Cluster success: fMRI inferences for spatial extent have acceptable false-positive rates

Scott D. Slotnick
Department of Psychology, Boston College, Chestnut Hill, MA, USA

ABSTRACT
In an editorial (this issue), I argued that Eklund, Nichols, and Knutsson’s ‘null data’ reflected resting-state/default network activity that inflated their false-positive rates. Commentaries on that paper were received by Nichols, Eklund, and Knutsson (this issue), Hopfinger (this issue), and Cunningham and Koscik (this issue). In this author response, I consider these commentaries. Many issues stemming from Nichols et al. are identified including: (1) Nichols et al. did not provide convincing arguments that resting-state fMRI data reflect null data. (2) Eklund et al. presented one-sample t-test results in the main body of their paper showing that their permutation method was acceptable, while their supplementary results showed that this method produced false-positive rates that were similar to other methods. (3) Eklund et al. used the same event protocol for all the participants, which artifactually inflated the one-sample t-test false-positive rates. (4) At $p < .001$, using two-sample t-tests (which corrected for the flawed analysis), all the methods employed to correct for multiple comparisons had acceptable false-positive rates. (5) Eklund et al. contrasted resting-state periods, which produced many significant clusters of activity, while null data should arguably be associated with few, if any, significant activations. Eklund et al.’s entire set of results show that commonly employed methods to correct for multiple comparisons have acceptable false-positive rates. Following Hopfinger along with Cunningham and Koscik, it is also highlighted that rather than focusing on only type I error, type I error and type II error should be balanced in fMRI analysis.

ARTICLE HISTORY
Received 22 March 2017
Revised 3 April 2017

KEYWORDS
fMRI; default network; clusterwise inference; false cluster; false positive; familywise error; multiple comparisons; type I error

Given that fMRI analysis involves thousands of statistical tests, the methods employed to correct for multiple comparisons are important to all fMRI researchers. Eklund, Nichols, and Knutsson (2016) claimed that the false-positive rates for commonly employed methods to correct for multiple comparisons were many times higher than the acceptable rate of 5%, which questioned the validity of thousands of published fMRI studies and the validity of the methods themselves.

In Slotnick (this issue), I argued that Eklund et al.’s ‘null data’ actually reflected default network activity that inflated their false-positive rates. Three commentaries on that paper were received by Nichols, Eklund, and Knutsson (this issue), Hopfinger (this issue), and Cunningham and Koscik (this issue). I welcome these commentaries and the continued discussion of this important topic.

Evaluation of the arguments for using resting-state fMRI data as null data

In their commentary, Nichols et al. (this issue) made three arguments as to why resting-state fMRI data reflected null data. First, Nichols et al. stated that ‘synchronized brain activity was not the source of the inflated false positives’ because there were ‘comparable false-positive rates for a given smoothing level and analysis tool.’ Nichols et al. are correct that only activity that is synchronized/consistent across participants will be significant; this is always the case in fMRI analysis. However, that the different designs at a given smoothing level and analysis tool produced comparable results only shows that their results were consistent across those variables and has no bearing on whether or not resting-state fMRI data reflect null data. That is, consistent false-positive rates across those variables would be expected for any stable
pattern of activity, whether or not that activity reflects null data. Second, Nichols et al. stated, ‘task residuals have been found to have resting-state networks …, suggesting the spatial covariance structure of fMRI data is similar in task and rest.’ Task residuals refer to the time periods in the experiment in which the task is not performed (i.e., rest periods between task trials). The pattern of task-residual activity, which corresponds to non-task/rest periods, would be expected to be similar to the pattern of resting-state activity and different from the pattern of task activity. Therefore, fMRI data should not be assumed to be ‘similar in task and rest.’ Third, Nichols et al. stated, ‘we conducted two-sample t-tests on task data … and found the same rates of false positive (sic) as on pure resting data.’ Similar false-positive rates between two-sample t-test results and pure resting data/one-sample t-test results do not indicate that resting-state fMRI data reflect null data. This only implies that the results were consistent across different types of tests (and the one-sample t-test results and two-sample t-test results were actually inconsistent, as discussed in the next section). As such, Nichols et al. did not provide convincing arguments that resting-state fMRI data reflect null data.

