

REPLY TO BROWN AND BEHRMANN, COX ET AL., AND KESSLER ET AL.: Data and code sharing is the way forward for fMRI

Anders Eklund^{a,b,c,1}, Thomas E. Nichols^{d,e}, and Hans Knutsson^{a,c}

We are glad that our paper (1) has generated intense discussions in the fMRI field (2–4), on how to analyze fMRI data, and how to correct for multiple comparisons. The goal of the paper was not to disparage any specific fMRI software, but to point out that parametric statistical methods are based on a number of assumptions that are not always valid for fMRI data, and that nonparametric statistical methods (5) are a good alternative. Through AFNI's introduction of nonparametric statistics in the function 3dttest++ (3, 6), the three most common fMRI softwares now all support nonparametric group inference [SPM through the toolbox SnPM (www2.warwick.ac.uk/fac/sci/statistics/staff/academic-research/nichols/software/snpm), and FSL through the function randomise].

Cox et al. (3) correctly point out that the bug in the AFNI function 3dClustSim only had a minor impact on the false-positive rate (FPR). This was also covered in our original paper (1): "We note that FWE [familywise error] rates are lower with the bug-fixed 3dClustSim function. As an example, the updated function reduces the degree of false positives from 31.0% to 27.1% for a CDT [cluster-defining threshold] of $P = 0.01$, and from 11.5% to 8.6% for a CDT of $P = 0.001$." It is unfortunate that several media outlets focused extensively on this bug, when the main problem was found to be violations of the assumptions in the statistical models.

The statement that AFNI had particularly high FPRs, compared with SPM and FSL, is for example supported by figure S1A in our original paper (1) (Beijing data, two-sample t test with 20 subjects, CDT $P = 0.01$). For 8-mm smoothing, the FPR for AFNI is 23–31%, whereas it is 13–20% for SPM and 14–18% for FSL OLS. To understand the higher FPRs, we investigated how the 3dClustSim function works, which eventually led us to finding the bug in 3dClustSim. However, we agree that AFNI did not produce higher FPRs for all parameter combinations.

The 70% FPR comes from figure S9C in our original report (1) (Oulu data, one-sample t test with 40 subjects, CDT $P = 0.01$, FSL OLS with 4-mm smoothing) and not, as some readers believed, from figure 2 in the original paper (1), which shows results for the ad hoc clustering

approach. The main reason for using the highest observed FPR was to give the reader an idea of how severe the problem can be, but we agree that it led to a too pessimistic view.

As pointed out by Cox et al. (3), the nonparametric approach also performed suboptimal for the one-sample t test, especially for the Oulu data. As discussed in our paper (1), the one-sample t test has an assumption of symmetrically distributed errors that can be violated by outliers in small samples. Our current research is therefore focused on how to improve the nonparametric test for one-sample t tests. Regarding the flexibility of the permutation testing, recent work has shown that virtually any regression model with independent errors can be accommodated (5), and even longitudinal and repeated-measures data can be analyzed with a related bootstrap approach (7).

Kessler et al. (4) extend our evaluations to (non-parametric) cluster-based false-discovery rate (FDR) on-task data, to better understand how existing parametric cluster P values based on the FWE should be interpreted. For the problematic CDT of $P = 0.01$, Kessler et al. conclude that a cluster FWE-corrected P value smaller than $P = 0.00001$ survives FDR correction at $q = 0.05$. Indeed, this information makes it easier to interpret existing results in the fMRI literature, but it should be noted that it is not straightforward to generalize these results to other studies. For example, the fMRI software used, the MR sequence used (EPI or multiband), the degree of smoothing, and the number of subjects are all likely to affect this cut-off. The only way to retrospectively evaluate existing results is, in our opinion, to reanalyze the original fMRI data [e.g., made available through OpenfMRI (8)] or to apply a new threshold to the statistical maps [e.g., made available through NeuroVault (9)].

Finally, we would like to note the importance of data and code sharing. Cox et al. (3, 6) replicated and extended our findings with the same open fMRI data (10) as in our original paper (1) (and made use of our processing scripts available on github, <https://github.com/wanderine/ParametricMultisubjectfMRI>), ultimately resulting in

^aDivision of Medical Informatics, Department of Biomedical Engineering, Linköping University, S-581 85 Linköping, Sweden; ^bDivision of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University, S-581 83 Linköping, Sweden; ^cCenter for Medical Image Science and Visualization, Linköping University, S-581 83 Linköping, Sweden; ^dDepartment of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom; and ^eWMG, University of Warwick, Coventry CV4 7AL, United Kingdom

Author contributions: A.E., T.E.N., and H.K. wrote the paper.

The authors declare no conflict of interest.

¹To whom correspondence should be addressed. Email: anders eklund@liu.se.

