Equitable Thresholding and Clustering

Robert W. Cox
Scientific and Statistical Computing Core, NIMH/NIH/DHHS, Bethesda, MD, USA

Abstract

We describe a novel hybrid method to threshold FMRI group statistical maps derived from voxelwise second-level statistical analyses. The proposed Equitable Thresholding and Clustering (ETAC) approach is grounded in two ideas: (i) reducing the dependence of clustering results on arbitrary parameter values by using multiple sub-tests—each equivalent to a “standard” FMRI clustering analysis—to make decisions about which groups of voxels are potentially “significant”, then combining the results of each sub-test to decide which voxels are "accepted" ; and (ii) adjusting the cluster-thresholding parameters of each sub-test from (i) in an equitable way—so that the individual false positive rates (FPRs) are balanced across sub-tests and voxels—to achieve a desired global FPR (e.g., 5%). ETAC is independent of parametric assumptions about the spatial correlation of FMRI noise, because resampling methods are used to simulate the null (noise-only) distribution required to compute the FPR evaluations required in (ii). Resting FMRI datasets, analyzed with pseudo-task timings to provide a null model, were used to show the accuracy of the ETAC FPR control. A task FMRI data collection was used to compare ETAC’s true positive detection power vs. a standard cluster detection method, with ETAC providing equivalent or favorable results. Additionally, an important general note on the use of one-sample t-tests in neuroimaging is also made, based on an examination of the variability of cluster results.

Introduction

One of the most common (if not *the* most common) methods used for group level FMRI statistical map decision making is *dual thresholding* of statistical parametric maps. In most software, this is (schematically) performed in two successive steps:

A. At each voxel, reject any voxel whose test statistic likelihood (*p*-value) is larger than some user-selected *p*-threshold (1- or 2-sided tests can be used; 1-sided are more common in FMRI practice);
B. Among the surviving voxels, accept only those that form neighborhoods with other surviving voxels in a cluster of some threshold size or larger.

The per-voxel statistic thresholding (step A) is easy to apply, since the voxelwise *t*-statistic with known degrees of freedom (for example) can be directly converted to a *p*-value. For the cluster thresholding (step B), determining a cluster-size threshold that results in a desired global false positive rate (FPR) for a given voxelwise *p*-threshold is nontrivial and does not have an exact closed-form solution. There are two main categories of standard approaches to this latter problem. The method used in SPM (Worsley & Friston, 1994;

Flandin & Friston, 2017) uses a formula based on a Gaussian-shaped spatial autocorrelation function (ACF) for the FMRI noise and is asymptotically accurate for large smoothness when the Gaussian-shaped ACF model is accurate. The method commonly used in AFNI (Cox, 1996) to date is a noise-only cluster-simulation technique, based on a Gaussian or (more recently) a longer-tailed model for the ACF of the simulated pseudo-random FMRI noise volumes (Cox et al., 2017a; Cox et al., 2017b). The cluster-simulation approach avoids potential inaccuracy from the approximate cluster-size threshold formula, at the cost of brute force computation.

Controversy flared in 2016 with the publication of a paper (Eklund et al., 2016) that put forth strong claims about the failure in controlling the FPR of the most common software tools used in FMRI group analysis. Although the authors withdrew or tempered some of their most tendentious claims (e.g., about the number of FMRI papers brought under suspicion), the main thrust of their work remained largely unrefuted: many parametric methods for cluster-thresholding showed higher-than-nominal FPR rates in their testing framework—though again, it is worth noting that their own results show that these overreaches tended to be moderate for the kinds of parameters most used in the literature; also see (Cox et al. 2017a; Cox et al., 2017b).

The principal causes of the inflated empirical FPRs they found in the examined parametric methods were, in our opinion, twofold. First, at that time, all of these methods used a Gaussian shape to model the spatial ACF of the noise in FMRI data, but it appears that the actual noise autocorrelation function often has significantly longer tails, which has substantial effects on the estimated FPR, especially at larger *p*-thresholds (Cox, 2017b). Second, these parametric methods implicitly assumed that the ACF is stationary (i.e., roughly constant) across the brain; however, the ACF can have very large variations across the brain, even within a given tissue type, which also greatly affects FPR. Additionally, from follow-up simulations and analyses, we would hypothesize that the ACF is also not the same at all temporal frequencies, which may partially explain why FPR control can be less accurate for slow block designs (since it is the noise in the subspace of the FMRI time series model that is the “enemy” of detection), at least in their testing framework.

In an earlier paper (Cox et al., 2017b), we addressed the first point directly in AFNI's parametric cluster-simulation approach by modeling the ACF with a non-Gaussian shape (termed a "mixed model" ACF), and were able to improve the FPR performance markedly for relatively stringent per voxel *p*-thresholds (≤0.002). In a separate nonparametric analysis that omits any explicit model for the ACF, we were also able to achieve robust FPR control for voxelwise *p*≤0.01 by using a pure resampling and cluster-simulation method. Unfortunately, the penalty for this latter approach to detection is that the resulting cluster-size thresholds can be very large, making the method very (and likely "overly," for many parts of the brain) conservative. In addition, the false positives are far from being uniformly distributed in the brain volume (Eklund et al., 2016), suggesting that biases still remain, in large part due to the nonstationarity of smoothness across the brain.

We developed the method presented herein, equitable thresholding and clustering (ETAC), in an attempt to overcome the related problems of overly-strict thresholds and highly non-uniform FPR density. As described here and below, the term "equity" is applied in several important contexts. The method has the added benefit of reducing the influence of parameters chosen semi-arbitrarily (e.g., smoothing radius or voxelwise *p*-value threshold; "semi-" because there is generally a window of acceptable values in the literature, but the exact value used may have a large impact on the final outcome).

Briefly, ETAC works by implementing a collection of related individual component tests—called sub-tests from here onwards—across a range of parameter values, such as smoothing radius and voxelwise *p*-threshold, for dual (voxelwise, then cluster-wise) thresholding, and then merging their results via set union. Each sub-test is a “standard” type of test carried out in FMRI group analyses, as illustrated in (Eklund et al., 2016). We maintain *equity* (or balance) among this collection of sub-tests (and therefore across otherwise arbitrarily chosen parameters) by constraining each individual sub-test’s cluster-threshold to produce the same FPR as every other sub-test. The global (or final) FPR is controlled by adjusting all sub-tests’ equitably constrained cluster-threshold parameters simultaneously in order to reach the desired target FPR in the simulations. Furthermore, each voxel in the brain mask is treated similarly—equity is applied among voxels, as well as equity among sub-tests—to have its own FPR approximately equalized. These abstract concepts are presented in detail in the Methods section below and demonstrated concretely in the Results.

The use of multiple sub-tests reduces the number of semi-arbitrary choices the FMRI data analyst must make. Effectively, multiple *p*-thresholds and multiple levels of data blurring are used: the data analyst only needs to choose ranges of these parameters, rather than a specific *p*-threshold and a specific spatial blur, and each combination comprises one sub-test. As a result, ETAC has the potential to detect both small, intense clusters (found using small *p*-value thresholds and small blurring) and large, weak clusters (found using large *p*-value thresholds and perhaps more blurring). This choice of sub-tests is intended to balance cluster detection across spatial location, spatial scale (cluster-size), and cluster intensity. As will be demonstrated, ETAC gives reasonably accurate control of false positives, and provides the same or slightly more statistical power than the use of a fixed cluster-size threshold that also controls global FPR.

