Influence of multiple hypothesis testing on reproducibility in neuroimaging research

Tuomas Puoliväli[a,*], Satu Palva[a,b], J. Matias Palva[a,b,*]

[a]*Helsinki Institute of Life Science, Neuroscience Center, University of Helsinki, Finland*
[b]*Centre for Cognitive Neuroimaging, Institute of Neuroscience and Psychology, University of Glasgow, UK*

Abstract

Reproducibility of research findings has been recently questioned in many fields of science but the problem of multiple hypothesis testing has received little attention in this context. The elevated false positive rate in multiple testing is well known and solutions to this problem have been extensively studied for decades. However, finding the balance between exceedingly liberal and too conservative rejection thresholds, i.e., avoiding false discoveries without excessive loss of power, has remained a challenge. We show here that this loss of power, rather than false discoveries per se, greatly aggravate the reproducibility problem in research where multiple hypotheses are tested simultaneously. We also advance here a Python-based open-source and freely available toolkit, "MultiPy", containing procedures for controlling the family-wise error rate (FWER) and the false discovery rate (FDR) by techniques based on random field theory (RFT), cluster-mass based permutation testing, adaptive FDR, and the core classic methods. We quantified the performance of these methods with simulated data and show here that rigorous control of false positives abolishes also the true positive rate (power) and reproducibility. Moreover, we further show that under-powered studies entailing multiple comparisons, even with liberal multiple comparison correction, are disproportionately poorly reproducible. Finally, we demonstrate the usage of this toolkit also in a magnetic resonance imaging (MRI) based assessment of the age-dependent decline of cortical thickness using data from the open access series of imaging studies (OASIS). We hope that this effort will help to improve current data analysis practices, facilitate building new software for group-level data analyses, and assist in conducting power and reproducibility analyses for upcoming studies.

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

1. Introduction

New research is motivated by questions that arise from previously acquired knowledge. Therefore, it is central to scientific progress that a researcher can repeat a peer's previous experiment. However, the reproducibility of reported research findings has been recently called into question in many fields, including psychology and neurosciences (Open Science Collaboration, 2015; Baker, 2016; Poldrack et al., 2017). This problem has a spectrum of causes but is known to be influenced by the testing of multiple hypotheses simultaneously (Ioannidis, 2005), which increases the number of false positive findings unless the corresponding p-values are carefully corrected. Hence, development of tools to evaluate multiple-testing data-analysis plans and to perform the required corrections using appropriate methods is a central task. However, much of the focus in the multiple comparisons considerations has

*Correspondence should be addressed to Tuomas Puoliväli or to Matias Palva.
Email addresses: `tuomas.puolivali@helsinki.fi` (Tuomas Puoliväli), `satu.palva@helsinki.fi` (Satu Palva), `matias.palva@helsinki.fi` (J. Matias Palva)

been placed on the false positive rate while it is important to also acknowledge that adopting too conservative correction thresholds or techniques risks significant discoveries remaining unnoticed, i.e., becoming false negatives. While this is a problem for scientific inference and gaining understanding the studied system, we emphasize here that false negatives, or inadequate true-positive rate, have fundamental consequences for the reproducibility of research and, in fact, drive the poor reproducibility of studies entailing multiple comparison. Here we compare techniques for controlling the family-wise error rate (FWER) and the false discovery rate (FDR) and present a software called MultiPy, which is a Python-based open-source collection of procedures for controlling the FWER and FDR.

Python has become in the recent past a programming language of choice for many scientists across several disciplines. Accordingly, there already exists packages for scientific computing (van der Walt et al, 2011), data visualization and manipulation (McKinney, 2010), machine learning (Pedregosa et al, 2011), and pre-processing, analysis, and source reconstruction of electroencephalography (EEG) and magnetoencephalography (MEG) data (Gramfort et al, 2013; 2014) to name a few. While these and other packages allow efficient analysis of single-subject neurophysiological and neuroimaging data, most provide a limited number of options for correcting group-level results for multiple comparisons. This limits the researchers' ability to compare for example parametric and non-parametric multiple testing approaches in analyzing the same dataset. In addition, some advanced techniques are implemented in only one software as part of a larger data analysis pipeline, which renders them difficult to access in practice for many. In a broader perspective, the analysis of single-subject data has received perhaps more attention than the analysis of group-level data, and even started to take first steps towards standardization (Gross et al, 2013). Several factors obviously contribute to the heterogeneity of group-level analyses but one of them is a lack of specialized multiple testing software that is designed to be accessible, comprehensive, compatible with existing neuroscience data analysis tools, and open for the scientific community to develop further. MultiPy is a software that attempts to help addressing this issue.

To highlight differences between methods that control the FWER or FDR, we first compare their performance using simulated data generated under two related realistic models. The first one is the classic two-group model but with a spatial structure that consists of distinct signal and noise regions (Bennett et al., 2009). The second one is a combination of two spatial two-group models, which makes it otherwise identical to Efron's (2008) separate-classes model except for the spatial dependency structure. These models can be thought to represent effects in neurophysiological data in the time-frequency domain or effects in a single slice of functional or anatomical magnetic resonance imaging (MRI) data. In addition to allowing comparison of methods, researchers can use these models for simulation-based power analyses to facilitate planning of new experiments. The models can be also used to estimate sample and effect sizes required to observe effects reproducibly.