**Eklund et al. supplemental results show all methods have acceptable false-positive rates**

As mentioned above, Nichols et al. (this issue) stated that they found the same false-positive rates using one-sample t-tests and two-sample t-tests. Eklund et al. (2016) summarized their findings using one-sample t-test results (see their Figure 1). For all methods, except for their permutation method, the false-positive rates exceeded the nominal 5% level. However, this figure presented selected results and highlighted findings in which the false-positive rates were inflated due to a flawed analysis. The results were presented from the two datasets (Beijing and Cambridge) in which the false-positive rates of their permutation method were acceptable. Figure 1(a) shows the one-sample t-test results from the other dataset (Oulu) in which the false-positive rates for their permutation method also exceeded the nominal 5% level. It should be highlighted that only results at an individual voxel threshold of $p < .001$ are considered, as this is the most common individual voxel threshold employed in fMRI analysis.

With regard to the inflated false-positive rates, Eklund et al. conducted a flawed analysis where the same event protocol was employed for all the participants. As pointed out by Flandin and Friston (2016), ‘arguably, this was a mistake because any systematic fluctuation in resting state timeseries—that correlates with the regressor—will lead to significant one-sample t-tests against the null hypothesis of zero (e.g., magnetic equilibration effects)’ (pp. 2–3). Using resting-state data, one regressor (with magnitude of 0 or 1 across time) was contrasted with another regressor that was anticorrelated with the first regressor (i.e., the magnitude was 1 when the magnitude of other regressor was 0 and vice versa). Because the magnitude of fMRI activity is transiently high at the beginning of each run due magnetic equilibration effects, the regressor that had a magnitude of 1 at the beginning of the time series would be associated with greater activity than the other regressor that had a magnitude of 0 at the beginning of the time series. Thus, for the one-sample t-tests, the contrast of these regressors produced artifactual activity that inflated false-positive rates. For the two-sample t-tests, since both groups of participants had the same regressors, the artifactual activity was subtracted out such that the false-positive rates were not inflated. Figure 1 illustrates that the false-positive rates for the one-sample t-tests (which ranged up to 20–30%) were often many times higher than the false-positive rates for the two-sample t-tests (which ranged up to about 10%; compare Figures 1(a) and 1(b)). The same pattern of results was also observed for another dataset (Beijing, compare Eklund et al.’s Supplementary Figures 4(a) and 2(a); the Cambridge dataset did not show these differential results, perhaps because magnetic equilibration effects had been corrected in this dataset). These findings contradict the claim by Nichols et al. that the same rate of false positives were observed with two-sample t-tests and one-sample t-tests.

Figure 1(b) also shows that at $p < .001$ (the most commonly employed threshold in fMRI analysis) and using two-sample t-tests (which did not suffer from artifactually inflated false-positive rates), the false-positive rates ranged up to only about 10%. Eklund et al. also reported the highest false-positive rate across thousands of simulations. As pointed out by Cox, Chen, Glen, Reynolds, and Taylor (in press), ‘a single case from a host of simulations does not provide grounds for representative characterization or
generalization’ (manuscript proofs, p. 4). Cox et al. conducted the same set of simulations as Eklund et al., but plotted each false-positive rate as a mean ± 1 standard error (see their Figure 2, right). The mean values were nearly identical to the values shown in Figure 1(b) and, of primary importance, the lower standard errors for all methods spanned well below the 5% false-positive rate. These results indicate that the commonly employed methods to correct for multiple comparisons have acceptable false-positive rates.

Should resting-state fMRI data be used as null data?

