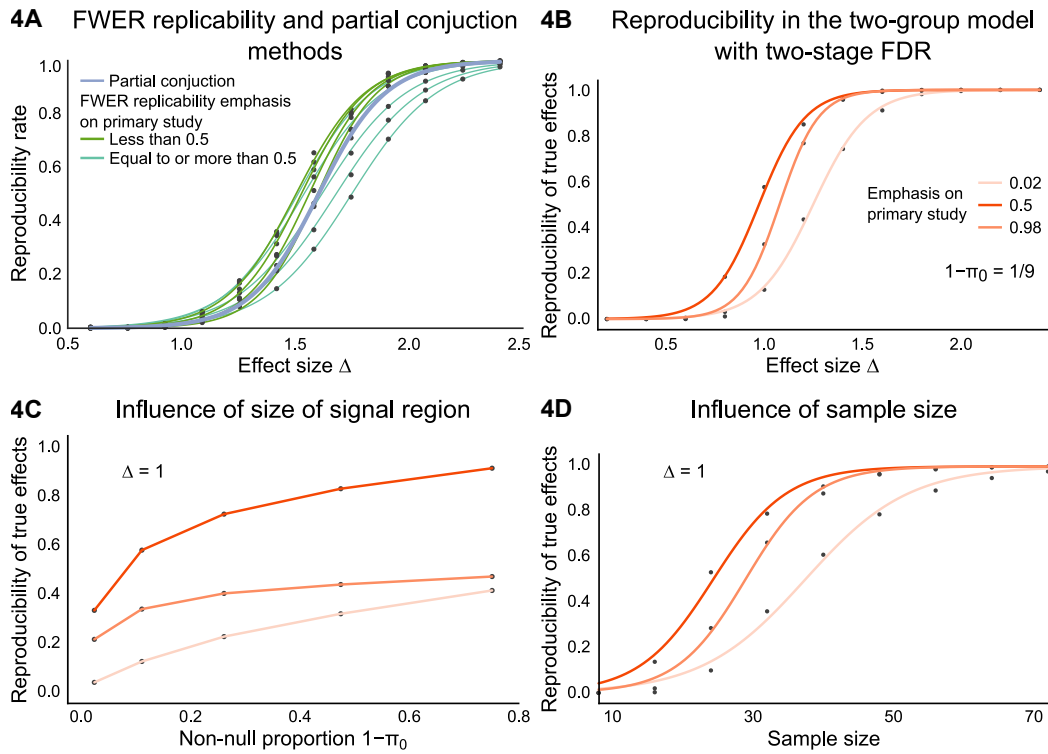


Figure 4: (A) A comparison of the partial conjunction and FWER replicability methods while performing a conservative multiple testing correction using the Hochberg’s method. The relative performance of the two methods depends on the emphasis of the primary study over the follow-up study. (B) Primary and follow-up experiments were simulated using the spatial two-group model and the FWER replicability method was used to decide which hypotheses were reproducible across the two experiments. Now, the multiple testing correction was performed using the two-stage FDR procedure. The optimal result is obtained when the primary and follow-up studies are given equal importance, and not when strict corrections are performed in the primary study. (C) The result seen in panel (B) is not specific to a particular null density, or in other words, size of the signal region in the two-group model. (D) Sample sizes that are substantially larger than those typically employed in neuroscience experiments are needed to observe true effects reproducibly even at large effect sizes.