To illustrate differences in controlling the FWER and FDR in empirical data, we perform mass-univariate analyzes on MRIs extracted from the open-access series of imaging studies (OASIS) dataset (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 subcortical structures and morphology. The same data 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 numerous methodological studies for benchmarking different approaches and testing new classification algorithms (Khajehnejad et al, 2017; Alam et al, 2017). Here we compare the number of voxel-level cortical thickness estimates significantly affected by age in healthy participants when the multiple testing correction is based on controlling the FWER and FDR.

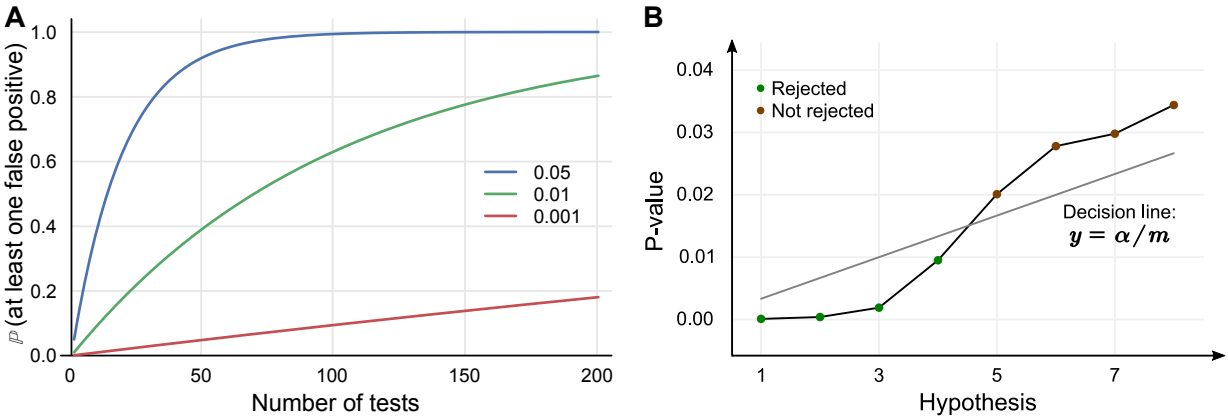

Figure 1: (A) For independent tests, false discoveries quickly start to occur when more than few dozens of tests are performed simultaneously. Each line corresponds to a particular conventional critical level indicated in the figure legend. (B) The Benjamini-Hochberg FDR procedure can be interpreted graphically as rejecting p-values sorted from the smallest to the largest that fall below the line $y = \alpha/m + 0$ until the first non-significant p-value is found.

2. Methods

2.1. Null hypothesis significance testing, family-wise error rate, and false discovery rate

In statistical null hypothesis significance testing (NHST), there are four possible distinct 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. The false positives and false negatives are also known as type I and type II errors respectively. Since this study is about multiple hypothesis testing, our discussion is largely focused on the number of incurred false positives while testing m hypotheses simultaneously under a specified critical level α, and how the count can be controlled using various procedures. For clarity, we use the variable names defined in Table 1 throughout the manuscript. Bold upper-case letters are used to denote random variables. Parts of the text refer to 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 respectively.

Table 1: Possible outcomes in NHST

	Declared non-significant	Declared significant	Total
Null hypothesis is true	$\mathbf{U}$	$\mathbf{V}$	m_0
Null hypothesis is false	$\mathbf{T}$	$\mathbf{S}$	$m - m_0$
	$m - \mathbf{R}$	$\mathbf{R}$	m

FWER is the probability of making one or more false positive conclusions when testing m hypotheses simultaneously. For independent tests, it can be derived from basic principles in a few steps as follows. Let α be the probability of making a false positive conclusion while performing a single test; then $1 - \alpha$ is the probability of no such mistake occurring. Further, by the product rule, $\prod_m 1 - \alpha = (1 - \alpha)^m$ is the probability of making zero such mistakes while performing the m tests. Hence the opposite event is the probability of making at least one false positive conclusion

$$\texttt{FWER} = \mathbb{P}(\mathbf{V} \geq 1) = 1 - (1 - \alpha)^m. \tag{1}$$

This identity is illustrated in Figure 1A for the three conventional critical levels.