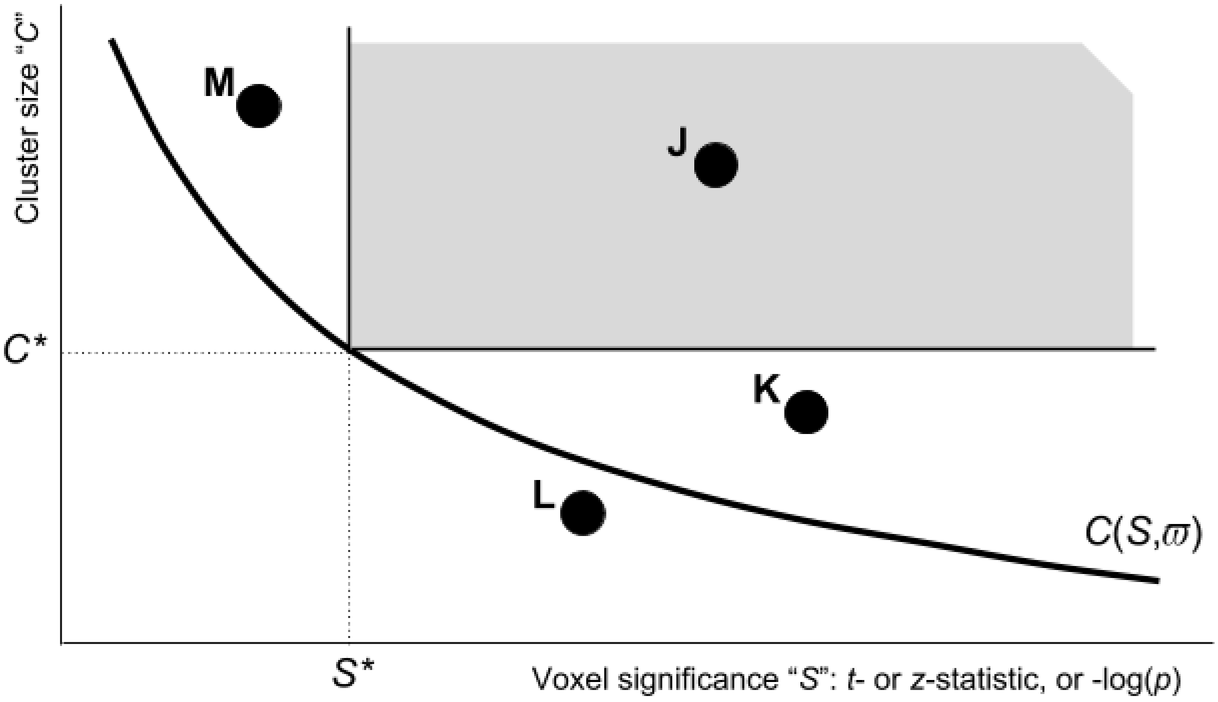
Above, we appealed to various features of the FMRI time series noise, especially its spatial smoothness, in our path of reasoning leading to the new thresholding method; however, we emphasize that our ETAC approach does not directly model these features. We do not attempt to use *ab initio* simulation to create synthetic random noise fields for cluster-threshold emulation. Instead, our approach here is to: 1) use the residuals from the voxelwise statistics (e.g., *t*-test or linear regression results) as exemplars of the FMRI noise at the group level, and then 2) build the thresholding method on randomizing/permuting these residuals to generate further realizations of the FMRI noise field, still at the group level. Our reasons for taking this approach are twofold: (a) using this randomization/permutation technique to select the cluster-size threshold in the dual threshold method has been more precise (in the tests undertaken) thus far in controlling the false positive rate across a wide range of *p*-thresholds than has been parametric modeling and simulation of the FMRI noise field (Eklund et al., 2016; Cox et al., 2017b); and (b) simulations of 3D random fields with complex spatial correlation structures is very compute-intensive and is bound up with choosing a reasonably accurate parametric model for these correlations.

In the next section, we outline the concepts underlying the ETAC approach. In the Appendix, we provide a detailed summary of how ETAC is implemented in AFNI. To examine the global FPR performance of ETAC, we used simulations *à la* (Eklund et al., 2016), with resting state FMRI as “null data” for task-based analyses. To examine the ability of ETAC to detect signal changes (thereby investigating its "power"), we analyzed an publicly available collection of task FMRI datasets.

Equitable Generalizations of Dual Thresholding

Figure 1 shows a schema for judging all possible clusters in a brain mask (i.e., globally). A point (*S*,*C*) in this configuration space represents a cluster of *C* neighboring voxels, each of which is individually "more significant" than *S* (representing *t*, *z*, -log(*p*), or other similar statistic). The dual thresholding approach for finding significant clusters is also shown: first, a voxel threshold is chosen (a value along the abscissa, e.g., *S**), which delimits part of the configuration space; one then determines a cluster-size *C** which further delimits the configuration space to the dark shaded region, for a desired global (i.e., within brain mask) FPR ϖ (cursive Greek "pi"). Thus, *C** depends on both *S* and ϖ; in general, for a given FPR, the cluster-size threshold $C^*=C(S^*,\varpi)$ is a decreasing function of per-voxel threshold significance *S**, as shown (the solid black curve). There is a tradeoff implicit in the dual thresholding approach: a more stringent per-voxel threshold (moving the shaded region rightwards) allows for detection of smaller clusters (the shaded region extends further downwards). The function $C(S,\varpi)$ is determined by simulation in AFNI and by an approximate asymptotic formula in SPM.

Figure 1. Graphical view of the dual thresholding tradeoff. All voxels that meet a voxelwise threshold and a cluster-size (contiguity) threshold are kept—the gray region indicates the combinations that pass a particular instance of this procedure: each voxel's statistic must pass a first significance threshold *S**, and the number of contiguous voxels (in a single cluster) must be above a second threshold *C**. The thick black curve $C(S,\varpi)$ indicates the cluster-size threshold that gives a fixed global FPR ϖ (e.g., 5%) as a function of the per-voxel threshold *S*; ϖ is the probability that a noise-only cluster falls into the gray region. See the main text for discussion of named clusters J-M.



