Significance tests for Leave-One-Out Inter-Subject Correlations in fMRI

L. De Angelis[a], V. Gazzola[a,b,**], C. Keysers[a,b,**]

[a]*Social Brain Lab, Netherlands Institute for Neuroscience, Royal Netherlands Academy of Arts and Sciences, Meibergdreef 47, 1105 BA, Amsterdam, The Netherlands*
[b]*Department of Psychology, University of Amsterdam, 1018 WV Amsterdam, The Netherlands*

Abstract

The inter-subject correlation of fMRI data of different subjects performing the same fMRI task (ISC) is in principle a powerful way to localize and differentiate neural processes caused by a presented stimulus from those that spontaneously or idiosyncratically take place in each subject. The wider adoption of this method has however been impeded by the lack of widely available tools to assess the significance of the observed correlations. Several non-parametric approaches have been proposed, but these approaches are computationally intensive, challenging to implement, and sensitive methods to correct for multiple comparison across voxels in these approaches are not yet well established. More widely available, and computationally simple, parametric methods have been criticized on the basis that dependencies in the data could inflate false positives. Here, using three independent resting state fMRI datasets, we demonstrate that conventional parametric tests actually do provide appropriate control for false positives for inter-subject correlation analyses. This finding paves the way to a wider adoption of ISC, and empowers a wider range of neuroimagers to use ISC to tackle the challenges of naturalistic neuroscience.

Keywords: Inter-Subject Correlation, fMRI, Resting State, Statistics

1. Introduction

To overcome some of the limitations of traditional block- and event-related designs in functional magnetic resonance imaging (fMRI), the past decade has seen the emergence of Inter-Subject Correlation (ISC) as an alternative method to localize brain regions involved in the processing of complex stimuli. This method leverages that voxels containing neurons that are time-locked to a particular stimulus will show blood oxygenation level dependent (BOLD) signals that synchronize across multiple viewers of the same stimulus. Along the same logic, voxels that respond more reliably to one stimulus or condition than another will synchronize more to that stimulus or condition than to the other [1, 2, 3]. This method allows experimenters to use complex, long lasting stimuli, such as movies or stories, to probe the neural mechanisms underlying emotional processing and social cognition [3, 4, 5, 1, 6, 7]. For traditional fMRI analyses, the field has converged onto highly validated statistical inference methods that correct for multiple comparisons across voxels in traditional contrast-based designs [8]. These "best-practice" methods are available in user-friendly software packages such as SPM and FSL [9, 10]. The adoption of these packages has tremendously im-

*Corresponding author
**These authors contributed equally to this work
Email address: c.keysers@nin.knaw.nl (C. Keysers)

proved the rigour and replicability of fMRI data-analyses by protecting neuroimagers from the dangers of writing their own analysis codes and of choosing sub-optimal methods. An issue hampering the wider adoption of ISC is that how best to determine whether a given pattern of correlations across participants and voxels is above chance or not remains a matter of debate. A successful approach to statistical inference in ISC has been to use pair-wise correlations. Specifically, one correlates the brain activity of every possible pair of participants in each voxel, and then tries to infer whether these pair-wise correlations in one condition are larger than zero, or whether they are higher in one condition than in another. Because an experiment consisting of N participants generates a total of $N_p = N(N-1)/2$ possible combinations of pairs, with each participant's data contributing to $(N-1)$ pairs, assuming that all of these pairs are statistically independent is untrue. As described in [11, 12], this is a case in which conventional parametric t-tests with $\mathrm{DOF} = N(N-1)/2-1$ do not appropriately control for Type I error, i.e., random data will generate significant t-tests at $p < 0.05$ in more than 5% of cases. For such pair-wise correlations, more appropriate non-parametric approaches should be pursued. Reasonably, since these non-parametric tests need to be performed in hundreds of thousands of voxels, corrections for multiple comparisons need to be applied [13]. Frameworks to apply such corrections in a way that takes the spatial auto correlation of fMRI data into account however have not yet been optimized, and such tests are not available in the software packages in which most cognitive neuroscientists have been trained. Moreover, to perform such tests is computationally explosive because the number of possible pairs increases quadratically with sample size.

A second approach has been advocated, in which each participants time-course is not individually correlated with the time course of each other participant, but instead with the average time-course of all other participants (leave-one-out ISC). This approach considers the average brain activity of all other participants as a reference time-course that approximates the stimulus-triggered response in that voxel. The correlation of a participant's time course with that reference time-course is then considered a measure of stimulus-triggered processing in that participant. This approach generates a single ISC map per participant, avoiding the inflation of degrees of freedom encountered in pair-wise approaches, and leave-one-out ISC results have been analysed using traditional parametric t-tests after z-transforming the data by leveraging the power of standard fMRI analysis packages [3, 5, 14]. However, it has been argued, that such parametric approaches to leave-one-out data still run the risk of inflated false positive risks for two reasons. First, each subject does contribute to more than one ISC map, because it contributes to the average time course of all other participants, albeit in a way that is diluted by averaging [11]. Second, fMRI data follow a power law, and such data can generate spurious correlations [15]. However, whether these theoretical considerations actually translate into inflated Type I errors in ISC when applying parametric second level t-test to real fMRI data has never been explored. This is unfortunate, because using parametric t-tests would enable a wide community of neuroimagers to draw statistical inferences using software packages that have well validated corrections for multiple comparisons and lower the hurdle to the adoption of ISC analyses and the risk of errors associated with in-house analyses codes.

Here, we therefore assess the risk of false positive ISC results in parametric leave-one-out ISC analyses using three publicly available resting state fMRI datasets. The rationale is that if we measure brain activity at rest, without participants viewing a common stimulus, we should not find significant ISC re-

sults. If parametric t-tests control Type I error despite the above-mentioned concerns, we would expect that when using corrections for multiple comparisons at $p < 0.05$ for family wise error corrections (FWE) or $q < 0.05$ for false discovery rate corrections (FDR), a t-test on leave-one-out ISC on resting state fMRI should not produce significant results in more than 5% of cases. We thus apply these analyses on 1000 subsamples of N participants from larger resting state datasets, and use SPM (`www.fil.ion.ucl.ac.uk/spm`), one of the most widely used and available fMRI analysis software to perform parametric t-tests on the ISC maps using the aforementioned corrections for multiple comparison and measure the false-positive rate for a number of combinations of p/q-thresholds and minimum cluster-sizes. From our data we conclude that both FDR and FWE corrections with a minimum cluster size of 5 voxels provide robust protection against Type I error. In addition, we compared the sensitivity of these methods in detecting ISC in a dataset in which participants did watch the same movie, and synchronization is present [16] and compare these results with established non-parametric bootstraping methods [11]. We find that although all methods concord on the most highly correlated voxels, FDR is the most sensitive method, and is more sensitive that non-parametric bootstrapping despite its ability to control Type I error.

2. Methods

2.1. Datasets for false-positive analysis

To quantify the false-positives that arise when performing parametric tests on ISC data we analyze resting state fMRI data from three publicly available datasets.

Two datasets are from the Human Connectome Project (HCP) [17]: a first one consisting of 100 unrelated subjects and a second one of 518 subjects including twins and non-twins siblings (S500 release). The resting state data consists of 3T MR imaging data from healthy adult participants. These recordings include 1200 continuously acquired volumes for each participant, acquired using multi-band acceleration at a TR = 0.72 s (multi-band factor = 8) with a spatial resolution of 2 mm × 2 mm × 2 mm [17]. Additional behavioral and demographic measures on the individual participants can be downloaded from the project website [17].

We include a third dataset from the Autism Brain Imaging Data Exchange (ABIDE) [18]. From this dataset we selected 100 resting state recordings from healthy participants used as control group. The data was collected for ABIDE by the New York University Langone Medical Center. This dataset includes a minimum of 176 continuously acquired volumes for each participant acquired at TR = 2 s with a spatial resolution of 3 mm × 3 mm × 3 mm [19]. The inclusion of datasets aquired at different TR values (0.72 s for HCP and 2 s for ABIDE) ensures that our findings apply to a broader range of datasets.