In practice, controlling the FWER using the classical methods may be too conservative when the number of simultaneously performed hypothesis tests is large and there exists dependencies between the tests. In such circumstances, it can be advantageous to control the FDR instead (Benjamini & Hochberg, 1995), which is the expected proportion of discoveries that are false. Alternatively, the

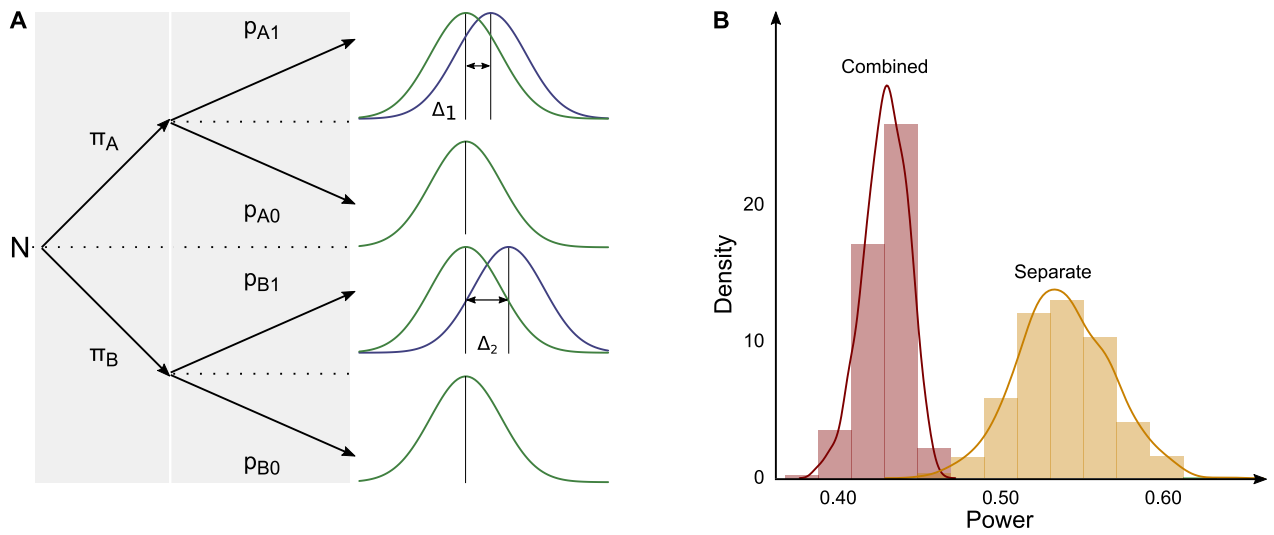

Figure 2: Schematic of the separate-classes model. The left panel (A) is partly adapted from Efron (2008). If one knew what the separate classes were a priori, performing two separate corrections would give a higher power than performing one combined analysis (B). The histograms and kernel density estimates were constructed by simulating data under the separate-classes model for 1000 iterations with $\Delta_1 = 1.2$ and $\Delta_2 = 0.7$.

correction procedure may include estimation of autocorrelations as part of the multiplicity correction procedure, which is the approach taken in cluster-based permutation testing and RFT. Before discussing these and other available methods at length, we present two models that can be used to compare their performance.

2.2. Models for evaluating multiple hypothesis testing methods and planning data analyses

2.2.1. The two-group model

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, we first generate simulated data using the well-known two-group model with the spatial dependency structure suggested by Bennett and co-authors (2009). Specifically, our simulation consists of a 30×30 variable signal region centered on eight 30×30 variable noise regions, which collectively form a 90×90 variable two-dimensional grid. For each location or variable on the grid, there are 25 samples in each of the two groups denoted by A and B. The samples of both groups in the noise regions are distributed as $\mathcal{N}(0, 1)$. In the signal region, the samples of A and B are distributed as $\mathcal{N}(0, 1)$ and $\mathcal{N}(\Delta, 1)$ respectively. The parameter Δ controls the effect size (Cohen's d) and was set to 0.8 in our specific simulation. In other words, there is a true effect at every location within the signal region, and no true effects within the noise regions. This allows comparing the different multiple testing procedures numerically in terms of their power and false discovery rate, or more generally using any metrics that can be derived from the resulting error matrices.

2.2.2. The separate-classes model

In addition to being continuous in space, time, or frequency, fundamental effects in neuroscience data are also often distributed (Heller et al., 2006; Chumbley et al., 2010). However, only few previous studies investigated properties of multiple testing procedures under models that considered two or more distinct effects with different effect sizes simultaneously. Yet, such models could be informative for answering questions such as whether large effects are prone to mask smaller ones, or how do the effects' spatial dependency structures influence their detectability. Efron (2008) analyzed inference under a combination of two two-group models, called a "separate-classes model", which was motivated by diffusion tensor imaging data from dyslexic and control participants indicating distinct

Table 2: Controlling the FWER using Šidák's procedure in MultiPy.

```
from multipy.data import neuhaus
from multipy.fwer import sidak

pvals = neuhaus()
significant_pvals = sidak(pvals, alpha=0.05)
print zip(['{:.4f}'.format(p) for p in pvals], significant_pvals)
```

effects in anterior and posterior parts of the brain. The findings indicated that since the underlying data had distinct effects and structures, performing two separate corrections was better than a combined analysis (Efron, 2008).

Our separate-classes model is essentially a combination of two distinct two-group models similar to Efron's earlier study (2008) but with the spatial structure suggested by Bennett and co-authors (2009). There are now groups A and B in the first class and groups C and D in the second class. Similar to the two-group model, all samples are distributed as $\mathcal{N}(0,1)$ in the noise regions. In 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 allows testing how differences in the effect sizes Δ_1 and Δ_2 influence the multiple testing results. The model is visualized schematically in Figure 2.