Resting-state periods are associated with activity in the default network, which includes the dorsolateral prefrontal cortex, the medial prefrontal cortex, the inferior parietal cortex, and the medial parietal cortex. Eklund et al. (2016) conducted contrasts between different resting-state periods in an effort to produce null physiological data. Theoretically, contrasting the same event type, even rest, should produce null/nonsignificant results, which was pointed out by Nichols et al. (this issue) and web commentaries (Neuro skeptic, 2016; Ridgway, 2017). However, contrasting different default network regressors did not produce the expected null results. As discussed in the previous section, Eklund et al.’s one-tailed t-tests produced significant artifactual activity that inflated false-positive rates. As shown in Figure 1(b) of Slotnick (this issue; adapted from Supplementary Figure 18 in Eklund et al.), Eklund et al.’s two-tailed t-tests also produced many significant clusters of activity (e.g., the large cluster in the posterior cingulate within the medial parietal cortex). This is at odds with the null/nonsignificant results that would be expected from contrasting the same event type.

The employment of resting-state/default network activity as null data is also questionable when one considers activity that is typically produced by an fMRI contrast. A contrast generally aims to isolate a specific cognitive process that is associated with a very limited number of activations. For instance, in a spatial memory experiment, the contrast of old-left-hits and old-left-misses isolates the process of explicit memory for items in the left visual field. In a spatial attention experiment, the contrast of left visual field-attended and left visual field-unattended isolates the process of attention to items in the left visual field. Both of these contrasts produce activity that is largely restricted to right hemisphere striate and extrastriate cortex. Default network activity should not be employed as null data for these types of contrasts, as the spatial extent of task activity is very limited and default network activity spans large portions of the brain.

Null data should be associated with few, if any, significant activations. This was the point of Slotnick (this issue) when it was suggested that one way to create null physiological data is to contrast odd versus even trials of one event type (randomized trials should work just as well). To go even further, the event type should be an isolated cognitive process (e.g., old-left-hits or left visual field-attended) to avoid variability between regressors that may produce significant

Figure 1. Eklund et al.’s (2016) one-sample t-test and two-sample t-test results. (a) One-sample t-test familywise error/false-positive rates for SPM, FSL, AFNI, and a permutation method with a cluster-defining threshold (CDT)/individual voxel threshold of $p = .001$, 4–10 mm of smoothing (key to the upper right), and four paradigms (B1, B2, E1, E2; adapted from Supplementary Figure 4c in Eklund et al.). (b) Two-sample t-test familywise error/false-positive rates for SPM, FSL, AFNI, and a permutation method with a cluster-defining threshold (CDT)/individual voxel threshold of $p = .001$, 4–10 mm of smoothing (key to the upper right), and four paradigms (B1, B2, E1, E2; adapted from Supplementary Figure 2c in Eklund et al.).
activations. Related to this, default network regions (e.g., the posterior cingulate and the medial prefrontal cortex) with the highest magnitude of activity have the most variable magnitude of activity (Mao et al., 2015), which may explain why Eklund et al. observed significant activations in these regions. An isolated cognitive process should produce a very stable magnitude of activity in a limited number of regions, and contrasting this event with itself should produce true null physiological data with few, if any, significant activations.

Cox et al. (in press) identified many methodological problems with Eklund et al. However, Cox et al. also assumed that resting-state data reflected null data and modified their analysis protocol to account for long-tailed/non-Gaussian activity. For all the reasons outlined above, I do not believe the modified protocol of Cox et al. should be employed to compute false-positive rates. If true physiological null data are evaluated, it is anticipated that the spatial distribution of activity will be Gaussian, as is assumed by the current methods employed to correct for multiple comparisons.

Although I have argued that null physiological data should be associated with few, if any, significant activations, this may not always be the case. Null physiological data might be associated with long-tailed activity under certain conditions, which is a topic for future research. The point is that it should not simply be assumed that resting-state data reflect null data.