For both the HCP and the ABIDE data, we downloaded and used the preprocessed datasets. Details on the specific preprocessing can thus be found in [19, 20]. We did not apply additional processing prior to the inter-subject correlation analysis. Briefly, the preprocessing that had been applied to these datasets involved standard pipelines combining FSL [21], FreeSurfer [22], Connectome Workbench [23], AFNI [24] and ANTS [25]. The imaging data from the ABIDE dataset had been registered to a MNI template with isotropic voxel size of 3 mm and spatially smoothed with a 6 mm FWHM Gaussian kernel. The HCP dataset had been registered on a MNI template with isotropic voxel size of 2 mm without smoothing. It should be noted, that the HPC500 dataset contains siblings that could have brain activity that is more similar than that of non-siblings. We however ignored this fact in our analyses.

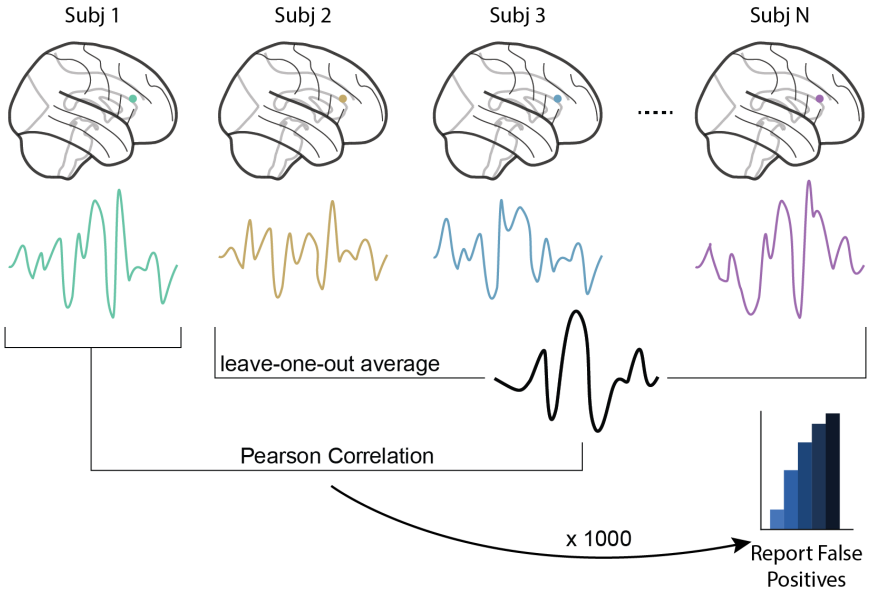

Figure 1: The basic procedure we followed for each subsample of N participants was to calculate leave-one-out ISC values throughout the brain based on resting state fMRI data. Within each voxel and subsample, this involved taking one subject out (here Subj 1), calculate the average time course of all the other subjects (here Subj 2 to N), and calculate the person correlation between the time course of the one subject against the average time course of all others. This is done for each of the subjects, correlating their time course against the average of all others. A parametric t-test is then performed to assess whether the N r-values differ from zero, after z-transformation. Because this is done for each voxel in the brain, we then apply a variety of thresholding methods typically used in fMRI ($p_{\mathrm{FWE}} < 0.05$, $q_{\mathrm{FDR}} < 0.05$, and $p_{\mathrm{unc}} < 0.001$). This procedure is repeated for 1000 random sub samples of N subjects to yield an estimate of the false positive rate for that threshold.

2.2. False-positive analysis strategy

For each of these three sets, we perform 1000 analyses by randomly subsampling a certain number of subjects. We subsample groups of $N = 10, 20, 30, 40, 50$ subjects, with the aim of emulating realistic numbers of participants in an fMRI experiment. Having defined a sample of N subjects, we extract a time series from each participant. The HCP dataset consists of functional files of approximately 1 Gb per subject containing 1200 volumes. To limit the memory resources required to carry our false positive analyses, we extracted 200 contiguous volumes and perform leave-one-out ISC across these 4D volumes. From the 1200 volumes available for each subject we extract the 200 in the middle (volumes [500:699]). To ascertain that our findings are not restricted to this arbitrary choice of 200 volumes, we also replicated our analyses for the sample size most typical of contemporary fMRI studies ($N = 20$) using the following numbers of consecutive volumes: 10, 20, 50, 100, 200 and 400 (see Supplementary Material 1). This analysis came to the same conclusions as those presented in the main paper. For the ABIDE dataset we used the full 4D volume, i.e., volumes [0:175] from the 176 volumes available for each participant.

The calculation of leave-one-out ISC consists in computing the Pearson correlation ρ between the time course recorded in subject i at voxel $\mathbf{v}$, with the average time course of the remaining $N - 1$ subjects at the same voxel $\mathbf{v}$. We perform this computation with python code based on the package BrainIAK [3]. As a result of the ISC computation, we obtain N brain maps containing the leave-one-out ISC of each subject. To assess the Type I error rate of parametric tests on this group-level data, we use SPM, one of the most widely used software for statistical parametric mapping [9], to perform second level t-tests based on the r-to-z transformed ISC maps of each participant. For automatizing the analysis for different sample-sizes of subjects and different datasets, we use the python framework provided by nipype [26]. Importantly, for a given dataset users could perform the same analysis using the SPM GUI, and would only require python code for a single step of the analysis: the ISC computation, which is very simple when using BrainIAK, as explained in [3].

2.2.1. One sample t-test

To simulate a situation in which a researcher investigates what brain regions systematically respond to a long stimulus (e.g., a movie) in a given group of participants, we first perform a t-test against the null hypothesis $\rho = 0$, separately for all the different group sizes ($N = 10, \ldots, 50$). We r-to-z transform the ISC maps before applying the t-test to ensure normality. We perform this statistical test with and without correction for multiple comparison, and with different minimum cluster sizes. Specifically we used $\alpha_{FWE} < 0.05$ corrected at the voxel level and $q_{FDR} < 0.05$ corrected at the voxel level. We also present data thresholded at $p_{unc} < 0.001$. For all of these we test minimum cluster sizes of 1, 5 and 20 voxels.

2.2.2. Paired sample t-test

To simulate a situation in which a scientist investigates what brain regions respond more systematically to one stimulus than to another, we conducted a matched-pair t-test between the r-to-z transformed ISC calculated in a first segment and the ISC calculated in a non-overlapping second segment of the resting state data of each participant. We tested this separately for all the different group sizes ($N = 10, \ldots, 50$). For the HCP datasets, we extract two non-overlapping 200 volume segments from each participant (segment 1 included volumes [100:299], segment 2 volumes [500:699]). For ABIDE, we extracted two 80 volume segments [1:80] and [97:176] from each participant's fMRI recording.

2.2.3. Two sample t-test

Finally, to simulate a situation in which a scientist investigates what brain regions respond more systematically to a stimulus in one group A of N participants ($N = 10, \ldots, 50$) than in another group B of N participants (e.g., patients vs. controls) we also performed a two sample t-test. We randomly selected two non-overlapping subgroups A and B of N participants from each dataset, computed the leave-one-out ISC as described for the one-sample t-test, and then perform a two-sample t-test testing the null hypothesis $\rho_A = \rho_B$ across the two subgroups, again after r-to-z transforming the data for normality.

For each type of test, for each of the 1000 random subsamples, and for each combination of sample size, correction type and cluster size threshold, we stored every cluster of voxels surviving the chosen statistical threshold. The proportion of subsamples with at least one significant cluster is then used to estimate the effective Type I error associated with each scenario, and is compared against the 5% to be expected if parametric tests do control adequately for Type I error.