The separate-classes model was used to generate data for comparing FWER and FDR algorithms where both signal regions were 15×15 variables and the overall size was 45×90 variables. The effect sizes (Cohen's d's) in the first and second signal regions were $\Delta_1 = 0.9$ and $\Delta_2 = 0.4$ respectively. The sample size was set at 25 in each group.

2.3. Multiple hypothesis testing methods

2.3.1. Classic methods for controlling FWER

Well-known and widely used 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 in that they compute new critical levels $\alpha_{\texttt{Bonferroni}} = \alpha/m$ and $\alpha_{\texttt{Sidak}} = 1-(1-\alpha)^{\frac{1}{m}}$ by directly adjusting the original α with the number of performed comparisons. Table 2 shows an example of performing Šidák's correction using MultiPy. In turn, it is educational to compare the Holm-Bonferroni and Hochberg's procedures, since they apply the exact same threshold $\alpha/(m-k+1)$ while correcting the kth ascendingly sorted p-value. The distinction between the 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. False discovery rate

False discovery rate (FDR) controls the expected proportion of discoveries that are false, that is $\mathbb{P}(\mathbf{R}>0)\mathbb{E}(\mathbf{V}/\mathbf{R}|\mathbf{R}>0)$ (Benjamini & Hochberg, 1995). Since often there are no prior expectations regarding each individual test, controlling the FDR instead of FWER is a sound choice. Practically, controlling the FDR implies willingness to accept a small fraction of false positives in exchange for a substantially improved power. The Benjamini-Hochberg FDR procedure is perhaps the most widely used FDR procedure, which is easy to understand graphically: p-values sorted from the smallest to the largest that fall below the line $y = \alpha/m + 0$ are rejected until the first non-significant p-value (see Figure 1B).

2.3.3. Adaptive false discovery rate methods

The classic FDR method assumes that $\pi_0 = 1$ which can make it too conservative when the number of hypotheses is large (Storey & Tibshirani, 2003). Therefore, newer 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 multiplicity correction procedure.

Table 3: Example of simulating an experiment using the two-group model and computing Storey-Tibshirani q-values and the associated diagnostics plots.

```
from multipy.data import two_group_model
from multipy.fdr import qvalue
from multipy.viz import plot_qvalue_diagnostics

tstats, pvals = two_group_model(N=25, m=1000, pi0=0.5, delta=1)
_, qvals = qvalue(pvals)
plot_qvalue_diagnostics(tstats, pvals, qvals)
```

While a successful estimation of π_0 can allow greater power than the classic Benjamini-Hochberg method, there is a possible caveat: under some specific circumstances, applying the adaptive procedures results in *more* significant discoveries than there were in uncorrected data (Reiss et al, 2012). In the study by Reiss and colleagues (2012), this kind of paradoxical results were found to occur in a 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). Since such paradoxical outcomes are easy to detect by counting the number of declared significant effects before and after the correction has been applied, we have implemented an automatic warning mechanism to MultiPy to draw user's attention to such occurrences. An example of using MultiPy to generate data under the spatial two-group model, apply the q-value method, and show diagnostic plots is given in Table 3.

2.3.4. Permutation testing

In contrast to the other approaches of controlling the FWER, permutation tests provide a non-parametric but 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 relabeling the observations and computing the corresponding test statistics, and then used to calculate the significance of the correct labeling (Nichols & Holmes, 2001). Maris and Oostenveld (2007) were among the first to apply permutation testing systematically to analyzing EEG and MEG data using permutation tests.

The only assumption in permutation testing is the exchangeability of observations under the null hypothesis. Practically, this means that the analyzed data should be free of autocorrelations, which conflicts with the previous argument of fundamental effects in systems-level neuroscience data being temporally, spatially, or spectrally continuous. To resolve the discrepancy, it possible to incorporate the underlying dependency structure as part of the non-parametric permutation test. Maris and Oostenveld (2007) did this by using a cluster-mass test statistic (Bullmore et al., 1999), which is also the approach taken in implementing MultiPy.

An important point in permutation testing is that permutation test p-values should not be allowed to take zero values (Phipson & Smyth, 2010). In the procedure described by Maris and Oostenveld (2007), this situation arises when all permutation test statistics are less extreme than the observed one (i.e., when the effect size is large). The reason is that the p-value is defined as $p = b/m$ where b is the number of times the permutation test statistic was 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, the issue can be entirely 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). Previous studies have also explored its applicability for the analysis of inverse-modeled 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; Flanding & Friston, 2017). Application of RFT to neuroimaging data proceeds in three subsequent steps. First, the smoothness of the analyzed statistical map is estimated, giving the number of resolution elements or resels (denoted by R) it consist of. Second, the expected Euler characteristic is computed for a range of thresholds. For two-dimensional data, the expected Euler characteristic is $\mathbb{E}[\mathtt{EC}] = R(4\log_e 2)(2\pi)^{-\frac{3}{2}}\exp(-\frac{1}{2}z_t^2)$ where R is the number of resels and Z_t is the z-score threshold (Worsley et al., 1992). Finally, the empirical data is compared to the theoretical expectation.

