Reevaluating "Cluster Failure" Using Nonparametric Control of the False Discovery Rate

Daniel Kessler[a,1,2], **Michael Angstadt**[a,1], **and Chandra Sripada**[a,1]

[a]Department of Psychiatry, University of Michigan, Ann Arbor

This manuscript was compiled on October 18, 2016

In a substantial contribution to the functional magnetic resonance imaging (fMRI) field, Eklund et al. [1] use nonparametric methods to demonstrate that random field theory (RFT)-based family-wise error (FWE) correction techniques for cluster-level inference do not control errors as they are supposed to, and this discrepancy is more pronounced for lenient cluster defining thresholds (CDT). Moreover, they point to violations of RFT assumptions as the culprit for this discrepancy.

Given these results, what advice can we offer to the reader exploring existing fMRI literature when faced with a table of cluster-wise RFT-based FWE-corrected p-values ($p_{\text{RFT-FWE}}$)? To suggest caution is reasonable but incomplete; we require concrete, quantitative guidelines to enable appropriate calibration of skepticism.

Here, we undertake an initial attempt to construct such guidance. We heed Eklund et al.'s warning and prefer null distributions obtained through nonparametric methods rather than RFT. However, we focus on the False Discovery Rate (FDR; [2]), which is a more natural target for multiple testing control (as recognized by Nichols in previous work; [3]): A researcher is naturally more concerned with the proportion of reported clusters that are false positives (FDR) than whether *any* are false positives (FWE). Given these considerations, a reader faced with a table of clusters significant under RFT-FWE correction might ask: Which of these results would have survived had the study instead employed a nonparametric FDR-based method?

We address this question using task fMRI data [4, 5] analyzed by Eklund et al. (available from openfMRI [6]).

For each contrast, we generate 5,000 realizations of the data through sign-flipping (our code, data, Extended Methods: http://github.com/mangstad/FDR_permutations). To obtain a null distribution of cluster extents (for an arbitrarily chosen cluster), we combine normalized frequencies of extents at each realization. This distribution is used to assign uncorrected p-values to each observed cluster. We next submit the vector of uncorrected p-values for each contrast to Benjamini and Hochberg's [2] FDR procedure with $\alpha_{\text{FDR}} = .05$ (cf. [7] for a parametric implementation of cluster-wise FDR).

We compare $p_{\text{RFT-FWE}}$-values to q_{FDR}-values and note whether they survive FDR-correction under $\alpha_{\text{FDR}} = .05$. We generate separate plots for this analysis conducted at CDT=\{.001, .01\}.

Based on our results (see Figure 1), we can suggest that nearly all clusters identified as significant when using CDT=.001 and RFT-FWE correction are trustworthy by the nonparametric FDR benchmark. For clusters identified as significant with CDT=.01 and RFT-FWE correction, the guidance depends on the corrected p-value: clusters with corrected-p less than .00001 appear consistently trusthworthy by the nonparametric FDR benchmark, whereas clusters with corrected-p greater than this value are not reliabily trustworthy. These findings have promising implications for past fMRI studies using RFT-based cluster-level inference that used this stricter CDT, estimated to be upwards of 8,500 reports [8, 9]. While the story is more mixed for CDT=.01 (used in approximately 3,500 studies [8, 9]), our findings suggest that not all clusters reported in the relevant studies are unreliable: we identify .00001 as a potential cutoff for trustworthiness.

These results offer initial guidance on interpreting past literature that employed RFT-based FWE, providing a more granular appreciation of the relationship between $p_{\text{RFT-FWE}}$ and trustworthiness of the result. A more comprehensive examination of fMRI task data sets that used RFT-based FWE may further refine this guidance.

ACKNOWLEDGMENTS. We thank Anders Eklund and Thomas Nichols for providing us with processed data and for very helpful comments on earlier versions of this letter.

1. Eklund A, Nichols TE, Knutsson H (2016) Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences* p. 201602413.
2. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1):289–300.
3. Genovese CR, Lazar NA, Nichols T (2002) Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate. *NeuroImage* 15(4):870–878.
4. Duncan KJ, Pattamadilok C, Knierim I, Devlin JT (2009) Consistency and variability in functional localisers. *NeuroImage* 46(4):1018–1026.
5. Tom SM, Fox CR, Trepel C, Poldrack RA (2007) The Neural Basis of Loss Aversion in Decision-Making Under Risk. *Science* 315(5811):515–518.
6. Poldrack RA et al. (2013) Toward open sharing of task-based fMRI data: the OpenfMRI project. *Frontiers in Neuroinformatics* 7:12.
7. Chumbley JR, Friston KJ (2009) False discovery rate revisited: FDR and topological inference using Gaussian random fields. *NeuroImage* 44(1):62–70.
8. Nichols TE (2016) Bibliometrics of Cluster Inference, 06/07/16, Neuroimaging Statistics Tips & Tools. http://blogs.warwick.ac.uk/nichols/entry/bibliometrics_of_cluster/.
9. Woo CW, Krishnan A, Wager TD (2014) Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage* 91:412–419.

