Cluster Failure Revisited: Impact of First Level Design and Data Quality on Cluster False Positive Rates

Anders Eklund[a,b,c], Hans Knutsson[a,c], Thomas E. Nichols[d,e,f]

[a] *Division of Medical Informatics, Department of Biomedical Engineering, Linköping University, Linköping, Sweden*
[b] *Division of Statistics & Machine Learning, Department of Computer and Information Science, Linköping University, Linköping, Sweden*
[c] *Center for Medical Image Science and Visualization (CMIV), Linköping University, Linköping, Sweden*
[d] *Big Data Institute, University of Oxford, Oxford, United Kingdom*
[e] *Wellcome Trust Centre for Integrative Neuroimaging (WIN-FMRIB), University of Oxford, Oxford, United Kingdom*
[f] *Department of Statistics, University of Warwick, Coventry, United Kingdom*

Abstract

Methodological research rarely generates a broad interest, yet our work on the validity of cluster inference methods for functional magnetic resonance imaging (fMRI) created intense discussion on both the minutia of our approach and its implications for the discipline. In the present work, we take on various critiques of our work and further explore the limitations of our original work. We address issues about the particular event-related designs we used, considering multiple event types and randomisation of events between subjects. We consider the lack of validity found with one-sample permutation (sign flipping) tests, investigating a number of approaches to improve the false positive control of this widely used procedure. We also discuss the implications of our work on the fMRI literature as a whole, estimating that at least 10% of the fMRI studies have used the most problematic cluster inference method (P = 0.01 cluster defining threshold), and how individual studies can be interpreted in light of our findings. These additional results underscore our original conclusions, on the importance of data sharing and thorough evaluation of statistical methods on realistic null data.

Keywords: fMRI, false positives, cluster inference, permutation, physiological noise

1. Introduction

In our previous work (Eklund et al., 2016) we used freely available resting state functional magnetic resonance imaging (fMRI) data to evaluate the validity of standard fMRI inference methods, in particular finding that some variants of cluster size inference were troubled by inflated false positive rates. Our paper has generated intense discussions regarding cluster inference in fMRI (Eklund et al., 2017; Cox et al., 2017a,b; Kessler et al., 2017; Flandin & Friston, 2017; Kaundinya et al., 2018a), on the validity of using resting state fMRI data as null data (Slotnick, 2016, 2017; Nichols et al., 2017), how the spatial resolution can affect parametric cluster inference (Mueller et al., 2017), how to obtain residuals with a Gaussian spatial autocorrelation function (Kaundinya et al., 2018b), as well as how different MR sequences can change the spatial autocorrelation function and thereby cluster inference (Wald & Polimeni, 2017). Furthermore, some of our results have been reproduced and extended (Cox et al., 2017b; Kessler et al., 2017; Flandin & Friston, 2017; Mueller et al., 2017), using the same freely available fMRI data (Biswal et al., 2010; Poldrack et al., 2013) and our processing scripts available on github[1]. Cluster based methods have now also been evaluated for surface-based group analyses of cortical thickness, surface area and volume (using FreeSurfer) (Greve & Fischl, 2018), with a similar conclusion that the non-parametric permutation test showed good control of the familywise error rate (FWE) for all settings, while traditional Monte Carlo methods fail to control FWE for some settings.

Non-parametric group inference is now available in the AFNI function 3dttest++ (Cox et al., 2017a,b), meaning that the three most common fMRI softwares now all support non-parametric group inference (SPM users can use the SnPM toolbox (http://warwick.ac.uk/snpm), and FSL users can use the randomise function (Winkler et al., 2014)). Permutation tests cannot only be applied to simple designs such as one-sample and two-sample t-tests, but to virtually any regression model with independent errors (Winkler et al., 2014). To increase statistical power, permutation tests enable more advanced thresholding approaches (Smith & Nichols, 2009) as well as the use of multivariate approaches with complicated null distributions (Friman et al., 2003; Stelzer et al., 2013).

Here, we focus on ironing out the wrinkles in our original study. Specifically, we investigate if randomizing the

*Corresponding author
Email address: anders.eklund@liu.se (Anders Eklund)

[1] https://github.com/wanderine/ParametricMultisubjectfMRI

first level design over subjects affects the group level FWE, and explain why the non-parametric one-sample t-test failed for a number of parameter combinations (especially for the Oulu data). Finally, we offer a discussion of the interpretation of our findings and their impact on the literature as a whole.

2. Methods

2.1. New paradigms

The event related paradigms (E1, E2) used in our study were criticized by some for not being realistic designs, as only a single regressor was used (Slotnick, 2016) and the rest between the events was too short. The concern here is that this design may have a large transient at the start (due to the delay of the hemodynamic response function) and then only small variation (due to the short interstimulus interval), which may be overly-sensitive to transients at the start of the acquisition (Figure 1 (a), however, shows this is not really the case). Another criticism was that exactly the same task was used for all subjects (Flandin & Friston, 2017), meaning that our "false positives" actually reflect consistent pretend-stimulus-linked behavior over subjects. Yet another concern was if the first few volumes (often called dummy scans) in each fMRI dataset were included in the analysis or not[2], which can affect the statistical analyses. This last point we can definitively address, as according to a NITRC document[3], the first 5 time points of each time series were discarded for all data included in the 1000 functional connectomes project release.

To address the remaining concerns, we have made new analyses using two new event related paradigms, called E3 and E4. For both E3 and E4, two pretended tasks are used instead of a single task, and each first level analysis tests for a difference in activation between the two tasks. Additionally, the rest between the events is longer, see Figure 1 for a comparison between E1, E2, E3 and E4. For E4, the regressors are randomized over subjects, such that each subject has different random tasks (but the number of events is the same for all subjects), while for E3 the same regressors are used for all subjects.

First level analyses as well as group level analyses were performed as in the original study (Eklund et al., 2016), using the same data (Beijing, Cambridge, Oulu) from the 1000 functional connectomes project (Biswal et al., 2010). Familywise error rates were estimated for different levels of smoothing (4 mm - 10 mm), one-sample as well as two-sample t-tests, and two cluster defining thresholds ($p = 0.01$ and $p = 0.001$). Group analyses using 3dMEMA in AFNI were not performed, as the results for 3dttest++ and 3dMEMA were very similar in the original study (Eklund et al., 2016). Another difference is that cluster thresholding for AFNI was performed using the new ACF (autocorrelation function) option in 3dClustSim (Cox et al., 2017b), which uses a long-tail spatial ACF instead of a Gaussian one. To be able to compare the AFNI results for the new paradigms (E3, E4) and the old paradigms (B1, B2, E1, E2), the group analyses for the old paradigms were re-evaluated using the ACF option (note, that this ACF AFNI option still assumes a stationary spatial autocorrelation structure).

2.2. Valid FWE control for one-sample t-tests via permutation

The non-parametric permutation test (Winkler et al., 2014) produced nominal results for all two sample t-tests, but not for the one sample t-tests (Eklund et al., 2016), and the Oulu data were more problematic compared to Beijing and Cambridge. We investigated numerous ways to achieve nominal familywise error rates;

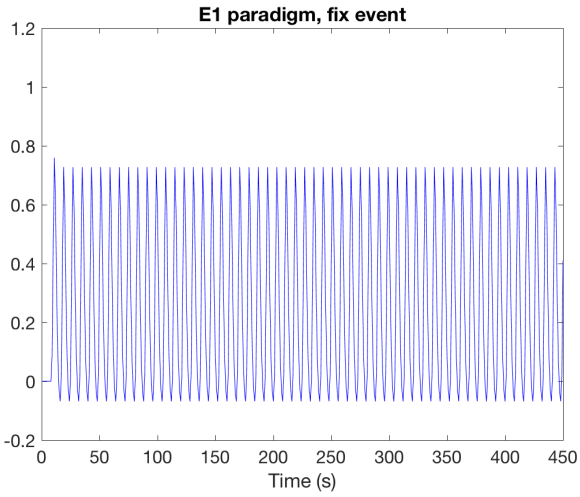
1. Applying the Yeo & Johnson (2000) transform (signed Box-Cox) to reduce skew (as the sign flipping test is based on an assumption of symmetric errors)
2. Using robust regression (in every permutation) to suppress the influence of outliers (Wager et al., 2005; Woolrich, 2008; Mumford, 2017)
3. Using two-sided tests instead of one-sided
4. Increasing the number of head motion regressors from 6 to 24
5. Using bootstrap instead of sign flipping, and
6. Including the global mean as a covariate in each first level analysis (Murphy et al., 2009; Murphy & Fox, 2017) (which is normally not done for task fMRI).

While some of these approaches resulted in nominal familywise error rates for a subset of the parameter combinations, no approach worked well for all settings and datasets.

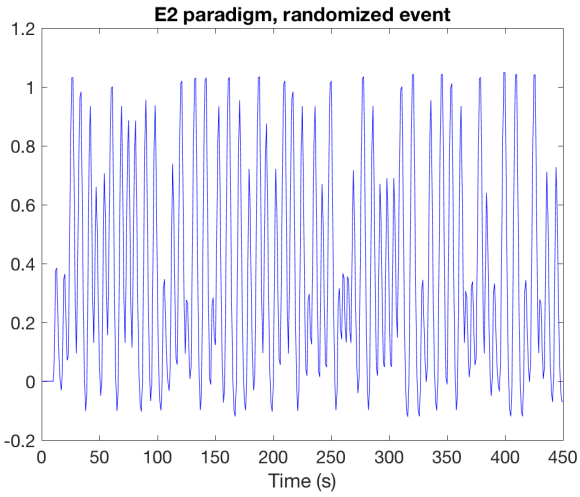
To understand the spatial distribution of clusters we created images of prevalence of false positive clusters, computed by summing the binary maps of FWE-significant clusters over the random analyses. In our original study, we found a rather structured spatial distribution for the two-sample t-test (supplementary Figure 18 in Eklund et al. (2016)), with large clusters more prevalent in the posterior cingulate. We have now created the same sort of maps for one-sample t-tests, with a small modification: to increase the number of clusters observed, we created clusters at a CDT of $p = 0.01$ for both increases and decreases on a given statistic map. As discussed in the Results, there appears to be physiological artefacts which ideally would be remediated by respiration or cardiac time series modelling (Glover et al., 2000; Lund et al., 2006; Birn et al., 2006; Chang & Glover, 2009; Bollmann et al., 2018), but unfortunately the 1000 functional connectomes datasets (Biswal et al., 2010) do not have these physiological recordings.

