Received: 16 March 2018 | Revised: 30 July 2018 | Accepted: 1 August 2018

DOI: 10.1002/hbm.24350

WILEY

Cluster failure revisited: Impact of first level design and physiological noise on cluster false positive rates

Anders Eklund[1,2,3] | Hans Knutsson[1,3] | Thomas E. Nichols[4,5,6]

[1]Division of Medical Informatics, Department of Biomedical Engineering, Linköping University, Linköping, Sweden

[2]Division of Statistics & Machine Learning, Department of Computer and Information Science, Linköping University, Linköping, Sweden

[3]Center for Medical Image Science and Visualization (CMIV), Linköping University, Linköping, Sweden

[4]Big Data Institute, University of Oxford, Oxford, United Kingdom

[5]Wellcome Trust Centre for Integrative Neuroimaging (WIN-FMRIB), University of Oxford, Oxford, United Kingdom

[6]Department of Statistics, University of Warwick, Coventry, United Kingdom

Correspondence
Anders Eklund, Division of Medical Informatics, Department of Biomedical Engineering, Linköping University, Linköping, Sweden.
Email: anders.eklund@liu.se

Funding information
Wellcome Trust, Grant/Award Number: 100309/Z/12/Z; NIH, Grant/Award Number: R01 EB015611; Wellcome Trust, Grant/Award Number: 100309/Z/12/Z; Knut och Alice Wallenbergs Stiftelse; Linköping University; Center for Industrial Information Technology (CENIIT); Swedish research council, Grant/Award Number: 2013-5229 and 2017-04889

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 randomization 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 found that the combination of a two-sided test and cleaning the data using ICA FIX resulted in nominal false positive rates for all data sets, meaning that data cleaning is not only important for resting state fMRI, but also for task fMRI. Finally, we 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 = .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

cluster inference, false positives, functional magnetic resonance imaging, ICA FIX, permutation, physiological noise

1 | INTRODUCTION

In our previous work (Eklund, Nichols, & Knutsson, 2016a), we used freely available resting state functional magnetic resonance imaging (fMRI) data to evaluate the validity of standard fMRI inference methods. Group analyses involving only healthy controls were used to empirically estimate the degree of false positives, after correcting for multiple comparisons, based on the idea that a two-sample t test using only healthy controls should lead to nominal false positive rates (e.g., 5%). By considering resting state fMRI as null task fMRI data, the same approach was used to evaluate the statistical methods for one-sample t tests. Briefly, we found that parametric statistical methods (e.g., Gaussian random field theory [GRFT]) perform well for voxel inference, where each voxel is separately tested for significance, but the combination of voxel inference and familywise error (FWE) correction is seldom used due to its low statistical power. For this reason, the false discovery rate is in neuroimaging (Genovese, Lazar, & Nichols, 2002) often used to increase statistical power. For cluster inference, where groups of voxels are tested together by looking at the size of each cluster, we found that parametric methods perform well for a high cluster defining threshold (CDT; p = .001) but result in inflated false positive rates for low CDTs (e.g., p = .01). GRFT is for

cluster inference based on two additional assumptions, compared to GRFT for voxel inference, and we found that these assumptions are violated in the analyzed data. First, the spatial autocorrelation function (SACF) is assumed to be Gaussian, but real fMRI data have a SACF with a much longer tail. Second, the spatial smoothness is assumed to be constant over the brain, which is not the case for fMRI data. The nonparametric permutation test is not based on these assumptions (Winkler, Ridgway, Webster, Smith, & Nichols, 2014) and, therefore, produced nominal results for all two-sample *t* tests, but in some cases failed to control FWE for one-sample *t* tests.

1.1 | Related work

Our article has generated intense discussions regarding cluster inference in fMRI (Cox, 2018; Cox, Chen, Glen, Reynolds, & Taylor, 2017a, 2017b; Eklund, Nichols, & Knutsson, 2017; Flandin & Friston, 2017; Gopinath, Krishnamurthy, & Sathian, 2018; Kessler, Angstadt, & Sripada, 2017), on the validity of using resting state fMRI data as null data (Nichols, Eklund, & Knutsson, 2017; Slotnick, 2016, 2017), how the spatial resolution can affect parametric cluster inference (Mueller, Lepsien, Möller, & Lohmann, 2017), how to obtain residuals with a Gaussian SACF (Gopinath, Krishnamurthy, Lacey, & Sathian, 2018), how to model the long-tail SACF (Cox et al., 2017b), as well as how different MR sequences can change the SACF and thereby cluster inference (Wald & Polimeni, 2017). Furthermore, some of our results have been reproduced and extended (Cox et al., 2017b; Flandin & Friston, 2017; Kessler et al., 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 nonparametric permutation test showed good control of the FWE for all settings, while traditional Monte Carlo methods fail to control FWE for some settings.

1.2 | Realistic first level designs

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 1a, 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 data set 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 1,000 functional connectomes project release. In the Methods section we, therefore, describe new analyses based on two new first level designs.

1.3 | Nonparametric inference

Nonparametric group inference is now available in the AFNI function 3dttest++ (Cox et al., 2017a, 2017b), meaning that the three most common fMRI softwares now all support nonparametric group inference (SPM users can use the SnPM toolbox (http://warwick.ac.uk/snpm), and FSL users can use the randomize 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, Borga, Lundberg, & Knutsson, 2003; Stelzer, Chen, & Turner, 2013).

The nonparametric permutation test produced nominal results for all two sample *t* tests, but not for the one sample *t* tests (Eklund et al., 2016a), and the Oulu data were more problematic compared to Beijing and Cambridge. As described in Section 2, we investigated numerous ways to achieve nominal results, and finally concluded that (physiological) artifacts are a problem for one-sample *t* tests, especially for the Oulu data. This is a good example of the challenge of validating statistical methods. One can argue that real fMRI data are essential since they contain all types of noise (Birn, Diamond, Smith, & Bandettini, 2006; Chang & Glover, 2009; Glover, Li, & Ress, 2000; Greve, Brown, Mueller, Glover, & Liu, 2013; Lund, Madsen, Sidaros, Luo, & Nichols, 2006) which are difficult to simulate. From this perspective, the Oulu data are helpful since they highlight the problem of (physiological) noise. On the other hand, one can argue that a pure fMRI simulation (Welvaert & Rosseel, 2014) is better, since the researcher then can control all parameters of the data and independently test different settings. From this perspective, the Oulu data should be avoided, because the assumption of no consistent activation over subjects is violated by the (physiological) noise, however data quality varies dramatically over sites and studies and we expect there is plenty of data collected that has quality comparable to Oulu.

1.4 | Implications

The original publication (Eklund et al., 2016a) 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 $p = .01$; this was clarified in a subsequent correction (Eklund, Nichols, & Knutsson, 2016b). In Section 4, we consider the interpretation of our findings and their impact on the literature as a whole. We estimate that at least 10% of 23,000 published fMRI studies have used the problematic CDT $p = .01$.