2.4. Test-retest reproducibility

A priori power calculations based on previously published research findings, pilot data, or smallest clinically or biologically meaningful effect size is the recommended practice for sample size selection. However, multiple testing is often omitted from consideration from these power calculations due to analytical intractability or lack of models for generating simulated data. Subsequently, reproducibility is considered a priori in even fewer studies, and it can be argued that it has remained poorly understood how much the choice of a multiple testing correction method influences reproducibility even in the relatively simple experiments. To take some steps in addressing this issue, we perform test-retest experiments at different effect sizes under the spatial two-group model. The critical level is set at $\alpha = 0.05$, the sample size at $N = 25$ in both groups, the signal region is 30×30 variables centered on a 90×90 grid containing no other effects, and the results are averaged over twenty realizations. The Benjamini-Hochberg FDR procedure and Šidák's correction are compared in analyzing the data to highlight the need for using modern multiple testing techniques.

2.5. Comparison of methods using large sample size empirical MRI data

OASIS is an open data initiative that distributes cross-sectional and longitudinal MRI datasets (`http://www.oasis-brains.org/`). Here we analyze data from the release OASIS-1, which is a cross-sectional sample of $N = 416$ structural T1-weighted MRIs from healthy and demented participants aged 18 to 96 years (Marcus et al, 2007). Given its broad coverage of the human lifespan, these data allow investigating how age and age-related diseases influence brain morphology with high statistical power. In this study were restricted our analyses to the healthy individuals to keep the statistical testing as simple as possible. Vertex-level cortical thickness estimates were extracted after spatial smoothing with a Gaussian 10 mm full-width at half-maximum (FWHM) kernel as implemented in FreeSurfer (Fischl & Dale, 2000). A linear regression was performed at each individual vertex between cortical thickness and age, yielding a large univariate statistical map. Šidák's correction and the Benjamini-Hochberg FDR procedure were compared in correcting the obtained statistics for multiple testing.

2.6. 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/multipy`, and published as open-source under a permissive license to facilitate further development and integration to other software and data analyses.

3. Results and discussion

Here we developed a new open-source Python software for multiple hypothesis testing which implements techniques for controlling the FWER and FDR. The software includes a variety of different parametric and non-parametric techniques which were first compared using simulations and then with empirical data from a cross-sectional large sample size MRI aging study. The simulations were based on spatial two-group and separate-classes models designed to capture features typically observed in neurophysiological and neuroimaging data. The MRI data was analyzed on vertex-level with mass-univariate testing to highlight differences between the implemented multiplicity correction procedures when the number of variables is large. Our results indicate that permutation testing was the most powerful method for the analysis of these data.

3.1. How to choose between controlling the FWER and FDR

The FWER based methods control the probability of making one or more false positive conclusions. This strict control is gained at the expense of false negatives, which may lead to important discoveries remaining unnoticed, especially when the number of tested hypotheses is large. Therefore, controlling the FWER is best justified when any false positive discoveries are expensive due to ethical, financial, time or other constraints. For example, a researcher might wish to conduct a subsequent confirmatory study corresponding to each discovery; the use of animals or high research costs could make the false leads too expensive.

Controlling the FDR at the conventional critical level implies that one is willing to accept on average 5% of false positives among the hypotheses declared significant. This is a sound choice for many types of data analyses, since often there are no theories or models concerning each individual test. For example, consider 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 result. 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). In most cases similar to this example, there would be no a priori expectation for a given time-frequency tuple such as (115 ms, 16 Hz) in contrast to some neighboring tuple, say (117 ms, 17 Hz). Instead, one would look for a consistent activity pattern over several frequencies and time points; a small number of false positives within such region would not compromise the result's interpretation.

3.2. Permutation testing, RFT, and adaptive FDR techniques outperform classic techniques

In the two-group model, the classic methods that control the FWER produced similar results. In contrast, the FDR based methods detected expectedly more true positives while incurring a small fraction of false positives. The adaptive FDR procedures yielded similar results, which makes the two-stage procedure more attractive due ease of implementation and more stable null density estimator. The RFT and permutation testing based methods outperformed the other approaches. The results are visualized in Figure 3.

3.3. Separating multiple hypothesis testing analyses into distinct parts can increase power

In the separate-classes model, detecting a second effect is difficult if a single combined analysis is performed (Figure 4). This suggests that researchers should be keen to consider whether their neurophysiological and neuroimaging data is most appropriately analyzed by combined or separate analyses. For example, it is well known that in EEG and MEG data the signal-to-noise (SNR) ratio decreases as a function of frequency, but yet most published time-frequency domain analyses were performed by combining tests from all time points and frequencies into a single analysis. We suggest future studies to leverage the available information about EEG and MEG measurement techniques as well as the distinct roles of oscillations, components of evoked responses, and other similar information while performing multiple testing corrections. Similar arguments can be put forward for functional and structural MRI data (consider e.g. hemispheric lateralization of language, speech, and auditory functions).