Don’t be afraid of harmless or imaginary bugs

Nichols et al. (this issue) claimed that there was a potential bug in my code that computes cluster threshold extent to correct for multiple comparisons (cluster_threshold_beta.m; Slotnick, n.d.). Nichols et al. stated, ‘while the cluster_threshold_beta program makes use of super-resolution for more accurate convolution, it does not appear to correct for edge effects, similar to the bug we found in 3dClustSim.’ Related to this, Eklund et al. (2016) stated, ‘the bug in 3dClustSim resulted in cluster extent thresholds that are much lower compared to SPM and FSL …, which resulted in particularly high familywise error rates’ (p. 7904). Cox et al. (in press) described the bug in their code as follows, ‘the bug … 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) (manuscript proofs, p. 2). Cox et al. also showed that, at \( p < .001 \), the bug only inflated the false-positive rates \(<1\%\), which shows that it was quite harmless and that Eklund et al.’s characterization of 3dClustSim as having particularly high familywise error rates’ was false. With regard to cluster_threshold_beta, this script does not include any zeros outside the grid/matrix, so there is no basis to the claim that there is a bug in the code that is similar to one previously found in 3dClustSim. Second, if there were uncorrected edge effects in cluster_threshold_beta, this would predict that the cluster extent threshold computed using the program would systematically change as a function of matrix edge size. To test this, a series of 10,000 Monte Carlo simulations was conducted using cluster_threshold_beta based on an individual voxel \( p = .001 \), a corrected \( p < .05 \), spatial autocorrelation of \( 4.54 \text{ mm} \), original voxel dimensions of \( 4 \text{ mm} \times 4 \text{ mm} \times 4 \text{ mm} \), and resampled voxel dimensions of \( 3 \text{ mm} \times 3 \text{ mm} \times 3 \text{ mm} \) (these parameters matched the simulation conducted in Slotnick, this issue). Matrix size \((x, y, z)\) was varied linearly in 10 steps from \((64, 64, 33)\), corresponding to whole-brain acquisition, to \((16, 16, 12)\). Figure 2 shows the cluster

![Figure 2. Cluster extent threshold as a function of matrix surface area to volume ratio. For each matrix size, which varied linearly from \((64, 64, 33)\) to \((16, 16, 12)\), cluster extent was determined by running 10,000 Monte Carlo simulations using cluster_threshold_beta based on an individual voxel \( p = .001 \), a corrected \( p < .05 \), spatial autocorrelation of \( 4.54 \text{ mm} \), original voxel dimensions of \( 4 \text{ mm} \times 4 \text{ mm} \times 4 \text{ mm} \), and resampled voxel dimensions of \( 3 \text{ mm} \times 3 \text{ mm} \times 3 \text{ mm} \).](image-url)
extent threshold (i.e., number of resampled voxels) as a function of matrix surface area to volume ratio. As expected, larger matrices were associated with larger cluster extent thresholds and smaller matrices were associated with smaller cluster extent thresholds, as larger clusters of activity will occur less often in smaller volumes. Of particular relevance, the matrix surface area to volume ratio, which is a measure of matrix edge size, varied widely while the cluster threshold remained unchanged. For instance, the surface area to volume ratio increased by over 50% while the cluster extent threshold remained unchanged at eight voxels. These results indicate that cluster_threshold_beta extent thresholds are not influenced by edge effects and that the purported bug does not exist.

It is also notable that the present simulation with a whole-brain matrix size (64, 64, 33) resulted in a cluster extent threshold of ten 3 mm × 3 mm × 3 mm resampled voxels (i.e., 270 mm³). By comparison, Eklund et al. employed a cluster extent threshold of ten 2 mm × 2 mm × 2 mm resampled voxels (i.e., 80 mm³) in their ad hoc SPM and FSL analysis, which yielded false-positive rates of 60–90% (see their Figure 2). As Eklund et al.‘s cluster extent was less than a third of what would be required to correct for multiple comparisons, it is not surprising that their reported false-positive rates were so high.