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

[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

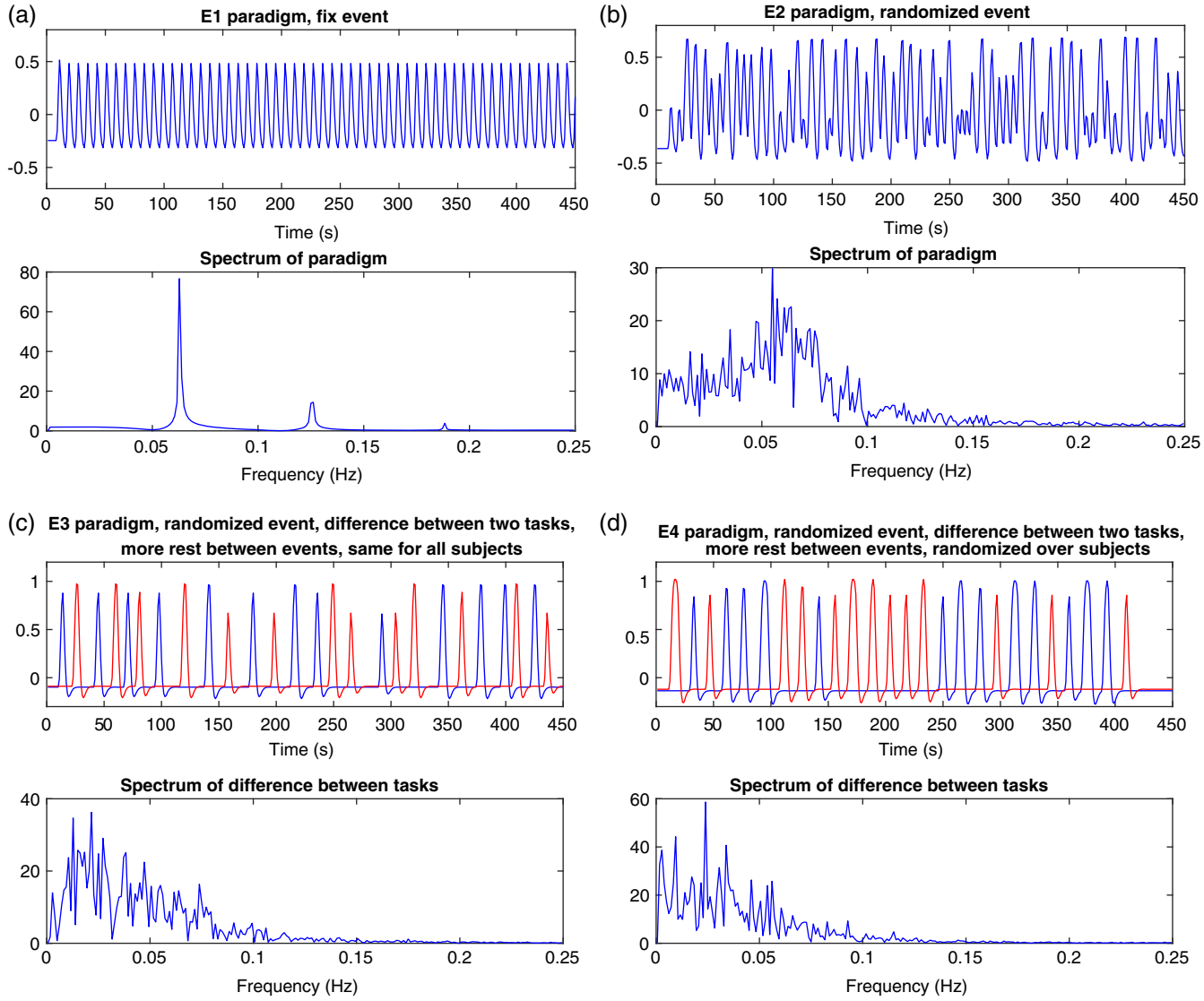


FIGURE 1 A comparison of the paradigms used in the original paper (a) E1, (b) E2), and the new paradigms used in this article (c) E3, (d) E4, for the Beijing data sets (sampled with a TR of 2 s). A single task was used for both E1 and E2, while two pretended tasks where used for E3 and E4 (and all first-level 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. For all paradigms, the default hemodynamic response function in SPM 8 (double gamma) was used to generate these plots

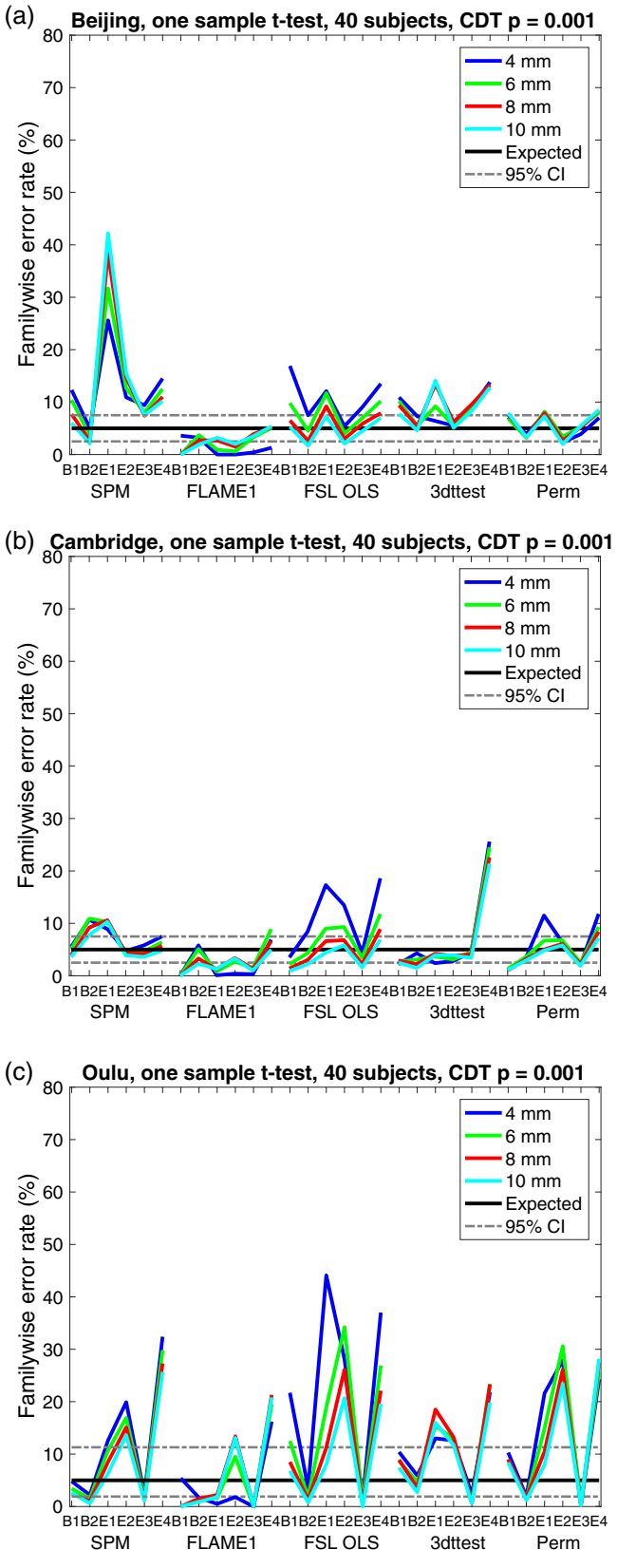
2 | METHODS

2.1 | New paradigms

To address the concerns regarding realistic first level designs, 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. For Beijing data, 13 events were used for each of the two tasks. Each task is 3–7 s long, and the rest between each event is 11–13 s. For Cambridge data, 11 events of 3–6 s were used for each task. For Oulu data, 13 events of 3–6 s were used for each task. 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 the same number of events for each task, but the order and the timing of the events is different for every subject. 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., 2016a), using the same data (Beijing, Cambridge, Oulu) from the 1,000 functional connectomes project (Biswal et al., 2010). Analyses were performed with SPM 8 (Ashburner, 2012), FSL 5.09 (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012) and AFNI 16.3.14 (Cox, 1996). FWE rates were estimated for different levels of smoothing (4–10 mm), one-sample as well as two-sample t tests, and two CDTs (p = .01 and p = .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., 2016a). 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 reevaluated using the ACF option (note, that this ACF AFNI option still assumes a stationary spatial autocorrelation structure). Interested readers are referred to our github account for further details.



2.2 | Using ICA-FIX for denoising

We investigated numerous ways to achieve nominal FWE rates for the one-sample (sign flipping) permutation test;

1. Applying the Yeo and 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 (Mumford, 2017; Wager, Keller, Lacey, & Jonides, 2005; Woolrich, 2008)
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, Birn, Handwerker, Jones, & Bandettini, 2009; Murphy & Fox, 2017; which is normally not done for task fMRI).

