BRAIN CONNECTIVITY
Volume 7, Number 3, 2017
© Mary Ann Liebert, Inc.
DOI: 10.1089/brain.2016.0475

FMRI Clustering in AFNI: False-Positive Rates Redux

Robert W. Cox, Gang Chen, Daniel R. Glen, Richard C. Reynolds, and Paul A. Taylor

Abstract

Recent reports of inflated false-positive rates (FPRs) in FMRI group analysis tools by Eklund and associates in 2016 have become a large topic within (and outside) neuroimaging. They concluded that existing parametric methods for determining statistically significant clusters had greatly inflated FPRs (''up to 70%,'' mainly due to the faulty assumption that the noise spatial autocorrelation function is Gaussian shaped and stationary), calling into question potentially ''countless'' previous results; in contrast, nonparametric methods, such as their approach, accurately reflected nominal 5% FPRs. They also stated that AFNI showed ''particularly high'' FPRs compared to other software, largely due to a bug in 3dClustSim. We comment on these points using their own results and figures and by repeating some of their simulations. Briefly, while parametric methods show some FPR inflation in those tests (and assumptions of Gaussian-shaped spatial smoothness also appear to be generally incorrect), their emphasis on reporting the single worst result from thousands of simulation cases greatly exaggerated the scale of the problem. Importantly, FPR statistics depends on ''task'' paradigm and voxelwise p value threshold; as such, we show how results of their study provide useful suggestions for FMRI study design and analysis, rather than simply a catastrophic downgrading of the field's earlier results. Regarding AFNI (which we maintain), 3dClustSim's bug effect was greatly overstated—their own results show that AFNI results were not ''particularly'' worse than others. We describe further updates in AFNI for characterizing spatial smoothness more appropriately (greatly reducing FPRs, although some remain >5%); in addition, we outline two newly implemented permutation/randomization-based approaches producing FPRs clustered much more tightly about 5% for voxelwise $p \leq 0.01$.

Keywords: autocorrelation function; clusters; false-positive rates; FMRI; thresholding

Introduction

Reports (Eklund et al., 2015, 2016) of greatly inflated false-positive rates (FPRs) for commonly used cluster threshold-based FMRI statistics packages (SPM, FSL, AFNI) caused a stir in both technical and semipopular publications. In large part, this was due to dramatic summary statements about the collective previous approaches that, ''Alarmingly, the parametric methods can give a very high degree of false positives (up to 70%, compared with the nominal 5%) for clusterwise inference,'' and[1] ''These results question the validity of some 40,000 fMRI studies and may have a large impact on the interpretation of neuroimaging results'' (Eklund et al., 2016). They hypothesized that a major factor for the problematic inflation was the mistaken assumption across the software that ''spatial autocorrelation functions … follow the assumed Gaussian shape.'' They contrasted these results with the output of nonparametric permutation tests, which they summarized simply as producing ''nominal results for voxelwise as well as clusterwise inference.'' Additionally, Eklund and associates stated that a 15-year-old bug in 3dClustSim of the AFNI software package (which we maintain) resulted in ''cluster extent thresholds that are much lower compared with SPM and FSL'' and ''particularly high FWE rates'' (FWE = family-wise error) (Eklund et al., 2016). Published responses to these findings ranged from saying that there is nothing new here (''tests based upon spatial extent become inexact at low thresholds''; Flandin and Friston, 2016), to cautious-but-concerned commentary (Stockton, 2016), to inflated and conflated hyperbole about invalidating 15 years of research due to a software bug (BEC Crew, 2016; Russon, 2016).

Here, we comment on several of these main points not only by rerunning simulations using some of the same

[1]This second statement was later palliated to, ''These results question the validity of a number of fMRI studies and may have a large impact on the interpretation of weakly significant neuroimaging results'' (Nichols, 2016).

Scientific and Statistical Computing Core, NIMH/NIH/DHHS, Bethesda, Maryland.

''null'' resting-state FMRI data from the 1000 Functional Connectomes Project (FCON-1000; Biswal et al., 2010) but also by examining closely the results from their own study [with article (Eklund et al., 2016) often referred in this text as ''ENK16,'' for convenience]. Eklund and associates did a vast amount of work in their study, running tens of thousands of simulations with several software packages, as well as making the data and scripts publicly available for further testing. Although they have raised an important issue that needs to be addressed, it must be noted that several of their extreme summarizing claims do not appear to be accurate characterizations of their own results.

The AFNI team takes the question of the inflated FPRs very seriously, but does not consider that the FMRI ''apocalypse'' has arrived. Rather than summarizing previous methods' results based on typical or average results, Eklund and associates repeatedly referred to the single worst result in over 3000 simulations when summarizing parametric output (''up to 70% FPR''). We note that their own permutation method, under similar criterion, should be characterized as providing ''up to 40% FPR'' (both results come from ENK16—Supplementary fig. S9). While noting an upper limit of results may have some utility, in no case is that value a representative characterization of the method's typical performance. It would be much clearer and more informative to use median values and/or ranges to summarize and compare distributions of results; however, in ENK16, this was only done for their own nonparametric method.

This report comprises three themes. First, looking to the past, we repeat a subset of the Eklund and associates simulations using AFNI to show that the effect of the erstwhile bug within 3dClustSim, while not negligible, was not large and did not lead to ''particularly high'' FPRs. Looking at the present, we discuss modifications that have been made in clustering functionality with AFNI (as of January 30, 2017; version AFNI_17.0.03) to address the inflated FPRs in these tests. First, assumptions of spatial smoothness are evaluated and a new approach for estimating a non-Gaussian spatial autocorrelation function (ACF) is implemented, leading to greatly reduced FPRs in 3dttest++'s parametric approach, although in many cases still above the nominal 5%; then, we present a new permutation/randomization-based approach (generating cluster-size thresholds from the residuals at the group level) that is now implemented in AFNI, which shows FPRs clustered tightly about 5% across all voxelwise $p \leq 0.01$ thresholds. These methods, first outlined at the 2016 OHBM meeting (Cox and Reynolds, 2016), are demonstrated on sets of simulations following those in ENK16. Finally, looking to the future, an extension and generalization of this new approach to allow for spatially variable cluster-size thresholding show promise in controlling FPRs while allowing for spatially variable smoothness in the FMRI noise. Throughout, we refer to various specific results of Eklund and associates from their article, Supplementary data, and tabular data they made available on GitHub. As in ENK16, all discussions of FPR refer to cluster counts at the whole brain level (as opposed to voxelwise p values).

Methods: Simulations

For simulations performed here, we repeated the steps carried out in Eklund et al. (2015, 2016) using the 198 Beijing-Zang data sets from the FCON-1000 collection (Biswal et al., 2010). The detailed AFNI processing for individual subjects was somewhat different than in ENK16, since we ran with our most up-to-date recommendations for preprocessing (e.g., despiking, using 3dREMLfit with generalized least squares and ARMA(1,1), prewhitening to allow for temporal correlation, AnatICOR for denoising, and 3dQwarp for nonlinear registration; see Supplementary Data (Supplementary Data are available online at www.liebertpub.com/brain) for the complete alignment and afni_proc.py commands), but the results are quite comparable. The group analyses, using 1000 random subcollections of the resting-state FMRI Beijing data sets, were carried out using 3dttest++ in the same way as described in ENK16.

In ENK16, various combinations of simulation parameters produced widely varying levels of agreement or disagreement with the nominal 5% setting for FPRs for all software tools tested. To present a broad description, the comparisons presented here are for the set of basic scenarios put forth in ENK16. This includes investigating the four values of Gaussian smoothing applied (full-width at half-maximum [FWHM] of 4, 6, 8, and 10 mm) and two different voxelwise p value thresholds (0.01 and 0.001; here, we also include the intermediate $p = 0.005$). Four separate pseudostimulus timings were used in analysis of these null (resting-state) FMRI data sets: blocks of 10-sec ON/OFF (''B1'') and 30-sec ON/OFF (''B2''), and event-related paradigms of (regular) 2-sec task with 6-sec rest (''E1'') and (random) 1–4-sec task with 6-sec rest (''E2''). Each subject had the same stimulus timing for each case (as in ENK16).

Results

The past, I: 3dClustSim, and "The Bug"

One problem with past results was particular to AFNI: there was a bug in 3dClustSim. This program works by generating a 3D grid of independent and identically distributed $N(0,1)$ random deviates, smoothing them to the level estimated from the residuals of the FMRI data model at the individual level, carrying out voxelwise thresholding, and finally clustering to determine the rate at which contiguous clumps of different sizes occur at the various voxelwise thresholds. The bug, pointed out by the ENK16 authors in an e-mail, was a flaw in how 3dClustSim rescaled the simulated 3D noise grid after smoothing to bring the variance of the values back to 1.0 (for ease of later p value thresholding). This rescaling was off due to improper allowance for edge effects (effectively, zeros ''outside'' the grid were included in the smoothing), with the result being that the cluster-size thresholds computed were slightly too small, so that the FPR would end up somewhat inflated. During part of the work leading to (Eklund et al., 2015, 2016), this bug was fixed in May 2015 and noted in the regular and publicly available log of AFNI software changes:

> 12 May 2015, RW Cox, 3dClustSim, level 2 (MINOR), type 5 (MODIFY) Eliminate edge effects of smoothing by padding and unpadding Simulate extra-size volumes then smooth, then cut back to the desired volume size. Can use new ''-nopad'' option to try the old-fashioned method. (H/T to Anders Eklund and Tom Nichols.)[2]

[2]https://afni.nimh.nih.gov/pub/dist/doc/program_help/history_all.html

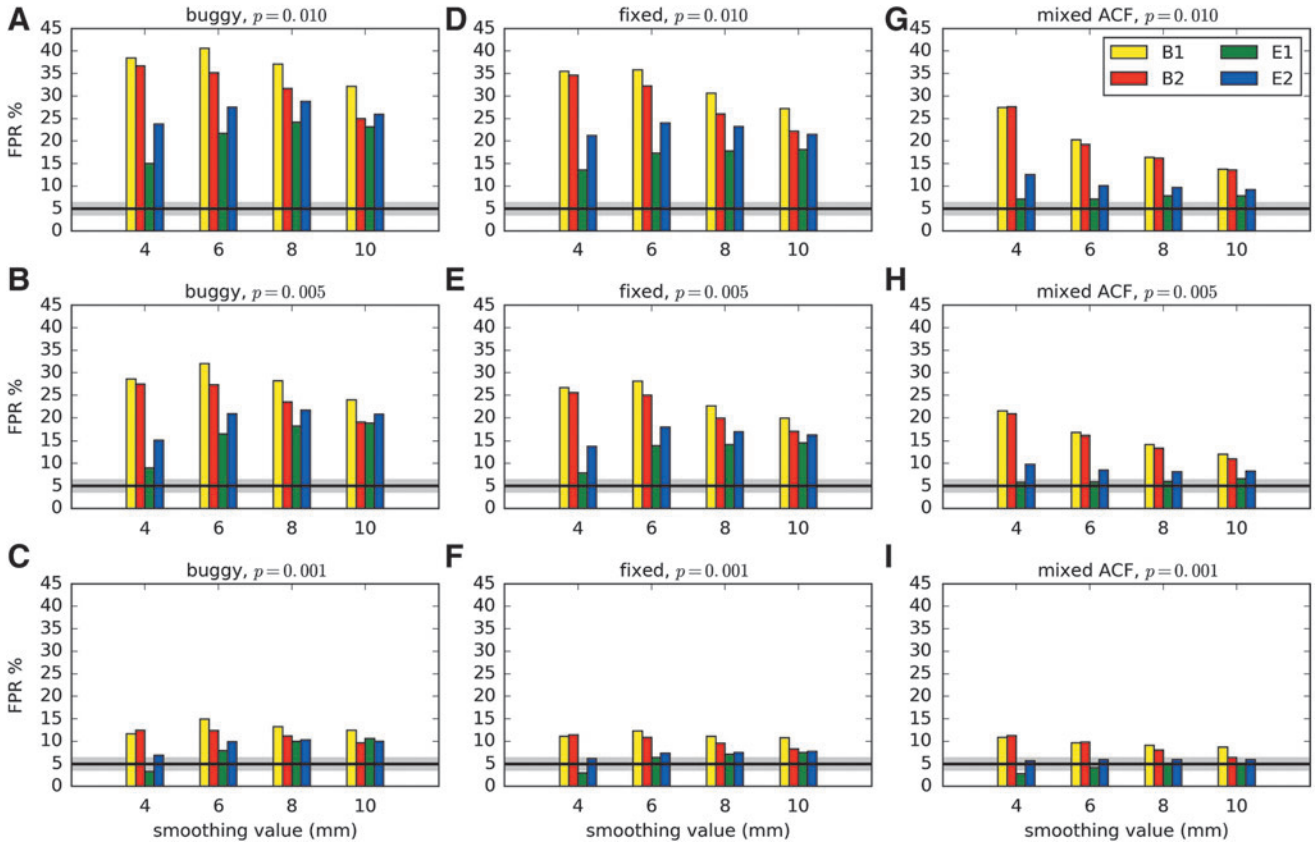


FIG. 1. FPRs for various software scenarios in AFNI, with 1000 two-sample 3D *t*-tests [as in Eklund and associates (2015, 2016)] using 20 subjects' data in each sample. ''Buggy'' **(A–C)** and ''fixed'' **(D–F)** mean that the cluster-size thresholds were selected using the Gaussian shape model with the FWHM being the median of the 40 individual subject values: ''buggy'' and ''fixed'' via 3dClustSim before and after the bug fix, respectively. ''Mixed ACF'' **(G–I)** means that the cluster-size threshold was selected using Eq. (3) for spatial correlation of the noise, with the *a,b,c* parameters being the median of the 40 individual subject's values (estimated via program 3dFWHMx). Three different voxelwise *p* value thresholds [one-sided tests, as used in Eklund and associates (2016)] are shown. The black line shows the nominal 5% FPR (out of 1000 trials), and the gray band shows its theoretical 95% confidence interval, 3.6–6.4%. As in ENK16, different smoothing values were tested (4–10 mm). B1 = 10-sec block; B2 = 30-sec block; E1 = regular event related; E2 = randomized event related. ACF, autocorrelation function; FPR, false-positive rate; FWHM, full-width at half-maximum.