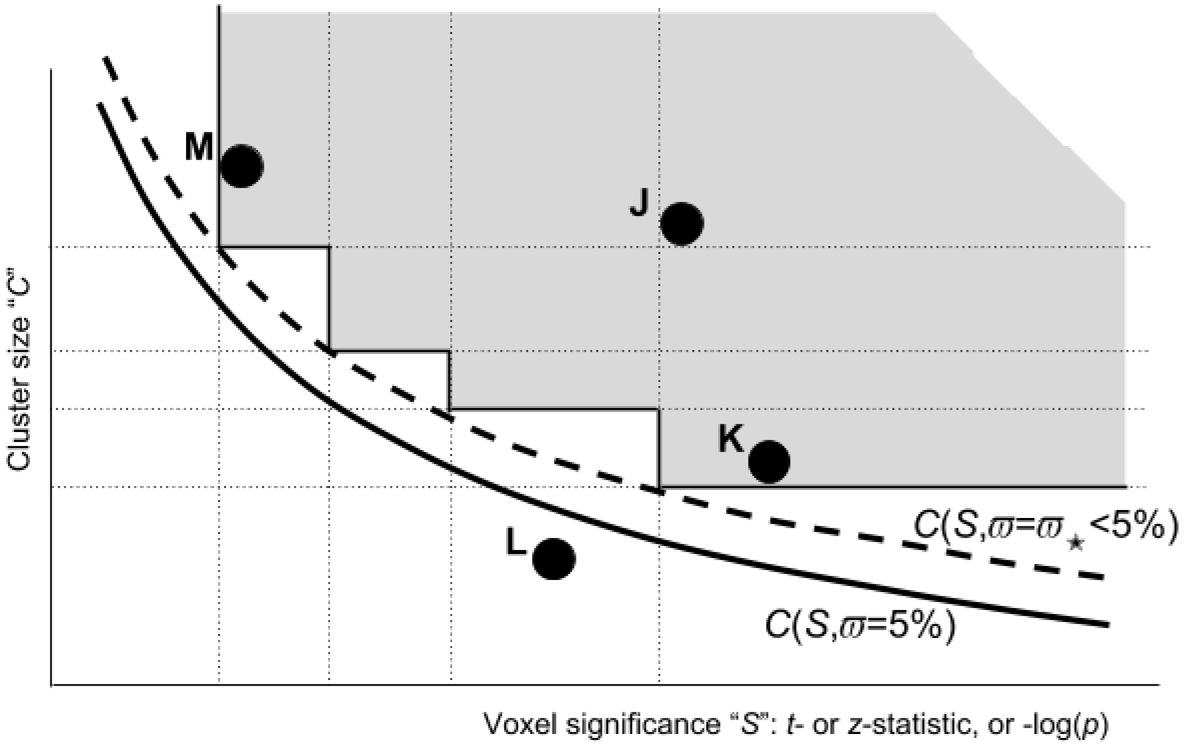
Here, the J cluster passes the thresholding test, by containing "enough" voxels for the given voxelwise threshold. The K cluster is made up of voxels that easily pass the threshold test, but it does not have enough voxels to “survive” in the present scenario; however, if the per-voxel threshold test were more stringent (S^* moved rightward), then the K cluster could pass the dual threshold test. For the “right” choice of per-voxel threshold, both clusters J and K would survive. The L cluster has voxels that pass the per-voxel threshold, but not enough of them and there is no per-voxel threshold that will permit L to survive at the desired FPR. Conversely, the M cluster has a lot of voxels above a slightly smaller per-voxel threshold. Again, one could make M survive by playing with the per-voxel significance threshold, as M lies above the $C(S,\varpi)$ tradeoff curve. Note that, for a given nominal FPR ϖ, the choice of voxelwise thresholding statistic S^* is essentially the primary quantity for determining which clusters are accepted or rejected, because the cluster size threshold C^* is just a function of S* parametrized by a (typically 5%) ϖ value.

Figure 1 and the reasoning above lead to the obvious question: If we are interested in obtaining clusters based on their dual properties of voxelwise significance and size, how should we choose the single voxelwise threshold that determines the result? As shown using the simulations of (Eklund et al., 2016) and (Cox et al., 2017b), at some larger *p*-thresholds our ability to control FPR decreases; in the present language, there did not appear to be an accurate relation $C(S,\varpi)$ for some large *p*-values (small *S*-values), such as *p*=0.10. But even within a reasonable range of *p*-values, the choice of threshold might affect the outcome dramatically. As illustrated in Fig. 1, we could shift this voxelwise threshold along the horizontal axis and obtain the following combinations of “statistically significant” clusters: only M; both J and M; only J; both J and K; or only K. When the results can depend strongly on the choice of an arbitrary parameter, any value of which might considered “reasonable” within a large interval, we are in a bad situation.

Figure 1 also leads to the idea of finessing the tradeoff between cluster-size and per-voxel thresholds; that is, simply accept every voxel configuration which lies above the solid black curve $C(S,\varpi)$. From the viewpoint of the final FPR of the dual threshold method, any place along that curve is as acceptable as any other. The principle of equity asks, “Why discriminate?”; however, a practicable algorithm must use a discrete set of *p*-thresholds combined with their corresponding cluster-size thresholds. Of course, one also has to make sure that the desired FPR is still achieved in the final results.

Figure 2 illustrates the use of four different per-voxel thresholds in a single test. In this example scenario, using the equitable thresholding method, clusters J, K, and M survive. When using the multiple voxelwise thresholds (proposed multi-*S* or multi-*p* case), it is important to note that we cannot use the original $C(S,\varpi)$ tradeoff curve (from the mono-S case) to calculate the cluster threshold for each sub-test; this would lead to a final FPR that is larger than the desired ϖ, since more potential voxel configurations are allowed with the use of multiple S-thresholds than with a single S-threshold. It is thus necessary to find an $\varpi_\star < \varpi_G$ (subscript "G" for "goal") such that using the $C(S,\varpi_\star)$ for the individual sub-tests will achieve the desired ϖ_G in the final combined simulations. That $\varpi_\star$ value then defines the multi-p-threshold clustering algorithm when applied to the original voxelwise statistical tests. We do not make the Bonferroni correction $\varpi_\star=\varpi_G/4$ among these four tests. This correction would be very conservative, as indicated by the strong overlap among the individual mono-p-threshold regions in Fig. 2. We describe in detail how $\varpi_\star$ is adjusted (via simulation) to make the final FPR equal to ϖ_G in the Appendix.

Figure 2. Thresholding as in Fig. 1, but using four different per-voxel p-thresholds at once (vertical dotted lines), each using its own cluster-size threshold (horizontal dotted lines). The FPR for the gray region (the union of four mono-p-threshold rectangles), if constructed above the solid black curve $C(S,\varpi)$, would be larger than the desired ϖ (FPR for the gray rectangle shown in Fig. 1), since this union region would include more possible voxel configurations. To compensate for that and maintain a nominal (5%) FPR, the cluster-size threshold curve from Fig. 1 (solid curve) must be raised to a higher level by picking a curve $C(S,\varpi_\star)$ with an $\varpi_\star$ < 5% (dashed). Which $\varpi_\star$ should be chosen is determined by simulating this multi-p-thresholding process—just as the curve $C(S,\varpi)$ for the original dual threshold method is determined (in AFNI) by simulating the mono-p-threshold procedure.


M
J
K
L
$C(S,\varpi=\varpi_\star<5\%)$
$C(S,\varpi=5\%)$
Cluster size "C"
Voxel significance "S": t- or z-statistic, or -log(p)

One advantage of this approach over the simpler mono-*S*-thresholding method is that it allows large clusters of low effect size (smaller *S*, larger *p*) to be detected along with small clusters of high effect size (larger *S*, smaller *p*) within a single analysis. In the more common mono-*p*-threshold approach, the user has to decide which type of cluster to favor. It would be difficult for a user to know *a priori* what single *p*-value would be appropriate or preferable for a study. This arbitrariness can all too easily lead to the user adjusting the per-voxel *p*-threshold to get more “desirable” results. Additionally, a user may be interested in both large and small clusters (e.g., responses in amygdala and PCC) within a single experiment, an outcome which might be arbitrarily excluded by the constraint of being allowed only one single voxelwise threshold. In the ETAC method, only a plausible range of *p*-thresholds needs to be chosen. In summary: where results initially depended on just a single parameter chosen from within a range of reasonable values, the ETAC method has allowed multiple parameter values to be chosen, tested, and have their results combined into a single result, while simultaneously balancing the weight of each sub-test and maintaining control of the final FPR.