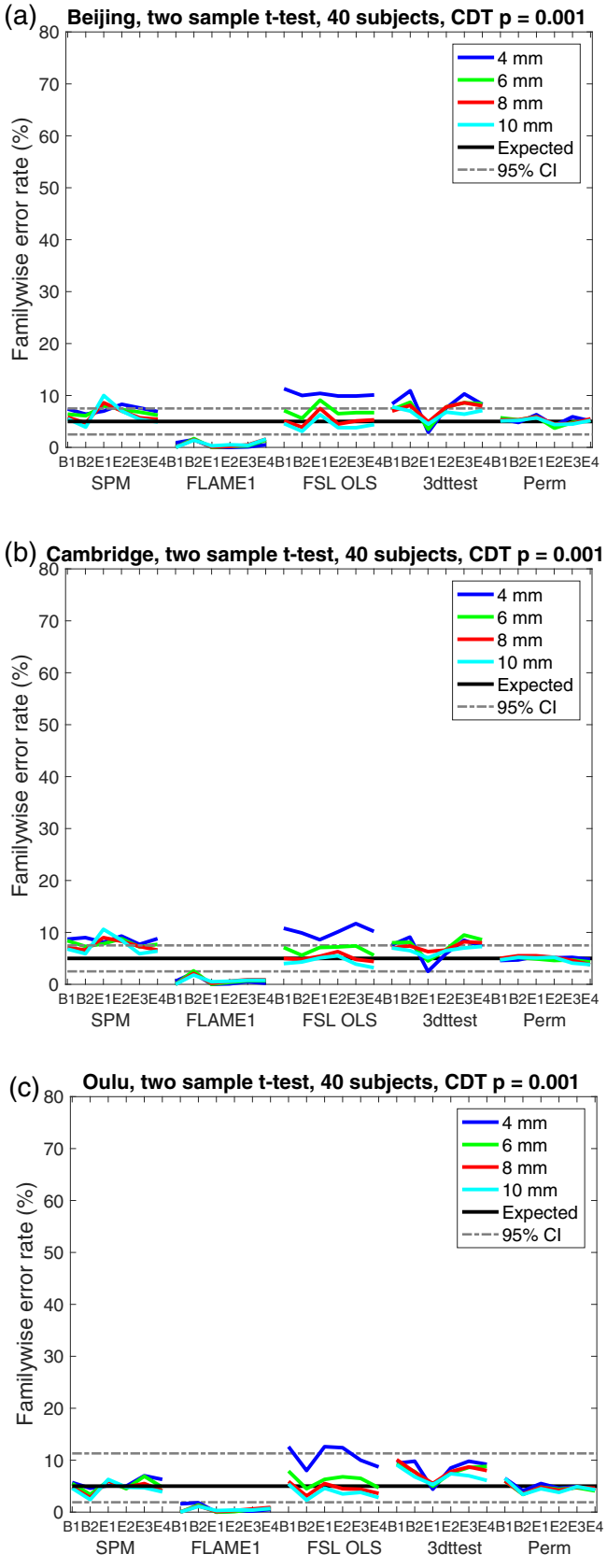
While some of these approaches resulted in nominal FWE rates for a subset of the parameter combinations, no approach worked well for all settings and data sets. In our original study, we only used one-sided tests, but this is based on an implicit assumption that a random regressor is equally likely to be positively or negatively correlated with resting state fMRI data. Additionally, most fMRI studies that use a one-sample *t* test take advantage of a one-sided test to increase statistical power (Chen et al., 2018).

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 fig. 18 in Eklund et al. (2016a)), 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 $p = .01$ for both increases and decreases on a given statistic map. As discussed in Section 3, there appears to be physiological artifacts which ideally would be remediated by respiration or cardiac time series modeling (Birn et al., 2006; Bollmann, Puckett, Cunnington, & Barth, 2018; Chang & Glover, 2009; Glover et al., 2000; Lund et al., 2006), but unfortunately the 1,000 functional connectomes data sets (Biswal et al., 2010) do not have these physiological recordings.

FIGURE 2 Results for one sample t-test and cluster-wise inference using a CDT of $p = .001$, showing estimated FWE rates for 4–10 mm of smoothing and six different activity paradigms (old paradigms B1, B2, E1, E2, and new paradigms 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 $p = .001$, uncorrected for multiple comparisons, and the surviving clusters were then compared to a FWE-corrected cluster extent threshold, $p_{FWE} = .05$. The estimated FWE 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

To suppress the influence of artifacts, we therefore instead applied ICA FIX (version 1.065) in FSL (Griffanti et al., 2014, 2017; Salimi-Khorshidi et al., 2014) 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 available in ICA FIX (trained for 5 mm smoothing). To use ICA FIX for 8 or 10 mm of smoothing would require retraining 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. To study the effect of retraining the ICA FIX classifier specifically for each data set (Beijing, Cambridge, Oulu), instead of using the pretrained weights available in ICA FIX, we manually labeled the ICA components of 10 subjects for each data set (giving a total of 350–450 ICA components per data set). Indeed, a large portion of the ICA components are artifacts that are similar across subjects. Interested readers are referred to the github account for ICA FIX processing scripts and the retrained classifier weights for Beijing, Cambridge, and Oulu.

First-level analyses for B1, B2, E1, E2, E3, and E4 were performed using FSL for all 499 subjects after ICA FIX, with motion correction and smoothing turned off. Group level analyses were finally performed using the nonparametric one-sample *t*-test available in BROCCOLI (Eklund, Dufort, Villani, & LaConte, 2014).



3 | RESULTS

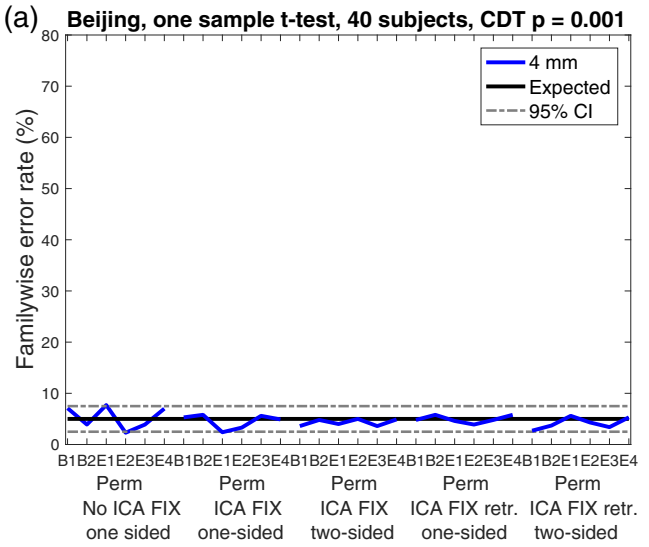
3.1 | New paradigms

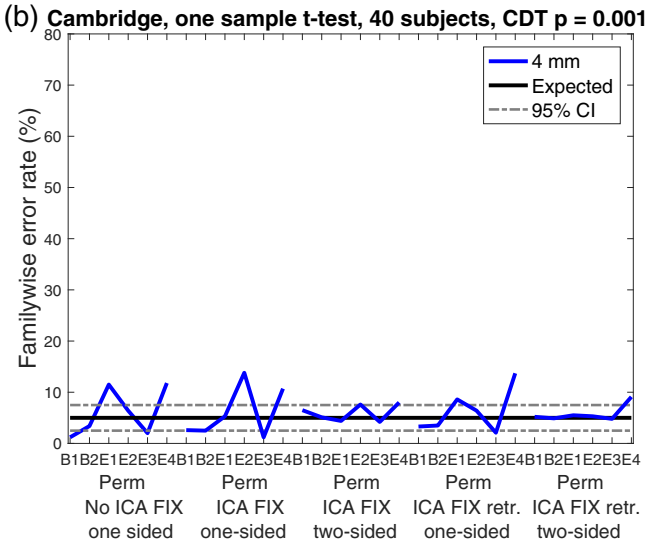
Figures 2 and 3 show estimated FWE rates for the two new paradigms (E3, E4), for 40 subjects in each group analysis and a CDT of $p = .001$. Figures A11 and A12 show the FWE rates for a CDT of $p = .01$. The four old paradigms (B1, B2, E1, E2) are included as well for the sake of comparison. In brief, the new paradigm with two pretend tasks (E3) does not lead to lower FWE rates, compared to the old paradigms. Likewise, randomizing task events over subjects (E4) has if anything worse FWE rates compared to not randomizing 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., 2016a)

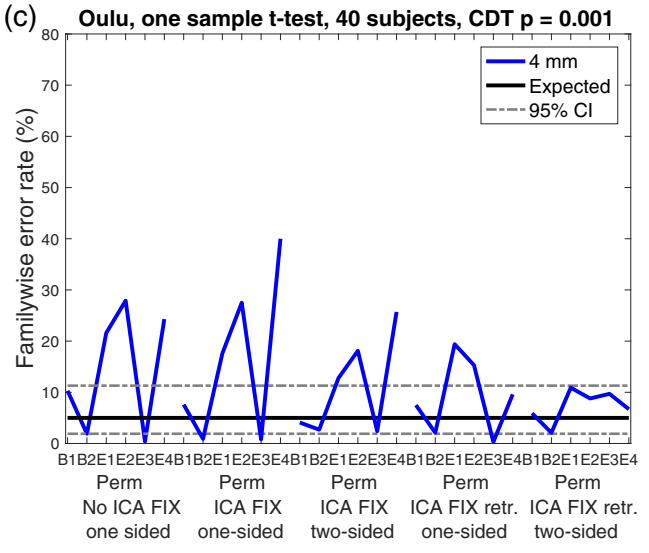
By looking at Figure 2, it is possible to compare the parametric methods (who are simultaneously affected by non-Gaussian SACF,

FIGURE 3 Results for two sample *t* test and cluster-wise inference using a CDT of $p = .001$, showing estimated FWE rates for 4–10 mm of smoothing and six different activity paradigms (old paradigms B1, B2, E1, E2, and new paradigms 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 $p = .001$, uncorrected for multiple comparisons, and the surviving clusters were then compared to a FWE-corrected cluster extent threshold, $p_{FWE} = .05$. The estimated FWE 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

nonstationary smoothness and physiological noise) and the nonparametric permutation test (only affected by physiological noise, as no assumptions are made regarding the SACF and stationary smoothness). For the Beijing data, the permutation test performs rather well, while all parametric approaches struggle despite a strict CDT. It is also clear that the Oulu data is more problematic compared to Beijing and Cambridge.