Results comparing the pre- and postfix versions of the standard 3dClustSim (''buggy'' and ''fixed,'' respectively) are shown in Figure 1 (first two columns), which presents FPRs from rerunning the two-sample *t*-tests (40 subjects total per each of 1000 3D tests) of ENK16. Comparing column heights for identical parameters, the size of inflation due to the bug ($\Delta\text{FPR} = \text{FPR}_{\text{buggy}} - \text{FPR}_{\text{fixed}}$) was modest, particularly for $p = 0.001$ where the inflation was $\Delta\text{FPR} < 1$–2%. At the less stringent voxelwise $p = 0.01$, where FPRs had been noticeably more inflated for most software packages, the difference was greatest for the largest smoothing (understandably, given the problem within the program), with approximately $\Delta\text{FPR} < 3$–5%. In each case, the difference between the ''buggy'' and ''fixed'' values was small compared to the estimated FPR, meaning the bug had only a relatively minor impact (contrary to the statements in ENK16). Similar magnitudes of changes were found with one-sample *t*-test simulations. (Some differences between the ''buggy'' results herein and those in ENK16 are likely due to the improved time series regression procedures that we used; however, we did not systematically investigate the magnitude or details of this potential effect.)

At $p = 0.001$, the ''fixed'' results for the event-related stimulus timings are not far from the nominal 5% FPR (Fig. 1, lower panel); however, the corresponding ''fixed'' results for the block design stimulus timings are still somewhat high. The $p = 0.01$ results are all still far too high in the ''fixed'' column (Fig. 1, upper panel). These results are revisited in the discussion of smoothness estimation below.

The past, II: Parametric method results in AFNI and other software

Looking at ENK16—figure 1, as well as at the Supplementary data; figures S1–12, it is difficult to see how any one set of parametric results (SPM, FLS-OLS,[3] AFNI-3dttest++ or

[3]Here and below, we do not compare to FSL-FLAME1 results, which often had much lower FPRs for voxelwise $p = 0.01$, and quite conservative results below the nominal FPR for $p = 0.001$; furthermore, we do not further investigate the other software tools, and merely note that FSL-FLAME1 had a much different behavior from the rest of the ''previous'' approaches examined in Eklund and associates (2016).

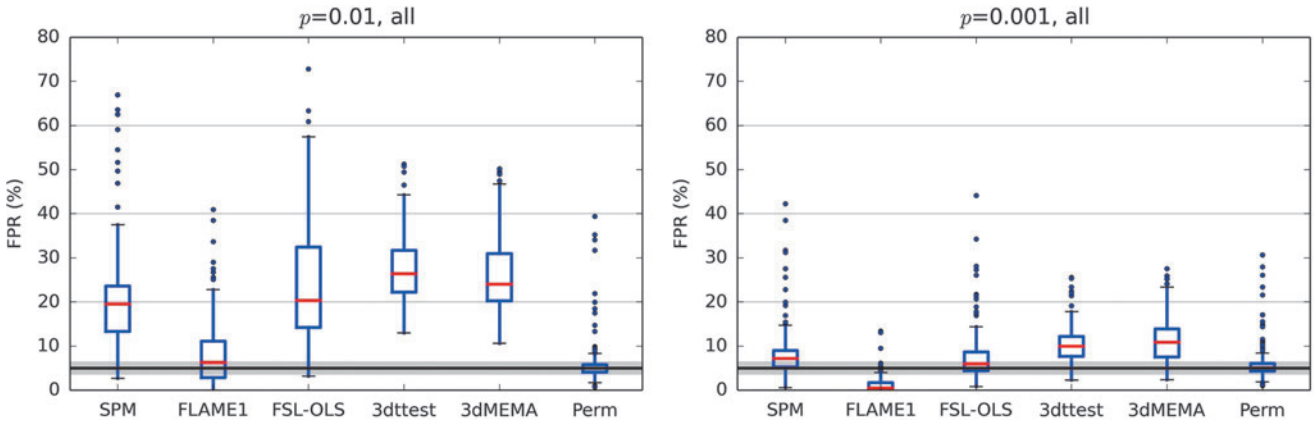


FIG. 2. Summary of the FPR results examined in Eklund and associates (2016), combining all their test results (available from their GitHub repository). The results of each software across all voxelwise $p = 0.01$ and $p = 0.001$ cases are shown separately. Red lines show the median; the box covers the 25–75% interquartile range; whiskers extend to the most extreme data point within 1.5× the interquartile range; and outliers are shown as dots. For a given voxelwise p, results are similar across parametric methods, with typical ranges of 15–30% FPR for $p = 0.01$ and 5–15% FPR for $p = 0.001$.

AFNI-3dMEMA) can be classified as having a ''particularly high'' FPR compared to the others. This general similarity of FPRs is true even though the ''buggy'' 3dClustSim was used in creating them, further showing its small effect on inflation. For the purpose of comparison across software, Figure 2 shows boxplots of their combined results across all their simulation cases (publicly available on GitHub; see Appendix 1 here for further decomposition of the results). For voxelwise $p = 0.01$, most parametric methods have results in the range of 15–40% FPR, and for $p = 0.001$, most tests have 5–15% FPR. We also note that the oft-cited ''up to 70% FPR'' result was not produced by AFNI.

The figure cited by Eklund and associates to demonstrate that AFNI ''cluster extent thresholds that are much lower compared with SPM and FSL'' (ENK16—Supplementary data; Appendix; Supplementary fig. S16) is not one based on general or summary results. Instead, it simply recapitulates the results of one set of simulations whose FPRs are already shown in one column of ENK16—Supplementary figure S1-a, with two-sample comparison ($n = 10$ subjects in each), one-sided testing, voxelwise $p = 0.01$, and 6 mm smoothing for the ''E2'' (random event) task. Given the FPRs shown for that one case (AFNI $\approx$30% FPR, SPM $\approx$20% FPR, Perm $\approx$5% FPR), it is perhaps not surprising that the AFNI cluster extent in this case is smaller than SPM by $\approx$50% and smaller than Perm by a factor of five.[4] ENK16—Supplementary figures S1-a and S16 show similar, consistent pieces of information for a single case, but this individual case is not a generic feature; certainly there are many simulations where AFNI had a lower FPR and likewise a greater cluster extent. As noted above, a single case from a host of simulations does not provide grounds for representative characterization or generalization.

In addition, we note that the comparison of cluster extents in ENK16—Supplementary figure S16, did not account for neighborhood differences across the software. The examined software packages use different voxel neighborhood definitions by default: AFNI 3dClustSim as run by Eklund and associates used a definition of nearest neighbor NN = 1 (face-wise bordering; $n = 6$ voxels, but the user could set it to any value—and now 3dClustSim always gives tables of results for NN = 1, 2, and 3); SPM uses NN = 2 (face + edge; $n = 18$ voxels); FSL and Perm use NN = 3 (face + edge + corner; $n = 26$ voxels). The nearest neighbor definition should not in itself significantly affect final FPR values, as long as comparisons are made self-consistently between threshold and cluster volume within each package. However, any comparison of threshold cluster volumes across software would have to take this into consideration, as the volume differences are typically in a range of roughly 5–20% (it is unclear what additional information the cluster extents would add to the FPR comparisons themselves, once the neighborhood choices are accounted for).

In addition, the parametric methods of AFNI, SPM, and FSL were evaluated with an ''*ad hoc*'' clustering method from Cunningham and Lieberman (2009), with results shown in ENK16—figure 2; ''Perm'' method results were not provided for this test. We note that we have *never* endorsed using this ''*ad hoc*'' approach, as it seems neither theoretically grounded nor experimentally generalizable. The initial article by Cunningham and Lieberman (2009) applied a voxelwise $p = 0.005$ with an additional extent threshold of +10 voxels to mimic the behavior of FPR = 5%, compared to running 1 million simulations; this was evaluated in only a single group of 32 subjects, where the additional value of ''+10 voxels'' was applied to a data set having $3.5 \times 3.5 \times 5\,\text{mm}^3$ voxels (for a total of 39,828 voxels within the whole brain mask) and 6 mm of blurring. Regardless of the results in that single data set, simply using a 10 voxel cluster threshold with any voxelwise p-threshold value in a data set of any arbitrary resolution and any processing stream cannot reasonably be expected to produce an appropriate FPR [e.g., ENK16 upsampled the FMRI data to $2 \times 2 \times 2\,\text{mm}^3$ voxels but still applied the same ''+10 voxel'' rule—which in this case would have the volume of only 1.3 voxels from Lieberman and Cunningham (2009). The fact that all presented software

[4]It is likely also consistent with FSL-OLS result, but we do not know why FSL-FLAME1 is so much more conservative, assuming the same cluster size; Eklund et al. discuss this in Eklund and associates (2016).

tools performed similarly in this test—and with such poor apparent FPRs—is not surprising.[5] This particular clustering method is "*ad hoc*" *at best*, and absolutely unfounded when carried out under differing conditions (particularly spatial resampling).

In summary, we simply cannot understand the conclusion from the Results section in the article by Eklund and associates that the AFNI FPR was "particularly high" compared to other software. The plotting of their own results (ENK16—fig. 2) shows similar performance across all parametric software packages. This is more comprehensively observable in Figure 2 here, which presents the medians and typical ranges of FPR values from all of the simulations performed in ENK16, and which provides a much better characterization of the results than using (some) extrema.

The present, I: Updating "The Flaw" and assumptions about spatial smoothness

The second problem in determining cluster-size threshold is much more widespread (to date) across the tools most used in the FMRI community: it is the flawed assumption that the shape of the ACF in the FMRI noise is Gaussian in form. That is, it has been generally assumed that, for voxels separated by Euclidean distance r, the spatial correlation between noise values has the form

$$f(r) = \exp\left[-r^2/\left(2b^2\right)\right], \text{ with } b > 0, \tag{1}$$

where it is traditional to specify the parameter b, and therefore, the full shape of Eq. (1) by the related width parameter:

$$\text{FWHM} = 8[\ln(2)]^{1/2} \times b = 2.35482 \times b. \tag{2}$$

In fact, as pointed out in ENK16, the empirical ACF shape (computed from the model residuals and averaged across the whole brain) has much longer tails than the Gaussian shape in Eq. (1). The heavy-tailed nature of spatial smoothness within the brain, which had been largely ignored previously, has significant consequences for thresholding clusters in FMRI analyses.

Figure 3 illustrates the problem along with the current solution adopted in AFNI. The empirical correlation falls off rapidly with r at first, but then tails off much slower than the Gaussian function. We found that the empirical ACF estimates are typically well fit by a function that mixes the Gaussian and monoexponential form

$$h(r) = a\exp\left[-r^2/\left(2b^2\right)\right] + (1-a)\exp[-r/c], \quad \text{with } 0 \le a \le 1 \text{ and } b, c > 0. \tag{3}$$

Given the demonstrated inadequacy of the pure Gaussian model in Eq. (1), 3dClustSim was modified to allow the generation of random 3D fields with autocorrelation given by Eq. (3). The mixed ACF model is now available in AFNI, and the (a,b,c) parameters are computed from each subject's time series regression model residuals in our FMRI processing stream tool (afni_proc.py).

[5]Results of their own nonparametric approach with this *ad hoc* clustering method were not presented by ENK16, although it is difficult to see how they would be significantly different than the shown parametric methods.

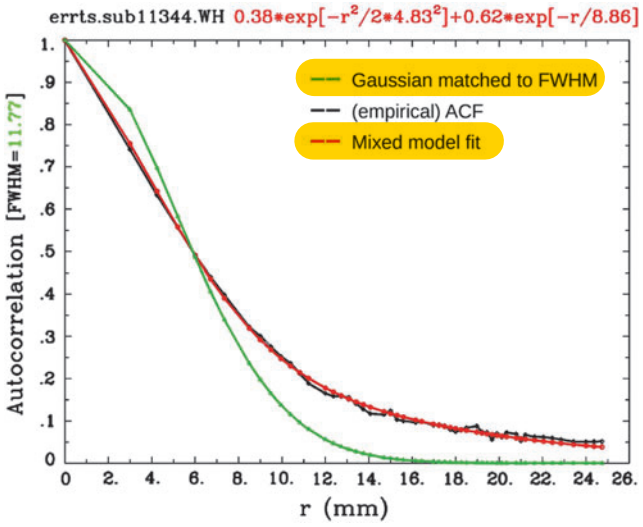


FIG. 3. An example comparison of the original Gaussian fit (green) and the globally estimated empirical ACF values (black) from a single subject, which have large differences (importantly, in the tail drop-off above $r \sim 8$ mm). The proposed mixed model (red) after fitting parameters as described in Eq. (3) provides a much better fit of the data in this case (and in all cases in the data sets used herein). This plot is automatically generated in program 3dFWHMx.

To illustrate that the long tails in the spatial ACF can make a difference in cluster thresholding, we used 3dClustSim to compute the cluster-size threshold for a nominal 5% FPR using the 198 sets of estimated ACF parameters from the Beijing cohort. The voxelwise thresholds were taken as $p = 0.010$ and 0.001 (left and right panels, respectively, of Fig. 4) from the NN = 2 one-sided thresholding table output by 3dClustSim. For each subject, 3dClustSim was run twice: once using a Gaussian ACF model with the FWHM estimated from the ACF mixed model (cf. Fig. 3), and once using the full mixed model ACF of Eq. (3). For $p = 0.010$, the cluster-size thresholds estimated from the longer tailed mixed model ACF are significantly larger than the Gaussian ACF model, showing that the long tails in the ACF have a large impact at larger p thresholds. For $p = 0.001$, the cluster-size thresholds from the two ACF cases differ far less, but are nontrivial. These facts are a major part of why the parametric software results in ENK16 are generally markedly "better" for the smaller p value threshold (cf. Fig. 2).

FPRs from rerunning the two-sample two-sided t-tests of ENK16, using this new "mixed ACF" model option in 3dClustSim, are also shown in Figure 1 (third column). For the voxelwise threshold of $p = 0.01$, the impact of the bug fix is much smaller than that of the long tail "fix" provided in the mixed ACF model. For the voxelwise threshold $p = 0.001$, the impact of the bug fix is about the same as that of the long tails in the mixed ACF model, as these estimates were closer to the nominal 5% rate already. In each case the bug fix reduced the FPR values, as did the change to the mixed ACF model. Extensive further results, using several variations on this type of cluster analysis, with one- and two-sample, one- and two-sided t-tests, are given in Appendix 2.