The idea of combining multiple dual threshold sub-tests for detection (in Fig. 2, each sub-test is specified by its voxelwise *S*- or *p*-threshold) can be directly abstracted and generalized to other parameter dependencies. For example, typically only one spatial smoothing length (full-width at half max, FWHM) is chosen for a given study, but commonly implemented values within the literature occupy a range of roughly 4-10 mm. No matter what the sub-test is, its individual FPR will be controlled by a cluster-thresholding parameter. The ETAC approach is to require each sub-test to have the same individual FPR $\varpi_\star$ and to accept a voxel if it passes *any* of the individual sub-tests; then, $\varpi_\star$ is adjusted to achieve the goal of FPR=ϖ_G for the final set of accepted voxels. In this way, equity across the parameter ranges of the sub-tests (in Fig. 2, defined by *S* threshold values) is maintained while the overall FPR is still achieved.

The presence of strong spatial inhomogeneity in the noise ACF (Cox, 2017b) leads to the requirement that large cluster-size thresholds be set in order to accommodate the highly smooth regions (e.g., the brain midline), which may be of interest in a study. In turn, these large cluster-size thresholds make detection of smaller clusters impossible, even in regions where the noise ACF decays quickly (i.e., where the noise is not “smooth”). This drawback led us to the next two generalizations of the ETAC method: applying equity across multiple cases of blurring, and applying equity across brain regions via spatially dependent cluster-thresholds.

Using multiple cases of blurring to enable FMRI detections across spatial scales is certainly not new (Worsley et al., 2001). The ETAC framework maintains equity across blur radii as follows: the clustering thresholds at the different scales of blurring are balanced by making each sub-test (i.e., combination of *p*-threshold and blurring FWHM) have the same individual FPR $\varpi_\star$. The effect of this multi-blur analysis is to remove the semi-arbitrary choice of blurring scale from the data analyst. For this approach to work, the first-level (individual subject) analyses should be done *without* spatial blurring, since un-blurring afterward is impracticable. Instead, the blurring of the first-level results (individual subject "betas") should be carried out by the ETAC software; alternatively, multiple cases of blurring at the first-level could be analyzed separately and those multiple results for each individual subject then entered into the ETAC algorithm. This latter approach would be useful for nonlinear time series blurring approaches (e.g., local principal singular vector), and for applying ETAC to resting-state FMRI seed-based correlation maps. However, this approach to supplying multiple blur cases has yet to be implemented in AFNI's ETAC software; instead, the software does the additional Gaussian-in-brain-mask blurring of the input datasets internally.

The sizes of cohesive brain regions vary across the brain (e.g., amygdala vs PCC) and so do the noise characteristics (e.g., smoothness of time series residuals). To be equitable, the cluster-threshold tradeoff curve should vary with brain region. Theoretically, this effect could be modeled with simulated noise, but in practice the difficulty lies in accurately modeling and efficiently creating realizations of 3D noise fields with non-uniform correlation. Instead, we have chosen to model this effect with a randomized *t*-test approach. The goal is to keep the *FPR* approximately the same across the brain (at least, to keep it more uniform than would result from using single global cluster-thresholds). As a corollary, regions that are less smooth will get smaller cluster-size thresholds, and regions that are more smooth will get larger cluster-size thresholds, and it will thus be possible to detect smaller clusters of activation in "naturally favorable" parts of the brain. In order to combine the sub-test information, each sub-test is now to be cluster-thresholded using a 3D cluster-threshold *map*, not just a single number. The cluster-thresholds to use at each voxel, for each sub-test, are determined again by "equitable" balancing—the Appendix describes the algorithmic implementation in more detail.

Finally, cluster-size is a commonly used figure of merit (FOM) for assessing significance in neuroimaging. The size of a cluster (voxel count) is the sum of the number 1 across all voxels within the cluster; that is, $\mathbf{\Sigma}$ 1. From this formulation, it is simple to generalize the cluster significance FOM to other sums; for example, incorporating the statistic associated with each voxel as a weight, such as $\mathbf{\Sigma}\, z^2$, where *z* denotes the $\mathcal{N}(0,1)$ normal deviate that matches the voxel-level test statistic (e.g., *t* or *p*) in tail probability. The TFCE method (Smith & Nichols, 2009) uses a related cluster FOM in its "threshold free cluster enhancement" algorithm. The current implementation of ETAC in AFNI allows for the cluster thresholding FOM to be computed as $\Sigma\,|z|^h$ where h=0, 1, and/or 2; the h=0 case corresponds to the unweighted cluster size, $h > 0$ provides for significance weighting, which allows for an increased possibility of finding small but relatively significant clusters to be detected.

Tests and Results

The first test of the proposed ETAC approach is a variation on the methodology in (Eklund et al., 2016). From the FCON1000 collection of resting-state FMRI datasets (Biswal et al., 2010), the 198 datasets in the Beijing sub-collection were processed with AFNI pipelines specified using afni_proc.py to produce the first-level (individual subject) activation maps. Three different stimulus durations were modeled as pseudo-tasks: 1 s, 10 s, and 30 s. For each duration, 5 sets of pseudo-random timings were used; thus, for each duration, 198×5=990 maps of fit coefficients (betas) were generated, with no spatial blurring used at this analysis level. The BOLD hemodynamic response function (HRF) regressor models for all 5 pseudo-task timings in each duration are nearly orthogonal—stimulus duration 30 s had the most highly correlated HRF models, with mean correlation 0.09 and standard deviation 0.21—so that the 5 estimated response magnitudes (betas) in each subject were approximately independent. It is this use of multiple randomized timings that distinguishes our testing methodology from (Eklund et al., 2016).

At the second (group) level, two-sample *t*-tests were carried out, with 20 randomly chosen subjects assigned to each sample; for each subject, 1 of the 5 results from the different task timings was selected randomly in each simulation. A total of 1,000 realizations of this procedure was analyzed with ETAC, for each of the 3 stimulus durations. When running ETAC, 3 blur cases were used in defining the sub-tests: 4 mm, 7 mm, and 10 mm (FWHM). Voxelwise *p*-thresholds were set at geometric intervals to: 0.010, 0.005, 0.002, and 0.001; the NN cluster-defining level was left at the default 2 (i.e., "neighbors" of a voxel are those sharing either face or edge), and the default FOM with h=2 (i.e., statistically weighted) was used. For each stimulus class (1 s, 10 s, 30 s), three results were calculated: the FPRs for 1-sided (positive or negative) or 2-sided *t*-thresholding; thus, there are 3×3=9 results for the two-sided *t*-tests for each goal FPR. Additionally, a very similar set of simulations was run separately using one-sample *t*-tests with 40 subjects per simulation. In all cases, the panoply of nominal (goal) FPRs was run, with ϖ_G=2%, 3%, ... , 9%. Links to AFNI scripts for the first and second level analyses are given in the Appendix.