[2] http://www.ohbmbrainmappingblog.com/blog/keep-calm-and-scan-on, comment by John Ashburner

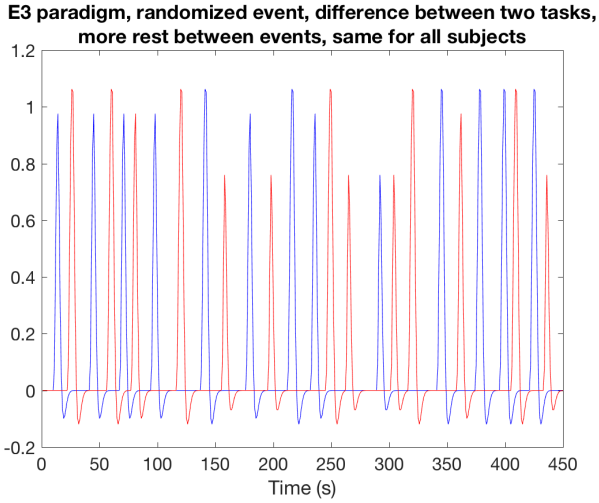
[3] http://www.nitrc.org/docman/view.php/296/716/fcon_1000_Preprocessing.pdf



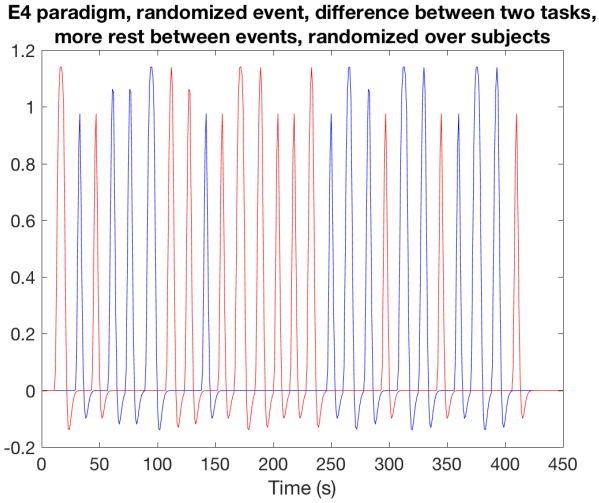
(a)



(b)



(c)



(d)

Figure 1: *A comparison of the paradigms used in the original paper (a) E1, b) E2) and the new paradigms used in this paper (c) E3, d) E4), for the Beijing datasets. A single task was used for both E1 and E2, while two pretended tasks where used for E3 and E4 (and all first level analyses tested for a difference in activation between these two tasks). Paradigms E1, E2 and E3 are the same for all subjects, while E4 is randomized over subjects.*

To suppress the influence of artifacts, we therefore instead applied ICA FIX in FSL (Salimi-Khorshidi et al., 2014; Griffanti et al., 2014, 2017) to all 499 subjects, to remove ICA components that correspond to noise or artifacts. We applied 4 mm of spatial smoothing for MELODIC (Beckmann & Smith, 2004), and used the classifier weights for standard fMRI data avalable in ICA FIX (trained for 5 mm smoothing). To use ICA FIX for 8 or 10 mm of smoothing would require more work, as it would be necessary to re-train the classifier. The cleanup was performed using the aggressive (full variance) option instead of the default less-aggressive option, and motion confounds were also included in the cleanup. First level analyses for B1, B2, E1, E2, E3 and E4 were performed for all 499 subjects after ICA FIX, with smoothing turned off. Group level analyses were finally performed using the non-parametric one-sample t-test available in BROCCOLI (Eklund et al., 2014).

3. Results

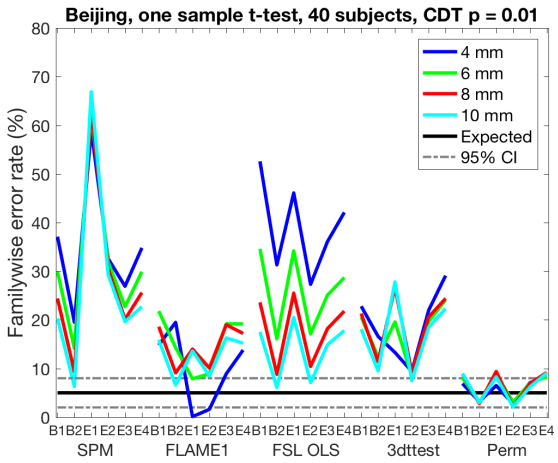
3.1. New paradigms

Figures 2 - 5 show estimated familywise error (FWE) rates for the two new paradigms (E3, E4), for 40 subjects in each group analysis. The four old paradigms (B1, B2, E1, E2) are included as well for the sake of comparison. In brief, the new paradigm with two pretend tasks (E3) does not lead to lower familywise error rates, compared to the old paradigms. Likewise, randomising task events over subjects (E4) has if anything worse familywise error rates compared to not randomising the task over subjects. As noted in our original paper, the very low FWE of FSL's FLAME1 is anticipated behavior when there is zero random effects variance. When fitting anything other than a one-sample group model this conservativeness may not hold; in particular, we previously reported on two-sample null analyses on task data, where each sample has non-zero but equal effects, and found that FLAME1's FWE was equivalent to that of FSL OLS (Eklund et al., 2016).

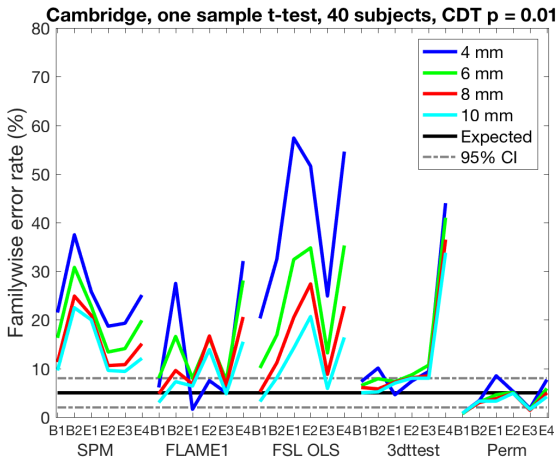
3.2. ICA denoising

Figure 6 shows familywise error rates for the non-parametric one-sample t-test, with and without first running ICA FIX for every subject, and the combination of ICA FIX with a two-sided test procedure. For the Beijing data, the familywise error rates are almost within the 95% confidence interval even without ICA FIX, and come even closer to the expected 5% after ICA FIX. For the Cambridge data, it is necessary to combine ICA FIX with a two-sided test to achieve nominal results (only using a two-sided test is not sufficient). For the Oulu data, neither ICA FIX in isolation nor in combination with two-sided inference was sufficient to bring false positives to a nominal rate. There is an ICA FIX threshold parameter, set by default to 20, we tried varying this threshold to extremes, and found that a value of 60 or 80 will finally produce nominal results for the Oulu data (not shown). We note, however, that this extreme setting is not recommended and if applied in practice would likely be removing a substantial amount of task-related variation.

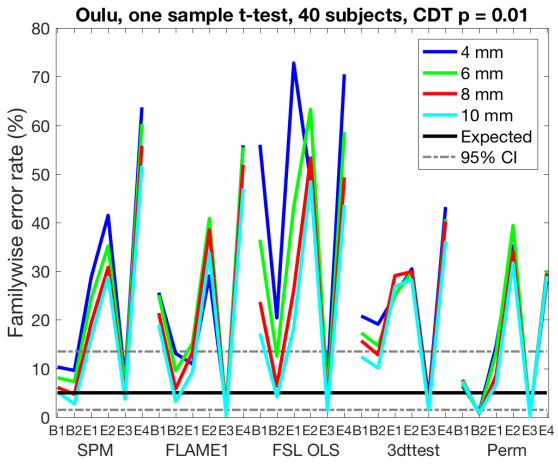
Figures 7 - 9 show cluster prevalence maps for group analyses with and without first running ICA FIX, for first level designs B2, E2 and E4. Using ICA FIX leads to false cluster maps that are more uniform across the brain, with fewer false clusters in white matter, and ICA FIX made the biggest difference for the Oulu data. While Beijing and Cambridge sites have a concentration of clusters in posterior cingulate, frontal and parietal areas, Oulu has more clusters and a more diffuse pattern. Further inspection of these maps suggested a venous artefact, and running a PCA on the Oulu activity maps for design E2 finds substantial variation in the sagittal sinus picked up by the task regressor (see Figure 10). The posterior part of the artefact is suppressed by ICA FIX, but a higher ICA FIX threshold is necessary to fully remove the artefact (not shown). Also see Figure 11 for activation maps from 5 Oulu subjects, analyzed with design E4. In several cases, significant activity differences between two random task regressors are detected close to the superior sagittal sinus, indicating a vein artefact.