2.3. True-positive analysis

To assess the sensitivity of the parametric tests, i.e., their ability to detect ISC where there is significant correlation, we used data from a previous experiment where we detected significant ISC values across a group of 19 subjects watching two video clips with an empathic or detached approach [16]. These were healthy, right-handed, native English speaking males recruited from the city of Amsterdam, The Netherlands. Individuals with a history of neurological/psychiatric symptoms or medication usage were excluded from participation. Here we limit our analysis to testing whether $\rho = 0$ for the first clip in the detached condition. The recording associated with this clip consists of 361 volumes with a TR of 2 s. We show the results from the same thresholding approaches described in the previous section for the one-sample t-test. We then compare the results with that of the non-parametric bootstrap method considered optimal for leave-one-out ISC metrics [11].

3. Results

3.1. False-positive analysis

Figure 2 illustrates that for the one-sample t-test testing whether ISC is significant in one condition, FDR and FWE corrections for multiple comparisons combined with a minimum cluster size of $k = 5$ voxels or more control false discovery rates to less than 5%. This is true across all three resting state datasets we explored. Inflated Type I errors only occur if no minimum cluster size correction is applied at all, showing that only very small clusters falsely survive this voxelwise threshold. Actually, when applying $k = 5$ as is standard practice in neuroimaging, both FDR ($q = 0.05$) and FWE ($\alpha = 0.05$) corrections are overly conservative, as less than 5% false positives are detected. The same holds true for the stricter requirement on the minimum cluster size $k = 20$, for which no false positives at all are detected. Unsurprisingly, applying no correction for multiple comparisons, and only applying an uncorrected $p_{unc} < 0.001$ threshold is the least conservative of the three analyzed cases. In this case both $k = 0$ and $k = 5$ do not sufficiently control for Type I error over the entire brain. However, $k = 20$ suffices to control Type I error, illustrating again, that only relatively small clusters survive.

The false positive rates in Fig. 2a-c increase as a function of sample size across the three different datasets. Because the number of participants in all datasets was finite, we wondered whether this increase in false positives as a function of N might relate to the fact that as N increases, the subsamples will increasingly overlap. However, if this were the case, the effect would be stronger in datasets with fewer participants (HCP S100 and ABIDE) than in the dataset with more participants

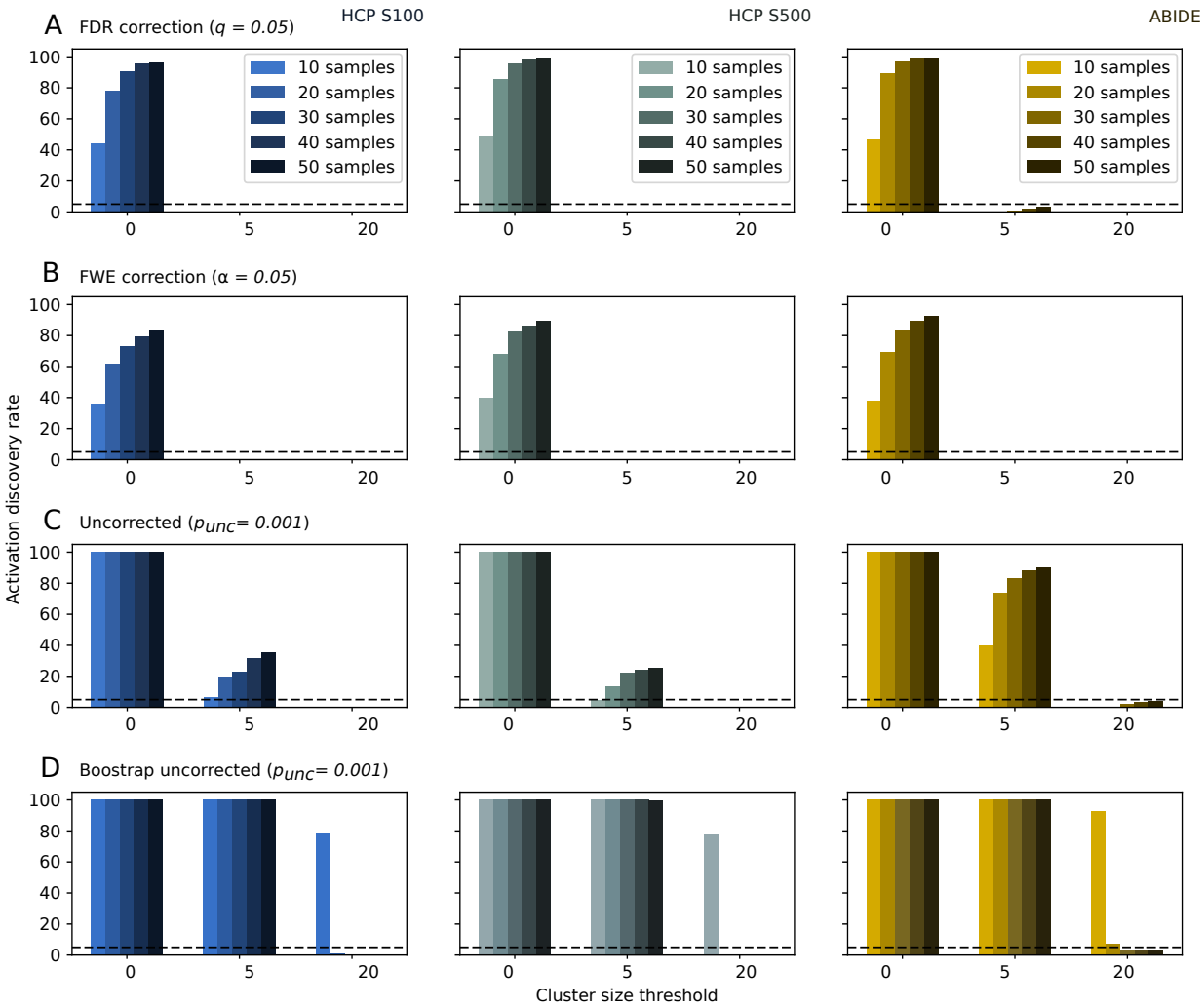

Figure 2: One-sample t-test: percentage of analyses displaying at least one significant cluster for FDR (a), FWE (b), uncorrected (c) and using a non-parametric bootstrap ISC (d). The first column displays the percentage of subsamples with at least one significant cluster as a function of sample size and threshold on the minimum cluster size for the HCP S100 dataset. The second column shows the same for HCP S500 and the third for ABIDE. Note that TR=0.72 s voxelsize=2 mm for the HCP datasets, TR=2 s voxelsize=3 mm for the ABIDE dataset.

(HCP S500), but this was not the case. In addition, we performed two simulations where 1000 samples are either drawn from the same data set of 100 simulated brains without systematic correlations, or are drawn independently. Neither of these simulations shows increased false discovery rates with increasing N (see Supplementary Material 2).

To assess whether this effect is specific for our parametric tests, we performed the same analyses using a non-parametric bootstrap algorithm in brainiak [11]. Because the issue of multiple comparisons has been less systematically addressed for such non-parametric approaches, here we present the results of the bootstrap method without correction for multiple comparison (Fig. 2d), at the same threshold $p_{unc} < 0.001$ used for uncorrected parametric analyses in (Fig. 2c). Directly comparing the non-parametric and parametric approaches at the same uncorrected threshold shows that the non-parametric test is slightly less conservative. In particular, with a cluster size threshold of $k = 5$ the fraction of analyses with at least one false positive is always close to 100% in the bootstrap method and consistently higher than in the parametric analyses, particularly for small samples ($N = 10$). Finally, when using larger k values $k = 20$, parametric tests always control Type I error appropriately below 5% while the bootstrap method leads to inflated Type I error rate for small samples ($N = 10$).

A salient observation is also that while the parametric tests experience increased false positive rates as N increases, the reverse is true for the non-parametric tests, which control Type I errors more effectively as sample size is increased [11]. This highlights the fact that also non-parametric tests such as the bootstrap algorithm used in this case have their weaknesses, particularly when only small samples are available.