For the block designs (B1 and B2, 10- and 30-sec blocks of "task"), the FPRs are still somewhat high even with the

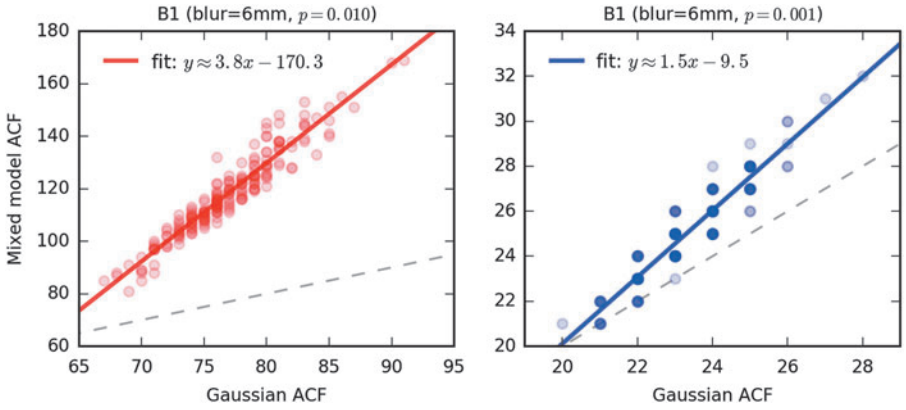


FIG. 4. Cluster-size thresholds from 3dClustSim ran over the estimated ACF for each of the 198 data sets in the Beijing-Zang cohort. The x-axis is the cluster-size threshold assuming a Gaussian-shaped ACF, with FWHM taken from the mixed model ACF estimate for each subject. The y-axis is the cluster-size threshold assuming the mixed model ACF of Eq. (3); the parameter estimates are computed from the residuals from the pseudostimulus B1, blur = 6 mm time series analyses. The left graph is for per voxel p threshold 0.010; the right graph is for $p = 0.001$. Approximate linear fits are shown overlaid; the dashed gray line shows $x = y$, providing a reference to indicate the disparity in cluster-size thresholds between the Gaussian and mixed-model ACF assumptions. Darker circles indicate points where multiple subjects had the same pair of thresholds (which are integer valued). Cluster-size thresholds are taken from the NN = 2, one-sided test table output from 3dClustSim (which also output tables for NN = 1 and NN = 3, and for two-sided tests).

mixed ACF model. Preliminary investigations indicate this bias is partly due to the fact that the spatial smoothness of the FMRI noise is a function of temporal frequency—that is, the FMRI noise at lower temporal frequencies is somewhat smoother than at higher frequencies. We do not know if that effect accounts for most of the disparity between the block- and event-related designs.

The present, II: A nonparametric approach to cluster-size thresholding

A second approach to adjusting the FPR in cluster-size thresholding has been implemented in the AFNI program 3dttest++ (which is also capable of incorporating between-subjects factors and covariates, in addition to carrying out the simple voxelwise t-tests implied by its name; perhaps it should have been named 3dOLStest). The procedure is straightforward:

- Compute the residuals of the model at each voxel at the group level.
- Generate a null distribution by randomizing among subjects the signs of the residuals in the test (and permuting subject data sets between groups, if doing a two-sample test without covariates), repeat the t-tests (with covariates, if present), and iterate 10,000 times.
- Take the 10,000 3D t-statistic maps from the randomization and use those as input to 3dClustSim (with no additional smoothing or random deviate generation): threshold the maps at a large number of p values (e.g., 0.01, 0.005, and 0.001), clusterize them, then record the maximum cluster size for each p value from each t-map, accumulate statistics across the 10,000 t-maps, and then determine the cluster-size threshold to achieve a given FPR, for each p value.

All these steps are carried out by using the 3dttest++ program with the command line option ''-Clustsim.''

The output is a table of cluster-size thresholds for a range of voxelwise p value thresholds and a range of cluster-significance values. Such a table is produced for each of the clustering methods that AFNI supports: nearest neighbors NN = 1,2,3, and one-sided or two-sided voxelwise thresholding. (In general, we prefer two-sided t-statistic thresholding in AFNI, as providing more transparency into the analysis results given the types of questions typically asked by researchers, but we do allow the user to opt for one-sided thresholding when appropriate.[6]) These tables are saved in text format, and also stored in the header of the output statistics data set for use in interactive thresholding in the AFNI GUI.

For comparison here, the 1000 two-sample t-tests described above were rerun for the 16 cases (4 blurring levels times 4 stimulus timings) with this new ''-Clustsim'' option, and tested against each of the six combinations of thresholding-sidedness and clustering-neighborliness possible in AFNI, over a range of voxelwise p value thresholds. The results were similar across all 96 cases (and across sets of other tests, including one sample, paired, and with covariates). The results for the one-sided t-test NN = 1 nearest neighbor clustering approach are shown graphically in Figure 5; all FPRs are within the ''nominal'' 95% confidence interval for the FPR (3.65–6.35%) over the collection of voxelwise p value thresholds tested. At this time, the use of this option is one of our recommendations, for cases where the group analysis can be carried out via a simple general linear model (GLM) with or without covariates.

[6]Performing a pair of one-sided tests in both directions under all circumstances would make it easier for a cluster to survive familywise error correction, but it would also increase the FPR when a two-sided test is more appropriate. This unfortunate technique is all too common in FMRI.

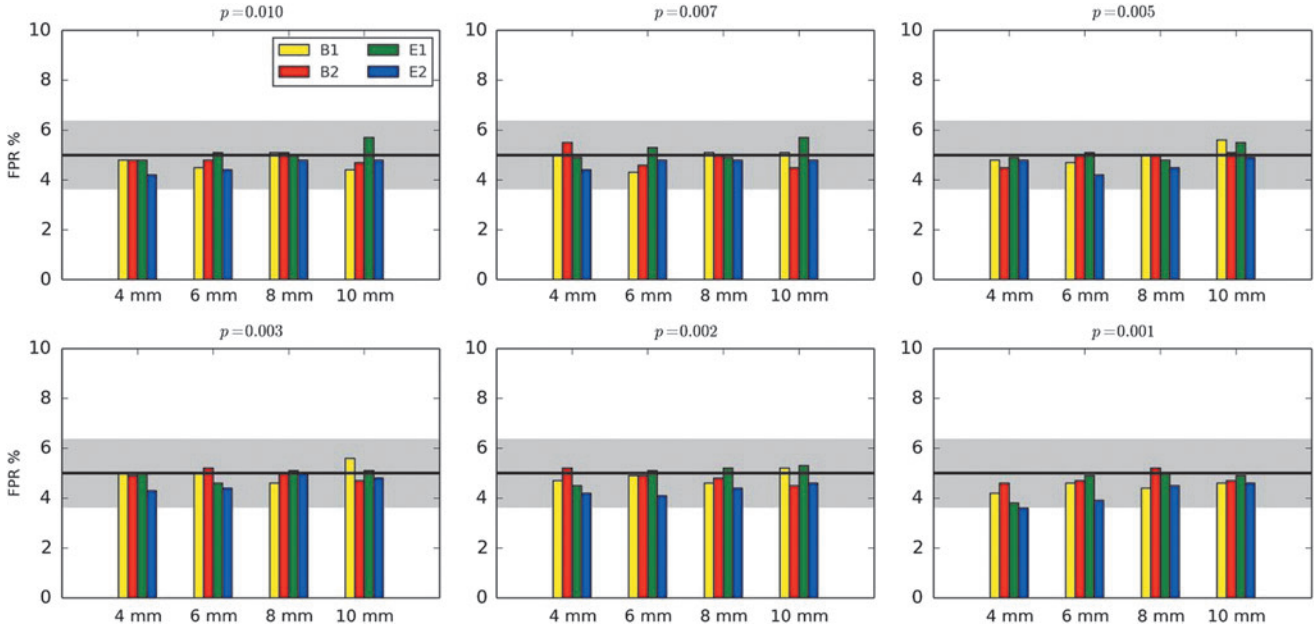


FIG. 5. FPRs with cluster-size thresholds now determined from the ''-Clustsim'' option of 3dttest++ (one-sided tests with NN = 1 clustering). See Figure 1 for description of labels, but note that the *y*-axis range has been significantly changed here for visual clarity.

The future, I: A third problem with cluster-threshold detection tools—inhomogeneous smoothness

A third standard assumption (present in AFNI, as well as the random field theory used in SPM and FSL; Worsley et al., 1996) also makes the idea of using a global cluster-size (or other cluster figure of merit) threshold somewhat nonoptimal. The spatial smoothness of the FMRI noise is not spatially stationary—it is significantly smoother in some brain regions (e.g., the precuneus and other large areas involved in the standard default mode network and also strongly affected by respiration artifacts) than in others; this inhomogeneity is also noted in Eklund and associates (2015, 2016). The presence of variable smoothness means that the density of false positives for a fixed cluster-size threshold will differ across the brain, especially since the FPR is strongly nonlinear in the cluster-size threshold and in the noise smoothness. Using the same cluster-size threshold everywhere in such brain data will result in higher FPRs than expected in the smoother areas and lower FPRs than expected in the less-smooth areas.

A new AFNI program, 3dLocalACF, has been written to estimate the (a,b,c) parameters from Eq. (3) locally (in a ball, constrained within a brain mask) around each brain voxel. The non-Gaussian smoothness can be partly characterized by a new parameter called the ''Full-Width at Quarter-Maximum'' (FWQM), which characterizes the scale of the model in Eq. (3) at a broader point than the FWHM used in the simple Gaussian case; in the limiting case that the ACF is Gaussian, then $\mathrm{FWQM} = 2^{1/2} \times \mathrm{FHWM}$. An example of the FWHM and FWQM smoothness estimates for one subject is shown in Figure 6. We can only speculate as to the precise causes of this nonuniformity in spatial smoothness. It is possible that high-resolution FMRI methods can reduce the size of this problem (Wald and Polimeni).

A better approach to cluster-level detection must take into account this inhomogeneity. One approach would be to adaptively blur the data to make the spatial smoothness more homogeneous. An alternative approach is to adapt the cluster thresholding technique to deal with the spatial smoothness as it appears. It is this latter approach that we have chosen

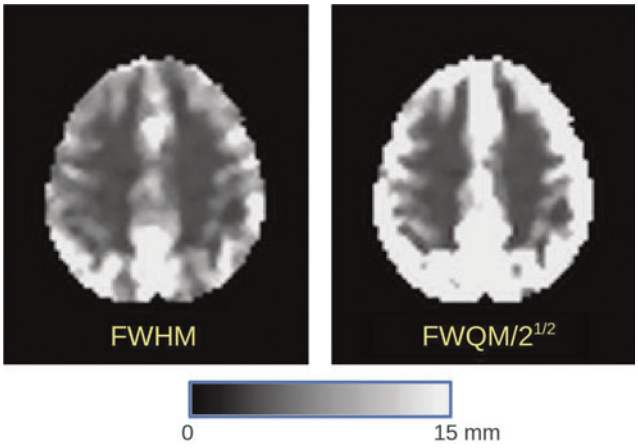


FIG. 6. Images of the FMRI noise FWHM and the FWQM from one subject (#11344) in the Beijing data set collection (after nominal smoothing with a Gaussian kernel of 4 mm FWHM during preprocessing). The scale in both images is linear from black = 0 to white = 15 mm (and above). If the ACF were Gaussian, $\mathrm{FWQM} = 2^{1/2} \times \mathrm{FHWM}$. The FWHM map shows that the noise smoothness is not uniform in space (even within gray matter), and the FWQM map shows that the non-Gaussianity of the noise smoothness is also nonuniform. The magnitude of this effect on the FPR and how to allow for it in thresholding are still under investigation. FWQM, full-width at quarter-maximum.

to develop first, through the randomization/permutation approach, as described in the next section. Another method, as yet unexplored, would be to generate synthetically nonstationary noise samples using the results from 3dLocalACF, and use those to compute spatially variable cluster-size threshold maps.

The future, II: Equitable thresholding and clustering

Thresholding test statistics in realistic cases requires judgment, not just mathematics. Even in a simple case of testing a one-parameter alternative hypothesis (e.g., "mean = 0" vs. "mean ≠ 0"), one must judge between a one-sided test ("mean > 0") or two sided ("mean < 0 or mean > 0"). In the latter case, one also has to judge whether it is necessary to give equal weight to both sides; that is, one could threshold so that under the null hypothesis of mean = 0, there is a 4% chance of a false positive (stating "mean > 0") and a 1% chance of a false negative (stating "mean < 0"). A principle of equity could reasonably force one to admit equal FPRs of either sign once the decision for a two-sided test has been made—but this principle is not required by the mathematics or statistics of the situation (and could be contravened, given unbalanced external costs between the two types of outcomes).

Here, we define equity to mean that one treats equally situations that do not have important *a priori* relative differences. In this way, the number of arbitrary choices is reduced. For example, why choose between $p < 0.01$ or $p < 0.001$ (or some other semiarbitrary p value) as a voxelwise threshold? Such a fixed threshold approach may fail to detect a region that is anatomically small with high statistic values or anatomically large with low statistic values. It would be more desirable to set the voxelwise threshold within a range so that a "holistic" or equitable FPR can be achieved across various cluster sizes. Why require large cluster-size thresholds in brain regions that are not very smooth? It would be more equitable to use smaller cluster-size thresholds in less smooth brain regions and use larger cluster-size thresholds in more smooth regions.