(a)

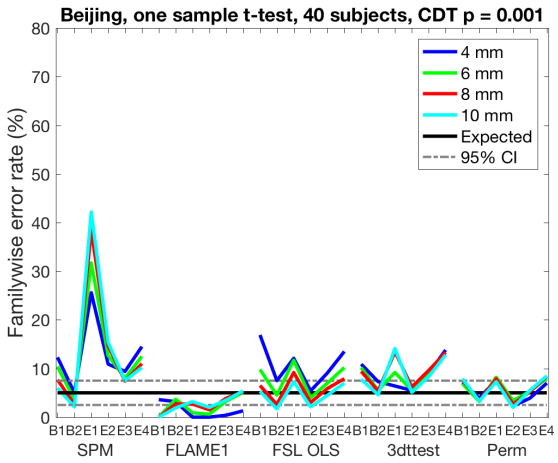


(b)

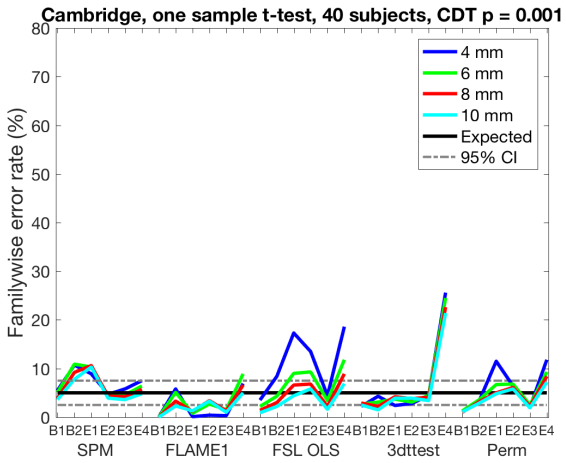


(c)

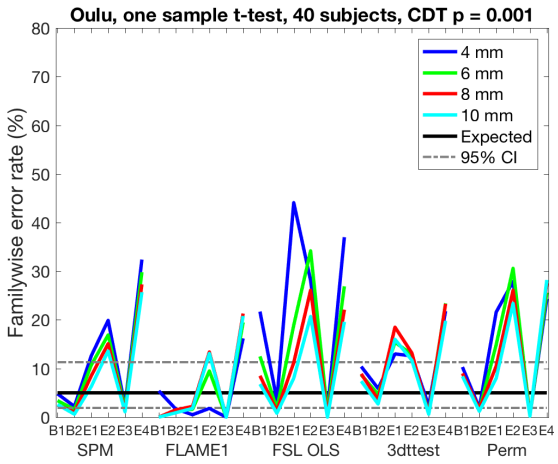
Figure 2: *Results for one sample t-test and cluster-wise inference using a cluster defining threshold (CDT) of p = 0.01, showing estimated familywise error rates for 4 - 10 mm of smoothing and six different activity paradigms (B1, B2, E1, E2, E3, E4), for SPM, FSL, AFNI and a permutation test. These results are for a group size of 40. Each statistic map was first thresholded using a CDT of p = 0.01, uncorrected for multiple comparisons, and the surviving clusters were then compared to a FWE-corrected cluster extent threshold,* $p_{FWE} = 0.05$. *The estimated familywise error rates are simply the number of analyses with any significant group activations divided by the number of analyses (1,000).* ***(a)*** *results for Beijing data* ***(b)*** *results for Cambridge data* ***(c)*** *results for Oulu data.*



(a)

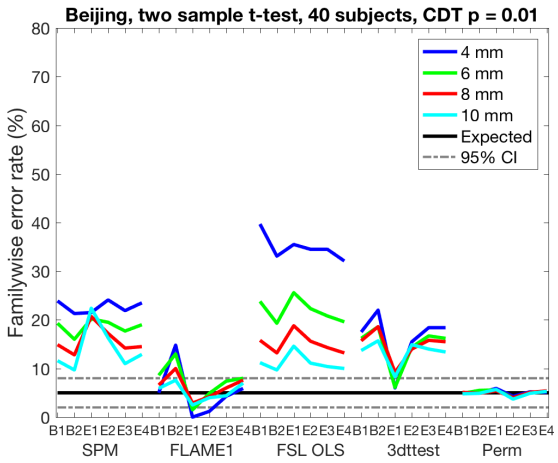


(b)

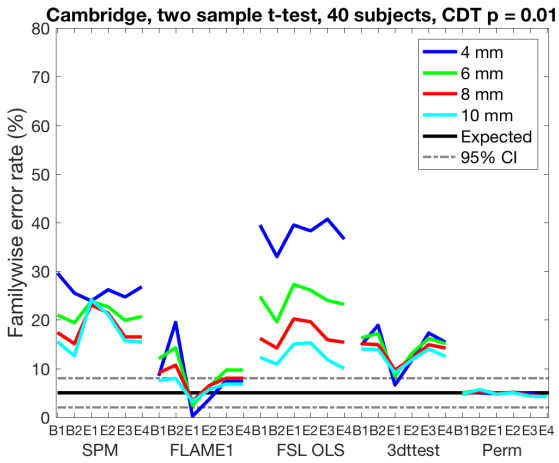


(c)

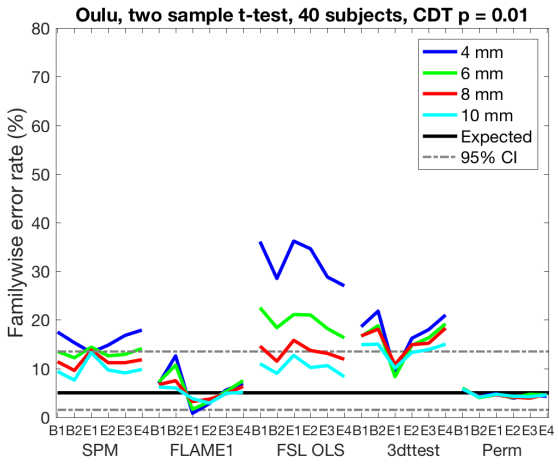
Figure 3: *Results for one sample t-test and cluster-wise inference using a cluster defining threshold (CDT) of p = 0.001, showing estimated familywise error rates for 4 - 10 mm of smoothing and six different activity paradigms (B1, B2, E1, E2, E3, E4), for SPM, FSL, AFNI and a permutation test. These results are for a group size of 40. Each statistic map was first thresholded using a CDT of p = 0.001, uncorrected for multiple comparisons, and the surviving clusters were then compared to a FWE-corrected cluster extent threshold,* $p_{FWE} = 0.05$. *The estimated familywise error rates are simply the number of analyses with any significant group activations divided by the number of analyses (1,000).* ***(a)*** *results for Beijing data* ***(b)*** *results for Cambridge data* ***(c)*** *results for Oulu data.*



(a)

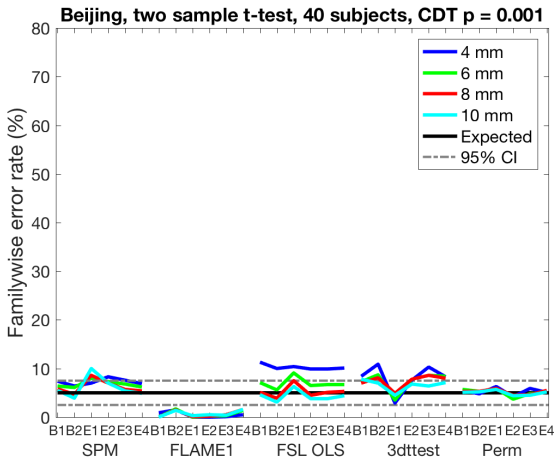


(b)

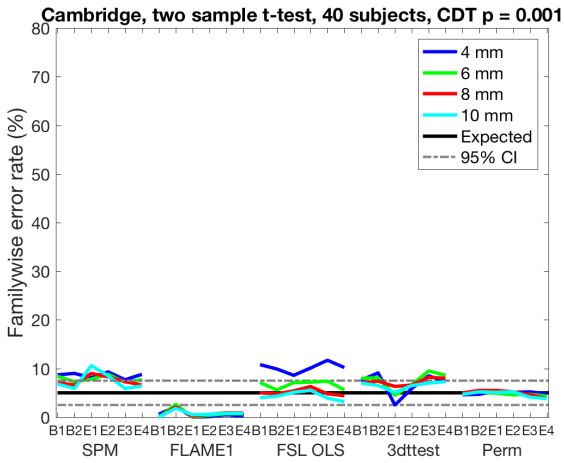


(c)

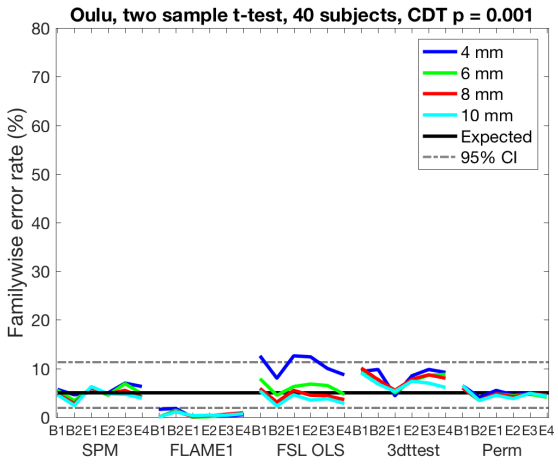
Figure 4: *Results for two sample t-test and cluster-wise inference using a cluster defining threshold (CDT) of $p = 0.01$, showing estimated familywise error rates for 4 - 10 mm of smoothing and six different activity paradigms (B1, B2, E1, E2, E3, E4), for SPM, FSL, AFNI and a permutation test. These results are for a group size of 20, giving a total of 40 subjects. Each statistic map was first thresholded using a CDT of $p = 0.01$, uncorrected for multiple comparisons, and the surviving clusters were then compared to a FWE-corrected cluster extent threshold, $p_{FWE} = 0.05$. The estimated familywise error rates are simply the number of analyses with any significant group activations divided by the number of analyses (1,000).* ***(a)*** *results for Beijing data* ***(b)*** *results for Cambridge data* ***(c)*** *results for Oulu data.*



(a)

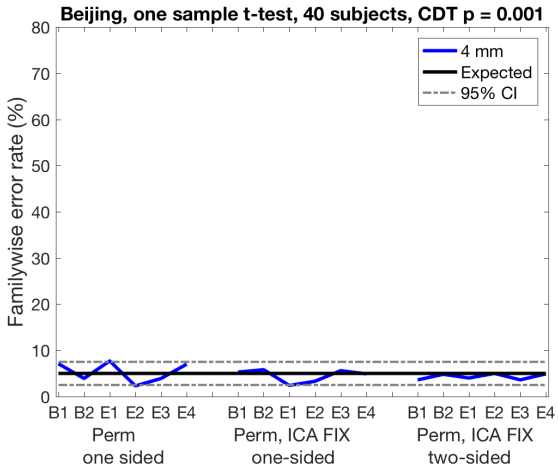


(b)