Figs. 3 and 4 illustrate the parametric results obtained for scenarios in which two conditions are compared within a given participant and when two different groups of participants are compared, respectively. In these scenarios, we observe that Type I errors are more frequent than for the one sample t-test. However, the thresholds that successfully control for Type I errors for the one-sample t-tests presented above still control Type I errors quite successfully. A FDR correction with $q = 0.05$ in combination with a minimum cluster size of $k = 5$ still controls false positive rates to levels close to the intended 5%. For the paired t-tests only the largest sample size $N = 50$ leads to a discovery rate slightly larger than 5% for the connectome datasets. For the two-sample t-tests it exceeds the 5% limit only for the ABIDE dataset with a sample size of 10, 20 or 30. When a cluster size threshold of $k = 20$ is used with FDR correction, no false positives are observed at all. For FWE correction with $\alpha = 0.05$ and $k = 5$, no false positives are ever observed. Again, even without correction for multiple comparisons, a $p_{unc} < 0.001$ threshold combined with $k = 20$ ensures that the false positive rate remains below 5% in all cases.

Overall, this additional data is consistent with the observations made for the one-sample t-test displayed in Fig. 2. However, while false positive rates increased monotonically with sample size for the one-sample t-test, this is less clear for the paired t-test and false positive rates actually decrease with increasing sample size in some instances for the two-sample t-test (Fig. 4).

3.2. True-positive analysis

After verifying that for a one-sample t-test FDR and FWE corrections at the voxel level with a minimum cluster size of $k = 5$, as well as $p_{unc} < 0.001$ with $k = 20$, do control Type I error rates to below 5%, we explored the relative sensitivity of these thresholding methods to detect ISC in a dataset in which participants did watch the same stimulus (using data published in [16] and described in the Methods Section). In this dataset, we expect time-locked brain

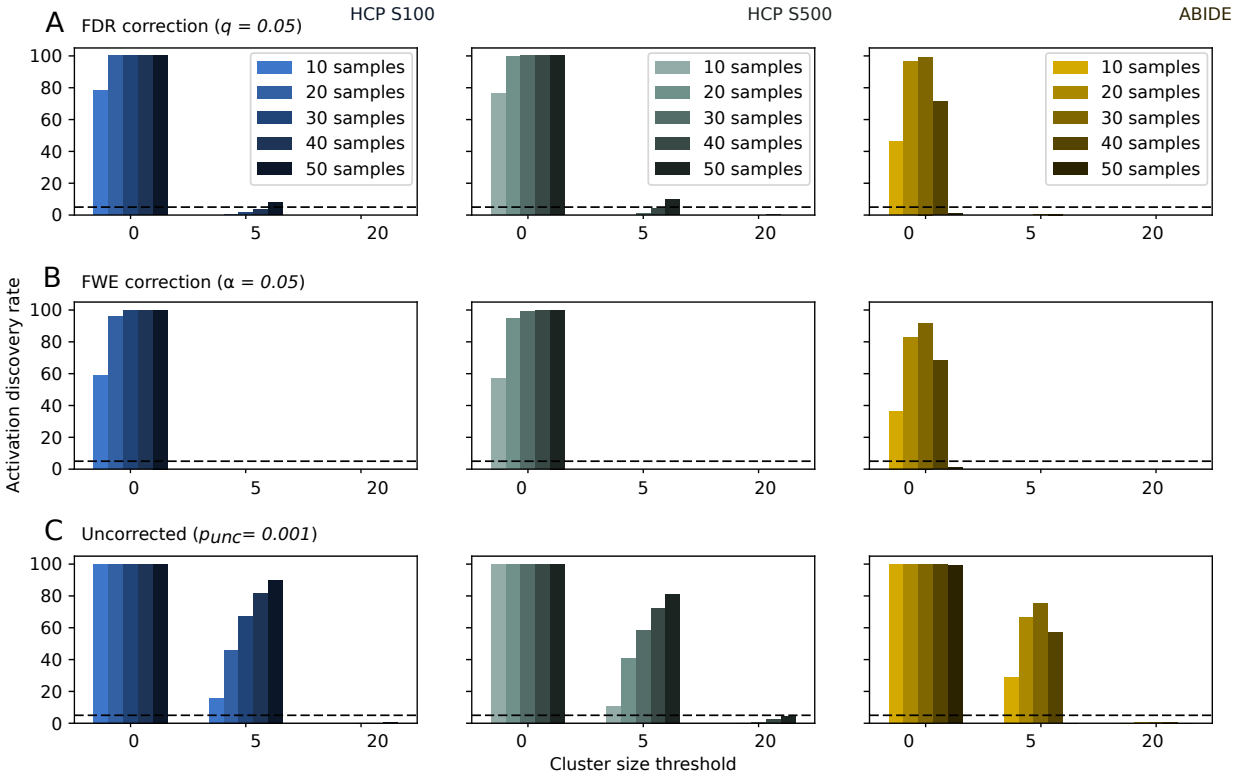

Figure 3: Paired t-test: percentage of analyses displaying at least one significant cluster for FDR, FWE and uncorrected thresholds. (a) Percentage of subsamples with at least one significant cluster as a function of sample size and threshold on the minimum cluster size for the HCP S100 dataset. (b) Same as (a) for HCP S500. (c) Same as (a) for ABIDE. Note that TR = 0.72 s voxelsize = 2 mm for a-b, and TR = 2 s voxelsize = 3 mm for c.

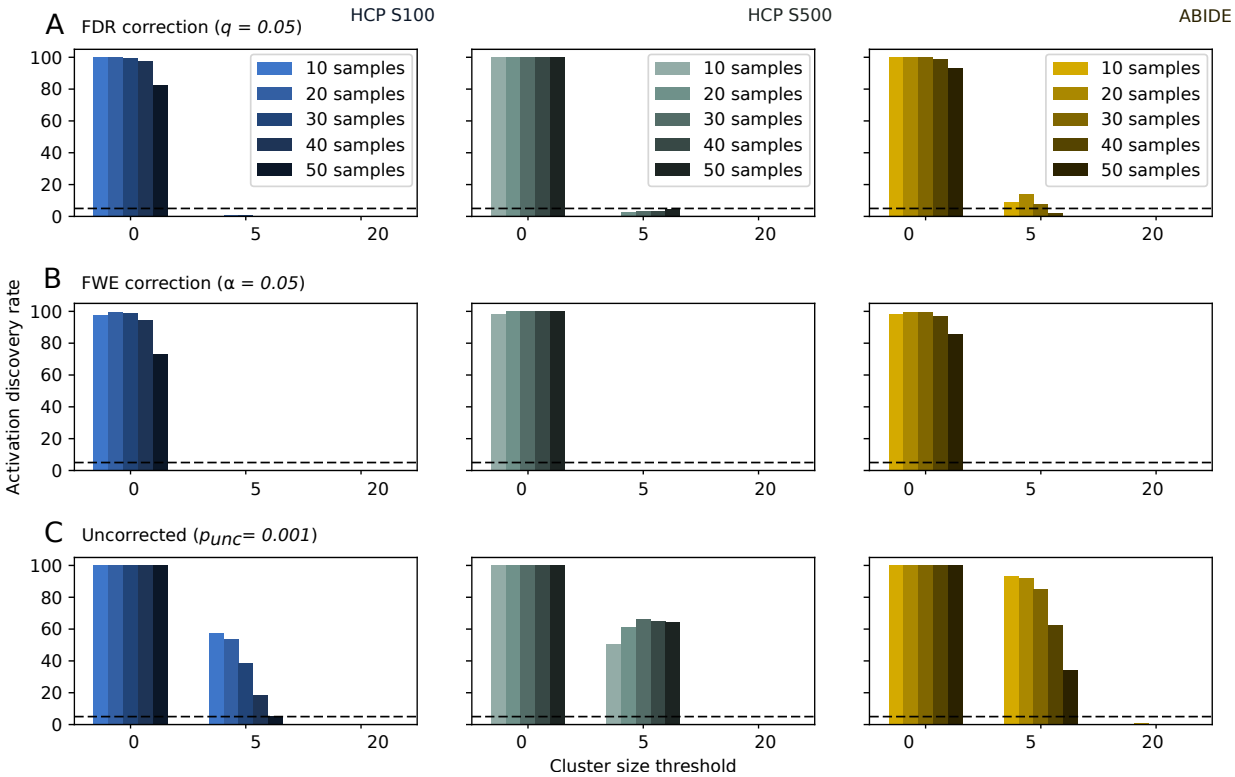

Figure 4: Two-sample t-test: percentage of analyses displaying at least one significant cluster for FDR, FWE and uncorrected thresholds. (a) Percentage of subsamples with at least one significant cluster as a function of sample size and threshold on the minimum cluster size for the HCP S100 dataset. (b) Same as (a) for HCP S500. (c) Same as (a) for ABIDE. Note that TR = 0.72 s voxelsize = 2 mm for a-b, and TR = 2 s voxelsize = 3 mm for c.