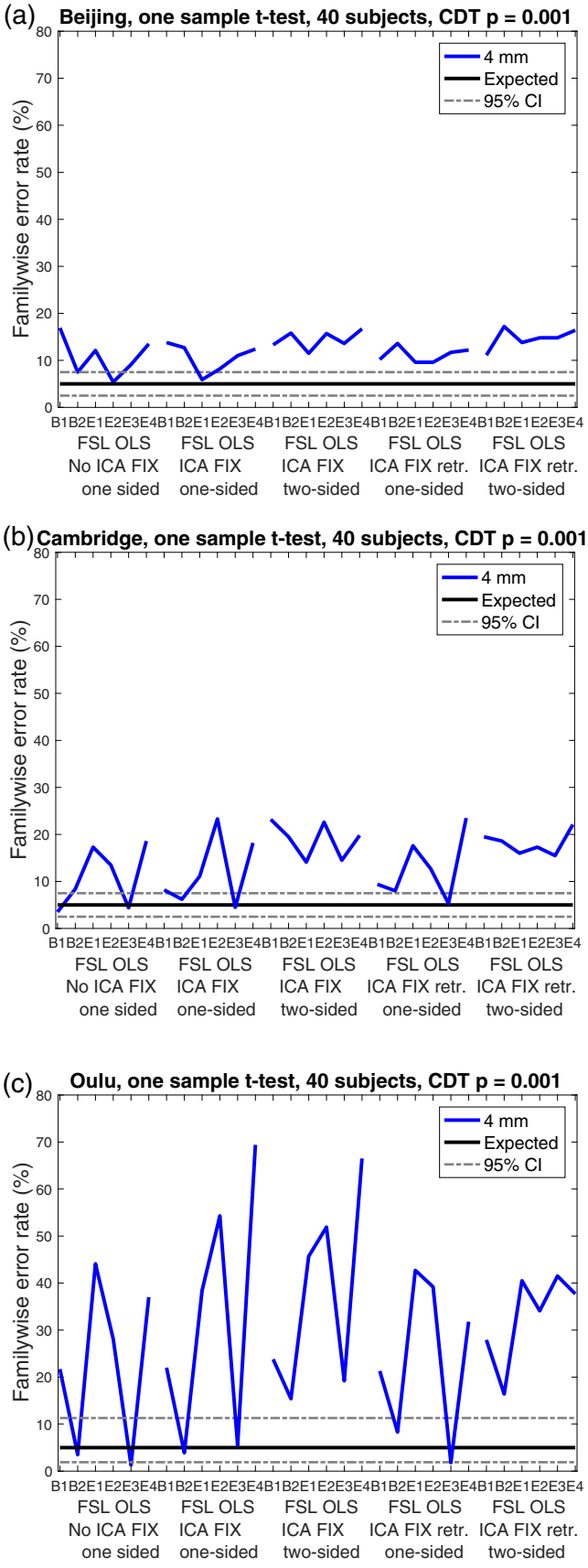
3.2 | ICA denoising

Figure 4 shows FWE rates for the nonparametric one-sample *t* test, for no ICA FIX, pretrained ICA FIX and retrained ICA FIX, for one-sided as well as two-sided tests. For the Beijing data, the FWE 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. However, retraining the ICA FIX classifier specifically for the Oulu data set finally resulted in nominal false positive rates.

To test if using ICA FIX also results in nominal FWE rates for FSL OLS, we performed group analyses for no ICA FIX, pretrained ICA FIX and retrained ICA FIX, for one-sided as well as two-sided tests, see Figure 5. As ICA FIX cleaning and all first-level analyses were performed using FSL, we only performed the group analyses using FSL. Clearly, using ICA FIX does not lead to nominal FWE rates for FSL OLS, and using a two-sided test leads to even higher FWE rates compared to a one-sided test. A possible explanation is that (two) parametric tests for $p = .025$ are even more inflated compared to parametric tests for $p = .05$. To test this hypothesis, we performed 18,000 one-sided one-sample group analyses (three data sets and six activity paradigms, 1,000 analyses each, for first-level analyses with no ICA FIX, CDT $p = .001$) with FWE significance thresholds of 2.5% and 1%. False positives at FWE 2.5% and 1% should occur 1/2 = 0.5 and 1/5 = 0.2 times as often with FWE 5% results. We found nominal FWE 2.5% false positives occurred at a rate 0.879× the 5% FWE results, and nominal FWE 1% false positives occurred at a rate 0.694× the 5% FWE results. That is, the relative inflation of false positives for parametric methods is much higher for more stringent significance thresholds. For permutation, inaccuracies arise due to a disparity between the upper tail of the sign flipping null and the actual null's upper tail. For the absolute value statistic used for two-sided tests, the upper tail is essentially an average of $f(x)$ and $f(-x)$ for $x > 0$ and is less sensitive to violations of symmetry assumption.

FIGURE 4 Results for nonparametric (sign flipping) one-sample *t* tests for cluster-wise inference using a CDT of $p = .001$, for no ICA FIX, pretrained ICA FIX and retrained ICA FIX. (a) Results for Beijing data (b) results for Cambridge data (c) results for Oulu data. Results are only shown for 4 mm smoothing, as other smoothing levels would require retraining the ICA FIX classifier. For both Beijing and Cambridge, the pretrained classifier weights for ICA FIX are sufficient to achieve nominal false positive rates, while it is necessary to retrain the ICA FIX classifier specifically for the Oulu data (a possible explanation is that the Oulu data have a spatial resolution of 4 x 4 x 4.4 mm^3, while ICA FIX for standard fMRI data is pretrained on data with a spatial resolution of 3.5 x 3.5 x 3.5 mm^3)

Figures 6–8 show cluster prevalence maps for group analyses without first running ICA FIX, with pretrained ICA FIX and with retrained ICA FIX, for first level designs 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 using 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 artifact, 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 9). The posterior part of the artifact is suppressed by the pretrained ICA FIX classifier, and the retrained ICA FIX classifier is even better at suppressing the artifact. Also see Figure 10 for activation maps from five 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 artifact.



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 behavior are the long-tail spatial autocorrelation data (also present in MR phantom data (Kriegeskorte, Bodurka, & Bandettini, 2008)) and spatially-varying smoothness. Recently, Greve and Fischl (2018) showed that group analyses of cortical thickness, surface area, and volume (using only the structural MRI data in the fcon1000 data set (Biswal et al., 2010)) also lead to inflated false positive rates in some cases, indicating that these issues affect structural analyses on the cortical surface as well, and thus is not specific to fMRI paradigms.

It should be noted that AFNI provides another function for cluster thresholding, ETAC (equitable thresholding and clustering; Cox, 2018), which performs better than the long-tail ACF function (Cox et al., 2017b) used here, but ETAC was not available when we started the new group analyses. AFNI also provides nonparametric group inference in the 3dttest++ function.

FIGURE 5 Results for FSL OLS one-sample *t* tests for cluster-wise inference using a CDT of $p = .001$, for no ICA FIX, pretrained ICA FIX and retrained ICA FIX. (a) Results for Beijing data (b) results for Cambridge data (c) results for Oulu data. Results are only shown for 4 mm smoothing, as other smoothing levels would require retraining the ICA FIX classifier. The two-sided tests show a higher degree of false positives compared to the one-sided tests. This is explained by the fact that a two-sided test involves two tests at $p = .025$, instead of one test at $p = .05$, and parametric methods are relatively more inflated at more stringent significance thresholds (as the statistical assumptions are more critical for the tail of the null distribution)