A method for such balanced or equitable thresholding is under development in AFNI. It is an extension of the randomization of the residual approach described earlier, in the "-Clustsim" option of program 3dttest++. The new method produces a set of maps of the cluster-size threshold to use, one for each member of a range of voxelwise p value thresholds. A voxel is "accepted" if it passes any one (or more) of the individual voxelwise p plus cluster-size threshold tests. The cluster-threshold levels are chosen to be balanced, so that each cluster-threshold map (for one given p) contributes individually at the same FPR at each voxel, say α. In this way, no particular p value and no particular location is specially favored: small high-intensity clusters are balanced with large low-intensity clusters; low-smoothness regions will get smaller cluster-size thresholds than high-smoothness regions. The global FPR is chosen by adjusting the individual mapwise α to get the desired final 5% rate. These calculations are implemented, using sign-randomization/permutation-simulated t-test volumes

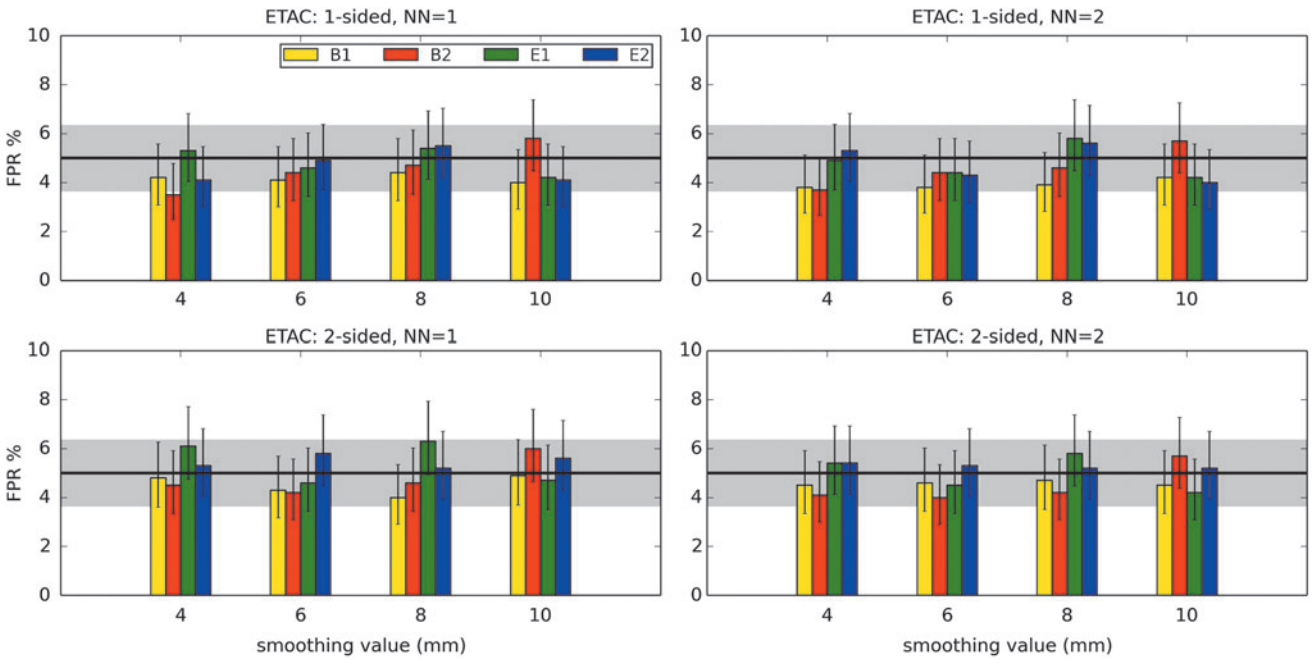


FIG. 7. FPRs from the ETAC method, with the Beijing subset of FCON-1000. See Figure 1 for description of labels, but note that the y-axis range has been changed here for visual clarity. Three p value thresholds (0.005, 0.002, 0.001) are used simultaneously, and p-specific spatially variable cluster-size threshold maps are created from sign-randomized (and intersample permuted for the two-sample cases) simulations. For each of the 16 cases, 1000 random subsets of 40 subjects were selected, and a two-sample t-test was run between the first 20 and second 20 data sets for each of the 1000 instances. As labeled in each panel caption, results were calculated using either NN = 1 or NN = 2 neighborhoods (see Results section: The future, II: Equitable thresholding and clustering) to define the clusters, and either one-sided or two-sided t-testing to define the p value thresholding. All FPRs fall within the 95% nominal confidence interval; error bars show the 95% confidence interval estimated for each result. FPRs from the Cambridge subset of the FCON-1000 (also 198 subjects) yielded similar results. ETAC, equitable thresholding and clustering.

generated via 3dttest++, in the new program 3dXClustSim, which generalizes the older 3dClustSim to derive voxel-specific cluster-size thresholds, as well as allowing the simultaneous use of several voxelwise p-thresholds.

Preliminary results from the Beijing subset (198 subjects) of the FCON-1000 collection are shown in Figure 7. In this example, three p value thresholds (0.005, 0.002, 0.001) are used simultaneously, and p-specific spatially variable cluster-size threshold maps are created from sign-randomized (and intersample permuted for the two-sample cases) simulations. Note that the calculated FPRs for each set of parameters (NN, sidedness, blurring, and stimulus) lie within the 95% confidence band of the nominal 5% value. Further equitable thresholding and clustering (ETAC) analyses with 3dXClustSim are needed before it will be released for general use; some comparisons of intermediate, single p value ETAC results with other methods are shown in Appendix 2. Several generalizations are planned, and testing with task-based FMRI data collections will be carried out to compare the statistical power versus alternative methods.

Discussion and Conclusions

A note on "The Bug" and on bugs in general

Of the many points considered here, we first commented on one of the least important and most publicly highlighted ones: the bug in the older versions of 3dClustSim (and its precursor AlphaSim). As shown here and noted before, this is actually a minor feature and the effect of the bug on FPR is relatively small. Correcting the underlying problem did indeed reduce the FPRs in these tests, but the change in results cannot be considered a major factor in the overall FPR inflation. Both before and after the bug fix, 3dClustSim performed comparably to the other software tools being investigated. This is not to say that the presence of the bug was not unfortunate, but by itself it could not ''have a large impact on the interpretation'' (Eklund et al., 2016) of FMRI results. To make significant FPR changes in the results of 3dClustSim, new methods were required, which were also presented herein and which are discussed further below.

Reproducibility has been a major topic in the field of FMRI, with several proposals of ''best practices'' put forth in various forms. It is obvious that the presence of bugs in software (as well as misuses of software settings inappropriate in the context and incorrect method implementations) damages the validity and reproducibility of reported results, and there is no greater concern for those writing software—particularly when it is intended for public use—than preventing bugs. Much of the discussion surrounding ENK16, particularly in comments to and takeaways chosen by the popular press, focused on the bug that was present in 3dClustSim. The discovery of this bug was highlighted and hyped as a major component for rejecting 15 years of brain studies and (up to) 40,000 peer-reviewed publications on the brain, under the tacit or explicit assumption that the reported results would be unreproducible.

However, rather than being evidence for ''a crisis of reproducibility'' within the field of FMRI, the advertisement of the bug is itself an important verification of the reproducibility of FMRI analysis. In this imperfect world, the philosophy for maintaining AFNI has always been to correct any bugs and to update the publicly available software as soon as practicable, often posting on the public Message Board for significant changes. The AFNI group maintains a permanent and public list of updates/changes/bugs online, which we view as an important resource for users and an aid for supporting reproducibility.

While certainly an annoying moment for the researchers who used 3dClustSim (and for those who maintain the software), the knowledge and dissemination of this bug is part of the reproducibility process. The existence of software bugs is unfortunate but inevitable. Even huge distributions such as Python, Windows, Mac, and Linux release bug fixes regularly. Clarity of description and speed of repair are the best tools for combating their effects once discovered.

The state of clustering

There were many valuable points raised in the work of ENK16. Several of these were important for general consideration within the FMRI field, such as the assumption of most clustering approaches that spatial smoothness was well-enough approximated by a Gaussian shape. To address this point, we have shown how an updated approach within AFNI using an estimated non-Gaussian ACF greatly improves the FPR controllability within the test data sets. In addition, there is also a new nonparametric method for clustering within AFNI that shows promise; however, this type of approach in general currently appears to be limited by practical considerations (that hold across software implementations) to relatively basic group analyses that can be performed through univariate GLM.

A permutation and/or randomization approach seems able to provide proper FPR control with few apparent assumptions; so why not use this approach for everything in FMRI group analysis?

Permutation tests of complex models (e.g., complex AN(C)OVA or LME) can become extremely computationally expensive, especially when coded in an interpreted language such as MATLAB, Python, or R, where not all statistics are easily recomputable hundreds or thousands of times (Chen et al., 2013). Nor is permutation/randomization easily implemented in complex situations with nesting and/or covariates. Issues of smoothness inhomogeneity of the noise structure in brain images present significant challenges for the development of a parametric method for spatial thresholding—but such a method would be very useful. Considering the overall complexity of the problem, it appears unlikely that a ''gold standard'' *prima facie* correct method for spatial thresholding of neuroimaging data will appear soon.

Also, a permutation test is neither always required nor uniformly advantageous, as it may sacrifice power unnecessarily in some cases. For example, a fixed number of permutations would set a lower bound for the p value that could be achieved (or require distributional extrapolation; Scholz), leading to failure to detect a small cluster with a potentially very high significance level that would survive through a parametric approach or a much larger number of permutations. While permutation testing may be useful and even necessary in some situations, a general rule for determining those cases is not clear, and, as noted above, it may be computationally or methodologically prohibitive to use (e.g., in the common case of including covariates, missing data, mixed effects). Further work is required on this important issue for the field.

Final (for now) thoughts on statistics in FMRI—and some recommendations

Certainly, when using a clusterwise FPR value, one would hope that a method would reliably reflect the nominal rates. However, in conjunction with other trending discussions in the statistics literature, p value thresholds are not sacred boundaries so that results around them live or die by tiny fractions above or below them; instead, thresholds are a convenience for focusing on reporting, but they are only part of the story. Our point here ties into discussions of reducing ''p-hacking'' and emphasizing effect sizes in results reporting for FMRI (Chen et al., 2016). One beneficial ''side effect'' of the equitable clustering method, as shown in Figure 7, is that results depend less sensitively on user-optional parameters such as blurring radius, NN value, and p value(s) used for decision-making.

Statistical testing and reporting are far from the end of a neuroscientific FMRI article; in fact, these steps are just the technical prelude to the neuroscientific interpretation. At present, the conclusions of a study depend strongly on previous work and knowledge, rather than relying solely on statistical arguments from the current data alone. It is very hard to decide without close examination if a weakly ''active'' (or ''connected'') cluster in a brain map is actually critical to forming the article's conclusions.

Despite being more than 25 years since the first BOLD FMRI experiments were carried out, the mapping of human brain activity and connectivity is still evolving in both methodology and interpretation. Many nontrivial mathematical models have been designed to describe the diverse phenomena observed in neuroimaging research, yet most remain without ''gold standard'' verification. While improvements to methodology can help the situation (e.g., by increasing the reliability and specificity of characterization), authors, editors, and reviewers should continue to use statistical thresholding as a source of information, but not as the final authority.

One option for somewhat reducing the cluster-size threshold in group analyses is to use accurate nonlinear alignment to a template and then perform clustering within a slightly inflated gray matter mask instead of a whole brain mask. In the analyses shown herein, a whole brain mask comprising 74,397 $3\times3\times3\,mm^3$ voxels was used. In a few simulations, we ran parallel analyses within a gray matter mask of 43,824 voxels to get a feel for the subsequent change in cluster-size thresholds; on average, their sizes were reduced by about 25% (more at $p=0.010$, less at $p=0.001$). Since most FMRI-detectable activation or long-distance correlation occurs in gray matter, this approach has the potential for increasing statistical power without inflating FPR. It does depend strongly on the accuracy of the alignment to the template: affine or low-order nonlinear alignment is *not* adequate for this purpose. Use of such a mask needs to be carefully vetted in any particular application to ensure that all brain regions of potential interest are included, and that the subjects' alignments are all carried out to the needed precision.

At the time of writing, our recommendations for AFNI users are the following:

- When carrying out a group analysis via one- or two-sample testing, or GLM with between-subject factors/covariates, one can have cluster-size thresholds determined by nonparametric analysis (using the program 3dttest++ with the ''-Clustsim'' option); this has been tested extensively and gives very well-behaved FPR, as shown herein.
- Group analysis can also be carried out via Monte-Carlo simulations (using 3dClustSim) with the new mixed model ACF option (''-acf'') to account for the noise smoothness structure; for instance, in the case of more complex models [e.g., linear mixed effects (Chen et al., 2013), 3dLME], this likely must be the method used, at present, due to modeling limitations with GLM. The cluster parameters should be derived from the mean of the individual subject mixed model ACF parameters (which are computed by the standard afni_proc.py pipeline), and a voxelwise $p=0.001$ or $p=0.002$ is reasonably safe with this method, as in Appendix 2 (and compare, e.g., our Appendix Fig. 2B-9 with ENK16—Supplementary fig. S2 SPM results).

In the near future, we hope to release the ETAC (3dXClustSim) method for general use with 3dttest++; however, further testing, development, and documentation are required before ETAC is fully ready.

Acknowledgments

The authors thank Alfonso Nieto-Castanon for informative discussions and figure suggestions. The research and writing of the article were supported by the NIMH and NINDS Intramural Research Programs (ZICMH002888) of the NIH/DHHS, USA. This work extensively utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

Author Disclosure Statement

No competing financial interests exist.

References

BEC Crew. 2016. A bug in FMRI software could invalidate 15 years of brain research. www.sciencealert.com/a-bug-in-fmri-software-could-invalidate-decades-of-brain-research-scientists-discover Last accessed December 3, 2016.

Biswal B, et al. 2010. Toward discovery science of human brain function. Proc Natl Acad Sci USA 107:4734–4739.

Chen GC, Saad ZS, Britton JC, Pine DS, Cox RW. 2013. Linear mixed-effects modeling approach to FMRI group analysis. NeuroImage 73:176–190.

Chen GC, Taylor PA, Cox RW. 2016. Is the statistic value all we should care about in neuroimaging? NeuroImage [Epub ahead of print]; DOI:10.1016/j.neuroimage.2016.09.066

Cox RW, Reynolds RC. 2016. Improved statistical testing for FMRI based group studies in AFNI:). OHBM Geneva. https://afni.nimh.nih.gov/pub/dist/HBM2016/Cox_Poster_HBM2016.pdf

Eklund A, Nichols T, Knutsson H. 2015. Can parametric statistical methods be trusted for fMRI based group studies? https://arxiv.org/abs/1511.01863 Last accessed December 3, 2016.

Eklund A, Nichols T, Knutsson H. 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proc Natl Acad Sci USA 113:7900–7905.

Flandin G, Friston KJ. 2016. Analysis of family-wise error rates in statistical parametric mapping using random field theory. https://arxiv.org/abs/1606.08199 Last accessed December 3, 2016.