activity across a wide network of regions associated with sensory, motor and affective processing [16]. In addition, we wished to compare how the spatial distribution of voxels with significant ISC using these parametric methods compare with that observed using the non-parametric bootstrap method. By eye, the results of the different thresholding methods (see Fig 5) seem to agree on where ISC is highest (as shown in the darkest colours), but to disagree in the extent of the networks that are detected as synchronizing their activity to the stimulus. Clearly, the FDR threshold $q_{FDR} = 0.05$ with $k = 5$ is the most permissive among the four algorithms (Fig. 5b), followed by the uncorrected parametric t-test $p_{unc} < 0.001$ with $k = 20$, which looks very similar to the non-parametric bootstrap method using the same $p_{unc} < 0.001$ with $k = 20$ threshold. Finally, the FWE thresholding approach is the most conservative among the four (Fig. 5a).

To further quantify these qualitative observations, we quantified the correspondence between the voxels detected as suprathreshold according to those four thresholding methods. This was done using Cohen's κ to compare the spatial distribution of significant voxels, and using a Venn diagram to quantify the proportion of voxels that overlap across the methods. We found that every pair of methods yields a significant value of Cohen's κ (p-value < 0.001), however there are substantial differences among the methods. The bootstrap and uncorrected parametric method showed the highest agreement ($\kappa = 0.7$) while the FDR and FWE methods showed the largest differences ($\kappa = 0.1$). These values become easier to interpret in the context of the Venn diagram displayed in Fig. 5f. In particular, it becomes evident that bootstrap and parametric methods identify the same voxels, with highly overlapping maps obtained when applying the same threshold ($p_{unc} < 0.001$). At this threshold, only 3% of the voxels (purple in Fig. 5f) identified by the bootstrap are not identified by the parametric t-test. However, it is notable that the $q_{FDR} < 0.05$, $k = 5$, which controls false positive rates in our resting state datasets to the desired 5% proportion, is nevertheless substantially more sensitive than the bootstrap method, and detects approximately twice the voxels to synchronize with the stimulus than the bootstrap method.

Another important aspect to keep in mind when comparing the result of parametric and non-parametric methods is the metric used in the resulting maps. When performing parametric tests it is customary to plot t-values [9], which are a metric of evidence against the null hypothesis that increases with $\sqrt{N}$. In contrast, the default estimator for a bootstrap method is the median r-value [11], which reflects the synchrony of the median participant with the average time-course of the other participants rather than the evidence for that voxel to show above chance correlation across the group of participants. We compared the spatial distribution of the t-statistic computed by SPM and the median r-value computed with brainIAK in Fig. 6 in a glass brain representation. The different metrics exhibit overall similar features, but they do differ in several details, and the location of the highest values are notably different. When computing a Kendall's τ to quantify the agreement among the two unthresholded maps, we obtain a value of $\tau = 0.62$, which is significantly larger than zero ($p < 0.001$), but also quite far from 1, confirming that although these metrics are related, they do convey slightly different information. Darker regions in the t-map identify regions more likely to be significantly higher than zero based on their distribution across participants, regardless of their absolute value. The darker regions in the median-r map are the ones that exhibit the higher synchronization for a typical subject, regardless of the width of the distribution across participants. Of course, t-maps could be used to determine significant voxels, and the actual median or average r-