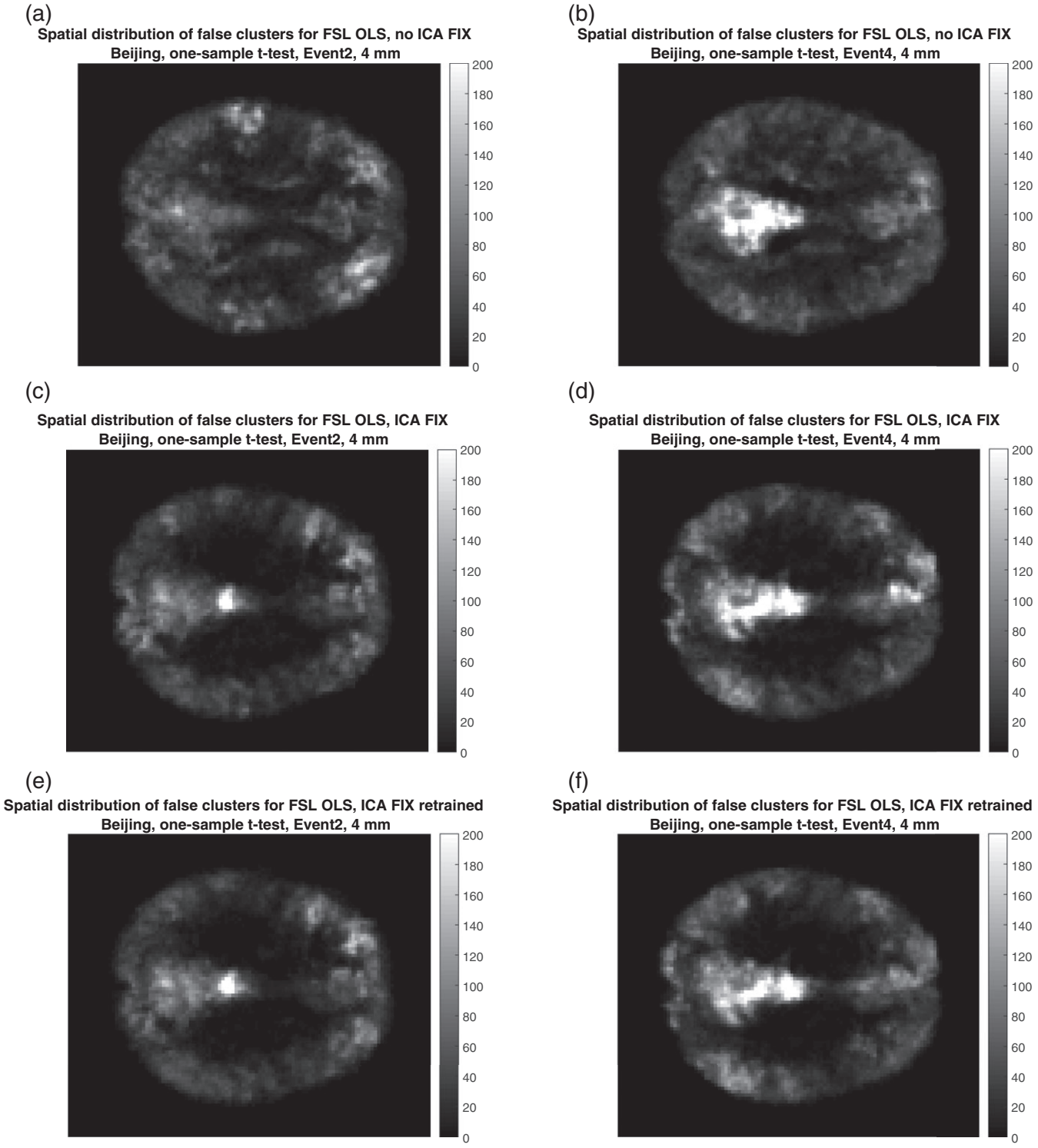


FIGURE 6 The maps show voxel-wise incidence of false clusters for the Beijing data, for two of the six different first level designs (a,b) no ICA FIX (c,d) ICA FIX pretrained (e,f) ICA FIX retrained for Beijing. Left: results for design E2, right: results for design E4. Image intensity is the number of times, out of 10,000 random analyses, a significant cluster occurred at a given voxel (CDT $p = .01$) for FSL OLS. Each analysis is a one-sample t test using 20 subjects. The maps represent axial slice 50 (MNI z coordinate = 26) for the MNI152 2 mm brain template used in FSL

4.1 | Influence of artifacts 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 modeling 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, Ye, Petridou, Murphy, & Bandettini, 2007); larger voxels lead to a lower temporal signal to noise ratio. The Oulu data have a

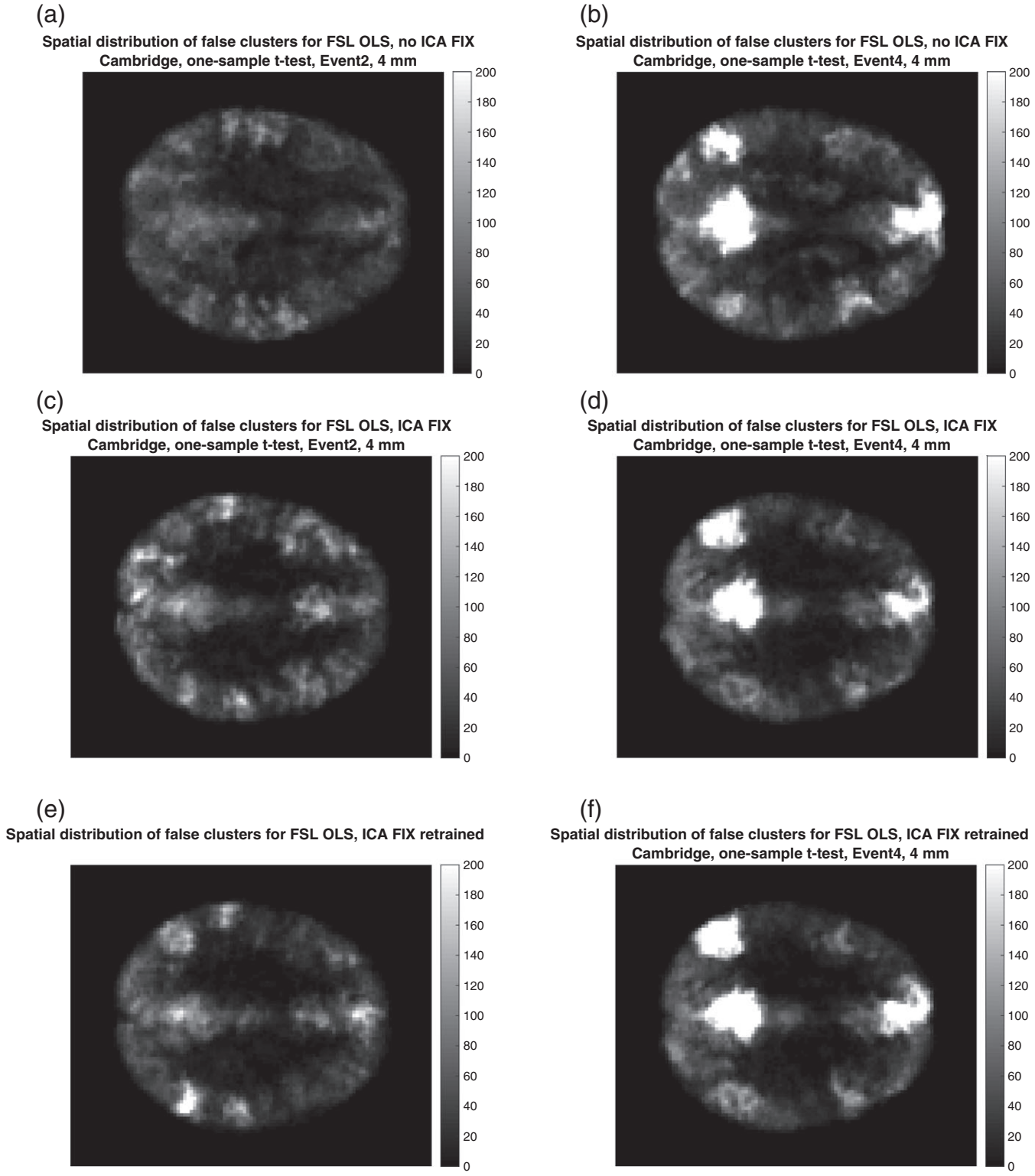


FIGURE 7 The maps show voxel-wise incidence of false clusters for the Cambridge data, for two of the six different first level designs (a,b) no ICA FIX (c,d) ICA FIX pretrained (e,f) ICA FIX retrained for Cambridge. Left: results for design E2, right: results for design E4. Image intensity is the number of times, out of 10,000 random analyses, a significant cluster occurred at a given voxel (CDT $p = .01$) for FSL OLS. Each analysis is a one-sample *t* test using 20 subjects. The maps represent axial slice 50 (MNI *z* coordinate = 26) for the MNI152 2 mm brain template used in FSL

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 two 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. As mentioned in Section 1, one can argue that a pure simulation (Welvaert & Rosseel, 2014) would avoid the problem of physiological noise, or that the Oulu data should be set aside, but we here opted to show results after denoising with ICA FIX, as many fMRI data sets have been collected without recordings of breathing and pulse.