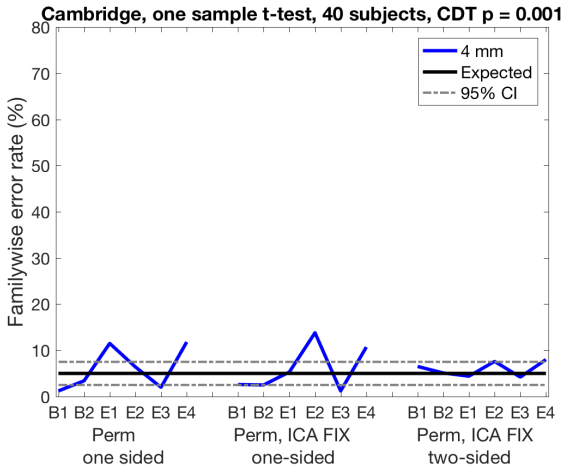


(c)

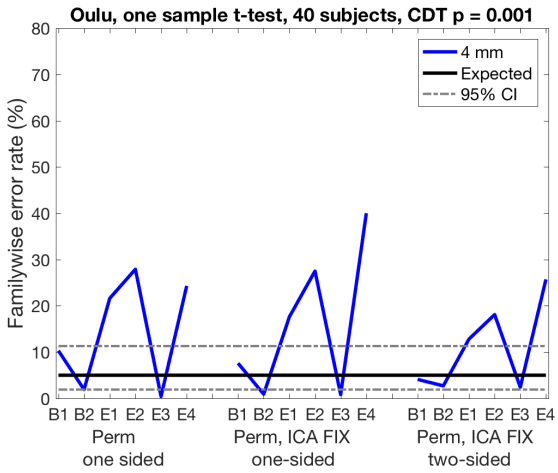
Figure 5: *Results for two sample t-test and cluster-wise inference using a cluster defining threshold (CDT) of $p = 0.001$, showing estimated familywise error rates for 4 - 10 mm of smoothing and six different activity paradigms (B1, B2, E1, E2, E3, E4), for SPM, FSL, AFNI and a permutation test. These results are for a group size of 20, giving a total of 40 subjects. Each statistic map was first thresholded using a CDT of $p = 0.001$, uncorrected for multiple comparisons, and the surviving clusters were then compared to a FWE-corrected cluster extent threshold, $p_{FWE} = 0.05$. The estimated familywise error rates are simply the number of analyses with any significant group activations divided by the number of analyses (1,000).* ***(a)*** *results for Beijing data* ***(b)*** *results for Cambridge data* ***(c)*** *results for Oulu data.*



(a)

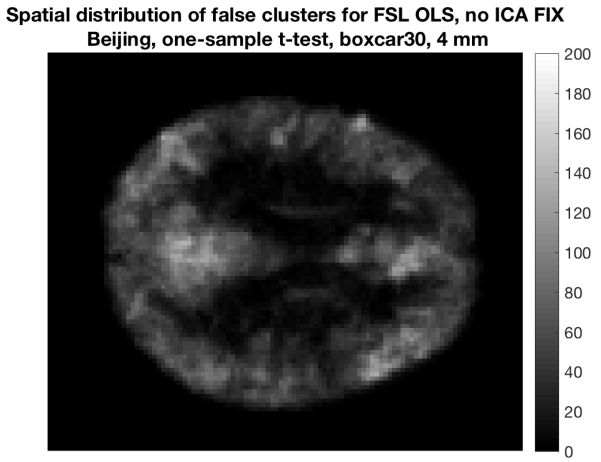


(b)

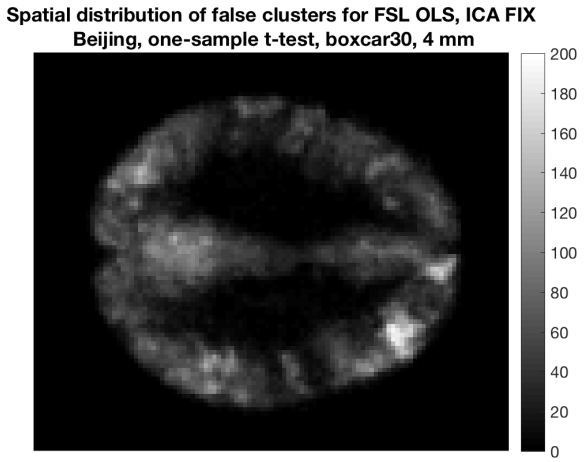


(c)

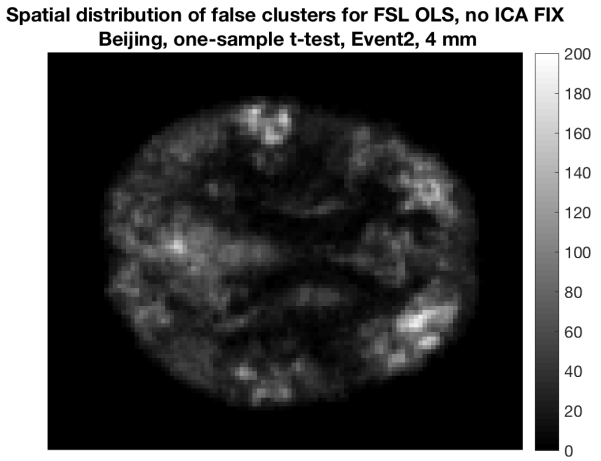
Figure 6: *Results for non-parametric one-sample t-tests for cluster-wise inference using a cluster defining threshold (CDT) of $p = 0.001$, after first running ICA FIX for every subject.* ***(a)*** *results for Beijing data* ***(b)*** *results for Cambridge data* ***(c)*** *results for Oulu data. Results are only shown for 4 mm smoothing, as the available classifier in ICA FIX is trained on data with 5 mm smoothing. Using ICA FIX for data with 8 or 10 mm smoothing would require re-training of the classifier. Only using a two-sided t-test is not sufficient for the Cambridge data. Increasing the ICA FIX threshold to 60 or 80, from the default 20, leads to nominal results for the Oulu data (not shown), but it is likely that such a high threshold removes signal components. Re-training the ICA FIX classifier specifically for Oulu may give better results.*



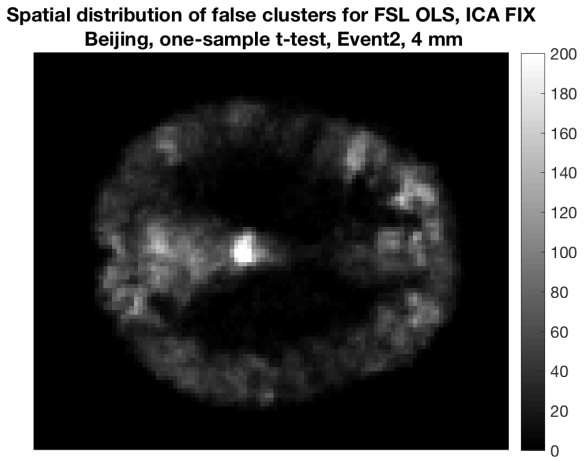
(a)



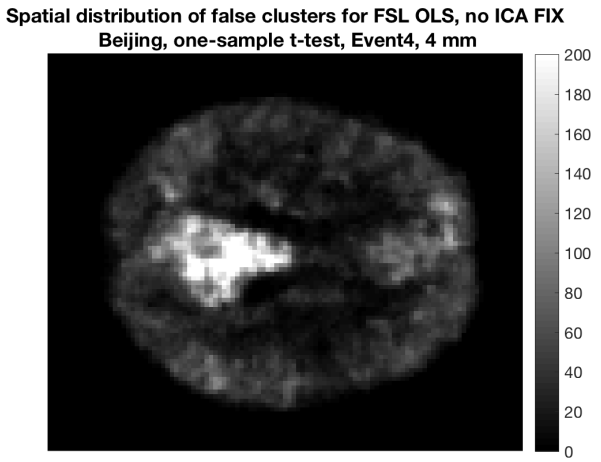
(b)



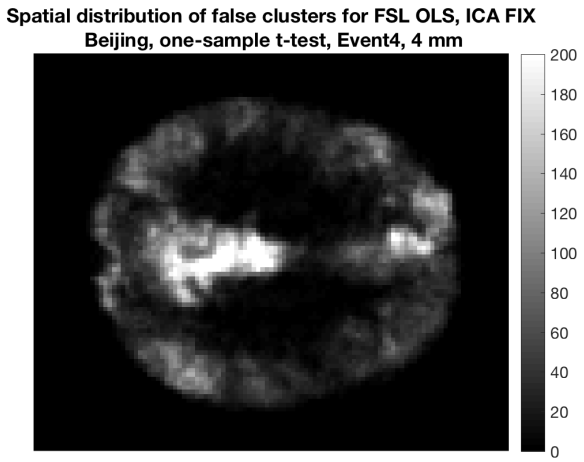
(c)



(d)

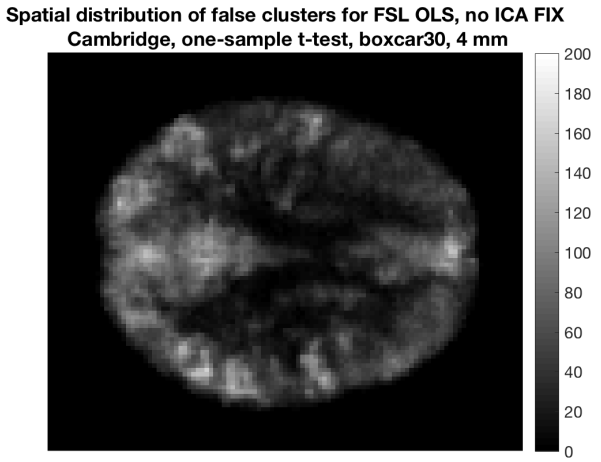


(e)

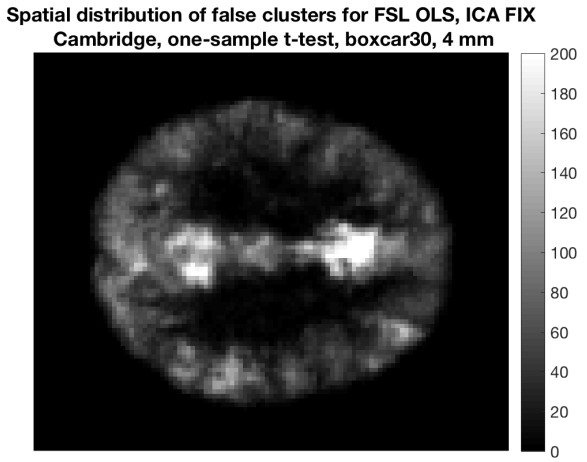


(f)