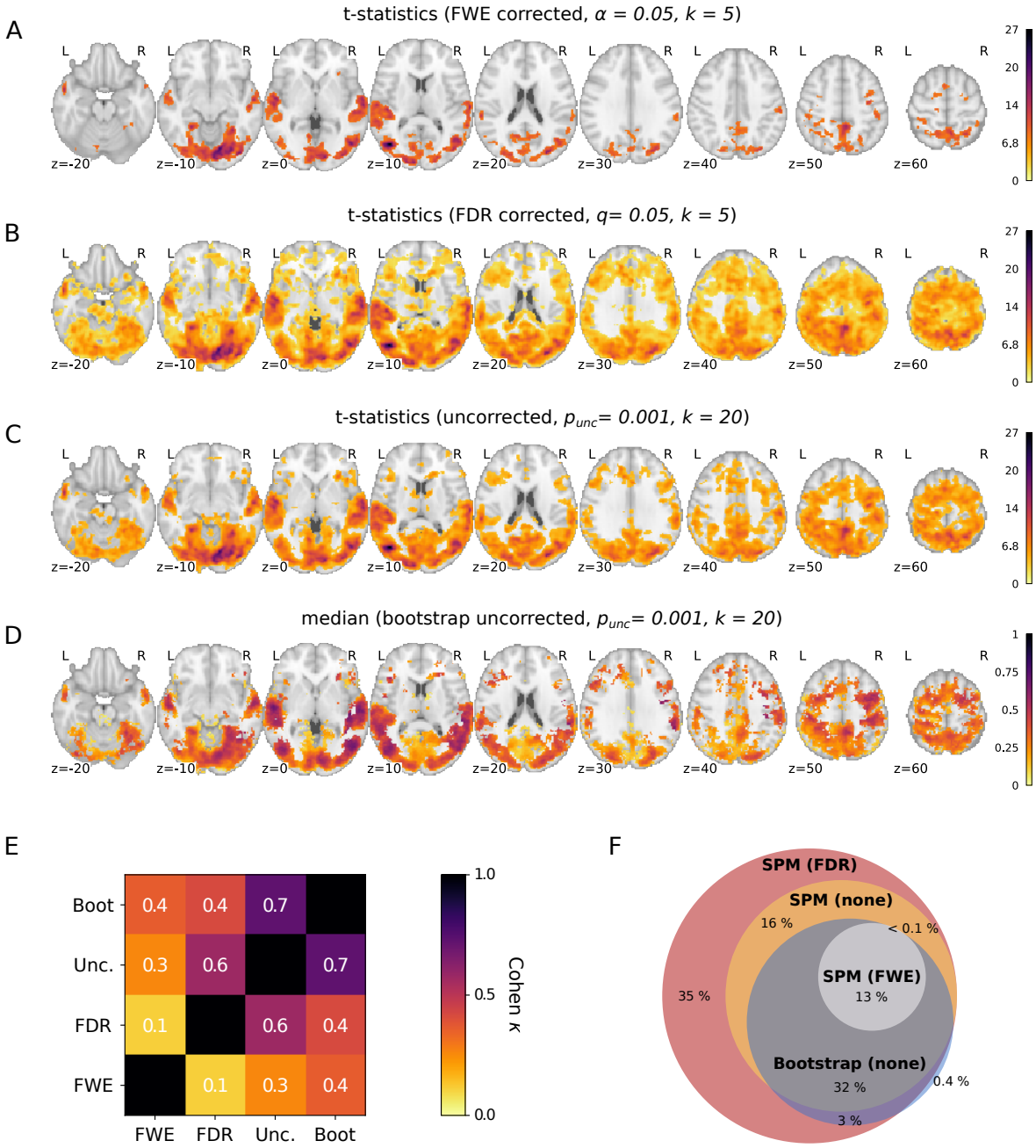

Figure 5: Significant ISC detected while participants watched a movie as a function of thresholding method. (a) Results from a parametric t-test thresholded using $\alpha_{FWE} < 0.05$, $k = 5$. (b) Results from a parametric t-test thresholded using $q_{FDR} < 0.05$, $k = 5$. (c) Results from a parametric t-test thresholded using $p_{unc} < 0.001$, $k = 20$. (d) Results from the non-parametric bootstrap technique thresholded at $p_{unc} < 0.001$, $k = 20$. (e) Pairwise comparison of the spatial pattern of the thresholded maps shown in a-d using Cohen's κ. For this analyses subthreshold voxels are coded as 0 and suprathreshold voxels based on their t or median r-value. (f) Venn diagram illustrating the overlap between the suprathreshold voxels in the parametric maps obtained using SPM and the bootstrap map obtained using BrainIAK. Percentages are expressed relative to those obtained using the most lenient threshold (i.e. $q_{FDR} < 0.05$ in SPM).

value in those voxels could then be illustrated in the colour maps. Conversely, also in the non-parametric framework one could estimate a quantity similar to a t-statistic, for example by dividing the average ISC value by the width of the confidence interval that can be estimated by the bootstrap method [11].

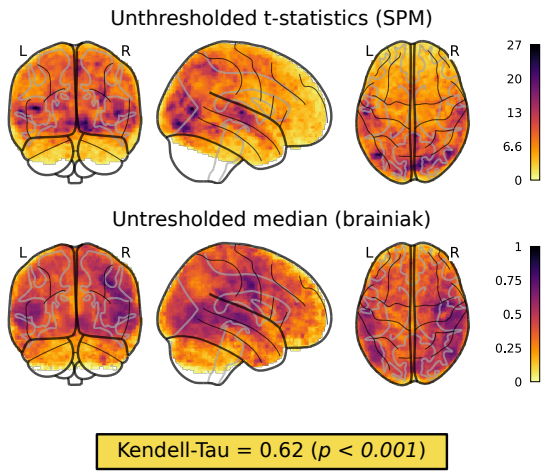

Figure 6: Visual comparison between the t-statistics computed with SPM (top) and the median r computed with brainIAK (bottom) on the leave-one-out ISC data of the dataset of empathy control. The Kendall's τ value indicated below the brains indicates the association between the values in the two unthresholded maps. Results are shown on a glass brain

4. Discussion

In this work we present a false positive rate analysis for Inter-Subject Correlation based on publicly available resting state datasets. We verified that parametric tests can be a valuable way of assessing significance in the context of leave-one-out ISC. In particular we explored whether parametric tests appropriately control Type I error, i.e., how often they would falsely detect significant synchronization in resting state datasets in which such synchronization should not be detected.

Our results indicate that parametric tests control Type I errors when testing hypotheses on ISC data in ways that are quite similar to the way they control Type I errors in standard subtraction designs. In particular, FDR corrections ($q_{FDR} < 0.05$) at the voxel level paired with a minimum cluster size of $k = 20$ voxels robustly protected against the false detection of synchrony in all the resting state datasets and types of comparisons we explored. This was true despite the actual differences in voxel-size and smoothing used in the HCP and ABIDE. Importantly, this thresholding method was the most sensitive in detecting synchrony in datasets in which participants watch a common stimulus. FWE corrections at the voxel level ($\alpha_{FWE} < 0.05$) also controls Type I error rates very effectively, even when a minimum cluster size of only k=5 voxels is used. However, FWE correction was the least sensitive of the methods we explored, when detecting synchrony in participants viewing the same movie. FDR correction may thus offer a better balance between Type I and Type II error, and may be preferable when scientists investigate smaller effect sizes.

Interestingly, using uncorrected ($p_{unc} < 0.001$) thresholds in combination with a $k = 20$ minimum cluster size also controls Type I error while providing reasonable sensitivity, but this approach is perhaps the least principled of the approaches, as it will fail to adapt to the size of the search-volume that is being investigated.

Our study has a number of limitations. Firstly, finding that a method controls against Type I error in a particular dataset does not ensure that it controls against Type I error in all datasets. That we have tested the method in two fundamentally different resting state datasets that have different TR values (0.72 s vs. 2 s), different isotropic voxel-size (2 mm vs. 3 mm), different smoothing and were acquired using different acquisition schemes (multiband acceleration in the HCP but not the ABIDE) together with the fact that we varied our sample sizes from 10 to 50 participants however suggests that parametric tests robustly control for Type I error across a reasonable range

of parameters. In addition, we also explored whether our results depend on the length of the neuroimaging data we consider, but our results are robust over a wide range of segment length (see Supplementary Material 1). Secondly, the assumption that resting state data does not include any synchronization between participants might be doubtful. Because data was taken at the same moment relative to the start of data acquisition in all participants, some synchronization may occur across participants in reaction to sensory and affective processes triggered by being scanned in the same fMRI scanner. That parametric tests control against finding synchronization across participants despite this possibility is reassuring. Thirdly, we interpret finding more voxels using FDR correction in a dataset in which participants view a movie as increased sensitivity. Unfortunately, we do not have external evidence that these additional voxels contain "true" activation. However, the pattern of additional voxels identified is similar to those reported in previous studies of natural vision.

Importantly, our results only apply to leave-one-out ISC. Naturally, parametric tests cannot be straightforwardly used on pairwise ISC data, where the numbers of degrees of freedom corresponding to the number of measured subjects is dramatically different from the number of pairs that one can generate with such subjects.

It is our experience that many neuroimagers stay away from using ISC-based neuroimaging paradigms because they shy away from the relatively complex coding necessary to perform robust statistical inference using non-parametric tools. We trust that our results will allow a broad community of neuroimagers trained in using and interpreting parametric second level neuroimaging statistics implemented in SPM or FSL to leverage the conceptual power of ISC. This is because when using parametric statistics, ISC data can be analysed using a pipeline almost identical to that of traditional fMRI analysis, including standard preprocessing, and standard second level analyses. Only the actual calculation of the ISC needs to be done using less standard code, but this step can be easily implemented using toolboxes we provide in [3]. Our findings also naturally extend to more sophisticated forms of ISC, including inter-subject functional connectivity, that rely on very similar, leave-one-out computations [3]. In addition to making analyses easier to run, the use of parametric statistics implemented in standard packages in particular paves the way to the systematic usage of sensitive but robust corrections for multiple comparisons (i.e., FWE and FDR) that adapt to the number of voxels and the spatial smoothness of a dataset. Such adaptive corrections are not currently readily available for non-parametric approaches. Finally, non-parametric approaches for ISC can become prohibitively demanding in terms of computations as sample sizes increase, and parametric tests can thus further improve the efficiency of data analyses. While we by no means wish to argue against non-parametric testing of ISC data, we hope that our validation of parametric tests will provide scientist with an easier-to-implement yet powerful and sensitive method to complement existing non-parametric tests.

Funding

The work was funded Dutch Research Council (NWO) VIDI grant (452-14-015) to VG and VICI grant (453-15-009) to CK.

Acknowledgments

We thank Christiaan Becker and Pierre-Louis Bazin for performing a pilot analyses that preceeded and informed this study, Ritu Bhandari for insights on the use of SPM, Alessandra Nostro and Leonardo Cerliani for help to identify appropriate resting state datasets.

References

[1] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, R. Malach, Intersubject synchronization of cortical activity during natural vision., Science (New York, N.Y.) 303 (5664) (2004) 1634–40. doi:10.1126/science.1089506.