DK, MA, and CS planned and executed the analysis and wrote the letter.

The authors have no conflicts of interest to declare.

[1]All authors contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: kesslerd@umich.edu

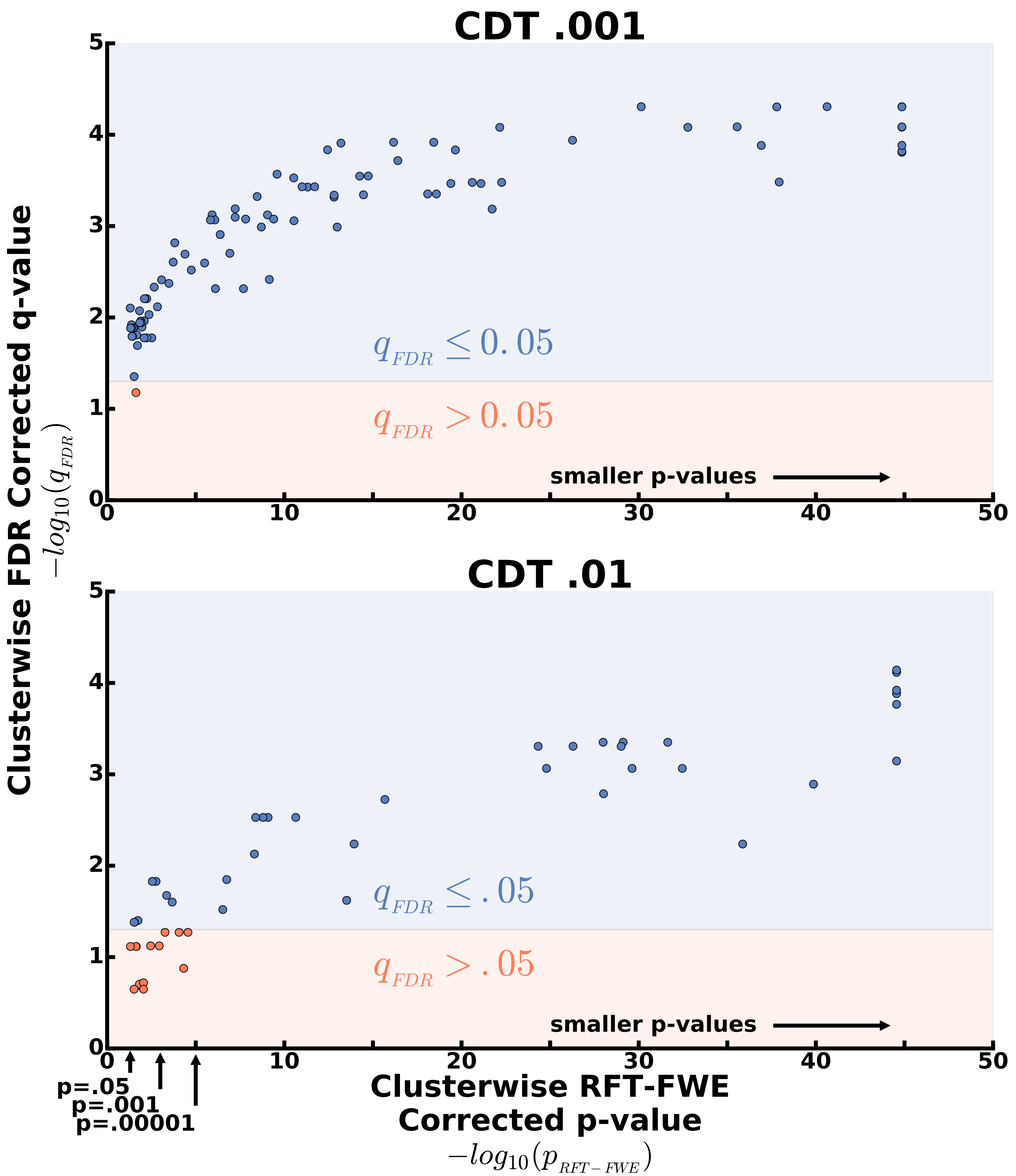

Fig. 1. Assessing RFT-Based FWE Using an FDR Benchmark. We submitted the same task data analyzed by Eklund et al. [1, 5, 6] to nonparametric clusterwise FDR analysis. For CDT $= .001$ (top), RFT-based FWE approximates effective FDR control with $\alpha_{\mathrm{FDR}} = .05$. For CDT $= .01$ (bottom), only clusters with $p_{\mathrm{RFT\text{-}FWE}} \leq .00001$ reliably survived correction at $\alpha_{\mathrm{FDR}} = .05$.