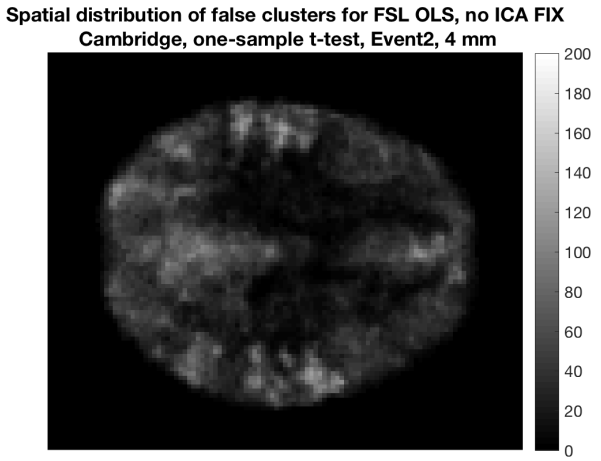
Figure 7: *The maps show voxel-wise incidence of false clusters for the Beijing data, for 3 of the 6 different first level designs (a,b) B2, (c,d) E2, (e,f) E4. Left: results without ICA FIX, Right: results after first running ICA FIX for every subject. Image intensity is the number of times, out of 10,000 random analyses, a significant cluster occurred at a given voxel (CDT p = 0.01) for FSL OLS. Each analysis is a one-sample t-test using 20 subjects.*



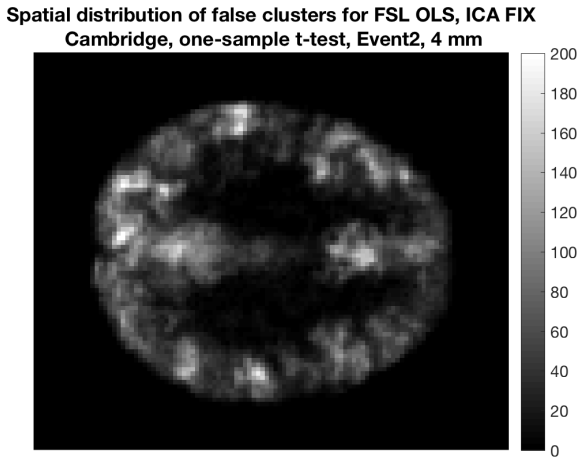
(a)



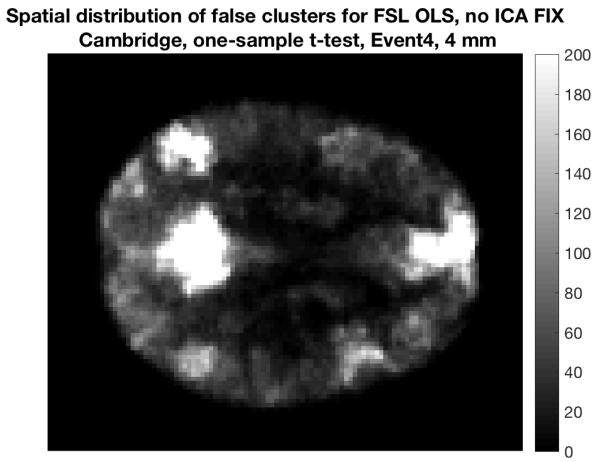
(b)



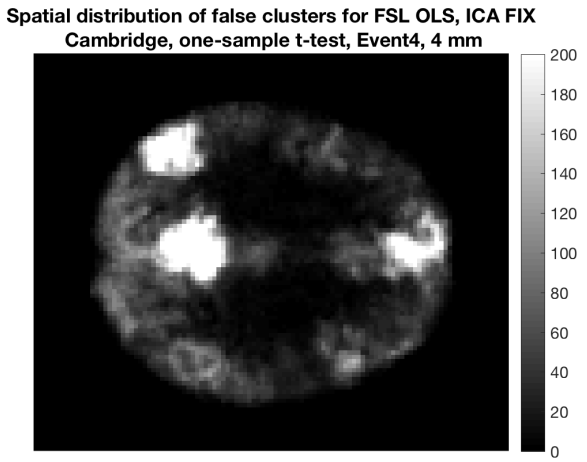
(c)



(d)

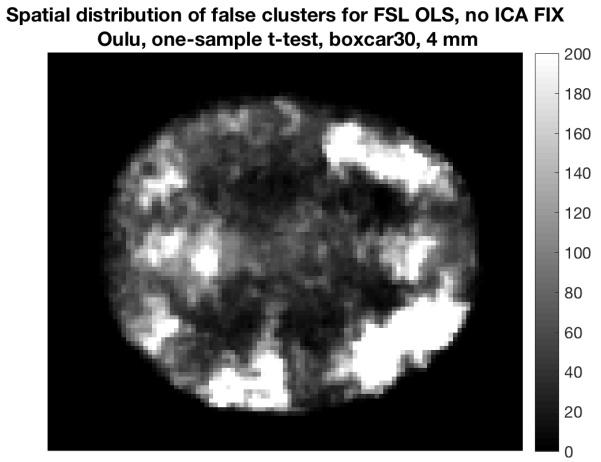


(e)

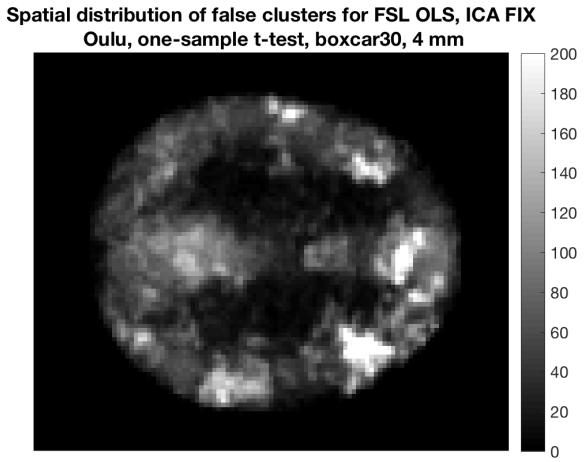


(f)

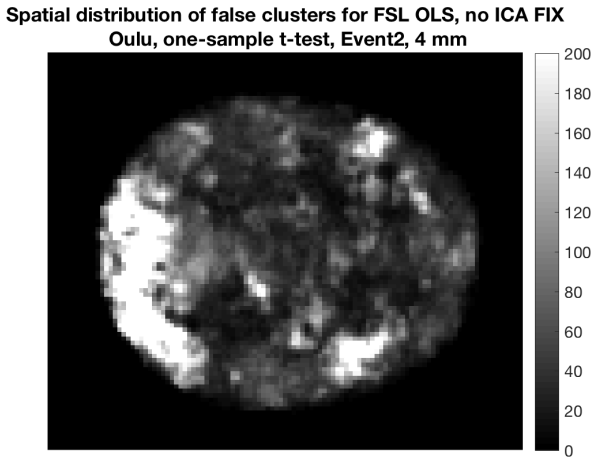
Figure 8: *The maps show voxel-wise incidence of false clusters for the Cambridge data, for 3 of the 6 different first level designs (a,b) B2, (c,d) E2, (e,f) E4. Left: results without ICA FIX, Right: results after first running ICA FIX for every subject. Image intensity is the number of times, out of 10,000 random analyses, a significant cluster occurred at a given voxel (CDT p = 0.01) for FSL OLS. Each analysis is a one-sample t-test using 20 subjects.*



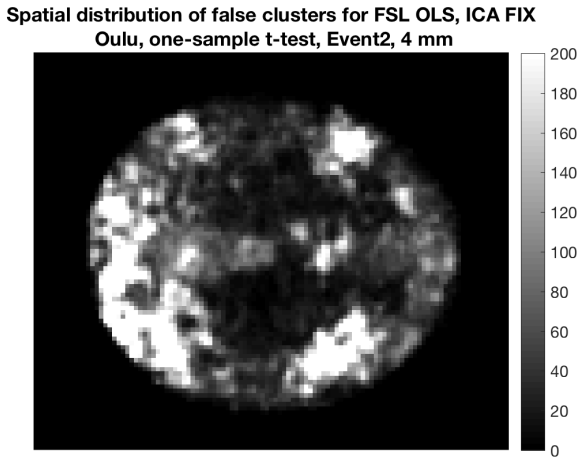
(a)



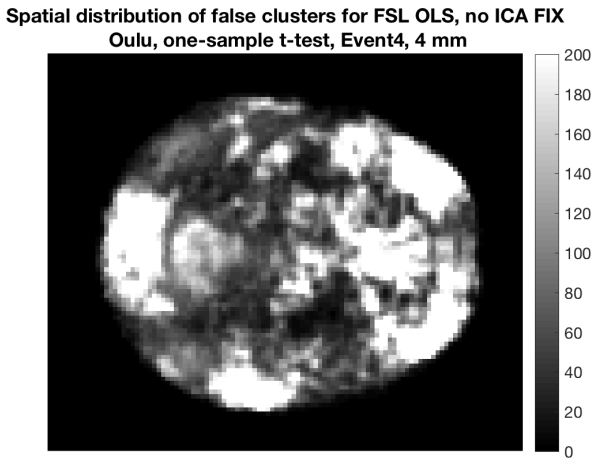
(b)



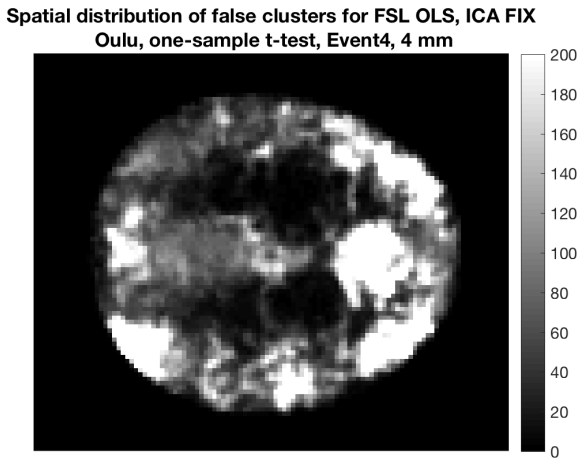
(c)