[2] C. Honey, T. Thesen, T. Donner, L. Silbert, C. Carlson, O. Devinsky, W. Doyle, N. Rubin, D. Heeger, U. Hasson, Slow Cortical Dynamics and the Accumulation of Information over Long Timescales, Neuron 76 (2) (2012) 423–434. doi:10.1016/j.neuron.2012.08.011.

[3] S. A. Nastase, V. Gazzola, U. Hasson, C. Keysers, Measuring shared responses across subjects using intersubject correlation, Social Cognitive and Affective Neuroscience 14 (6) (2019) 667–685. doi:10.1093/scan/nsz037.

[4] J. M. Lahnakoski, E. Glerean, I. P. Jääskeläinen, J. Hyönä, R. Hari, M. Sams, L. Nummenmaa, Synchronous brain activity across individuals underlies shared psychological perspectives, NeuroImage 100 (2014) 316–324. doi:10.1016/j.neuroimage.2014.06.022.

[5] R. M. Thomas, T. De Sanctis, V. Gazzola, C. Keysers, Where and how our brain represents the temporal structure of observed action., NeuroImage 183 (2018) 677–697. doi:10.1016/j.neuroimage.2018.08.056.

[6] G. J. Stephens, L. J. Silbert, U. Hasson, Speaker-listener neural coupling underlies successful communication., Proceedings of the National Academy of Sciences of the United States of America 107 (32) (2010) 14425–30. doi:10.1073/pnas.1008662107.

[7] U. Hasson, A. A. Ghazanfar, B. Galantucci, S. Garrod, C. Keysers, Brain-to-brain coupling: a mechanism for creating and sharing a social world., Trends in cognitive sciences 16 (2) (2012) 114–21. doi:10.1016/j.tics.2011.12.007.

[8] A. Eklund, T. E. Nichols, H. Knutsson, Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates, Proceedings of the National Academy of Sciences 113 (28) (2016) 7900–7905. doi:10.1073/pnas.1602413113.

[9] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, T. E. Nichols, Statistical parametric mapping: the analysis of functional brain images, Elsevier, 2011.

[10] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. Behrens, H. Johansen-Berg, P. R. Bannister, M. D. Luca], I. Drobnjak, D. E. Flitney, R. K. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. D. Stefano], J. M. Brady, P. M. Matthews, Advances in functional and structural mr image analysis and implementation as fsl, NeuroImage 23 (2004) S208 – S219, mathematics in Brain Imaging. doi:https://doi.org/10.1016/j.neuroimage.2004.07.051.

[11] G. Chen, Y.-W. Shin, P. A. Taylor, D. R. Glen, R. C. Reynolds, R. B. Israel, R. W. Cox, Untangling the relatedness among correlations, part i: Nonparametric approaches to inter-subject correlation analysis at the group level, NeuroImage 142 (2016) 248 – 259. doi:https://doi.org/10.1016/j.neuroimage.2016.05.023.

[12] G. Chen, P. A. Taylor, Y.-W. Shin, R. C. Reynolds, R. W. Cox, Untangling the relatedness among correlations, part ii: Inter-subject correlation group analysis through linear mixed-effects modeling, NeuroImage 147 (2017) 825 – 840. doi:https://doi.org/10.1016/j.neuroimage.2016.08.029.

[13] K. J. Worsley, S. Marrett, P. Neelin, A. C. Vandal, K. J. Friston, A. C. Evans, A unified statistical approach for determining significant signals in images of cerebral activation, Human Brain Mapping 4 (1) (1996) 58–73. doi:10.1002/(SICI)1097-0193(1996)4:1<58::AID-HBM4>3.0.CO;2-O.

[14] J.-P. Kauppi, J. Pajula, J. Tohka, A versatile software package for inter-subject correlation based analyses of fMRI, Frontiers in Neuroinformatics 8 (2014) 2. doi:10.3389/fninf.2014.00002.

[15] N. Schaworonkow, D. A. Blythe, J. Kegeles, G. Curio, V. V. Nikulin, Power-law dynamics in neuronal and behavioral data introduce spurious correlations, Human Brain Mapping 36 (8) (2015) 2901–2914. doi:10.1002/hbm.22816.

[16] K. B. Jimenez], A. Abdelgabar, L. D. Angelis], L. McKay, C. Keysers, V. Gazzola, Changes in brain activity following the voluntary control of empathy, NeuroImage (2020) 116529doi:https://doi.org/10.1016/j.neuroimage.2020.116529.

[17] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, et al., The human connectome project: a data acquisition perspective, Neuroimage 62 (4) (2012) 2222–2231. doi:https://doi.org/10.1016/j.neuroimage.2012.02.018.

[18] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, et al., The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism, Molecular psychiatry 19 (6) (2014) 659–667. doi:https://doi.org/10.1038/mp.2013.78.

[19] C. Craddock, Y. Benhajali, C. Chu, F. Chouinard, A. Evans, A. Jakab, B. S. Khundrakpam, J. D. Lewis, Q. Li, M. Milham, et al., The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives, Frontiers in Neuroinformatics 7 (2013). doi:10.3389/

conf.fninf.2013.09.00041.

[20] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. V. Essen], M. Jenkinson, The minimal preprocessing pipelines for the human connectome project, NeuroImage 80 (2013) 105 – 124, mapping the Connectome. doi:https://doi.org/10.1016/j.neuroimage.2013.04.127.

[21] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, S. M. Smith, Fsl, NeuroImage 62 (2) (2012) 782 – 790, 20 YEARS OF fMRI. doi:https://doi.org/10.1016/j.neuroimage.2011.09.015.

[22] A. M. Dale, B. Fischl, M. I. Sereno, Cortical surface-based analysis: I. segmentation and surface reconstruction, NeuroImage 9 (2) (1999) 179 – 194. doi:https://doi.org/10.1006/nimg.1998.0395.

[23] D. S. Marcus, M. P. Harms, A. Z. Snyder, M. Jenkinson, J. A. Wilson, M. F. Glasser, D. M. Barch, K. A. Archie, G. C. Burgess, M. Ramaratnam, M. Hodge, W. Horton, R. Herrick, T. Olsen, M. McKay, M. House, M. Hileman, E. Reid, J. Harwell, T. Coalson, J. Schindler, J. S. Elam, S. W. Curtiss, D. C. V. Essen], Human connectome project informatics: Quality control, database services, and data visualization, NeuroImage 80 (2013) 202 – 219, mapping the Connectome. doi:https://doi.org/10.1016/j.neuroimage.2013.05.077.

[24] R. W. Cox, Afni: software for analysis and visualization of functional magnetic resonance neuroimages, Computers and Biomedical research 29 (3) (1996) 162–173. doi:10.1006/cbmr.1996.0014.

[25] B. B. Avants, P. Yushkevich, J. Pluta, D. Minkoff, M. Korczykowski, J. Detre, J. C. Gee, The optimal template effect in hippocampus studies of diseased populations, NeuroImage 49 (3) (2010) 2457 – 2466. doi:https://doi.org/10.1016/j.neuroimage.2009.09.062.