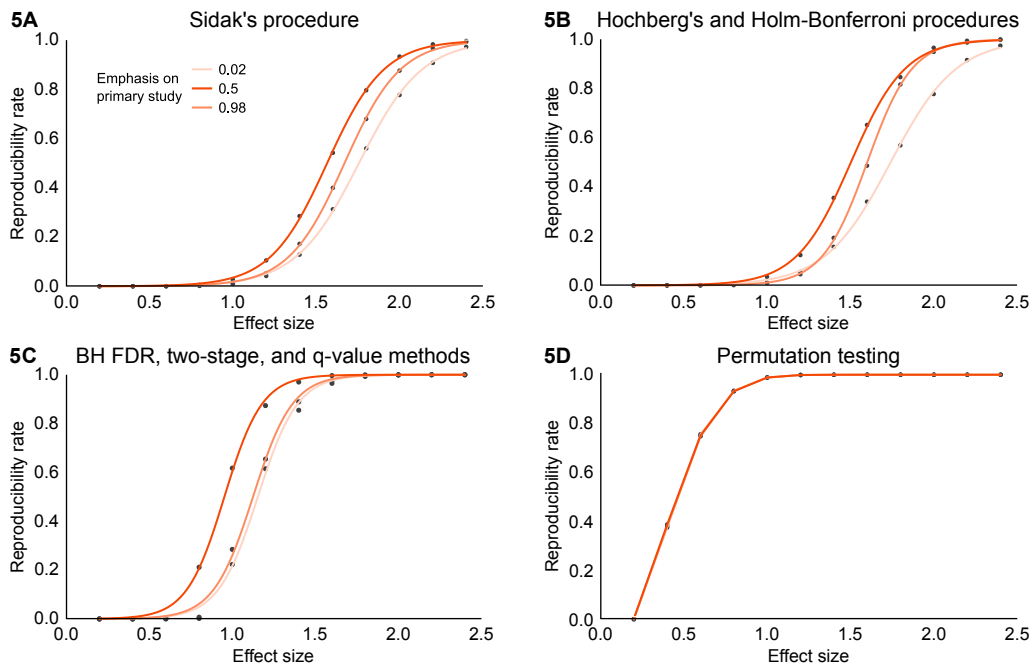


Figure 5: Placing an identical emphasis on the primary and follow-up studies produced the highest number of reproducible true effects in the two-group model regardless of the used correction procedure (A–C). However, for permutation testing (D), this difference was negligibly small since there is only one cluster of true effects in the two-group model.