Lieberman MD, Cunningham WA. 2009. Type I and type II error concerns in fMRI research: Re-balancing the scale. Soc Cogn Affect Neurosci 4:423–428.

Nichols T. 2016. Errata for cluster failure. http://blogs.warwick.ac.uk/nichols/entry/errata_for_cluster Last accessed December 3, 2016.

Russon M-A. 2016. 15 years of brain research has been invalidated by a software bug, say Swedish scientists. www.ibtimes.co.uk/15-years-brain-research-has-been-invalidated-by-software-bug-say-swedish-scientists-1569651 Last accessed December 3, 2016.

Scholz FW. Nonparametric Tail Extrapolation. (White paper from Boeing Information & Support Services). www.stat.washington.edu/fritz/Reports/ISSTECH-95-014.pdf Last accessed December 3, 2016.

Stockton N. 2016. Don't be so quick to flush 15 years of brain scan studies. www.wired.com/2016/07/dont-quick-flush-15-years-brain-scan-studies Last accessed December 3, 2016.

Wald LW, Polimeni JR. Impacting the effect of fMRI noise through hardware and acquisition choices—Implications for controlling false positive rates. NeuroImage [Epub ahead of print]; DOI: 10.1016/j.neuroimage.2016.12.057

Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC. 1996. A unified statistical approach for determining significant signals in images of cerebral activation. Hum Brain Mapp 4:58–73.

Address correspondence to:
Robert W. Cox
Scientific and Statistical Computing Core
NIMH/NIH/DHHS
10 Center Drive
Bldg. 10, Rm 1D73
Bethesda, MD 20892

E-mail: robertcox@mail.nih.gov

Appendix

Appendix 1. Additional Parsing and Plotting of Eklund and Associates False-Positive Rate Results

Figure 1A-1 shows the original results of Eklund and associates (as in Fig. 2 in the main text here; made public in their GitHub repository) dissected into their results from one- and two-sample testing and by task stimulus (block designs B1 and B2, and event-related E1 and E2). Several useful comparisons of the effect of both the statistical test and stimulus paradigm are observable. For event-related stimuli, two-sample tests show uniformly lower FPR distributions (lower mean, lower max, and fewer outliers), across software and for either voxelwise p. In particular, for $p = 0.001$, two-sample testing, and event-related stimuli, all methods are clustered closely near the nominal 5% FPR. For the (shorter) B1 blocks, a reduction of FPR is also seen with two-sample testing, although with the (longer) B2 it is much less apparent. Again we note that, while there are differences in software results for the various parameters, the parametric methods tend to perform quite similarly for most subsets of parameters. The notable patterns of FPR results for a given p value, statistical test, and stimulus may be useful in guiding further studies.

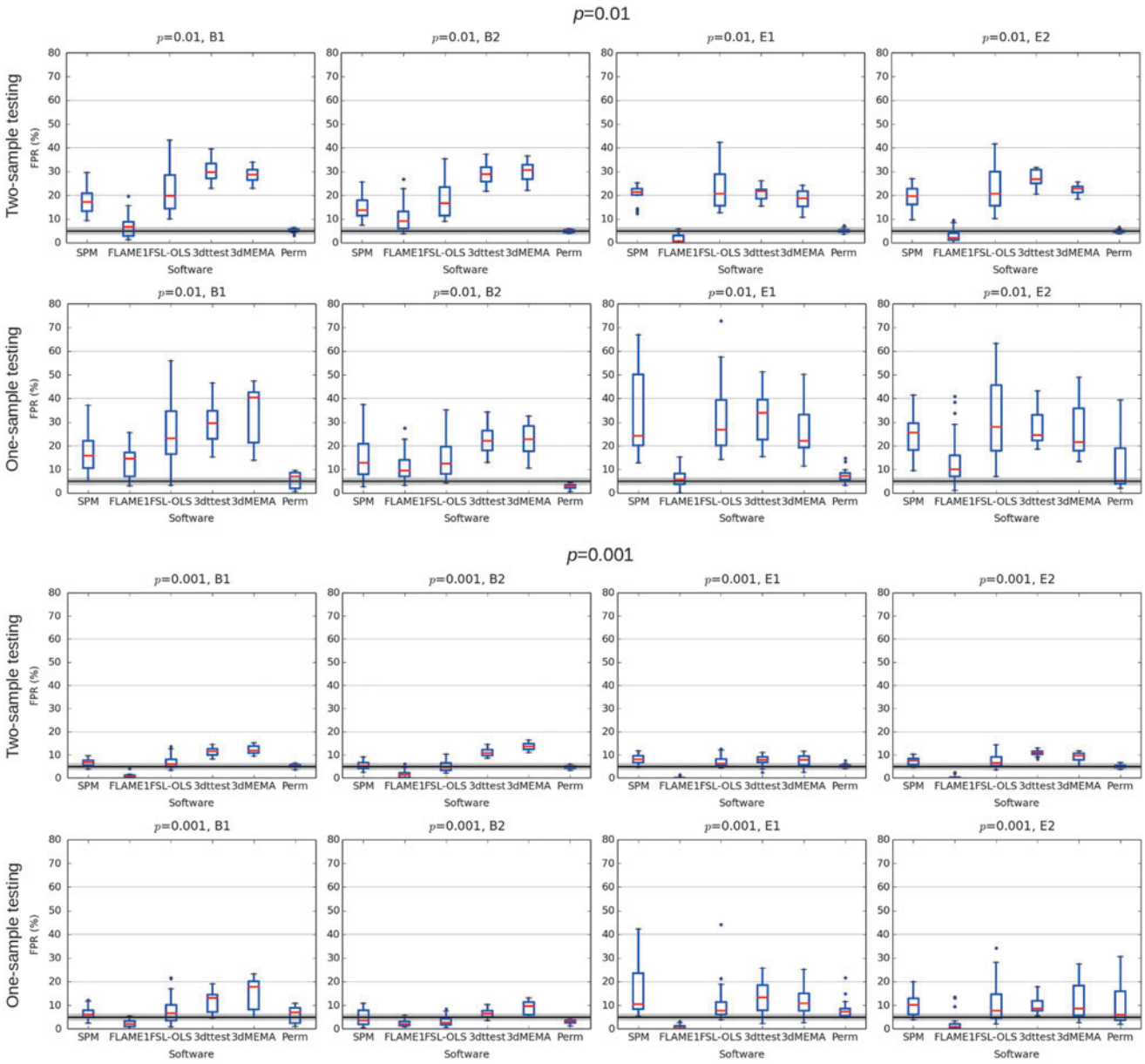
Appendix 2. Studies with 3dttest++ and Six Different Cluster-Size Thresholding Methods

Extensive null simulations were carried out on the 198 Beijing-Zang data sets from FCON-1000 collection. All the data (reorganized slightly from the original FCON-1000 download) and scripts used for the processing and analysis shown in this Appendix are available at the Globus shared Endpoint ''ClusteringDataBeijing,'' stored on the NIH HPC Data Transfer server. This collection can be downloaded by establishing a free Globus user account, installing the free Globus client software, then searching for this endpoint in the ''Transfer Files'' dialog. See www.globus.org for more information and to get started. All 198 subjects from the collection were processed successfully with the supplied scripts. The processing scripts are also available at https://afni.nimh.nih.gov/pub/dist/tgz/Scripts.Clustering.2017A.tgz (no data).

In brief, FPRs were estimated from 1000 3dttest++ runs, for:

- Four stimuli (B1, B2, E1, E2).
- Four blurs (4, 6, 8, 10 mm).
- Four voxelwise p thresholds (0.001, 0.002, 0.005, 0.010).
- One- and two-sided t-tests.
- One- and two-sample t-tests (20 subjects in each sample, pseudorandomly selected).
- Six updated cluster-size thresholding methods (each simulation used the same pseudorandom samples).
  - Equitable thresholding and clustering (ETAC[7]) (spatially variable cluster-size thresholds).
  - CS-RT-res (3dClustSim using the randomized/permuted t-tests from the t-test residuals).
  - CS-ACF-res (3dClustSim using the mixed model autocorrelation function (ACF) parameters estimated from the t-test residuals).
  - CS-FWHM-res (3dClustSim using the Gaussian model ACF with the full-width at half-maximum (FWHM) derived from the mixed model parameters estimated from the t-test residuals).
  - CS-ACF-sub (3dClustSim using the mean across the tested subjects of the mixed model ACF parameters estimated from the subject's time series residuals—this is the method used in Fig. 1, column 3).
  - CS-FWHM-sub (3dClustSim using the mean across the tested subjects of the FWHM estimated from the mixed model applied to the subject's time series residuals).

[7]The full ETAC method (as described in the main text and shown in Fig. 7) integrates results over a range of voxelwise p values. Each figure presented in this Appendix shows ETAC results for a single p value only, so that this intermediate step can be evaluated as well as compared with other existing approaches.



APPENDIX FIG. 1. **(A-1)** Summary of the FPR results examined in ENK16, combining all their test results (available from their GitHub repository). The results of each software are shown separately based on voxelwise p (=0.01 or 0.001), statistical test (one or two sample), and task stimulus (blocks B1 or B2, or event-related E1 or E2). Red lines show the median; the box covers the 25–75% interquartile range; whiskers extend to the most extreme data point within 1.5× the interquartile range; and outliers are shown as dots. While results are fairly similar across parametric approaches, there is notable variation in FPR distribution among cases.

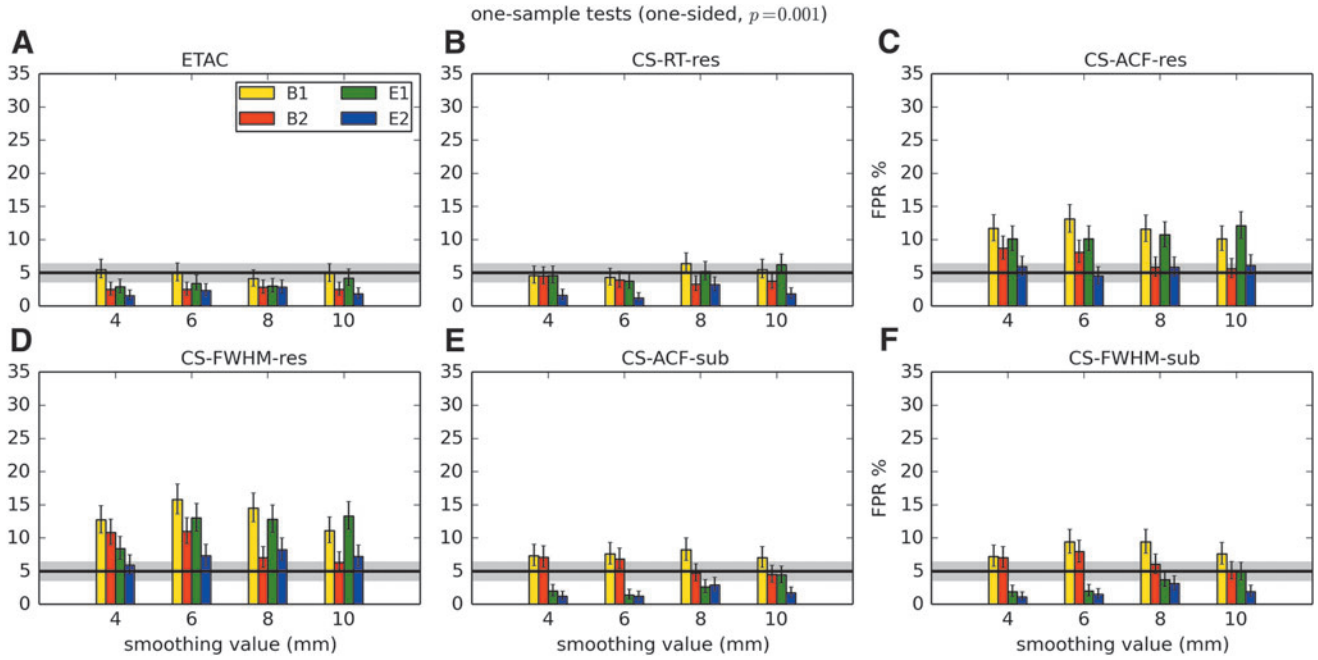
In total, 1536 ($=4\times4\times4\times2\times2\times6$) FPR estimates were generated by these simulations. These results are shown in Appendix Figures 2B-1 through B-16.

In general, the permutation/randomization approaches (ETAC, CT-RT-res) provide good FPR control at the desired 5% rate. Some cases are too conservative (e.g., $p=0.010$, one sample, one sided for the E2 stimulus), and some are too liberal (ETAC for $p=0.010$, one sample, two sided for various stimuli/blur combinations); these results reiterate the benefits of using lower voxelwise p values, as well as the preference for two-sample testing (particularly in this simulation setup, where the resting-state "null" is *known* to not be structure-free noise). However, none of these cases for these two methods is dramatically wrong, and most are consistently accurate for a range of experimental designs, blurring, p value, and so on.