(d)

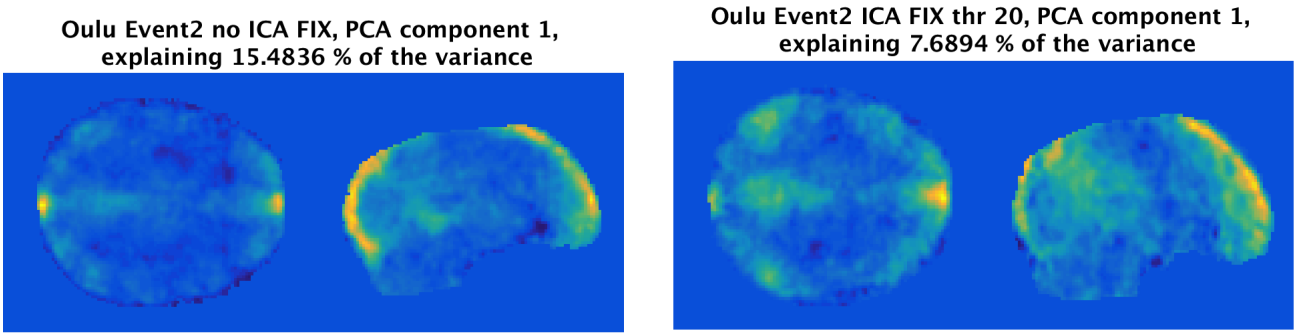


(e)



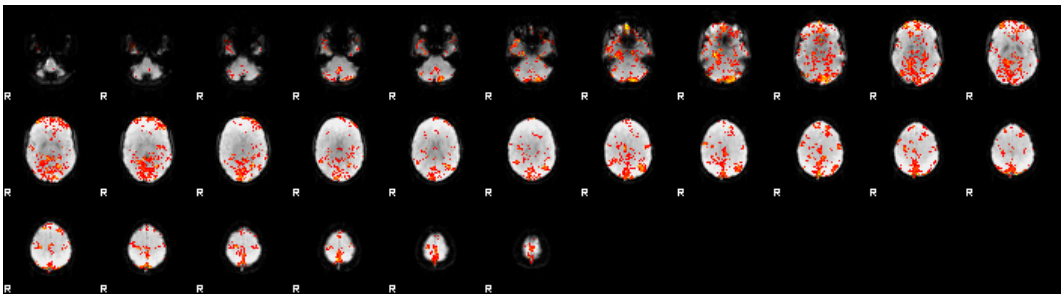
(f)

Figure 9: *The maps show voxel-wise incidence of false clusters for the Oulu data, for 3 of the 6 different first level designs (a,b) B2, (c,d) E2, (e,f) E4. Left: results without ICA FIX, Right: results after first running ICA FIX for every subject. Image intensity is the number of times, out of 10,000 random analyses, a significant cluster occurred at a given voxel (CDT p = 0.01) for FSL OLS. Each analysis is a one-sample t-test using 20 subjects.*

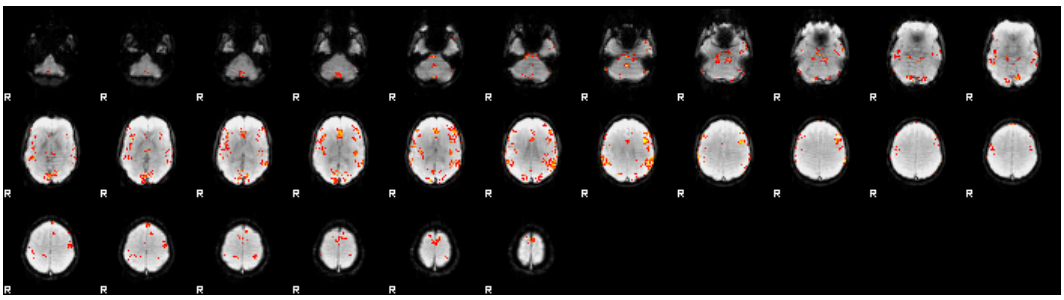


(a) (b)

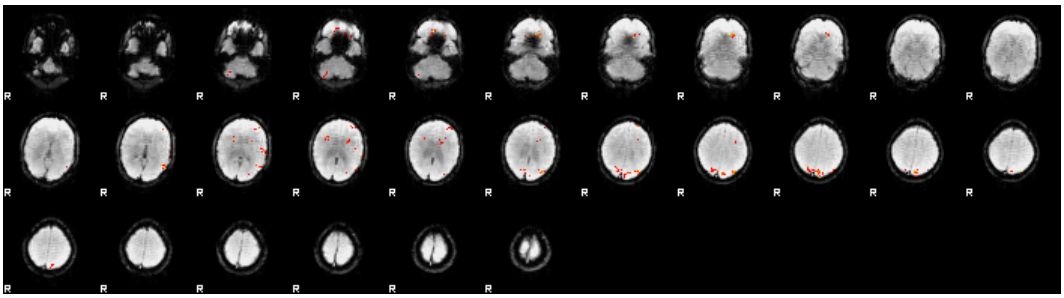
Figure 10: *The maps show an axial and a sagittal view of the first eigen component after running PCA on the 103 activity maps for Oulu E2, a) without ICA FIX b) with ICA FIX. Using ICA FIX clearly suppresses the posterior part of the vein artefact in the superior sagittal sinus, but a higher ICA FIX threshold is required to completely remove the artefact (not shown).*



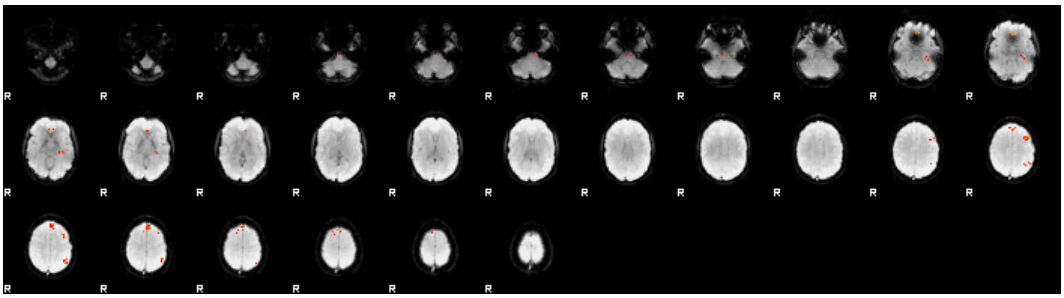
(a)



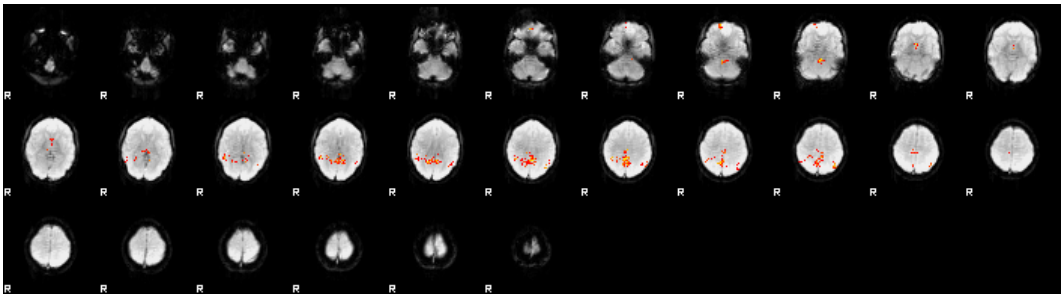
(b)



(c)



(d)



(e)

Figure 11: *Activity maps (thresholded at CDT $p = 0.01$ and cluster FWE corrected at $p = 0.05$, FSL default) for 5 Oulu subjects analyzed with 4 mm of smoothing and first level design E4. Despite testing for a difference between two random regressors, which are for design E4 also randomized over subjects, significant voxels are in several cases detected close to the superior sagittal sinus (indicating a vein artefact). Since many subjects have an activation difference in the same spatial location, this caused inflated false positive rates for the one-sample t-test. The two-sample t-test is not affected by these artefacts, since they cancel out when testing for a group difference.*

4. Discussion

We have presented results that support our original findings of inflated false positives with parametric cluster size inference. Specifically, new random null group task fMRI analyses, based on first level models with two fix regressors and models with two intersubject-randomized regressors, produced essentially the same results as the previous first level designs we considered. This argues against the charge that idiosyncratic attributes of our first level designs gave rise to our observed inflated false positives rates for cluster inference. Instead, we maintain that the best explanations for this behaviour are the long-tail spatial autocorrelation data (also present in MR phantom data (Kriegeskorte et al., 2008)) and spatially-varying smoothness. It should be noted that AFNI provides another function for cluster thresholding, ETAC (equitable thresholding and clustering), which performs better than the long-tail ACF function (Cox et al., 2017b) used here, but it was not available when we started the new group analyses. AFNI also provides non-parametric group inference in the 3dttest++ function.

4.1. Influence of artefacts on one-sample t-tests

Another objective of this work was to understand and remediate the less-than-perfect false positive rate control for one-sample permutation tests. We tried various alternative modelling strategies, including data transformations and robust regression, but none yielded consistent control of FWE. It appears that (physiological) artifacts are a major problem for the Oulu data, although the MRIQC tool (Esteban et al., 2017; Gorgolewski et al., 2017) did not reveal any major quality differences between Beijing, Cambridge and Oulu. The contribution of physiological noise in fMRI depends on the spatial resolution (see e.g. Bodurka et al. (2007)); larger voxels lead to a lower temporal signal to noise ratio. The Oulu data have a spatial resolution of 4 x 4 x 4.4 mm^3, compared to 3.13 x 3.13 x 3.6 mm^3 for Beijing and 3 x 3 x 3 mm^3 for Cambridge. Oulu voxels are thereby 2 times larger compared to Beijing voxels, and 2.6 times larger compared to Cambridge voxels, and this will make the Oulu data more prone to physiological noise.