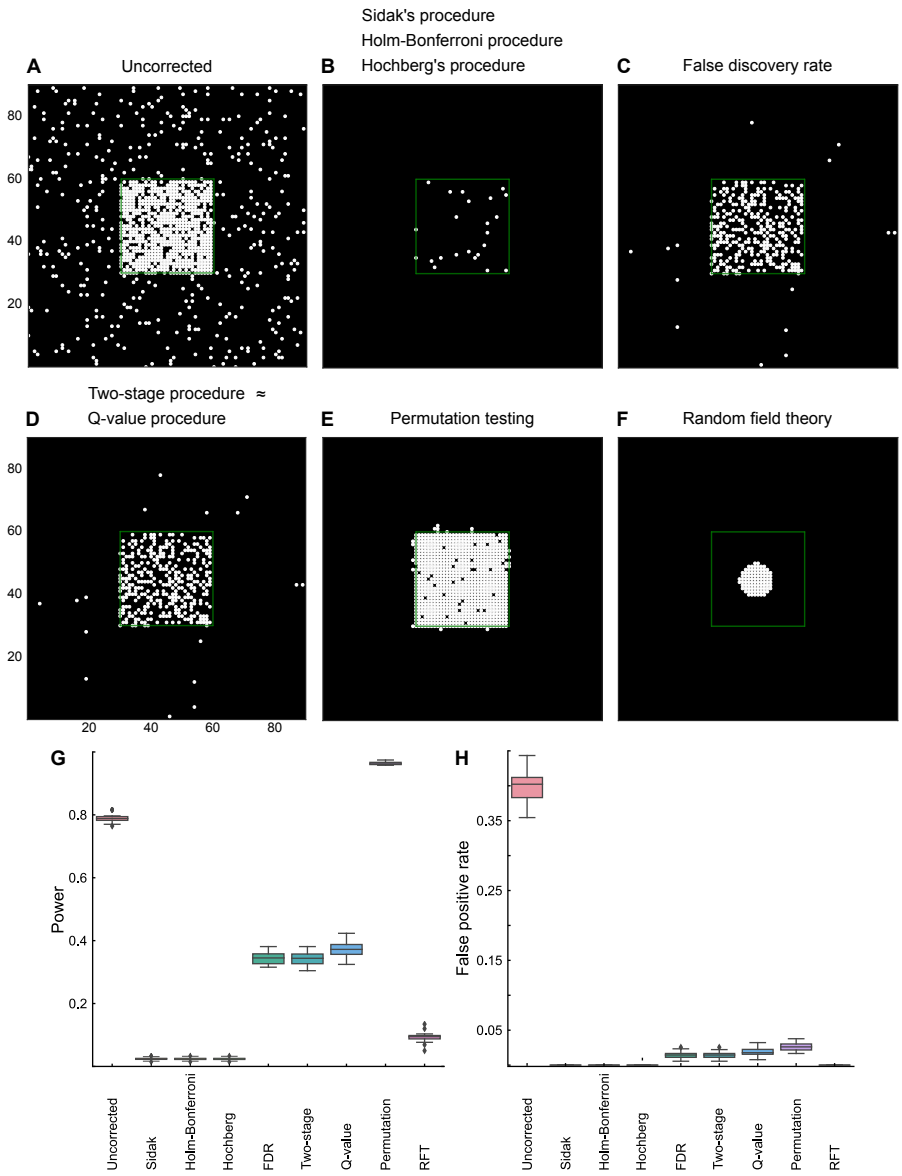

Figure 3: In the two-group model, permutation testing detects the largest number of true effects while incurring a minimal number of false positives under the chosen sample and effect sizes. In each panel, there is a true effect at every location within the green box and no true effects in the surrounding space.

3.4. Choice of correction technique influences interpretation of study results

A total of 55% of vertices showed a statistically significant effect of aging on cortical thickness after the p-values were corrected for multiple testing using Šidák'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 critical level was set at $\alpha = 0.05$ in both analyses. The results are illustrated in Figure 5. Qualitatively, there is a large difference between the interpretation of the FWER and FDR corrected results. When the FWER is controlled using Šidák's correction, it would appear that aging influences only some parts of the cortex. In contrast, the FDR based control using the Benjamini-Hochberg procedure suggests that essentially the whole cortex is affected.

3.5. Planning for reproducibility

The recommended practice for sample size selection is to perform a priori power calculations. However, multiple testing is usually omitted from consideration, and reproducibility is considered even less seldom. Hence, it can be argued that it has remained largely unknown how well-powered