The empirical FPRs from all these 1-sample and 2-sample *t*-test simulations are presented in Figure 3, along with the 95% (binomial) confidence intervals that would be expected from 1,000 *independent* simulations if the algorithm exactly achieved the nominal FPR. For the two-sample *t*-tests (with 40 subjects total), the results are tightly clustered, very near the nominal FPR and typically within the confidence interval, though biased slightly high. For the one-sample tests with 40 subjects, the results are more variable, as well as also being biased a little high.

To understand the variability of the one-sample results from ETAC, we carried out numerical simulations with pseudo-random normally distributed datasets. Two sets of simulations were run, the first using independent collections of datasets for each *t*-test, and the second using collections resampled from a finite sample of datasets (and thus, the *t*-tests are not independent). The results are in Figure 4 (where the "FP Count" of 50 corresponds to an FPR of 5%). The upper plot with independent collections exhibits both a near-zero bias and a small spread; The lower plot, with non-independent collections, has a much larger spread in estimated FPR values. This comparison demonstrates that the assumption of *independent* simulations is critical (and often not met) for accurately assessing the FPR distribution of the one-sample tests. In the above analyses with real data (cf. Fig. 3), the tests are clearly not independent, as there are only 198 subjects (or 990 beta datasets) from which to draw 40 subjects per test. The effect of this non-independence of the one-sample *t*-test collections is to greatly increase the variance of the empirical FPR results. In short, the methodology used in (Eklund et al., 2016), in (Cox et al., 2017ab), and herein, can only provide a rough evaluation of FPR accuracy for one-sample *t*-tests. While this point has arisen as a methodological sidenote to ETAC and cluster analysis methodologies, it is an important one for the neuroimaging community as a whole, as such one-sample tests are commonly used in published studies.

Figure 3. ETAC false positive rates from simulations of two-sample *t*-tests (20 subjects per sample) and one-sample *t*-tests (40 subjects in the sample). The nominal (desired) FPR ϖ_G is shown along the horizontal axis, and is also shown as a dashed line within the gray boxes, which represent the 95% confidence intervals for the FPR (from a binomial distribution, assuming the nominal FPR as the binomial parameter and assuming all 1000 replicates are independent). Thus, symbols that fall inside their gray box are within the 95% CI of what would be expected if the nominal FPR was the actual ETAC FPR. Medians for the 9 values in each partition are shown with solid black lines. The two-sample ETAC results are fairly close to the nominal FPR and tightly clustered, but the one-sample results are biased higher and have a great deal more variability; see Fig. 4 for an explanation of this increased dispersion.

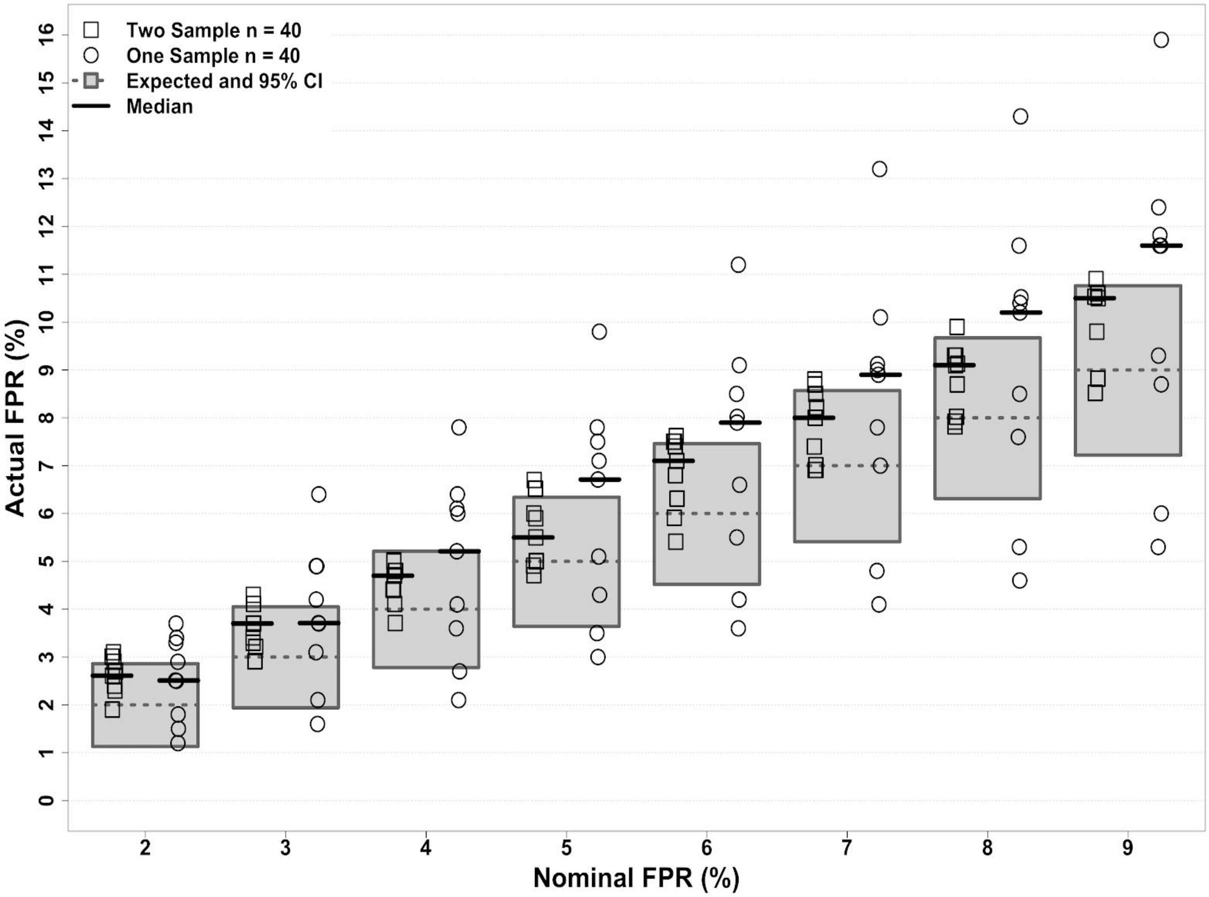
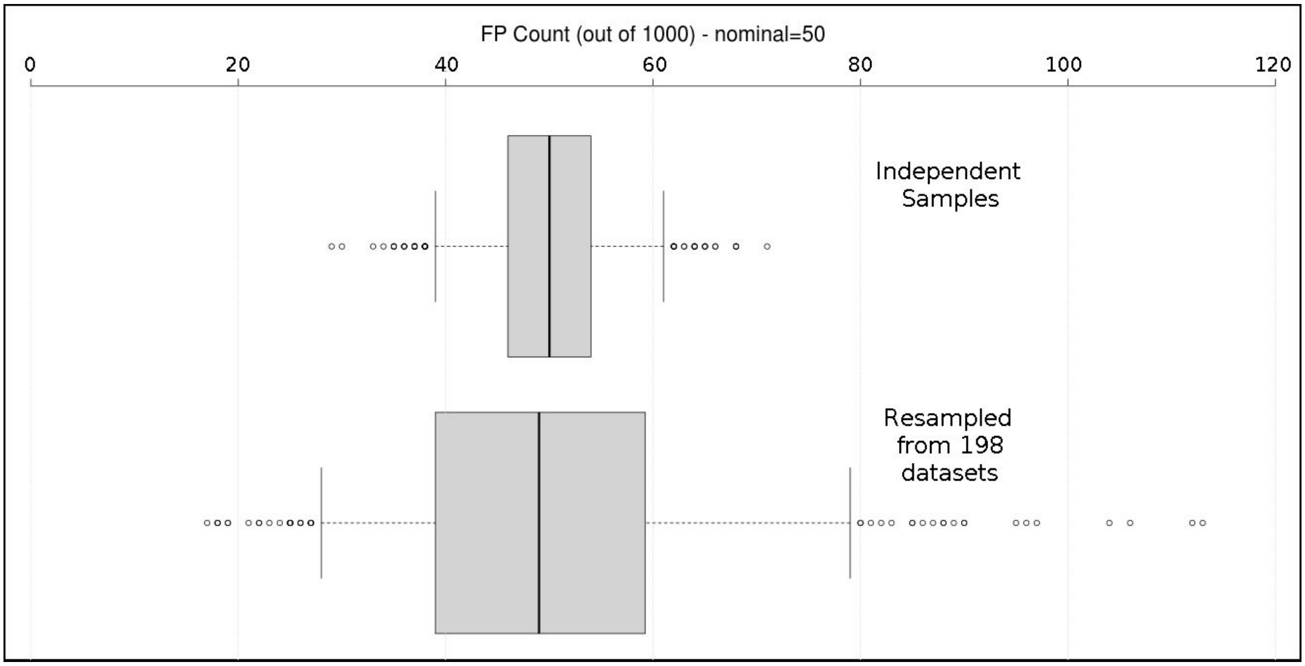
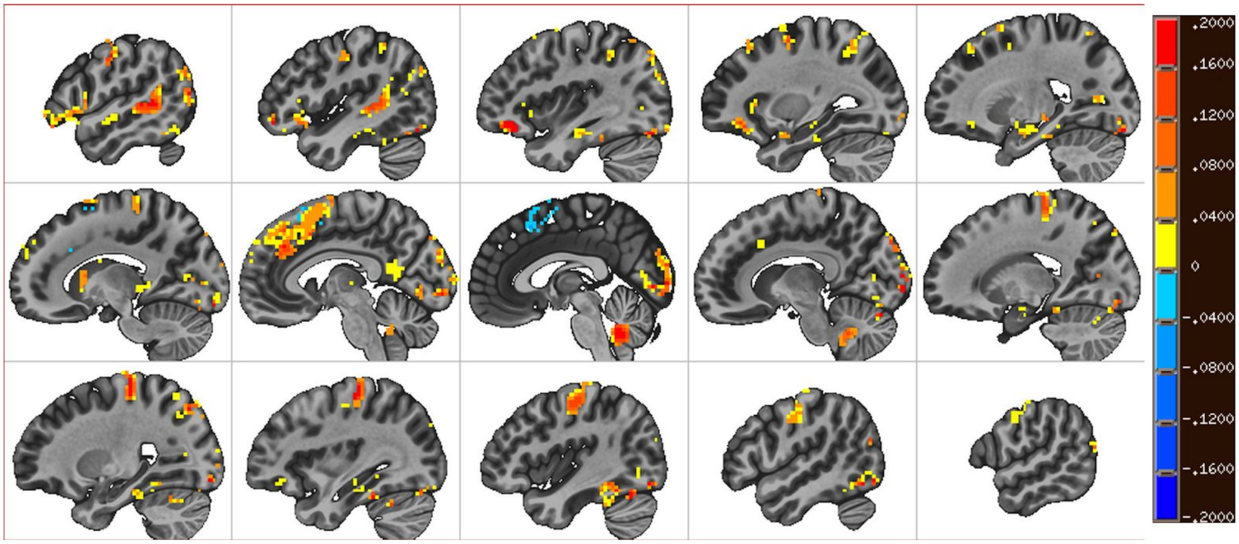