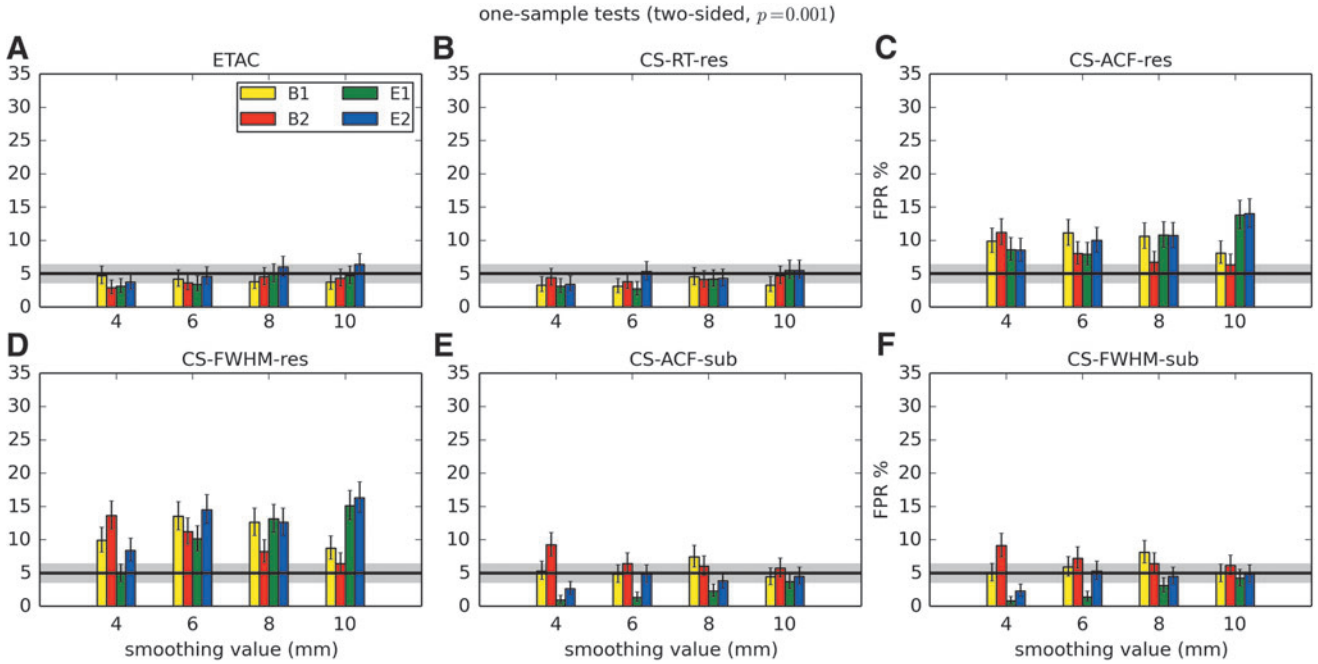
For the methods where a parametric (Gaussian or mixed model ACF) approach is used to calculate the cluster-size threshold, the results vary much more dramatically. It is clear that the Gaussian ACF is always more liberal than the mixed model ACF, as it should be since the two

approaches have the same FWHM by construction and the mixed model allows for longer range correlations. Somewhat surprisingly, it appears to be generally true that using the subject mean mixed model ACF parameters is more conservative than using the ACF parameters estimated directly from the *t*-test residuals. For this method (CS-ACF-sub), the FPR estimates are ''reasonable'' for the smaller *p* values (0.001 and 0.002), and for the larger smoothing levels (8 and 10 mm). These results form the basis for our recommendations at the end of the main text of this article.

APPENDIX FIG. 2.

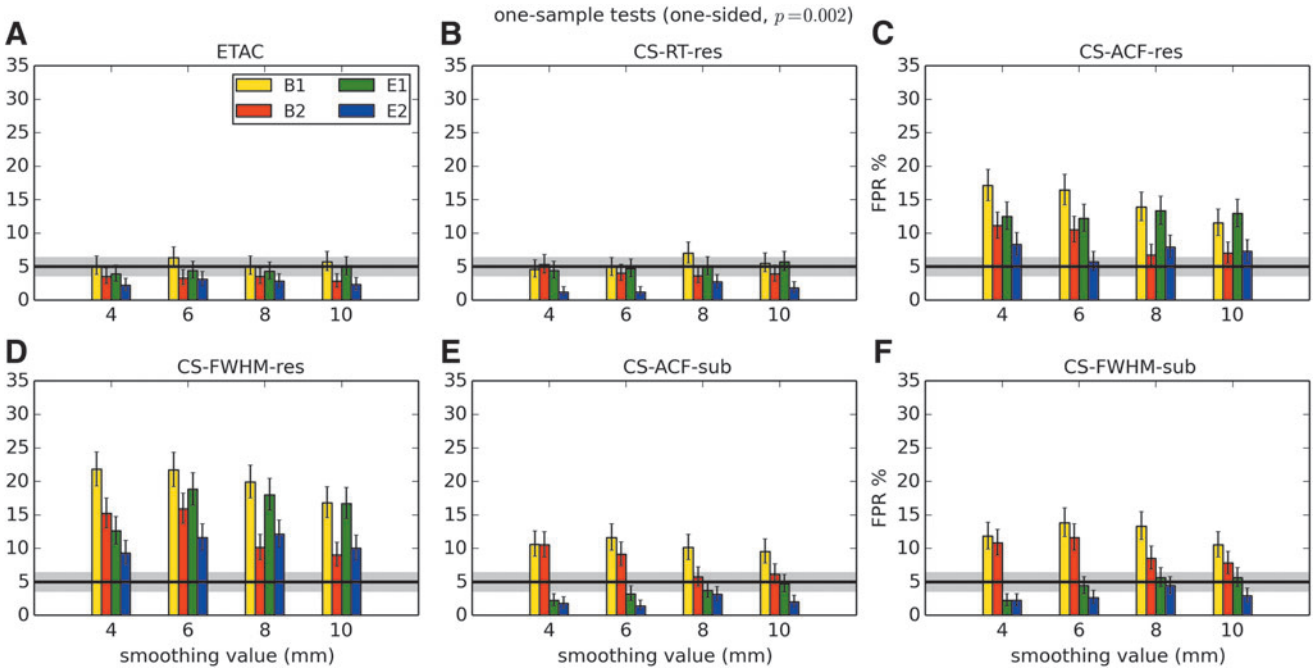


(B-1) One-sample, one-sided *t*-tests with voxelwise $p=0.001$.

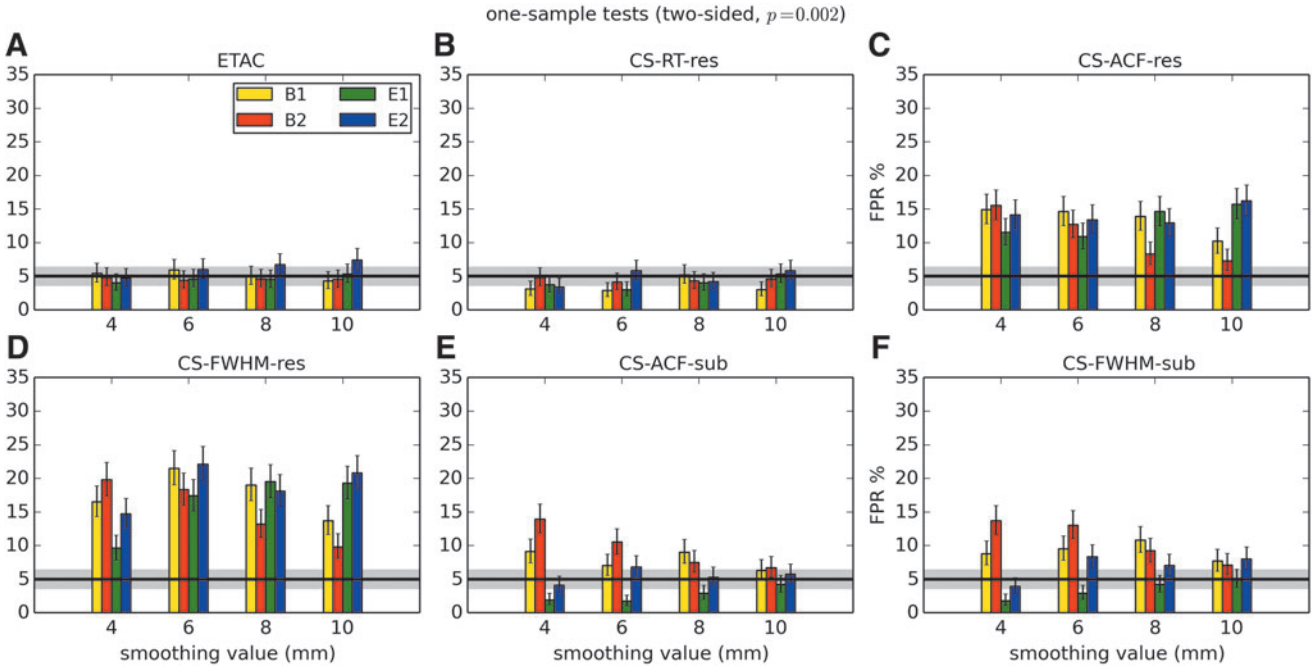


(B-2) One-sample, two-sided *t*-tests with voxelwise $p=0.001$.

(*Appendix Fig. 2 Continued*→)

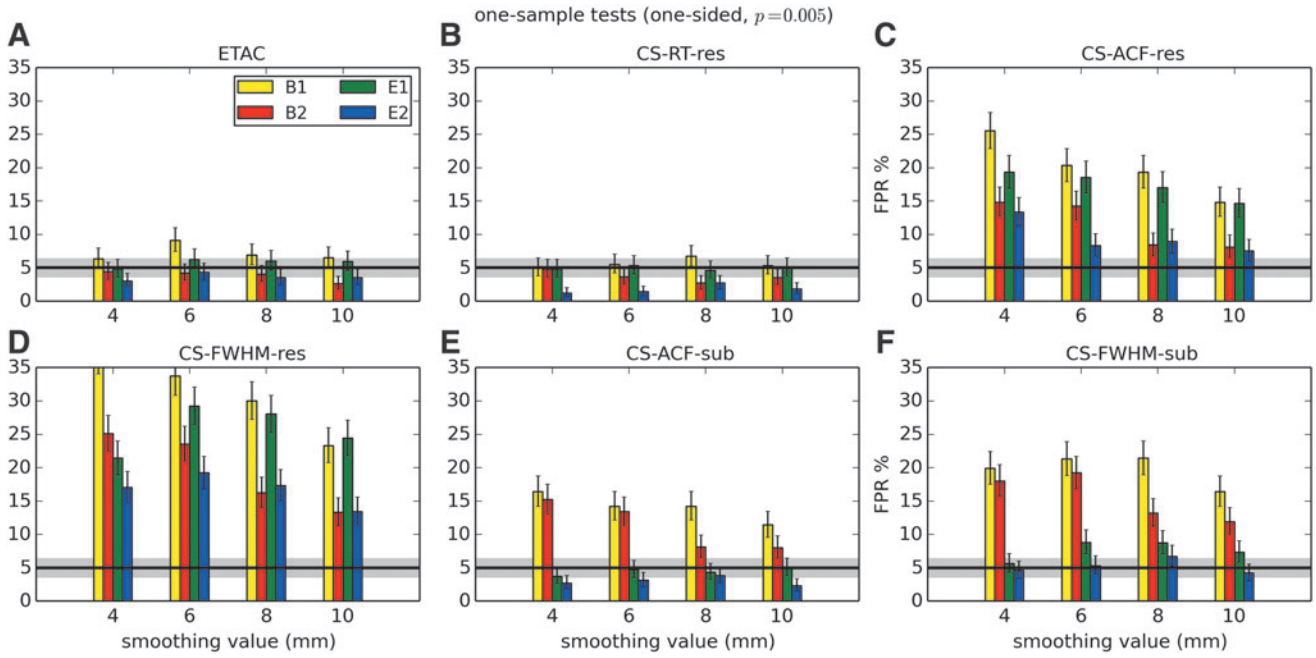


(B-3) One-sample, one-sided *t*-tests with voxelwise $p=0.002$.

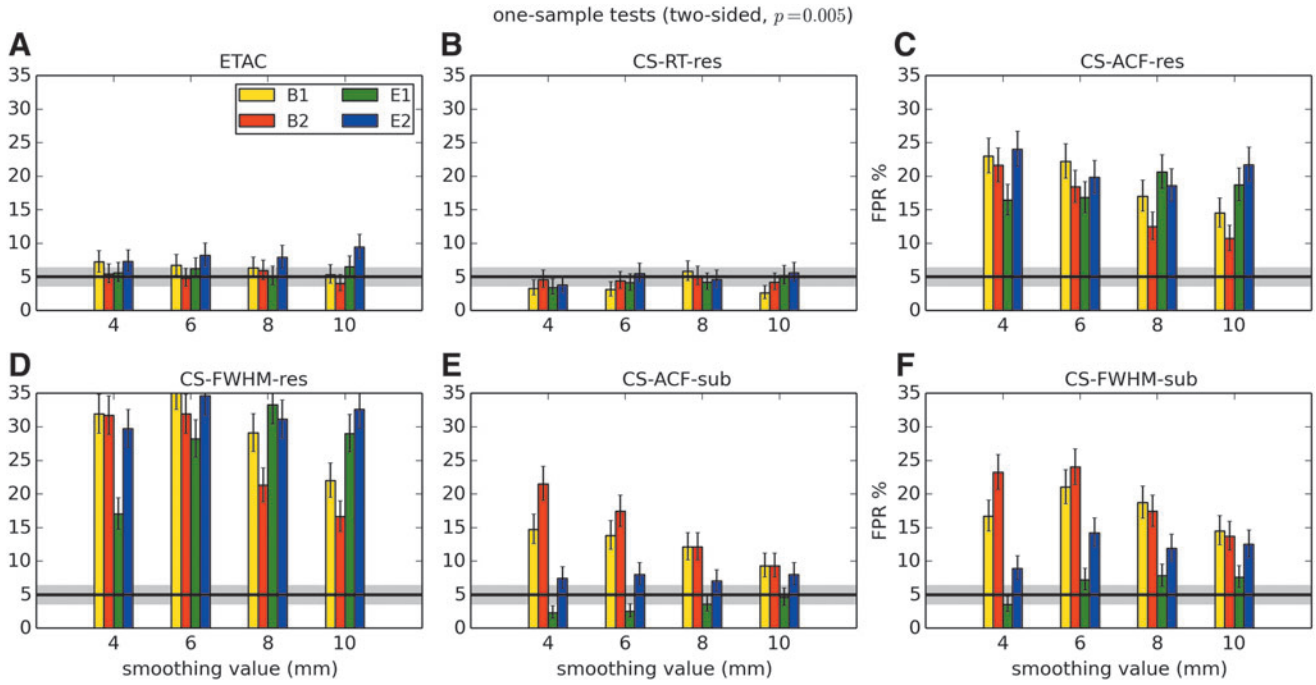


(B-4) One-sample, two-sided *t*-tests with voxelwise $p=0.002$.