Some of our random regressors are strongly correlated with the fMRI data in specific brain regions (especially the superior sagittal sinus, the transverse sinus and the sigmoid sinus), which lead to inflated false positive rates. Other artifacts, such as CSF artifacts and susceptibility weighted artifacts, are also present in the data (compared to examples given by Griffanti et al. (2017)). For a two-sample t-test, artifacts in the same spatial location for all subjects cancel out, as one tests for a difference between two groups, but this is not the case for a one-sample t-test. Combining ICA FIX with a two-sided test led to nominal FWE rates for Beijing and Cambridge, but not for Oulu. As can be seen in Figures 7 - 9, using ICA FIX clearly leads to false cluster maps which are more uniform accross the brain, with a lower number of false clusters in white matter. Increasing the ICA FIX threshold from the default 20 to 60 or 80 resulted in nominal false positive rates for Oulu (not shown), but it is likely that such a high threshold removes signal components. To instead train the ICA FIX classifier specifically for the Oulu data may give better results. Another possibility is to combine ICA FIX with the Yeo-Johnson transform, as skew can still be present in the denoised data. Evaluating all possible analysis combinations (ICA FIX, Yeo-Johnson transform, robust regression, global mean covariate, etc) is beyond the scope of this paper, due to the computational requirements.

Using ICA FIX for resting state fMRI data is rather easy (but it currently requires a specific version of the R software), as pretrained weights are available for different kinds of fMRI data. However, using ICA FIX for task fMRI data will require more work, as it is necessary to first manually classify ICA components (Griffanti et al., 2017) to provide training data for the classifier (Salimi-Khorshidi et al., 2014). An open database of manually classified fMRI ICA components, similar to NeuroVault (Gorgolewski et al., 2015), could potentially be used for fMRI researchers to automatically denoise their task fMRI data. A natural extension of MRIQC (Esteban et al., 2017; Gorgolewski et al., 2017) would then be to also measure the presence of artifacts in each fMRI dataset, by doing ICA and then comparing each component to the manually classified components in the open database. We also recommend researchers to collect physiological data, such that signal related to breathing and pulse can be modeled (Glover et al., 2000; Lund et al., 2006; Birn et al., 2006; Chang & Glover, 2009; Bollmann et al., 2018). This is especially important for 7T fMRI data, for which the physiological noise is often stronger compared to the thermal noise (Triantafyllou et al., 2005; Hutton et al., 2011).

4.2. Effect of multiband data on cluster inference

We note that multiband MR sequences (Moeller et al., 2010) are becoming increasingly common to improve temporal and/or spatial resolution, for example as provided by the Human Connectome Project (Essen et al., 2013) and the enhanced NKI-Rockland sample (Nooner et al., 2012). Multiband data have a potentially complex spatial autocorrelation (see, e.g. Risk et al. (2018)), and an important topic for future work is establishing how this impacts parametric cluster inference. The non-parametric permutation test (Winkler et al., 2014) does not make any assumption regarding the shape of the spatial autocorrelation function, and is therefore expected to perform well for any MR sequence.

4.3. Interpretation of affected studies

Finally, we consider the issue of the precise impact of our original findings. The original publication (Eklund et al., 2016) inadvertently implied that a large, unspecified proportion of the fMRI literature was affected by our findings, principally the severe inflation of false positive risk for a CDT of p = 0.01; this was clarified in a subsequent correction. In Appendix A we provide a rough bibliographic analysis to provide an estimate of how many articles used this particular CDT p = 0.01 setting. For a review conducted in January 2018, we estimated that out of 23,000 fMRI publications about 2,500, over 10%, of all studies have used this most problematic setting with parametric inference. While this calculation suggests how the literature as a whole can be interpreted, a more practical question is how one individual affected study can be interpreted. When examining a study that uses CDT p = 0.01, or one that uses no correction at all, it is useful to consider three possible states of nature:

State 1: Effect is truly present, and with revised methods, significance is retained

State 2: Effect is truly present, but with revised methods, significance is lost

State 3: Effect is truly null, absent; the study's detection is a false positive

In each of these, the statement about 'truth' reflects presence or absence of the effect in the population from which the subjects were drawn. When considering heterogeneity of different populations used for research, we could also add a fourth state:

State 4: Effect is truly null in population sampled, and this study's detection is a false positive; but later studies find and replicate the effect in other populations.

These could be summarised as "State 1: Robust true positive," "State 2: Fragile true positive," "State 3: False positive" and "State 4: Idiosyncratic false positive."

Unfortunately we can never know the true state of an effect, and, because of a lack of data archiving and sharing, we will mostly never know whether significance is retained or lost with reanalysis. All we can do is make qualitative judgments on individual works. To this end we can suggest that findings with no form of corrected significance receive the greatest skepticism; likewise, CDT p = 0.01 cluster size inference cluster p-values that *just* barely fall below 5% FWE significance should be judged with great skepticism. In fact, given small perturbations arising from a range of methodological choices, *all* research findings on the edge of a significance threshold deserves such skepticism. On the other hand, findings based on large clusters with P-values far below 0.05 could possibly survive a re-analysis with improved methods.

5. Conclusions

To summarize, our new results confirm that inflated familywise error rates for parametric cluster inference are also present when testing for a difference between two tasks, and when randomizing the task over subjects. Furthermore, the inflated familywise error rates for the non-parametric one-sample t-tests are due to random correlations with artifacts in the fMRI data, which for Beijing and Cambridge we found could be suppressed using ICA FIX. The Oulu data were collected with a lower spatial resolution, and are therefore more prone to physiological noise.

Acknowledgements

The authors would like to thank Jeanette Mumford for fruitful discussions. This study was supported by Swedish research council grants 2013-5229 and 2017-04889. Funding was also provided by the Center for Industrial Information Technology (CENIIT) at Linköping University. Thomas E. Nichols was supported by the Wellcome Trust (100309/Z/12/Z) and the NIH (R01 EB015611). The Nvidia Corporation, who donated the Nvidia Titan X Pascal graphics card used to run all permutation tests, is also acknowledged. This study would not be possible without the recent data-sharing initiatives in the neuroimaging field. We therefore thank the Neuroimaging Informatics Tools and Resources Clearinghouse and all of the researchers who have contributed with resting-state data to the 1,000 Functional Connectomes Project.

Appendix A. Bibliometrics of Cluster Inference

Appendix A.1. Number of affected studies

Here we conduct a biblibographic analysis to obtain an estimate of how much of the literature depends on our most troubling result, the severe inflation of FWE for a cluster-defining threshold (CDT) of p = 0.01.

We use the results of a systematic review of the fMRI literature conducted by Carp (2012) and Woo et al. (2014), which provides essential statistics on prevalence of cluster inference techniques. Carp (2012) defined a search for fMRI publications that today finds about $N(\text{fMRI}) = 23,000$ publications[4]. Drawing on a sample of 300 publications[5] published 2007 – 2012, Carp found $P(\text{HasData}) = 241/300 =$ 80% contained original data, and among these $P(\text{Corrected}|\text{HasData}) = 59\%$ used some form of correction for multiple comparisons[6]. Woo et al. (2014), considering a sample of 815 papers[7] published 2010 – 2011; of these, they found $P(\text{ClusterInference}|\text{HasData}) = 607/814 =$ 75%; noting that 6% (Fig. 1, Woo et al.) of the 814 include studies with no correction, we also compute $P(\text{ClusterInference}|\text{HasData}, \text{Corrected})$ $= 607/(814 - 0.06 \times 814) = 79\%$. Finally, with data from Figure 2(B) of Woo et al. (2014) (shown in Table Appendix A.1, kindly supplied by the authors) from the 480 studies that used cluster inference with correction (and had sufficient detail) we can compute

$$P(\text{CDT}(P \geq 0.01)|\text{ClusterInference}, \text{HasData}, \text{Corrected}) = (35{+}80)/480 = 24\%.$$

Thus we can finally estimate the number of published fMRI studies using corrected cluster inference as

$$N(\text{HasData}, \text{Corrected}, \text{ClusterInference}) = 23,000 \times 0.59 \times 0.79 = 10,720,$$

and, among those, $10,720 \times 0.24 = 2,573$ used a CDT of P = 0.01 or higher, or about 10% of publications reporting original fMRI results.

There are many caveats to this calculation, starting with different sampling criterion used in the two studies, and the ever-changing patterns of practice in neuroimaging. However, a recent survey of task fMRI papers published in early 2017 found that $270/388 = 69.6\%$ used cluster inference, and $72/270 = 26.7\%$ used a CDT of p = 0.01 or higher, suggesting that the numbers above remain representative (Yeung, 2018).

[4] 22,629 hits for Pubmed search text "(((((((fmri[title/abstract] OR functional MRI[title/abstract]) OR functional Magnetic Resonance Imaging[title/abstract]) AND brain[title/abstract]))) AND humans[MeSH Terms])) NOT systematic [sb]", conducted 30th January, 2018.

[5] Carp (2012) further constrained his search to publications with full text available in the open-access database PubMed Central (PMC).

[6] "Although a majority of studies (59%) reported the use of some variety of correction for multiple comparisons, a substantial minority did not."

[7] Woo et al. used "fmri" and "threshold" as keywords on original fMRI research papers published between in *Cerebral Cortex*, *Nature*, *Nature Neuroscience*, *NeuroImage*, *Neuron*, *PNAS*, and *Science*, yielding over 1500 papers; then following exclusion criterion were applied "(1) non-human studies, (2) lesion studies, (3) studies inwhich a threshold or correction method could not be clarified, (4) voxel-based morphometry studies, (5) studies primarily about methodology, and (6) machine-learning based studies"

Table A.1: Upon request, the authors of Woo et al. (2014) provided a detailed cross-tabulation of the frequencies of different cluster defining thresholds (the data presented in Figure 2(B) of their paper). Among the 607 studies that used cluster thresholding, they found 480 studies for which sufficient detail could be obtained to record the software and the particular cluster defining threshold used.

CDT	AFNI	BrainVoyager	FSL	SPM	Others	TOTAL
>.01	9	5	9	8	4	35
.01	9	4	44	20	3	80
.005	24	6	1	48	3	82
.001	13	20	11	206	5	255
<.001	2	5	3	16	2	28
TOTAL	57	40	68	298	17	480

References

Beckmann, C., & Smith, S. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, *23*, 137 – 152.

Birn, R., Diamond, J., Smith, M., & Bandettini, P. (2006). Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fMRI. *NeuroImage*, *31*, 1536 – 1548.