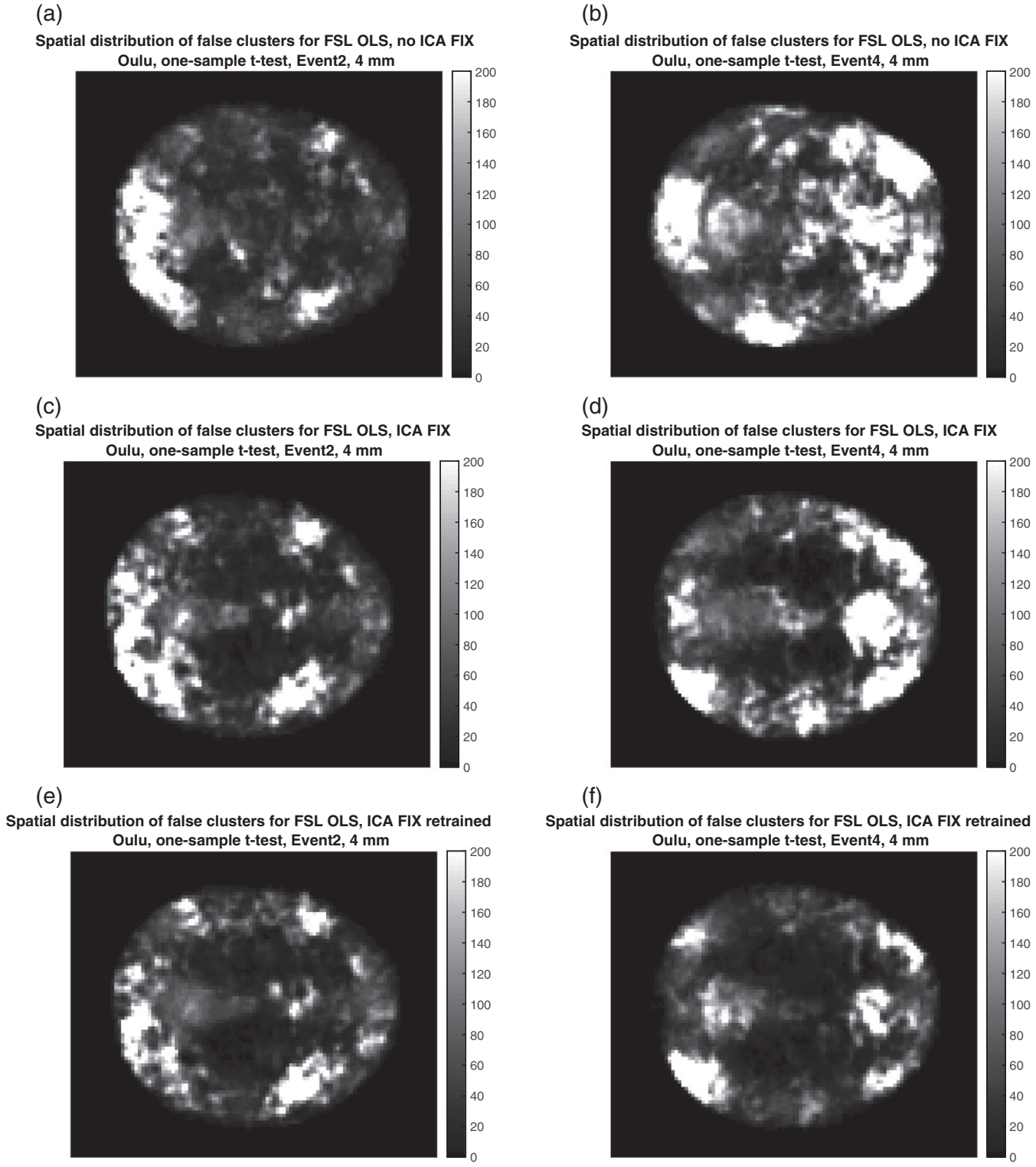


FIGURE 8 The maps show voxel-wise incidence of false clusters for the Oulu data, for two of the six different first level designs (a,b) no ICA FIX (c,d) ICA FIX pretrained (e,f) ICA FIX retrained for Oulu. Left: results for design E2, right: results for design E4. Image intensity is the number of times, out of 10,000 random analyses, a significant cluster occurred at a given voxel (CDT $p = .01$) for FSL OLS. Each analysis is a one-sample *t* test using 20 subjects. The retrained ICA FIX classifier is clearly better at suppressing artifacts compared to the pretrained classifier, especially for design E4. The maps represent axial slice 50 (MNI *z* coordinate = 26) for the MNI152 2 mm brain template used in FSL

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

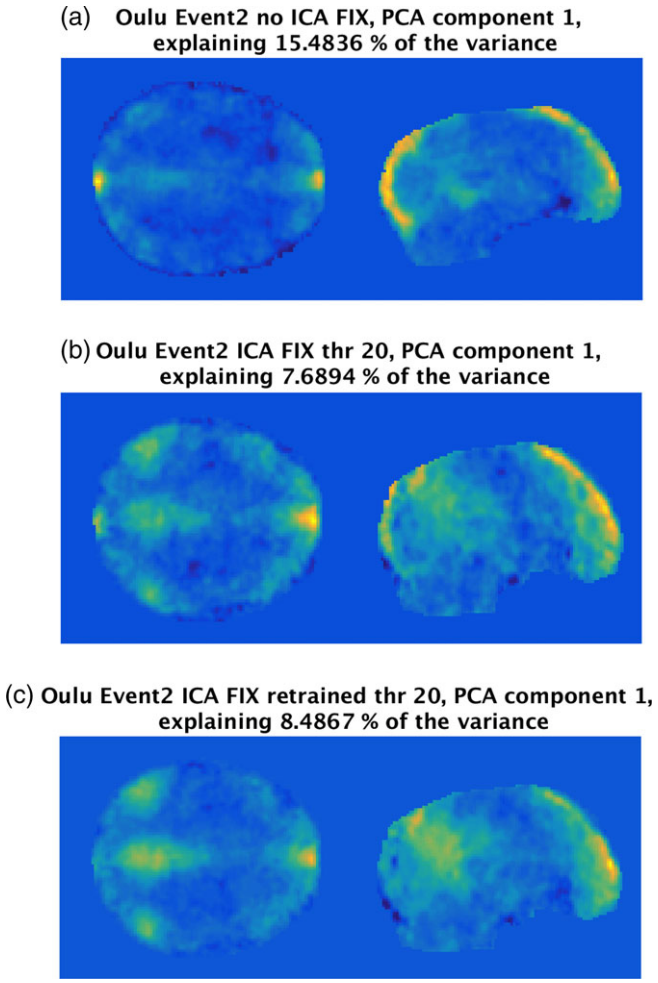


FIGURE 9 The maps show an axial and a sagittal view of the first Eigen component after running PCA on the 103 activity maps for Oulu E2, (a) without ICA FIX (b) with ICA FIX, using the pretrained classifier, (c) with ICA FIX, after retraining the ICA FIX classifier specifically for Oulu data. Using ICA FIX clearly suppresses the posterior part of the vein artifact in the superior sagittal sinus, but a portion of the artifact is still present. The retrained ICA FIX classifier is clearly better at suppressing the artifact. The axial maps represent axial slice 50 (MNI *z* coordinate = 26) for the MNI152 2 mm brain template used in FSL. The sagittal maps represent sagittal slice 48 (MNI x coordinate = −4)

not for Oulu. As can be seen in Figures 6–8, using ICA FIX clearly leads to false cluster maps which are more uniform across the brain, with a lower number of false clusters in white matter. Retraining the ICA FIX classifier finally lead to nominal results for the Oulu data. A possible explanation is that the pretrained classifier for standard fMRI data in ICA FIX is trained on fMRI data with a spatial resolution of 3.5 x 3.5 x 3.5 mm^3 (i.e., 1.6 times smaller voxels than Oulu). Figure 8 shows that the retrained classifier leads to more uniform false cluster maps, compared to the pretrained classifier, for design E4 for Oulu. As can be seen in Figure 9, the retrained classifier is better at suppressing the artifact in the sagittal sinus, compared to the pretrained classifier. We here trained the classifier for each data set (Beijing, Cambridge, Oulu) using labeled ICA components from 10 subjects, as recommended by the ICA FIX user guide, and labeling components from more subjects can lead to even better results.

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 data set, 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 (Birn et al., 2006; Bollmann et al., 2018; Chang & Glover, 2009; Glover et al., 2000; Lund et al., 2006). This is especially important for 7T fMRI data, for which the physiological noise is often stronger compared to the thermal noise (Hutton et al., 2011; Triantafyllou et al., 2005). Alternatives to collecting physiological data, or using ICA FIX, include ICA AROMA (Pruim et al., 2015), DPARSF (Yan & Zang, 2010), and FMRIPrep (Esteban et al., 2018). DPARSF and FMRIPrep can automatically generate nuisance regressors (e.g., from CSF and white matter) to be included in the statistical analysis.