(*Appendix Fig. 2 Continued* →)

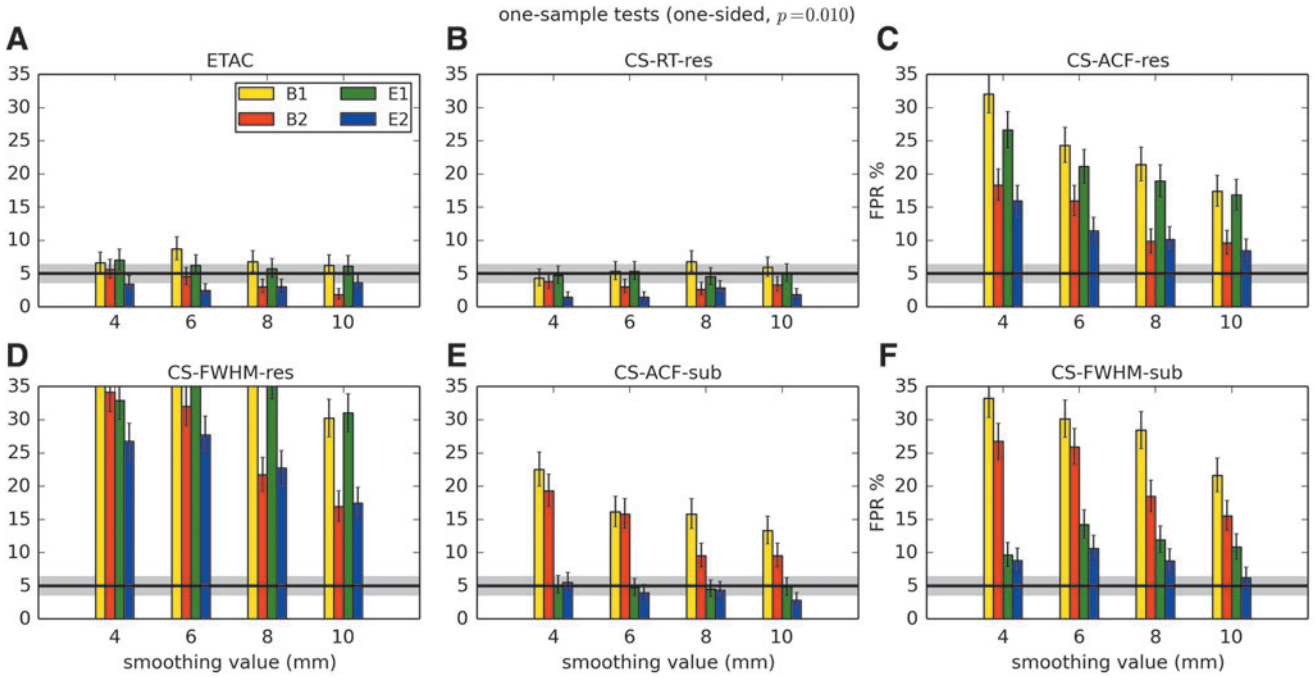


(B-5) One-sample, one-sided *t*-tests with voxelwise $p=0.005$.

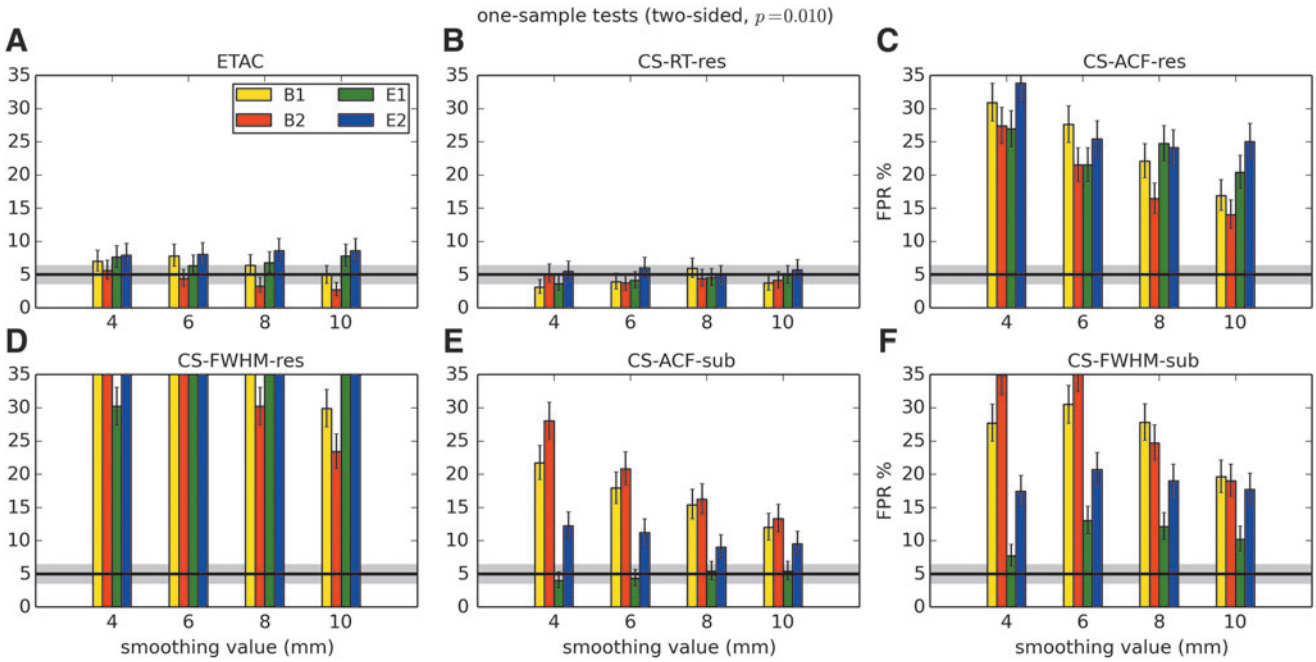


(B-6) One-sample, two-sided *t*-tests with voxelwise $p=0.005$.

(*Appendix Fig. 2 Continued*→)

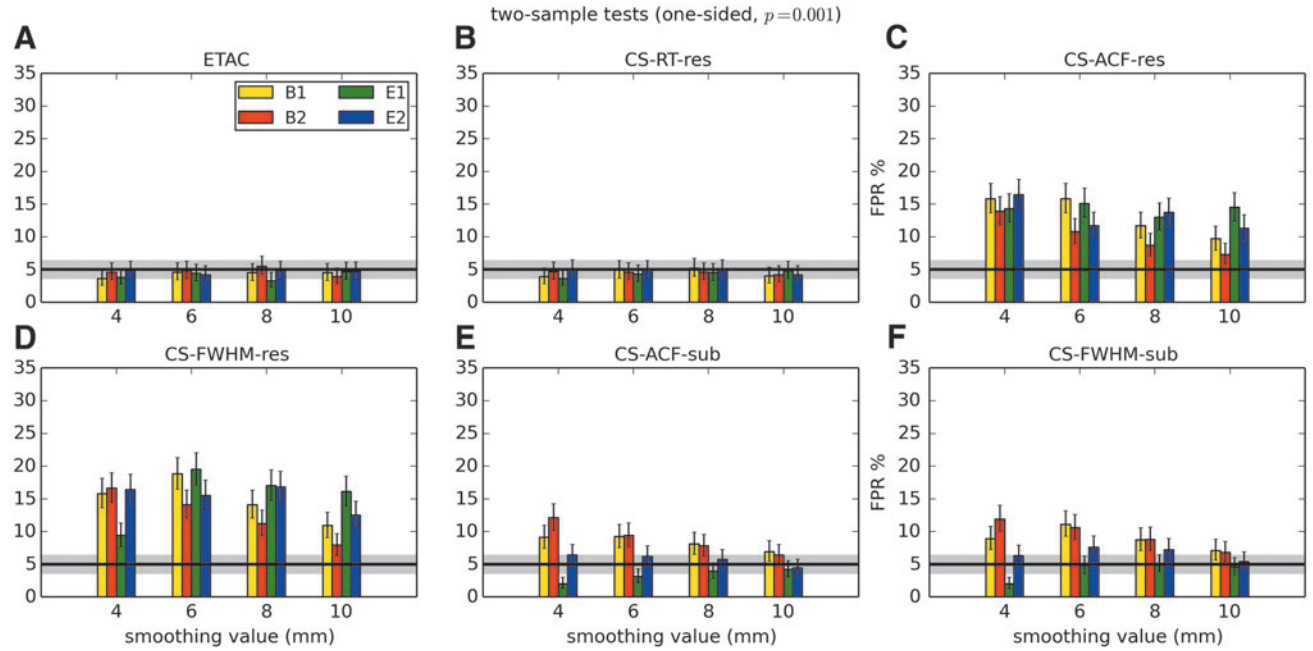


(B-7) One-sample, one-sided *t*-tests with voxelwise $p=0.010$.

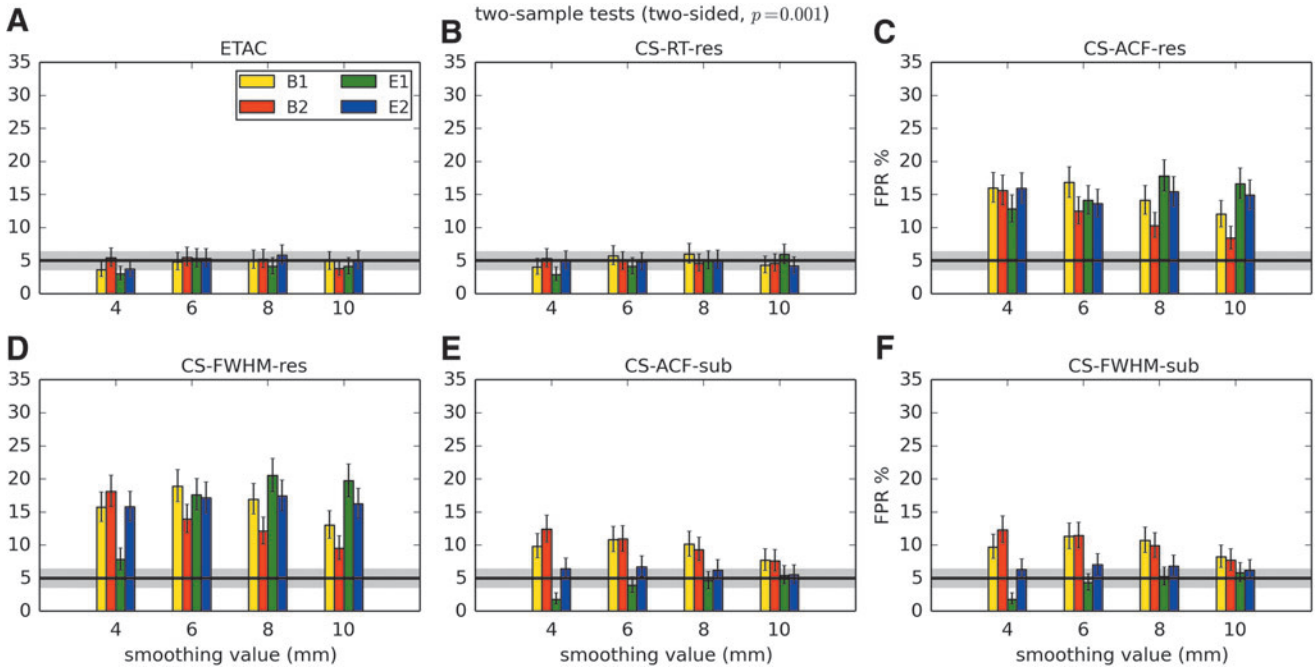


(B-8) One-sample, two-sided *t*-tests with voxelwise $p=0.010$.

(*Appendix Fig. 2 Continued* →)

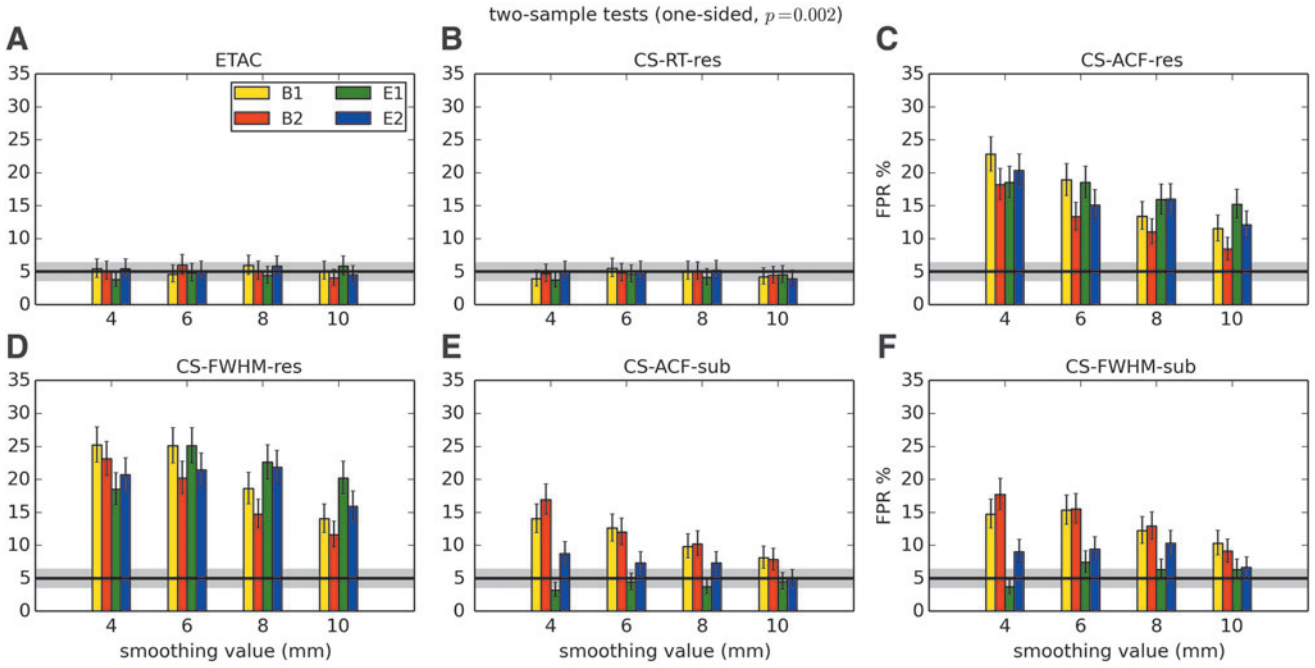


(B-9) Two-sample, one-sided *t*-tests with voxelwise $p=0.001$.