Commonly employed methods to correct for multiple comparisons are valid

As outlined in the present paper, there are many problems with Nichols et al. (this issue) and Eklund et al. (2016), including the following: (1) Nichols et al. did not provide convincing arguments that resting-state fMRI data reflect null data. (2) Eklund et al. presented one-sample t-test results in the main body of their paper showing that their permutation method was acceptable, while their supplementary results showed that this method produced false-positive rates that were similar to other methods (see Figure 1(a)). (3) Eklund et al. used the same event protocol for all the participants, which artifactualy inflated the one-sample t-test false-positive rates. (4) At p < .001, using two-sample t-tests (which corrected for the flawed analysis), all the methods employed to correct for multiple comparisons had acceptable false-positive rates (see Figure 1(b)). (5) Eklund et al. contrasted resting-state periods, which produced many significant clusters of activity, while null data should arguably be associated with few, if any, significant activations.

It is particularly surprising that even though Eklund et al. employed resting-state/default network activity as null data, which would be expected to substantially inflate false-positive rates, their results showed that all methods employed to correct for multiple comparisons had acceptable false-positive rates.

Type II error should not be ignored in fMRI analysis

To publish an fMRI study, the method employed to correct for multiple comparisons must limit the rate of false-positives (i.e., type I errors) to p < .05. Eklund et al. (2016) claimed that this type I error rate was not being achieved using commonly employed methods, and they showed no concern about type II errors in which true activations are not reported due to an overly stringent threshold. The field of researchers that employ fMRI has been moving toward an ever-increasing emphasis on limiting type I error by employing stricter thresholds and increasing sample sizes. As stated in a recent review paper (Poldrack et al., 2017), with regard to advances in fMRI acquisition and analysis, ‘these advances promise to offer important insights into the workings of the human brain but also generate the potential for a “perfect storm” of irreproducible results. In particular, the high dimensionality of fMRI data, the relatively low power of most fMRI studies and the great amount of flexibility in data analysis contribute to a potentially high degree of false-positive findings’ (p. 115). The problem with focusing on only decreasing type I error, which is the majority view in the field (as illustrated by Eklund et al., 2016, and Poldrack et al., 2017), is that employing stricter statistical thresholds necessarily increases type II error.

In his commentary, Hopfinger (this issue) pointed out several problems that occur when a strict
statistical threshold is employed to limit type I error while ignoring type II error. Hopfinger highlighted that even at a less stringent threshold, real findings will replicate across labs, that there is a danger in assuming only one analysis technique can uncover the truth (as all methods have limitations), and that if studies with only large samples are employed this will slow scientific progress and negatively impact innovation. Making some of the same points, in an excellent paper that considered the sole focus on type I error in fMRI analysis, Lieberman and Cunningham (2009) ‘consider four negative consequences: (i) increased type II errors, (ii) a bias toward publishing large and obvious effects, (iii) a bias against observing effects associated with complex cognitive and affective processes, and (iv) deficient meta-analyses’ (p. 424). Cunningham and Koscik (this issue) also highlighted that it is important to balance type I and type II errors, and introduced a new method that increases power for regions of interest while maintaining an acceptable false-positive rate. These commentaries make it clear that type II error should not be ignored.

In a related issue, the focus in the field on deceasing type I error has produced a systematic increase in sample size over time. From 2005 to 2015, the median sample size increased from about 13 participants to 28.5 participants, while the standardized effect sizes have improved by about only 25% (Poldrack et al., 2017). Is it better to run two smaller studies with moderate effect sizes or one large study with a slightly improved effect size? If type II error, replication, and innovation are ignored, larger and larger studies with stricter and stricter thresholds will be favored. If type I error and type II error are balanced, it is not so clear that larger studies with stricter thresholds are necessarily better. This issue warrants further consideration.

**Disclosure statement**

The author reports no conflicts of interest.

**ORCID**

Scott D. Slotnick [http://orcid.org/0000-0003-2948-6174](http://orcid.org/0000-0003-2948-6174)

**References**


Slotnick, S. D. (n.d.). Scripts and Stimuli. Retrieved from [https://www2.bc.edu/sd-slotnick/scripts.htm](https://www2.bc.edu/sd-slotnick/scripts.htm)