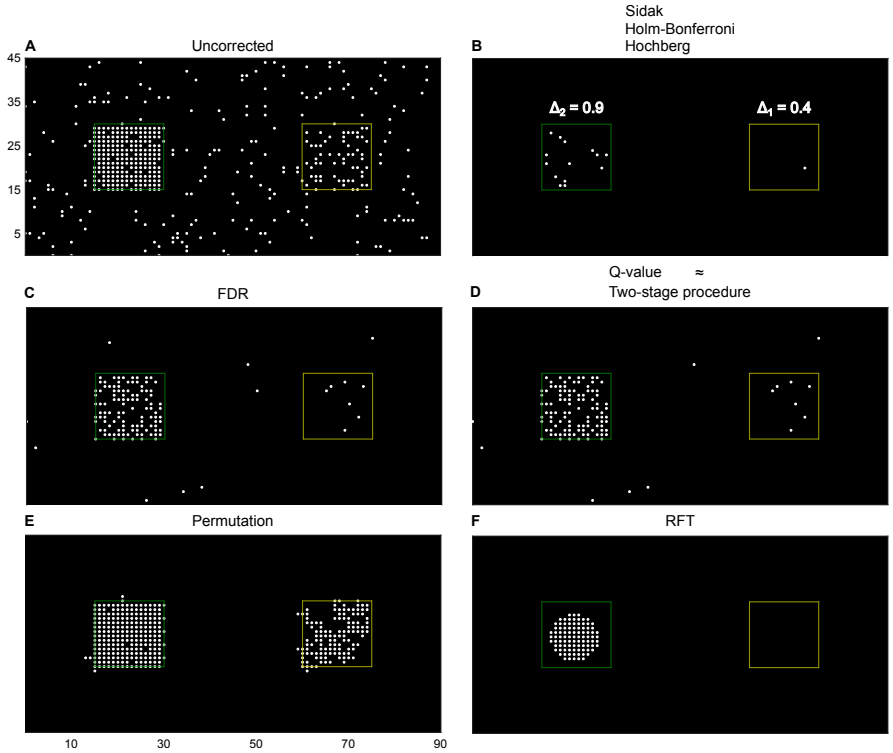

Figure 4: Simulation using the separate-classes model with two distinct effects. There is a true effect at every location within the green and yellow boxes and no true effects in the surrounding space. Permutation testing detects both true effects almost entirely whereas RFT detect only the larger effect. Adaptive and Benjamini-Hochberg FDR procedures find most parts of the larger effect but the smaller effect would likely be interpreted as false positives if the locations of the signal boxes were unknown to the observer.

studies are needed to observe reproducible effects. The issue is pressing 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 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 start addressing this issue, we performed test-retest simulations using the two-group model to draw attention on what the reproducibility rates might be at effect sizes typically observed in neuroscience data (Figure 6). Briefly, the results indicate that in this model the choice of multiple testing method has a large influence on observing the same effects twice, and therefore we suggest similar analyses to be carried out while planning new experiments. In the wider context, these results could also mean that reproducing existing results with a conservative multiple testing method in the original publication, when retained in the replication, could set the significance threshold at an excessively low value.

3.6. Future directions

In the future, the software can be expanded into several directions. There already exists for example several specialized techniques for multiple testing of neuronal synchrony estimates (Maris et al., 2007; Singh & Phillips, 2010; Singh et al., 2011) that could be implemented as part of the software. Other 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 (Ignatiadis et al, 2016). The software can be also extended to include more data-generating models, which will help researchers to better evaluate their data analysis plans and perform simulation-based power analyses.

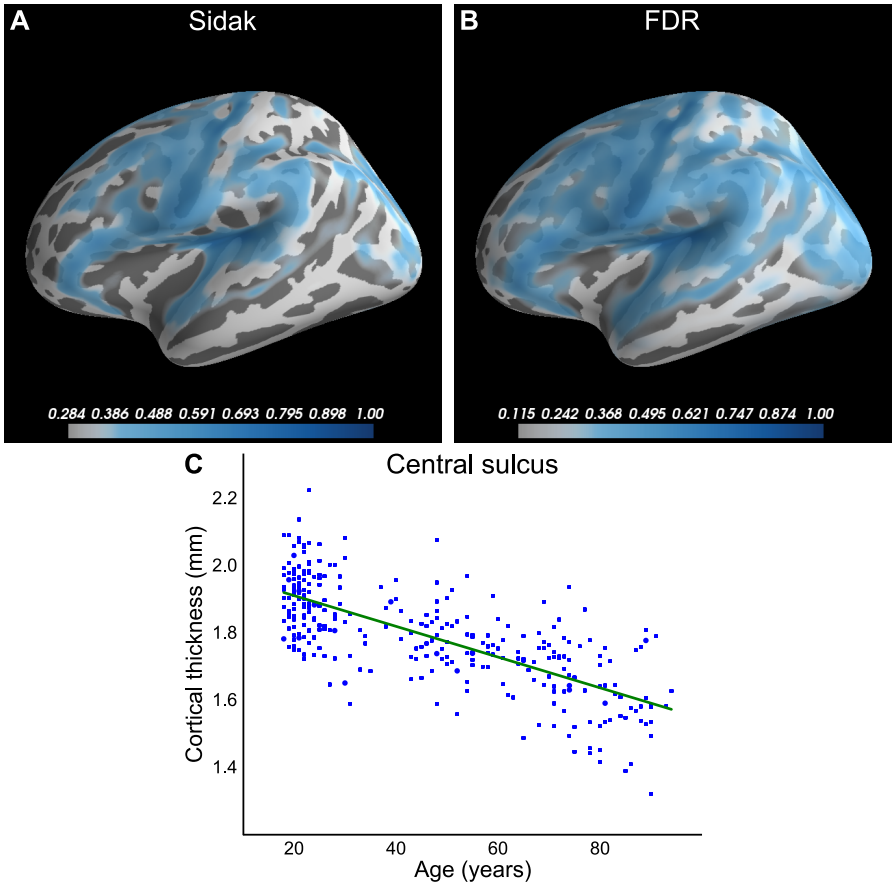

Figure 5: Controlling the FDR reveals more widespread association between cortical thickness and aging compared to controlling the FWER (A and B). Mean cortical thickness of the central sulcus decreases markedly as a function of age (C).