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, Kociuba, and Rowe (2018)), and an important topic for future work is establishing how this impacts parametric cluster inference. The nonparametric permutation test (Winkler et al., 2014) does not make any assumption regarding the shape of the SACF, and is therefore expected to perform well for any MR sequence.

4.3 | Interpretation of affected studies

In Appendix A, we provide a rough bibliographic analysis to provide an estimate of how many articles used this particular CDT $p = .01$ 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 $p = .01$, 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.

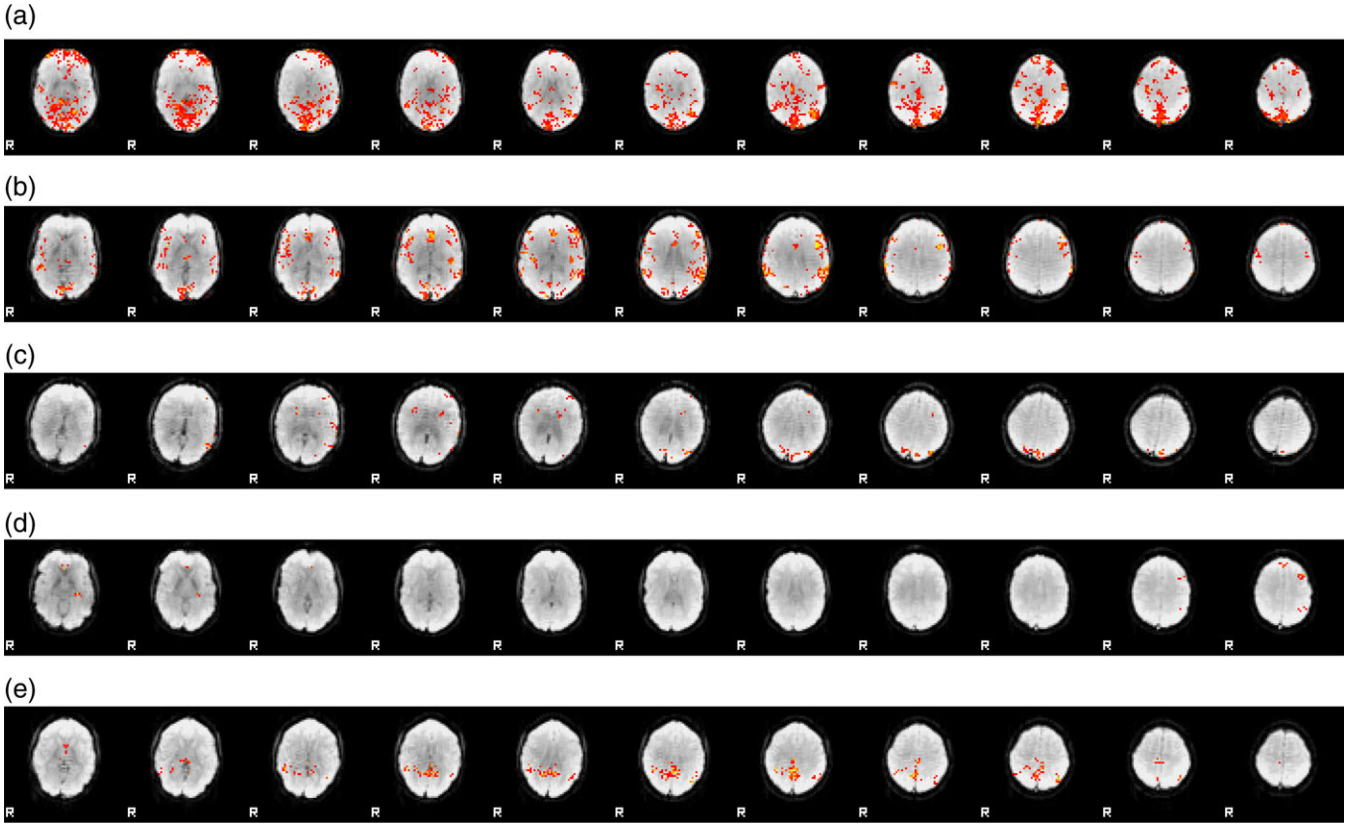


FIGURE 10 Activity maps (thresholded at CDT $p = .01$ and cluster FWE corrected at $p = .05$, FSL default) for five Oulu subjects analyzed with 4 mm of smoothing and first level design E4. Despite testing for a difference between two random regressors, which are for design E4 also randomized over subjects, significant voxels are in several cases detected close to the superior sagittal sinus (indicating a vein artifact). As 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 artifacts, since they cancel out when testing for a group difference

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 summarized 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 $p = .01$ 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 .05 could possibly survive a reanalysis with improved methods.

5 | CONCLUSIONS

To summarize, our new results confirm that inflated FWE 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 FWE rates for the nonparametric 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 the pretrained ICA FIX classifier for standard fMRI data. The Oulu data were collected with a lower spatial resolution, and are therefore more prone to physiological noise. By retraining the ICA FIX classifier specifically for the Oulu data, nominal results were finally obtained for Oulu as well. Data cleaning is clearly important for task fMRI, and not only for resting state fMRI.

ACKNOWLEDGMENTS

The authors have no conflict of interest to declare. 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, and the Knut och Alice Wallenbergs Stiftelse project "Seeing organ function." 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.

ORCID

Anders Eklund https://orcid.org/0000-0001-7061-7995

REFERENCES

Ashburner, J. (2012). Spm: A history. *NeuroImage*, *62*, 791–800.

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., Zuo, X., Gohel, S., Kelly, C., Smith, S. M., … Milham, M. P. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences of the United States of America*, *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.

Chen, G., Cox, R. W., Glen, D. R., Rajendra, J. K., Reynolds, R. C., & Taylor, P. A. (2018). A tail of two sides: Artificially doubled false positive rates in neuroimaging due to the sidedness choice with t-tests. https://doi.org/10.1101/328567.

Cox, R., Chen, G., Glen, D., Reynolds, R., & Taylor, P. (2017a). FMRI clustering and false positive rates. *Proceedings of the National Academy of Sciences of the United States of America*, *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.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*, 162–173.

Cox, R. W. (2018). Equitable thresholding and clustering, https://doi.org/10.1101/295931.

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. (2016a). Cluster failure: Why fMRI inferences for spatial extent have inflated false positive rates. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 7900–7905.

Eklund, A., Nichols, T., & Knutsson, H. (2016b). Correction for Eklund et al., cluster failure: Why fMRI inferences for spatial extent have inflated false positive rates. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, E4929.

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. *Proceedings of the National Academy of Sciences of the United States of America*, *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.

Esteban, O., Markiewicz, C., Blair, R. W., Moodie, C., Isik, A. I., Erramuzpe Aliaga, A., Kent, J., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S., Wright, J., Durnez, J., Poldrack, R., & Gorgolewski, K. J. (2018). FMRIPrep: A robust preprocessing pipeline for functional MRI, https://doi.org/10.1101/306951.

Flandin, G., & Friston, K. (2017). Analysis of family-wise error rates in statistical parametric mapping using random field theory. *Human Brain Mapping*, https://doi.org/10.1002/hbm.23839.

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

Genovese, C., Lazar, N., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, *15*, 870–878.

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.

Gopinath, K., Krishnamurthy, V., Lacey, S., & Sathian, K. (2018). 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.

Gopinath, K., Krishnamurthy, V., & Sathian, K. (2018). 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.

Gorgolewski, K., Alfaro-Almagro, F., Auer, T., Bellec, P., Capota, M., Chakravarty, M., … 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., … 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.

Greve, D. N., Brown, G. G., Mueller, B. A., Glover, G., & Liu, T. T. (2013). A survey of the sources of noise in fmri. *Psychometrika*, *78*, 396–416.

Griffanti, L., Douaud, G., Bijsterbosch, J., Evangelisti, S., Alfaro-Almagro, F., Glasser, M., … 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., … 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., … Weiskopf, N. (2011). The impact of physiological noise correction on fMRI at 7T. *NeuroImage*, *57*, 101–112.

Jenkinson, M., Beckmann, C., Behrens, T., Woolrich, M., & Smith, S. (2012). FSL. *NeuroImage*, *62*, 782–790.

Kessler, D., Angstadt, M., & Sripada, C. (2017). Reevaluating "cluster failure" in fMRI using nonparametric control of the false discovery rate. *Proceedings of the National Academy of Sciences of the United States of America*, *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., & Ugurbil, 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*, 345. https://doi.org/10.3389/fnhum.2017.00345

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., … 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., … Milham, M. (2013). Toward open sharing of task-based fMRI data: The OpenfMRI project. *Frontiers in Neuroinformatics*, *7*, 12. https://doi.org/10.3389/fninf.2013.00012.
Pruim, R. H., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J. K., & Beckmann, C. F. (2015). ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *NeuroImage*, *112*, 267–277.
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.
Welvaert, M., & Rosseel, Y. (2014). A review of fMRI simulation studies. *PLoS One*, *9*, 1–10.
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.
Yan, C., & Zang, Y. (2010). DPARSF: A MATLAB toolbox for "pipeline" data analysis of resting-state fMRI. *Frontiers in Systems Neuroscience*, *4*, 13.
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.