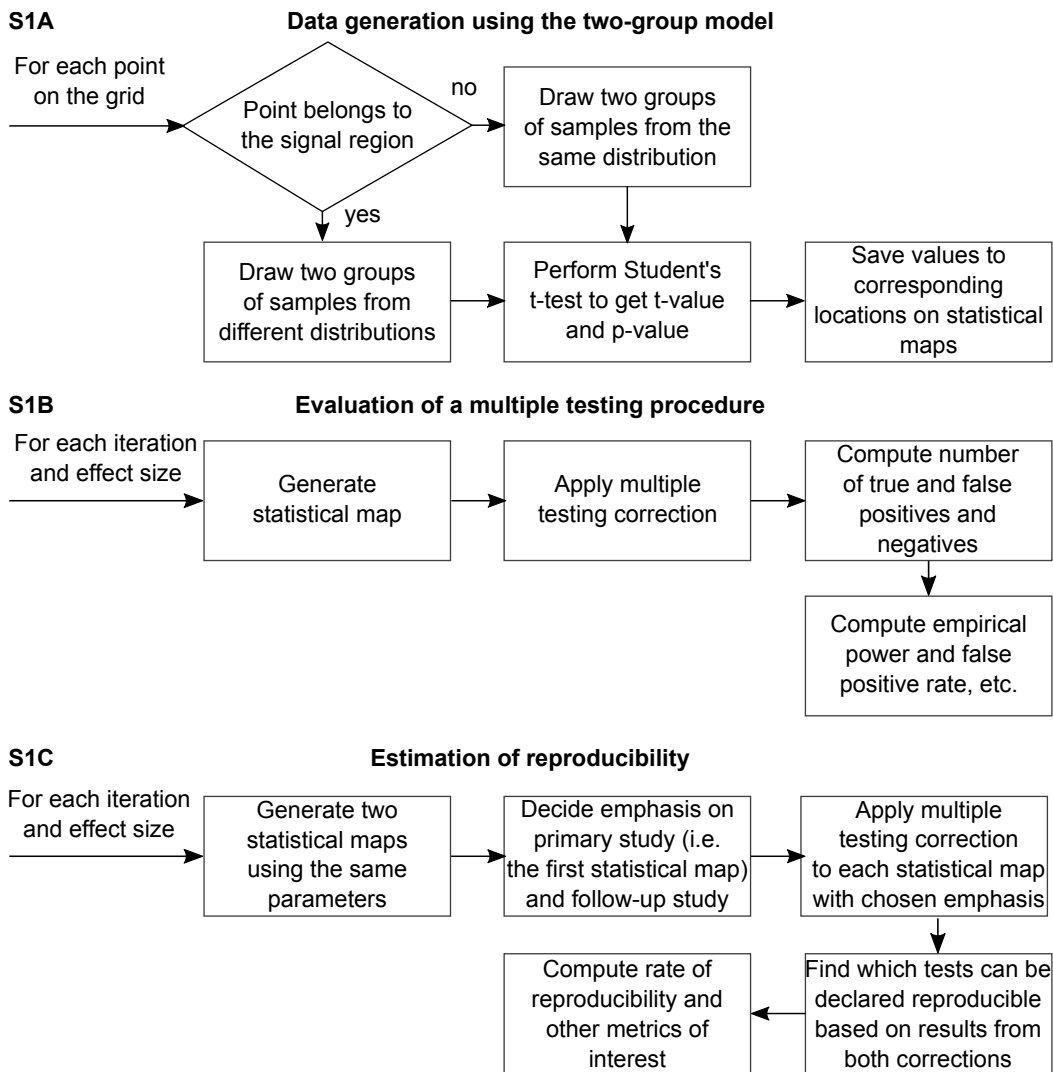


Figure S1: (A) Flowchart illustrating the main steps of generating data using the two-group model. The effect size is controlled by the difference in the distributions' location parameters. (B) The procedure used to evaluate the performance of a given multiple testing method in terms of true and false positives and negatives and metrics derived from them. (C) The procedure used to estimate reproducibility at various effect sizes when using a given multiple testing procedure.

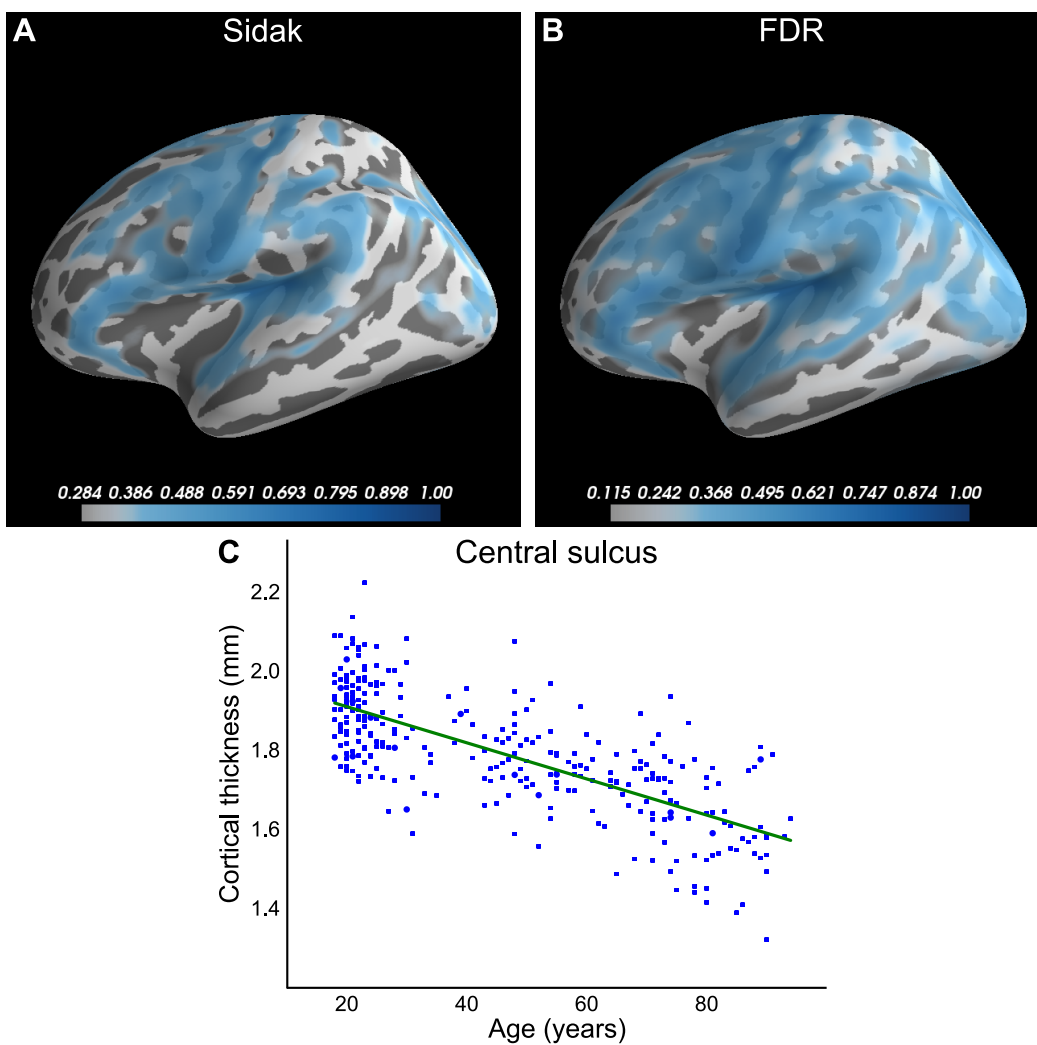


Figure S2: (A–B) A more widespread association between cortical thickness and aging was found when the FDR was controlled in comparison the controlling the FWER. (C) Example data from the central sulcus.