Biswal, B., Mennes, M., ..., X. Z., & Milham, M. (2010). Toward discovery science of human brain function. *PNAS*, *107*, 4734–4739.

Bodurka, J., Ye, F., Petridou, N., Murphy, K., & Bandettini, P. (2007). Mapping the MRI voxel volume in which thermal noise matches physiological noise - Implications for fMRI. *NeuroImage*, *34*, 542 – 549.

Bollmann, S., Puckett, A., Cunnington, R., & Barth, M. (2018). Serial correlations in single-subject fMRI with sub-second TR. *NeuroImage*, *166*, 152 – 166.

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, *63*, 289–300.

Chang, C., & Glover, G. (2009). Effects of model-based physiological noise correction on default mode network anti-correlations and correlations. *NeuroImage*, *47*, 1448 – 1459.

Cox, R., Chen, G., Glen, D., Reynolds, R., & Taylor, P. (2017a). FMRI Clustering and False Positive Rates. *PNAS*, *114*, E3370–E3371.

Cox, R., Chen, G., Glen, D., Reynolds, R., & Taylor, P. (2017b). FMRI Clustering in AFNI: False-Positive Rates Redux. *Brain Connectivity*, *7*, 152–171.

Eklund, A., Dufort, P., Villani, M., & LaConte, S. (2014). BROCCOLI: Software for fast fMRI analysis on many-core CPUs and GPUs. *Frontiers in Neuroinformatics*, *8:24*.

Eklund, A., Nichols, T., & Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false positive rates. *PNAS*, *113*, 7900–7905.

Eklund, A., Nichols, T., & Knutsson, H. (2017). Reply to Brown and Behrmann, Cox et al. and Kessler et al.: Data and code sharing is the way forward for fMRI. *PNAS*, *114*, E3374–E3375.

Essen, D. V., Smith, S., Barch, D., Behrens, T., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, *80*, 62–79.

Esteban, O., Birman, D., Schaer, M., Koyejo, O., Poldrack, R., & Gorgolewski, K. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLOS ONE*, *12*, 1–21.

Flandin, G., & Friston, K. (2017). Analysis of family-wise error rates in statistical parametric mapping using random field theory. *Human Brain Mapping*, .

Friman, O., Borga, M., Lundberg, P., & Knutsson, H. (2003). Adaptive analysis of fMRI data. *NeuroImage*, *19*, 837 – 845.

Glover, G., Li, T., & Ress, D. (2000). Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magnetic Resonance in Medicine*, *44*, 162–167.

Gorgolewski, K., Alfaro-Almagro, F., Auer, T., Bellec, P., CapotaÉ, M., Chakravarty, M., Churchill, N., Cohen, A., Craddock, C., Devenyi, G., Eklund, A., Esteban, O., Flandin, G., Ghosh, S., Guntupalli, S., Jenkinson, M., Keshavan, A., Kiar, G., Liem, F., Raamana, P., Raffelt, D., Steele, C., Quirion, P., Smith, R., Strother, S., Varoquaux, G., Wang, Y., Yarkoni, T., & Poldrack, R. (2017). BIDS Apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLOS Computational Biology*, *13*, 1–16.

Gorgolewski, K., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S., Maumet, C., Sochat, V., Nichols, T., Poldrack, R., Poline, J., Yarkoni, T., & Margulies, D. (2015). NeuroVault.org: A repository for sharing unthresholded statistical maps, parcellations, and atlases of the human brain. *NeuroImage*, *124*, 1242–1244.

Greve, D., & Fischl, B. (2018). False positive rates in surface-based anatomical analysis. *NeuroImage*, *171*, 6–14.

Griffanti, L., Douaud, G., Bijsterbosch, J., Evangelisti, S., Alfaro-Almagro, F., Glasser, M., Duff, E., Fitzgibbon, S., Westphal, R., Carone, D., Beckmann, C., & Smith, S. (2017). Hand classification of fMRI ICA noise components. *NeuroImage*, *154*, 188 – 205.

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C., Auerbach, E., Douaud, G., Sexton, C., Zsoldos, E., Ebmeier, K., Filippini, N., Mackay, C., Moeller, S., Xu, J., Yacoub, E., Baselli, G., Ugurbil, K., Miller, K., & Smith, S. (2014). ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *NeuroImage*, *95*, 232 – 247.

Hutton, C., Josephs, O., Stadler, J., Featherstone, E., Reid, A., Speck, O., Bernarding, J., & Weiskopf, N. (2011). The impact of physiological noise correction on fMRI at 7T. *NeuroImage*, *57*, 101 – 112.

Kaundinya, G., Venkatagiri, K., & Sathian, K. (2018a). Accounting for Non-Gaussian Sources of Spatial Correlation in Parametric Functional Magnetic Resonance Imaging Paradigms I: Revisiting Cluster-Based Inferences. *Brain Connectivity*, *8*, 1–9.

Kaundinya, G., Venkatagiri, K., Simon, L., & Sathian, K. (2018b). Accounting for Non-Gaussian Sources of Spatial Correlation in Parametric Functional Magnetic Resonance Imaging Paradigms II: A Method to Obtain First-Level Analysis Residuals with Uniform and Gaussian Spatial Autocorrelation Function and Independent and Identically Distributed Time-Series. *Brain Connectivity*, *8*, 10–21.

Kessler, D., Angstadt, M., & Sripada, C. (2017). Reevaluating "cluster failure"Äù in fMRI using nonparametric control of the false discovery rate. *PNAS*, *114*, E3372–E3373.

Kriegeskorte, N., Bodurka, J., & Bandettini, P. (2008). Artifactual time-course correlations in echo-planar fMRI with implications for studies of brain function. *International Journal of Imaging Systems and Technology*, *18*, 345–349.

Lund, T., Madsen, K., Sidaros, K., Luo, W., & Nichols, T. (2006). Non-white noise in fMRI: Does modelling have an impact? *NeuroImage*, *29*, 54 – 66.

Moeller, S., Yacoub, E., Olman, C., Auerbach, E., Strupp, J., Harel, N., & Ugüurbil, K. (2010). Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magnetic Resonance in Medicine*, *63*, 1144–1153.

Mueller, K., Lepsien, J., Möller, H., & Lohmann, G. (2017). Commentary: Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Frontiers in Human Neuroscience*, *11*.

Mumford, J. (2017). A comprehensive review of group level model performance in the presence of heteroscedasticity: Can a single model control Type I errors in the presence of outliers? *NeuroImage*, *147*, 658 – 668.

Murphy, K., Birn, R., Handwerker, D., Jones, T., & Bandettini, P. (2009). The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced? *NeuroImage*, *44*, 893 – 905.

Murphy, K., & Fox, M. (2017). Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *NeuroImage*, *154*, 169 – 173.

Nichols, T., Eklund, A., & Knutsson, H. (2017). A defense of using resting state fMRI as null data for estimating false positive rates. *Cognitive Neuroscience*, *8*, 144–149.

Nooner, K., Colcombe, S., Tobe, R., Mennes, M., Benedict, M., Moreno, A., Panek, L., Brown, S., Zavitz, S., Li, Q., Sikka, S., Gutman, D., Bangaru, S., Schlachter, R., Kamiel, S., Anwar, A., Hinz, C., Kaplan, M., Rachlin, A., Adelsberg, S., Cheung, B., Khanuja, R., Yan, C., Craddock, C., Calhoun, V., Courtney, W., King, M., Wood, D., Cox, C., Kelly, C., DiMartino, A., Petkova, E., Reiss, P., Duan, N., Thompsen, D., Biswal, B., Coffey, B., Hoptman, M., Javitt, D., Pomara, N., Sidtis, J., Koplewicz, H., Castellanos, F., Leventhal, B., & Milham, M. (2012). The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry. *Frontiers in Neuroscience*, *6*, 152.

Poldrack, R., Barch, D., Mitchell, J., Wager, T., Wagner, A., Devlin, J., Cumba, C., Koyejo, O., & Milham, M. (2013). Toward open sharing of task-based fMRI data: the OpenfMRI project. *Frontiers in Neuroinformatics*, *7*.

Risk, B., Kociuba, M., & Rowe, D. (2018). Impacts of simultaneous multislice acquisition on sensitivity and specificity in fMRI. *NeuroImage*, *172*, 538–553.

Salimi-Khorshidi, G., Douaud, G., Beckmann, C., Glasser, M., Griffanti, L., & Smith, S. (2014). Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage*, *90*, 449 – 468.

Slotnick, S. (2016). Resting-state fMRI data reflects default network activity rather than null data: A defense of commonly employed methods to correct for multiple comparisons. *Cognitive Neuroscience*, *8*, 141–143.

Slotnick, S. (2017). Cluster success: fMRI inferences for spatial extent have acceptable false-positive rates. *Cognitive Neuroscience*, *8*, 150–155.

Smith, S., & Nichols, T. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*, 83–98.

Stelzer, J., Chen, Y., & Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage*, *65*, 69 – 82.

Triantafyllou, C., Hoge, R., Krueger, G., Wiggins, C., Potthast, A., Wiggins, G., & Wald, L. (2005). Comparison of physiological noise at 1.5 T, 3 T and 7 T and optimization of fMRI acquisition parameters. *NeuroImage*, *26*, 243 – 250.

Wager, T., Keller, M., Lacey, S., & Jonides, J. (2005). Increased sensitivity in neuroimaging analyses using robust regression. *NeuroImage*, *15*, 99–113.

Wald, L., & Polimeni, J. (2017). Impacting the effect of fMRI noise through hardware and acquisition choices - implications for controlling false positive rates. *NeuroImage*, *154*, 15–22.

Winkler, A., Ridgway, G., Webster, M., Smith, S., & Nichols, T. (2014). Permutation inference for the general linear model. *NeuroImage*, *92*, 381–397.

Woo, C., Krishnan, A., & Wager, T. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage*, *91*, 412 – 419.

Woolrich, M. (2008). Robust group analysis using outlier inference. *NeuroImage*, *41*, 286–301.

Yeo, I., & Johnson, R. (2000). A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika*, *87*, 954 – 959.

Yeung, A. (2018). An updated survey on statistical thresholding and sample size of fMRI studies. *Frontiers in Human Neuroscience*, *12:16*.