Figure 4. Box-whisker plots of False Positives (FP) results from a fixed (not spatially variable) cluster-size thresholding method (Cox et al., 2017b) applied to simulated datasets. Each simulation (one data point in this Figure) used 40,000 independently generated normally distributed 3D volumes and carried out 1,000 one-sample *t*-tests using these as inputs; 40 datasets were used in each one-sample *t*-test. For the upper plot, the simulated datasets taken for the 1,000 *t*-tests were completely separate and thus independent. For the lower plot, only the first 198 datasets were used and 1,000 random subsets of 40 of these were taken for each FP calculation; thus, the 1,000 replicates were *not* independent. The result from each FP calculation using 1,000 *t*-tests is a single FP count. Then 500 iterations of the above processes were run to produce the results shown (20 million 3D volumes in total). For the resampled results, the variability in observed FP is much larger. Similar calculations (not shown) indicate that this higher variability due to non-independence is not a significant effect for two-sample *t*-tests.



To test the ability of ETAC to find true positives, we downloaded a collection of datasets from OpenFMRI.org (currently moving to OpenNeuro.org). We used the UCLA Consortium for Neuropsychiatric Phenomics LA5c Study (Poldrack et al., 2016; collection ds000030), and picked out the pattern matching and encoding “pamenc” task, using 81 control subjects (one subject’s data had to be discarded due to excessive head motion). One-sample *t*-tests were carried out with 20 subjects per test, randomly selected; 500 iterations of this analysis were carried out with ETAC and also with a fixed cluster-size threshold. Figure 5 shows the map of the *differences* in the detection rate between ETAC and fixed cluster-size thresholding (i.e., Fig. 5 is not an activation map, but shows the places where the two methods differed in likelihood of finding results). For the most part, ETAC was slightly superior to the fixed cluster-size threshold method, and did not significantly lose power in any region.

Figure 5. (A) This image is *not* an activation map, but rather the *difference* in the detection probabilities between ETAC and a fixed cluster-size threshold method, applied to the OpenFMRI.org ds000030 collection using 500 simulations of one-sample *t*-tests with 20 subjects per test. Yellow-red colors indicate ETAC was more likely to indicate (true) detection in a given voxel; cyan-blue colors indicate the fixed cluster-size threshold method was more likely. The point of this image is not that ETAC is greatly superior to the more standard fixed cluster-size threshold method, but that it does not lose power significantly. (Much more of the brain is “truly” active than shown here; in many places, the detection probabilities of the two methods were nearly identical.)



Discussion

The ETAC method has been shown to be reasonably effective at controlling the false positive rate for cluster-based detection in task FMRI. This performance is achieved while also reducing the influence of the arbitrary (but constrained) selection of parameters affecting the cluster results (which was the primary motivation for the method's inception). ETAC does not lose true detection power relative to simpler cluster-thresholding methods, and in some cases can be a little more powerful.