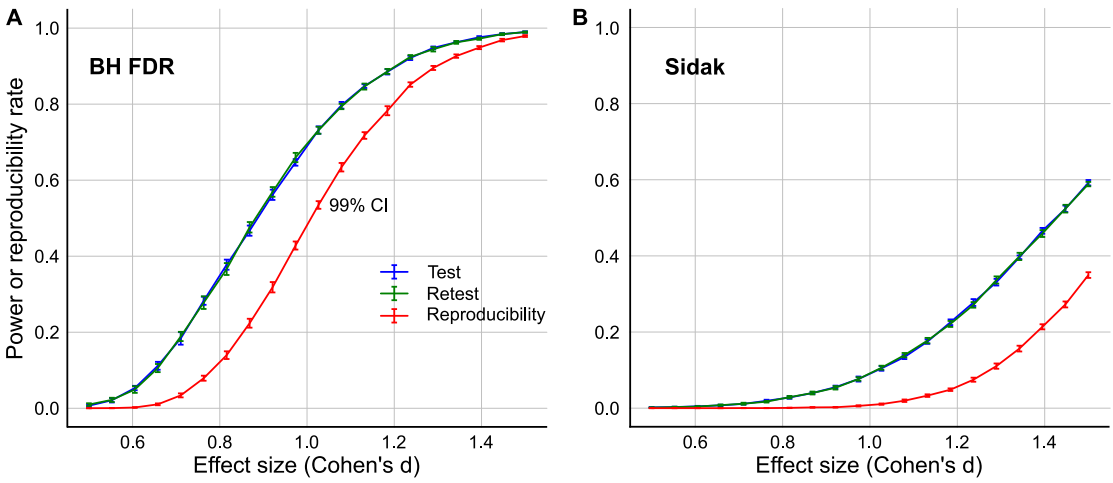

Figure 6: Test-retest data was simulated using the two-group model and the statistics were corrected using the BH FDR procedure (A) and Šidák's correction (B). Observing the same effects in the two repetitions was generally difficult, especially the when FWER was controlled using Šidák's correction. According to a recent review, the median power in neuroscience experiments is between 8–31%, which suggests that test-retest reproducible effects are observed only approximately 10% of the time. The error bars represent 99% confidence intervals.

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.

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.

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

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.

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

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 Transctions 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.

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.

Chen S, Wang C, Eberly LE, Caffo BS, Schwartz BS (2009): Adaptive control of the false discovery rate in voxel-based morphometry. *Human Brain Mapping* 30(7):2304–2311.

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

Design preclinical studies for reproducibility (2018). *Nature Biomedical Engineering* 2:789–790.

Destrieux C, Fischl B, Dale A, Halgren E (2010): Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage* 53(1):1–15.

Dickerson BC, Fenstermacher E, Salat DH, Wolk DA, Maguire RP, Desikan R, Pacheco J, Quinn BT, van der Kouwe A, Greve DN, Blacker D, Albert MS, Killiany RJ, Fischl B (2008): Detection of cortical thickness correlates of cognitive performance: Reliability across MRI scan sessions, scanners, and field strengths. *NeuroImage* 1(1):10–18.

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

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.

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.

Fotenos 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.

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.

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.

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.

Han X, Jovicich J, Salat D, van der Kouwe A, Quinn B, Czanner S, Busa E, Pacheco J, Albert M, Killiany R, Maguire P, Rosas D, Makris N, Dale A, Dickerson B, Fischl B (2006): Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage* 32(1):180–194.

Hayasaka S, Nichols TE (2003): Validating cluster size inference: Random field and permutation methods. *NeuroImage* 20(4):2343–2356.

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

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.

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.

Ioannidis JP (2005): Why most published research findings are false. *PLoS Medicine* 2(8):e124.

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

Madan CR, Kensinger EA (2017): Test-retest reliability of brain morphology estimates. *Brain Informatics* 4:107–121.

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.

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.

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

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

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.

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

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.

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.

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.

Reproducibility: let's get it right from the start (2018). *Nature Communications* 9:3716.

Salat DH, Buckner RL, Snyder AZ, Greve DN, Desikan RSR, Busa E, Morris JC, Dale AM, Fischl B (2004): Thinning of the cerebral cortex in aging. *Cerebral Cortex* 14(7):721–730.

Šidák Z (1967): Confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* 62(318):626–633.

Singh AK, Asoh H, Phillips S (2011): Optimal detection of functional connectivity from high-dimensional EEG synchrony data. *NeuroImage* 58(1):148–156.

Singh AK, Phillips S (2010): Hierarchical control of false discovery rate for phase locking measures of EEG synchrony. *NeuroImage* 50(1):40–47.

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.

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

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.

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.

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.

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.