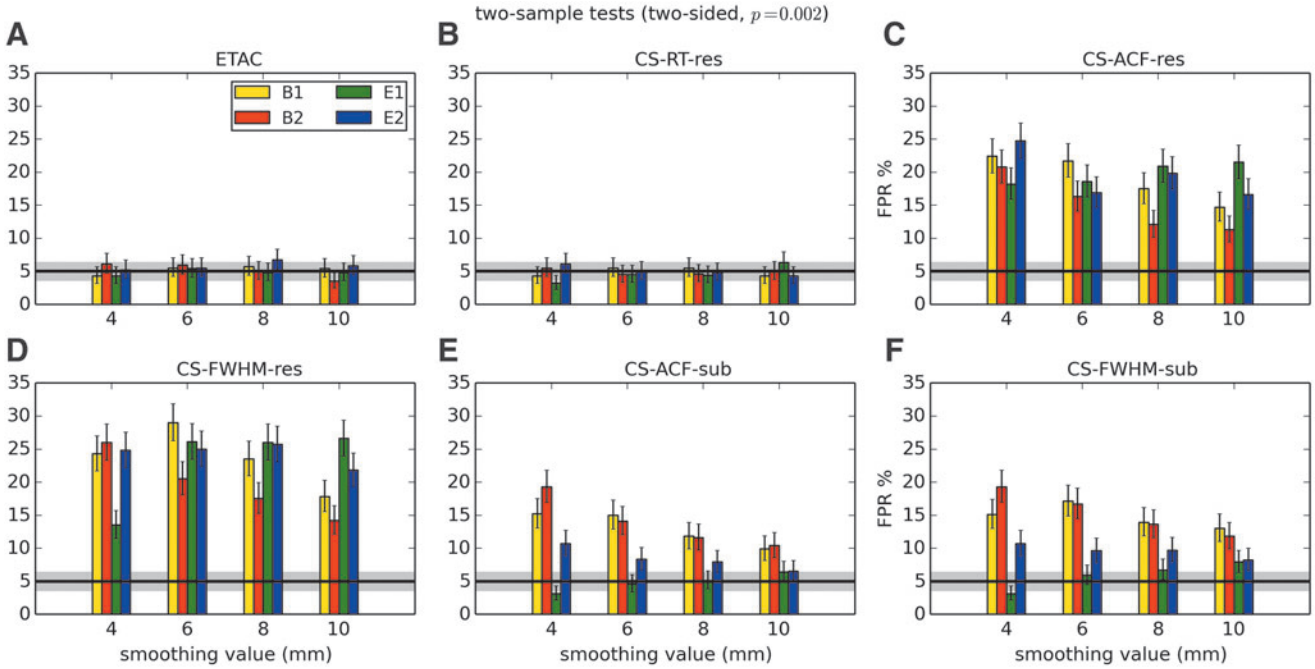


(B-10) Two-sample, two-sided *t*-tests with voxelwise $p=0.001$.

(*Appendix Fig. 2 Continued*→)

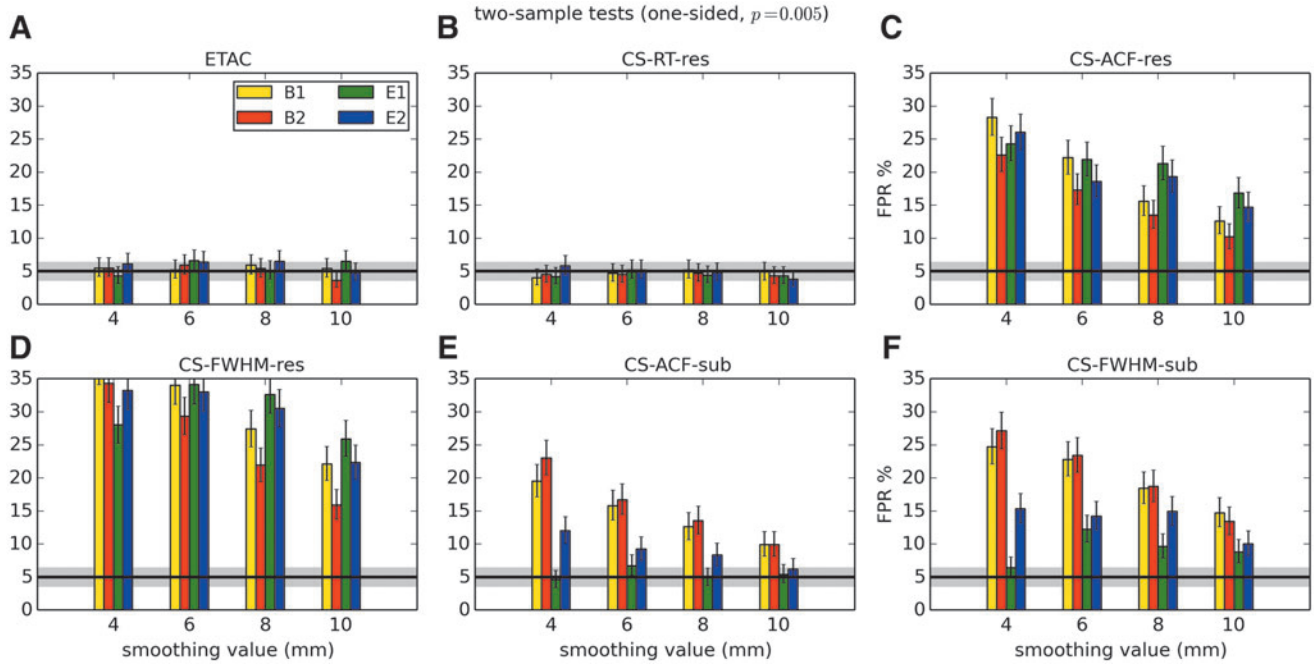


(B-11) Two-sample, one-sided *t*-tests with voxelwise $p=0.002$.

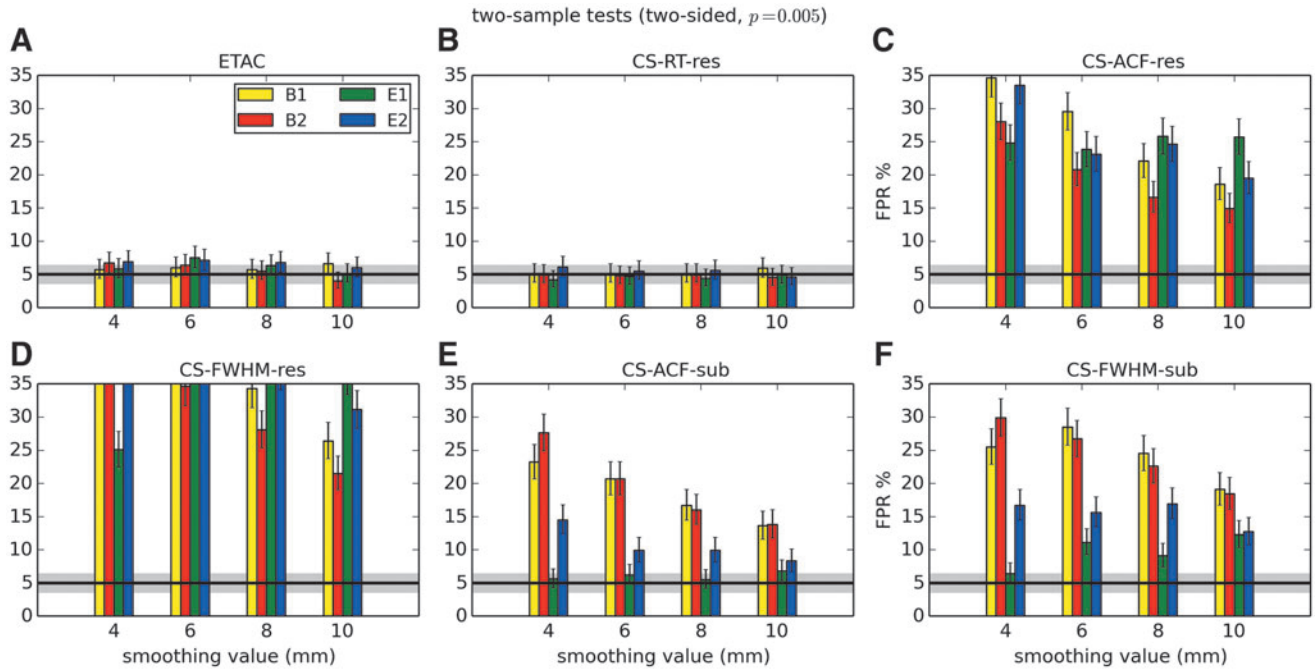


(B-12) Two-sample, two-sided *t*-tests with voxelwise $p=0.002$.

(*Appendix Fig. 2 Continued*→)

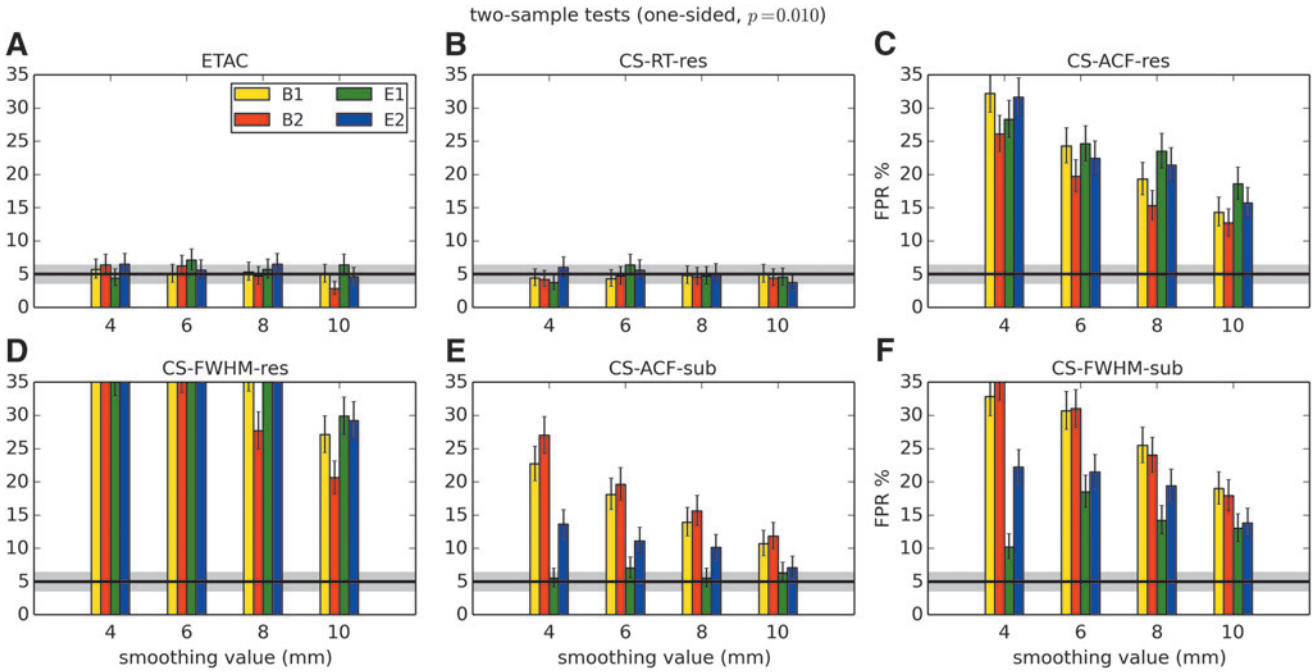


(B-13) Two-sample, one-sided *t*-tests with voxelwise $p=0.005$.

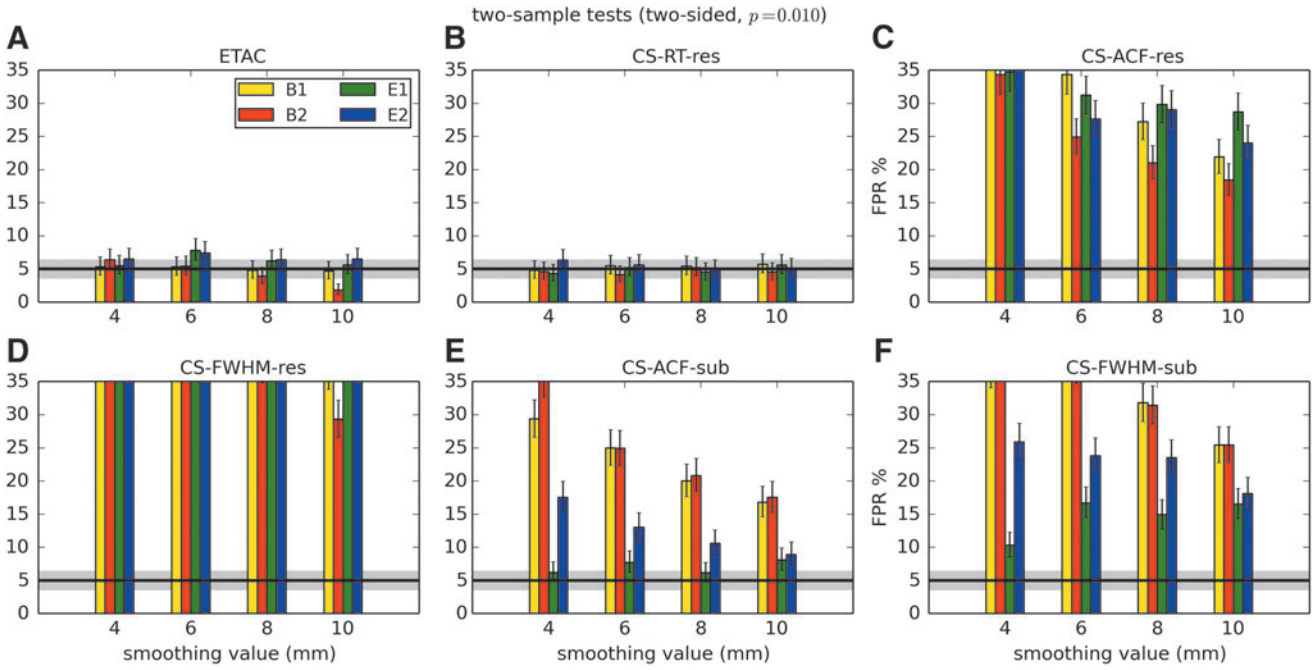


(B-14) Two-sample, two-sided *t*-tests with voxelwise $p=0.005$.

(*Appendix Fig. 2 Continued*→)



(B-15) Two-sample, one-sided t-tests with voxelwise $p=0.010$.



(B-16) Two-sample, two-sided t-tests with voxelwise $p=0.010$.