How to cite this article: Eklund A, Knutsson H, Nichols TE. Cluster failure revisited: Impact of first level design and physiological noise on cluster false positive rates. *Hum Brain Mapp*. 2019;40:2017–2032. https://doi.org/10.1002/hbm.24350

APPENDIX A: RESULTS FOR CDT p = .01

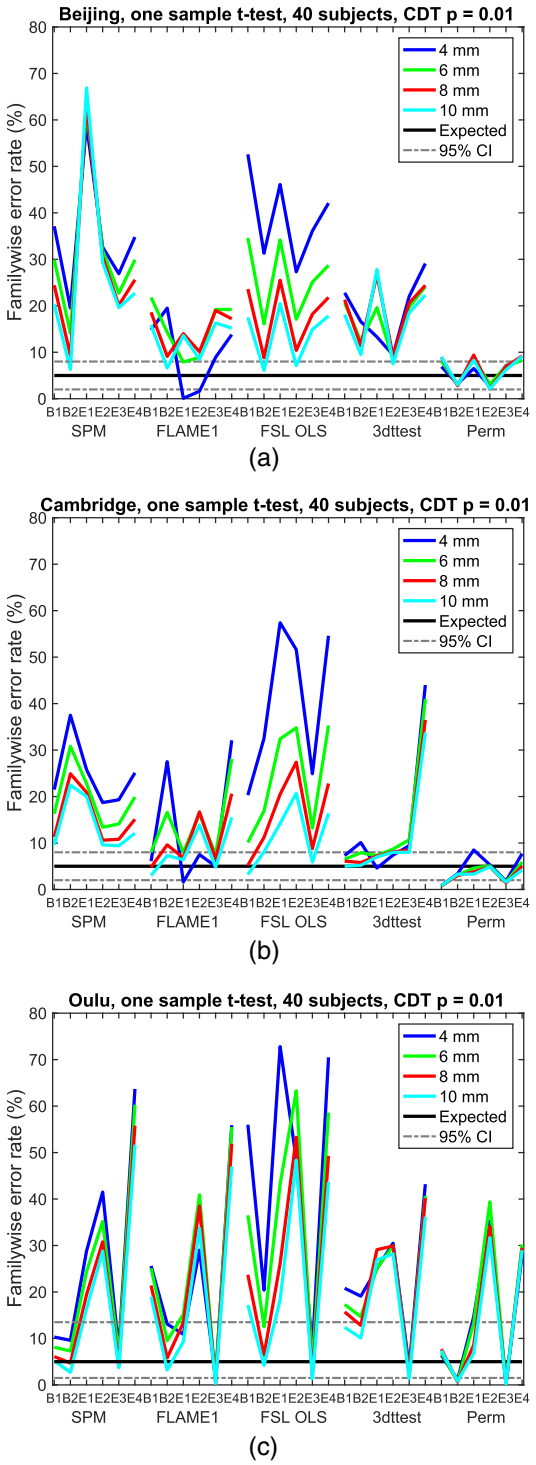
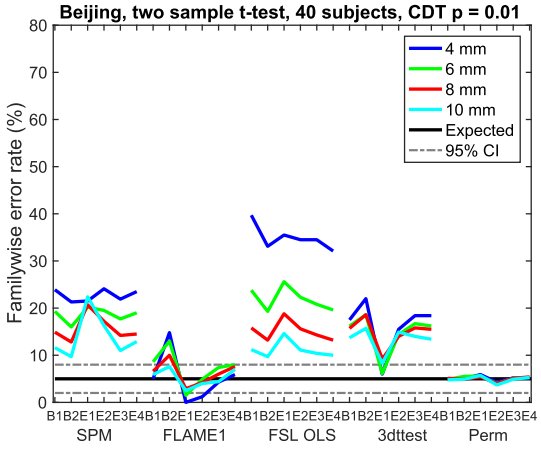
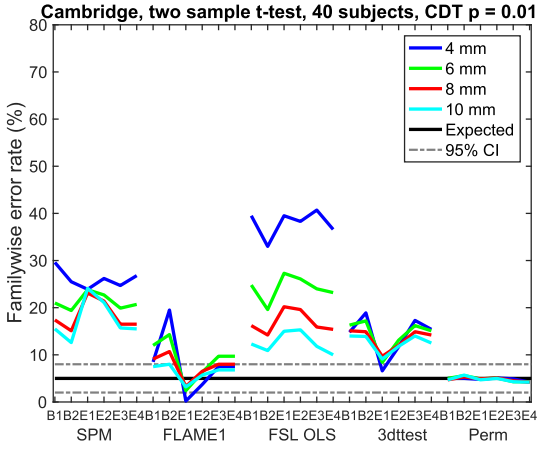


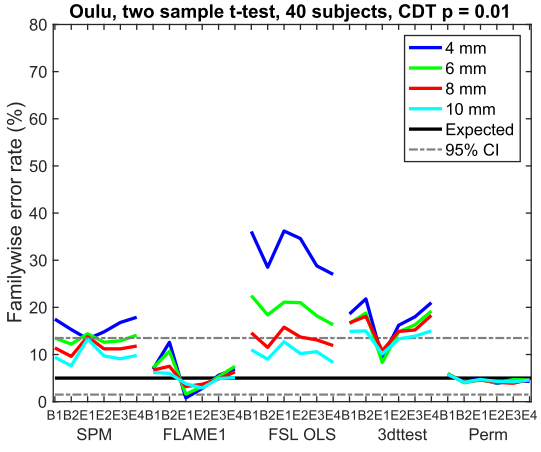
FIGURE A11 Results for one sample t test and cluster-wise inference using a CDT of p = .01, showing estimated FWE rates for 4–10 mm of smoothing and six different activity paradigms (old paradigms B1, B2, E1, E2, and new paradigms E3, E4), for SPM, FSL, AFNI, and a permutation test. These results are for a group size of 40. Each statistic map was first thresholded using a CDT of p = .01, uncorrected for multiple comparisons, and the surviving clusters were then compared to a FWE-corrected cluster extent threshold, p_{FWE} = .05. The estimated FWE 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 A12 Results for two sample t-test and cluster-wise inference using a CDT of $p = .01$, showing estimated FWE rates for 4–10 mm of smoothing and six different activity paradigms (old paradigms B1, B2, E1, E2, and new paradigms 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 $p = .01$, uncorrected for multiple comparisons, and the surviving clusters were then compared to a FWE-corrected cluster extent threshold, $p_{FWE} = .05$. The estimated FWE 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

APPENDIX B: BIBLIOMETRICS OF CLUSTER INFERENCE

Appendix B1. 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 CDT of $p = .01$.

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 (fMRI) = 23, 000 publications.[4] Drawing on a sample of 300 publications[5] published 2007–2012, Carp found P(HasData) = 241/300 = 80% contained original data, and among these P(Corrected| 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(ClusterInference| HasData) = 607/814 = 75%; noting that 6% (fig. 1, Woo et al.) of the 814 include studies with no correction, we also compute P(ClusterInference| HasData, Corrected) = $607/(814 - 0.06 \times 814) = 79\%$. Finally, with data from fig. 2(b) of Woo et al. (2014) (shown in Table A1, kindly supplied by the authors) from the 480 studies that used cluster inference with correction (and had sufficient detail) we can compute

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

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

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

and, among those, 10, 720 × 0.24 = 2, 573 used a CDT of $p = .01$ 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 $p = .01$ or higher, suggesting that the numbers above remain representative (Yeung, 2018).

[4]A total of 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 1,500 papers; then following exclusion criterion were applied "(a) non-human studies, (b) lesion studies, (c) studies in which a threshold or correction method could not be clarified, (d) voxel-based morphometry studies, (e) studies primarily about methodology, and (f) machine-learning based studies."

TABLE A1 Upon request, the authors of Woo, Krishnan, and Wager (2014) provided a detailed cross-tabulation of the frequencies of different CDTs (the data presented in fig. 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 CDT used

CDT	AFNI	BrainVoyager	FSL	SPM	Others	Total
>0.01	9	5	9	8	4	35
0.01	9	4	44	20	3	80
0.005	24	6	1	48	3	82
0.001	13	20	11	206	5	255
<0.001	2	5	3	16	2	28
Total	57	40	68	298	17	480