A significant advantage of ETAC is that it removes two arbitrary choices from the group analysis: the voxelwise *p*-threshold, and the amount of spatial blurring to apply. For example, there are no strong reasons to favor *p*=0.001 over *p*=0.01 (provided FPR is properly controlled), or 8 mm (FWHM) blurring over 4 mm blurring.

The computational burden of this method is not trivial. A single ETAC run with 40 subjects at 2 mm resolution took about 2 hours on a 16 core Intel node, and needed 40 gigabytes of memory to hold all the simulations and cluster tables. (This time could be cut by 50-70% if requesting results for only a single FPR, which is common in applied studies where FPR=5% is commonly desired, rather than the full range from 2%-9%, as done for methodological interest in Fig 3.) Of course, this analysis yields the final group map for a study; significant computational effort is a price worth paying to get more reliable and less arbitrary results.

The computational cost makes it difficult to extend the ETAC method to more complex voxelwise statistical mapping techniques, such as Linear Mixed Effects analyses (Chen et al., 2013). Such analysis methods are themselves computationally intensive, unlike simple voxelwise *t*-tests or linear regression. Thus iterating and randomizing such tests thousands of times is (at present) impracticable. This is a broader difficulty with all randomization/permutation methods at present, not just ETAC.

ETAC can be extended to surface domains (2D), or even to the mixed 2D+3D domains supported by the CIFTI format—for cortical surface models plus solid gray matter voxel sets (https://www.nitrc.org/projects/cifti/). There is no geo-metric model (based on distances) used in ETAC; rather, the whole notion is based on topology (neighborhoods), and adapts itself to different locations using the principle of spatial equity.

As it is implemented now, by using randomization and permutation, ETAC is of necessity a group analysis procedure. We have been asked about extending it to single-subject analysis. One approach would be to generate a model of the subject's EPI noise, repeatedly re-analyze the time series with synthetic EPI data, and then use those results to build up the noise-only statistical maps needed for the cluster simulations. The speed of linear regression (when programmed in a compiled language, such as C) should make this a practicable approach, but it would still not be a fast algorithm.

Applying the principle of equity in the spatial domain is one reason ETAC is slow: to get enough simulations to have a significant number of "hits" at most brain voxels requires a lot of work. An alternative approach would be to differentially smooth the first level brain activation maps to bring the noise ACF closer to spatial uniformity (i.e., blur less smooth regions more, and blur already smooth regions not at all). Then a single global cluster-size (or FOM) threshold might be a reasonable approach to control FPR and to keep the FPR density more spatially uniform. Fewer simulations (e.g., 2,000) are needed to compute the cluster thresholds when single values apply everywhere. However, this technique would have the drawback of making the entire brain volume as blurry as the most blurry region (typically retrosplenial or posterior parietal cortex), which is likely a price that most researchers would not wish to pay for the speedup in time.

Conclusion

We have described the ETAC (Equitable Thresholding and Clustering) methodology and compared it with an existing standard cluster technique that also controls false positive rates reasonably. ETAC's primary purpose is to reduce the dependence of arbitrary parameter choices on final cluster results, as well as to allow for the discovery of a greater range of cluster types (large with relatively low voxelwise significance; small with relatively high voxelwise significance; etc.) in a single analysis and with essentially equal footing. This method comes with higher computational cost, but it does not appear to sacrifice power. An additional finding of note in this study was a comment on one-sample tests: in most MRI analyses, the group sizes of data mean that the sets compared in the permutation/randomization tests are not fully independent, and as a result they tend to have much larger variability than a naïve binomial model would imply.

Acknowledgements

The research and writing of the paper were supported by the NIMH Intramural Research Programs (ZICMH002888) of the NIH/DHHS/USA. This work extensively utilized the outstanding computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). Data for the analyses illustrated herein included the FCON-1000 collection (https://www.nitrc.org/projects/fcon_1000/) and OpenFMRI collection ds000030 from (https://openfmri.org/dataset/ds000030/). Thanks go to Paul Taylor for many suggestions that improved the manuscript immeasurably. The ideas for equitable application of multiple sub-tests came to RWC while hiking in Grand Canyon National Park and in Gunnison National Forest; thanks go to the US National Park Service and US Forest Service for maintaining these inspirational public treasures.

References

Biswal BB, Maarten M, Zuo X-N, et alii (2010). Toward discovery science of human brain function. Proceedings of the National Academy of Sciences **107**:4734-4739. https://10.1073/pnas.0911855107

Chen G , Saad ZS, Britton JC, Pine DS, RW Cox. Linear mixed-effects modeling approach to FMRI group analysis. NeuroImage **73**:176-190. http://dx.doi.org/10.1016/j.neuroimage.2011.12.060

Cox RW (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Computers and Biomedical Research **29**:162-173. https://www.ncbi.nlm.nih.gov/pubmed/8812068

Cox RW, Chen G, Glen DR, Reynolds RC, Taylor PA (2017a). fMRI clustering and false-positive rates. Proceedings of the National Academy of Sciences **114**: E3370-E3371. https://doi.org/10.1073/pnas.1614961114

Cox RW, Chen G, Glen DR, Reynolds RC, Taylor PA (2017b). Brain Connectivity **7**:152-171. https://doi.org/10.1089/brain.2016.0475

Eklund A, Nichols TE, Knutsson H (2016). Cluster failure: Inflated false positives for fMRI. Proceedings of the National Academy of Sciences **113**:7900-7905. https://doi.org/10.1073/pnas.1602413113

Flandin G, Friston KJ (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

Friston KJ, Worsley KJ, Frackowiak RS, Mazziotta JC, Evans AC (1994). Assessing the significance of focal activations using their spatial extent. Human Brain Mapping **1**:210-220. https://doi.org/10.1002/hbm.46001030

Poldrack RA, Congdon E, Triplett W, Gorgolewski KJ, Karlsgodt KH, Mumford JA, Sabb FW, Freimer NB, London ED, Cannon TD, Bilder RM (2016). A phenome-wide examination of neural and cognitive function. Scientific Data **3**:article number 160110. https://doi.org/10.1038/sdata.2016.110

Smith SM, Nichols TE (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference, NeuroImage **44**:83-98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Worsley, K (2001). Testing for signals with unknown location and scale in a χ^2 random field, with an application to fMRI. Advances in Applied Probability **33**: 773-793. https://doi.org/10.1239/aap/1011994029

Appendix

A. Implementation Details

ETAC is implemented inside the AFNI group analysis program 3dttest++, for convenient use by the data analyst. All operations take place in a user-selected mask of voxels (e.g., a brain mask or a gray-matter mask). An algorithmic outline of the procedure is:

A. Generate *N* (default 40,000) noise-only random fields (volumes) of *z*-statistics for cluster analysis; the following steps are used to create each random field generated:
   a. The residuals in the voxelwise GLM (*t*-tests, possibly with subject-level covariates) are sign-randomized; if two subject samples are used, the

subjects' datasets are also randomly permuted between groups (unless covariates are present).

   b. The GLM is repeated using the randomized/permuted sample, saving the resulting voxelwise *z* statistics (converted from *t*).
   c. The program requires at least 17 input datasets, so that 40,000 distinct random realizations are possible.
   d. These *N* realizations can also be used for the global (non-spatially varying) nonparametric cluster-size threshold analysis described in (Cox et al., 2017), which will produce a table of cluster-size thresholds vs. *p*-thresholds; essentially, the $C(S,\varpi)$ curve from Fig. 1 (for a range of *p* and ϖ values).