[26] K. J. Gorgolewski, O. Esteban, D. G. Ellis, M. P. Notter, E. Ziegler, H. Johnson, C. Hamalainen, B. Yvernault, C. Burns, A. Manhães-Savio, D. Jarecka, C. J. Markiewicz, T. Salo, D. Clark, M. Waskom, J. Wong, M. Modat, B. E. Dewey, M. G. Clark, M. Dayan, F. Loney, C. Madison, A. Gramfort, A. Keshavan, S. Berleant, B. Pinsard, M. Goncalves, D. Clark, B. Cipollini, G. Varoquaux, D. Wassermann, A. Rokem, Y. O. Halchenko, J. Forbes, B. Moloney, I. B. Malone, M. Hanke, D. Mordom, C. Buchanan, W. M. PAULI, J. M. Huntenburg, C. Horea, Y. Schwartz, R. Tungaraza, S. Iqbal, J. Kleesiek, S. Sikka, C. Frohlich, J. Kent, M. Perez-Guevara, A. Watanabe, D. Welch, C. Cumba, D. Ginsburg, A. Eshaghi, E. Kastman, S. Bougacha, R. Blair, B. Acland, A. Gillman, A. Schaefer, B. N. Nichols, S. Giavasis, D. Erickson, C. Correa, A. Ghayoor, R. Küttner, C. Haselgrove, D. Zhou, R. C. Craddock, D. Haehn, L. Lampe, J. Millman, J. Lai, M. Renfro, S. Liu, J. Stadler, T. Glatard, A. E. Kahn, X.-Z. Kong, W. Triplett, A. Park, C. McDermottroe, M. Hallquist, R. Poldrack, L. N. Perkins, M. Noel, S. Gerhard, J. Salvatore, F. Mertz, W. Broderick, S. Inati, O. Hinds, M. Brett, J. Durnez, A. Tambini, S. Rothmei, S. K. Andberg, G. Cooper, A. Marina, A. Mattfeld, S. Urchs, P. Sharp, K. Matsubara, D. Geisler, B. Cheung, A. Floren, T. Nickson, N. Pannetier, A. Weinstein, M. Dubois, J. Arias, C. Tarbert, K. Schlamp, K. Jordan, F. Liem, V. Saase, R. Harms, R. Khanuja, K. Podranski, G. Flandin, D. Papadopoulos Orfanos, I. Schwabacher, D. McNamee, M. Falkiewicz, J. Pellman, J. Linkersdörfer, J. Varada, F. Pérez-García, A. Davison, D. Shachnev, S. Ghosh, nipy/nipype: Release 0.13.1 (May 2017). doi:10.5281/zenodo.581704.

- Supplementary Material -
Significance tests for Leave-One-Out Inter-Subject Correlations in fMRI

L. De Angelis[a,*], V. Gazzola[a,b,**], C. Keysers[a,b,**]

[a] *Social Brain Lab, Netherlands Institute for Neuroscience, Royal Netherlands Academy of Arts and Sciences, Meibergdreef 47, 1105 BA, Amsterdam, The Netherlands*
[b] *Department of Psychology, University of Amsterdam, 1018 WV Amsterdam, The Netherlands*

1. False positives as a function of temporal window

Our choice of using 200 TRs from the HCP datasets was somewhat arbitrary. To test the impact this decision has on our Type I error, we repeated our one-sample t-test analysis for a fixed group size of $N = 20$ but varying the number of volumes included from 10 to 400 TRs. Figure 1 shows that FWE ($p_{FWE} < 0.05$, $k = 0.05$) and FDR ($q = 0.05$, $k = 5$) protect against Type I error regardless of the number of volumes included. When detecting nonzero false positive rates (using no minimum cluster size), we notice an increase in Type I error up to approximately 100 TRs, followed by a decrease as larger segments are considered. These results are also important for applications in which short temporal windows are necessary, because analyses attempt to identify moments within a longer stimulus in which a particular network is selectively engaged.

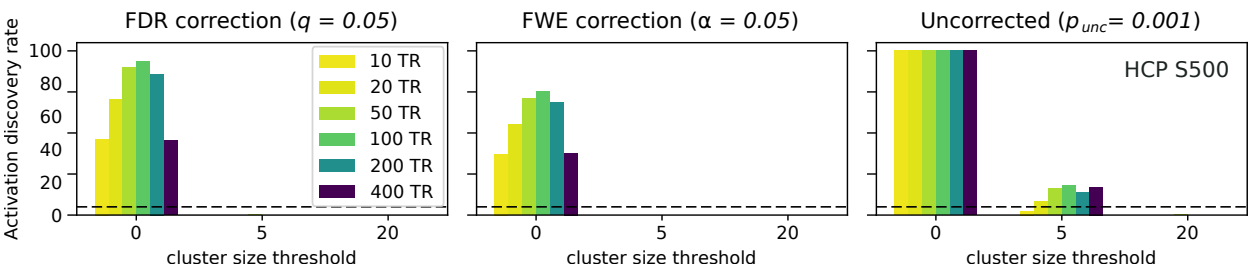

Figure 1: Analysis of the false-positive rate for one-sample t-test on intersubject correlation data at varying the temporal window used for the computation of the leave-one-out isc.

*Corresponding author
**These authors contributed equally to this work
Email address: c.keysers@nin.knaw.nl (C. Keysers)

2. Simulations for estimating the effects of sampling from a finite dataset

To check whether sampling 1000 times with different sample size from a finite dataset can lead to artifacts in the false positive rate measured in the main text, we performed numerical simulations. We created two scenarios. In the first scenario, similar to our experimental framework, we generate a dataset of 100 simulated brain data and compute the ISC by sampling this data 1000 times per each sample size. In the second scenario, we compute ISC starting from simulated data which is independently generated at each of the 1000 iterations. To mimic resting state data, we generate simulated data as Gaussian random noise with a finite temporal and spatial correlation. From Fig. 2 we deduce that having a limited dataset consisting of 100 independent measurements does not produce visible effects on the false positive analysis, as the results for the two different scenarios (top and bottom panels, respectively) are consistent with each other.

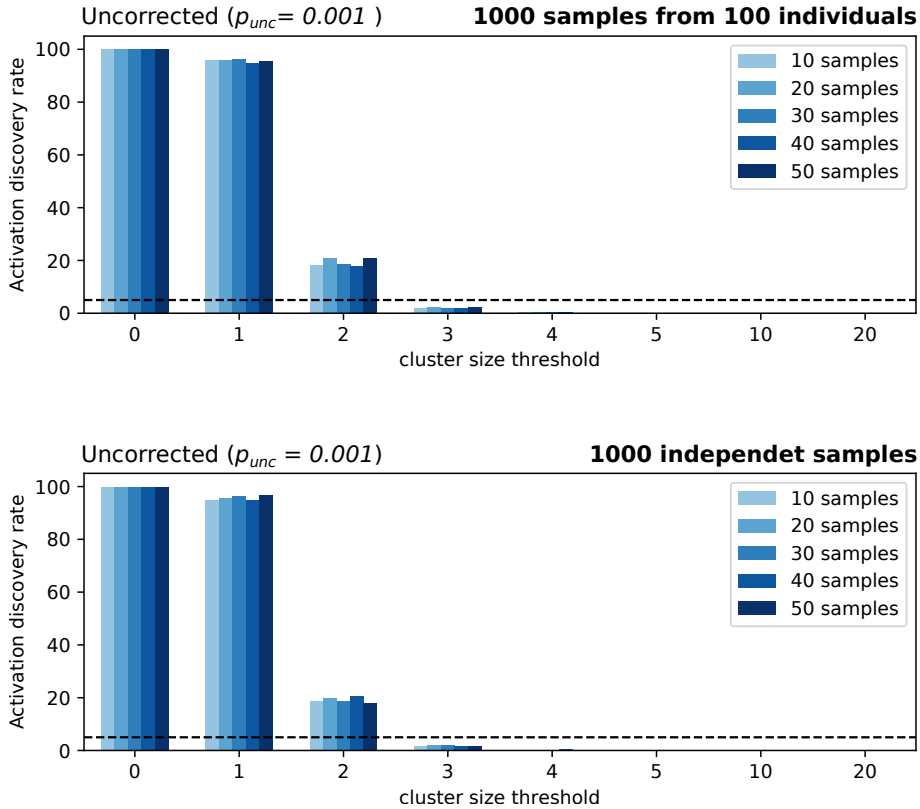

Figure 2: One sample t-test for simulated brain activity. We compute the ISC starting from simulated activity consisting in gaussian random noise with a finite temporal and spatial correlation. In the top panel a scenario similar to the experimental one, in which only 100 brains are available for random sampling. In the bottom panel a scenario in which the brain activities are independently generated for every step contributing to the overall statistics.