B. Apply each of the sub-tests (i.e., *p*-thresholds, blur cases, and *h* values):
   a. For each sub-test, form all clusters in all *N* random fields.
      i. The default *p*-thresholds are five values distributed geometrically between 0.01 and 0.001 (0.0100, 0.0056, 0.0031, 0.0018, 0.0010). The data analyst can choose a different set of *p*-thresholds.
      ii. There is no default set of blur cases; rather, the default is not to apply additional blurring, since the ETAC group analysis software doesn't "know" how much blurring was applied to its input data during preprocessing. If multiple blur cases are specified, each blur case has the entire set of *p*-thresholds applied. (In the simulation testing described later, blurring was not applied during preprocessing and single-subject task analysis, and values of 4, 7, and 10 mm were used at the ETAC group level to blur the subject-level betas.)
   b. Save a table of all clusters found for each sub-test. Each cluster includes a list of its voxels, and its computed FOMs; ETAC can also balance across different formulæ for figures of merit. The default FOM is $\Sigma\, z^2$.

C. "Spread out" the cluster-FOMs in regions where there are not a lot of "hits", so that all voxels have a significant number of FOM realizations for each sub-test, for use later in step D:
   a. For each voxel, make a count of how many clusters "hit" it from each sub-test's cluster table.
   b. For each cluster, find the median number of voxel hits in the corresponding sub-test.
   c. If the number of voxel hits in a cluster from a particular sub-test is below a target (default target is about 0.025*N*=1000), dilate that cluster outward one voxel—without changing the cluster's recorded FOM values (which are used in steps D and E).
   d. Repeat steps C.{a,b,c} until dilations are unnecessary as determined in step C.c (or at most 9 times).

D. For each sub-test and for each voxel, make a sorted (largest first) list of cluster FOMs which hit that voxel.

E. Pick a fraction τ (initial value 0.0006), and set the cluster-threshold for each sub-test, for each voxel, to be that sub-test's FOM list's (τN)'th largest entry; that is, a larger τ selects a smaller threshold (moving down in the sorted list). For example, when N=40,000, the initial cluster-threshold for each sub-test and for each voxel is the 24th entry in the corresponding list. (If τN is not an integer, the cluster-threshold is interpolated from the sorted FOM list.)

   a. This uniform selection in the FOM lists is how equity is applied, since the sorted list represents the reversed empirical cumulative distribution function of each sub-test's FOM in each voxel.

   b. Apply the resulting voxelwise cluster FOM threshold maps to produce the "significant voxels" map for each of the N random fields, merging the results from each sub-test, and count the fraction φ of the N random fields that have *any* surviving voxels; φ is the estimate of the global FPR. (See the text below for further implementation details for this step.)

   c. Adjust τ up (if $\varphi<\varpi_G$) or down (if $\varphi>\varpi_G$), until is φ approximately at the target ϖ_G (which may be any value from 2% to 9%; the default, of course, is 5%).

      i. If the two previous φ results bracket ϖ_G, then inverse linear interpolation in $\varphi(\tau)$ is used to adjust τ for the next trial.

      ii. Otherwise, τ is just scaled linearly by ϖ_G / φ for the next trial.

      iii. In usage thus far, usually only a few (2-3) iterations are necessary for convergence.

      iv. The user has the option to compute threshold maps for a range of FPR goals ϖ_G=2%, 3%, …, 9%, in order to allow perusal of the results at various levels of statistical stringency.

F. Apply the final multi-method cluster-FOM threshold maps to the actual *t*-statistics resulting from the original GLM analysis.

   a. If multiple blur cases are used, each blur case's multi-*p*-threshold maps are applied to the GLM tests on blurred copies of the original datasets, and the set of resulting detection maps (one from each blur case) are merged to make the final detection map.

   b. The outputs are a binary map (NIFTI dataset), indicating which voxels survived the process. This dataset can be used to mask the original GLM results to produce a final "activation map." A second dataset is also produced, indicating which of the sub-tests were passed for each voxel. The main use for this dataset is for analyzing the ETAC process itself, to determine which sub-tests might have contributed unique results. Datasets embodying the multi-threshold maps are also saved.

The software is written in the compiled language C to be able to run in an acceptable time frame for group analyses, and is parallelized across multiple CPU cores. In typical cases, the majority of the computational time is spent carrying out the 40,000 repeated *t*-tests to produce the simulated noise volumes.

We would like to expand upon a critical detail in the implementation of steps E.b and F.a above: how is a spatially variable cluster-FOM threshold map to be applied to a given cluster that results from the application of a given *p*-threshold? Unless a miracle occurs, the final cluster-FOM thresholds in every voxel of a given cluster from the real data will not be identical, requiring a further choice as to what cluster-FOM threshold should be applied to the particular cluster-FOM value calculated from the cluster in question. After some experimentation and even thought, we chose to take the cluster-FOM threshold values that overlap the given cluster, sort them, and take the 90% point (not quite the largest) in the cluster-FOM distribution as the threshold to apply (for each sub-test). This empirically motivated and validated choice resulted in reasonably robust final FPR control, and was simple to implement. (Other percentile points can be chosen by the intrepid user who wishes to experiment with the software.)

Cluster-contiguity can be defined in several ways. In AFNI, contiguity is defined by the nearest neighbor (NN) level, which can be 1, 2, or 3:

1. Voxels are defined as contiguous if faces touch (first nearest neighbors);
2. Voxels are defined as contiguous if faces or edges touch (second nearest neighbors);
3. Voxels are defined as contiguous if faces, edges, or corners touch (third nearest neighbors).

The ETAC software in AFNI does not allow for balancing across these different clustering possibilities; the NN level remains a parameter the analyst must choose (default value in AFNI's ETAC is 2). We do not think that providing equity across different clustering methods would be of major consequence, as simulations and comparisons are made consistently for a given NN value.

The application of different blur cases in ETAC is carried out by applying 3D Gaussian blurring to the input datasets, but the kernel is restricted to the voxel mask supplied by the user. This "blur in mask" procedure is carried out by a finite difference stepping method applied to the 3D diffusion equation, with Neumann (reflecting) boundary conditions at the edge voxels of the mask.

The cluster FOM chosen can be $\Sigma|z|^h$ for h=0, 1, and/or 2. The software can balance across any combination of these FOMs; however, the default choice is the single FOM with value h=2. In practice, we see little advantage in balancing across multiple cluster FOM formulæ. Other FOM formulæ could be added to the software with relatively little effort.

B. Processing Scripts and Data

All scripts are written in the Unix tcsh scripting language. For the most part, these scripts were run on the NIH Linux-based cluster ("Biowulf"). The scripts are available at the GitHub repository https://github.com/afni-rwcox/ETAC-scripts .

The 198 Beijing-Zang datasets can be downloaded from http://fcon_1000.projects.nitrc.org/fcpClassic/FcpTable.html .

The 81 UCLA Phenomics study datasets can be downloaded from https://openfmri.org/dataset/ds000030/ .