Controversy in statistical analysis of functional magnetic resonance imaging data

Emery N. Brown[a,b,1] and Marlene Behrmann[c]

To test the validity of statistical methods for fMRI data analysis, Eklund et al. (1) used, for the first time, large-scale experimental data rather than simulated data. Using resting-state fMRI measurements to represent a null hypothesis of no task-induced activation, the authors compare familywise error rates for voxel-based and cluster-based inferences for both parametric and nonparametric methods. Eklund et al.'s study used three fMRI statistical analysis packages. They found that, for a target familywise error rate of 5%, the parametric methods gave invalid cluster-based inferences and conservative voxel-based inferences.

Eklund et al. (1) attribute the invalid cluster-based inferences to the incorrect assumption of squared exponential structure in the spatial autocorrelation function of the parametric models. The authors suggest nonparametric methods as a more appropriate way to achieve targeted error rates, and conclude that statistical methods for fMRI data analysis should be validated. In addition, Eklund et al. state that their findings "question the validity of some 40,000 fMRI studies and may have a large impact on the interpretation of neuroimaging results" (1). This sentence from the Significance section of the original paper was picked up by the press and yielded the alarming negative headline that fMRI analyses produce incorrect results because of a bug in a widely used data analysis package (2–4). Eklund et al. revised their extrapolation regarding the implication of their findings in a correction to their article (5) and report that their analysis might apply to 3,500 rather than 40,000 fMRI studies (6). However, before this revision was published, the original statements created considerable debate about data analysis and the accuracy of fMRI findings (2–4).

The overstatements of the original paper and the subsequent media attention cast doubt on fMRI as a technique for studying brain function, and possibly even caused damage to the field of cognitive neuroscience (2–4). In PNAS, Cox et al. (7) and Kessler et al. (8) offer clarifications about the original paper and its revision. Eklund et al. have added their rejoinder (9). Several scientific points have now been mostly resolved. The remaining question is: What else can be learned from this controversy?

fMRI is a highly valued methodology for understanding brain function and its relationship to behavior. During the last 25 y, significant scientific advances have been made using this technique. To ensure continued progress, fMRI experimentalists want to be assured that the instruments, experimental protocols, and data analysis paradigms have been vetted by experts and work correctly. At the same time, experimentalists must be well informed about the fMRI process, and have a solid understanding of how to apply and interpret commonly used statistical methods (10–12). The ease of analysis afforded by some of the software programs belies the complexity of the methods. This ease of use does not release experimentalists from their responsibility to validate findings using established statistical principles (12, 13). Judicious use of nonparametric methods can, as Eklund et al. (1) suggest, improve the current analysis paradigm in certain cases. However, application of nonparametric methods cannot be the universal solution, nor did Eklund et al. suggest that it could be.

The current discussion shows that the validity of fMRI data analysis paradigms has not been uniformly established and needs continued in-depth investigation. fMRI is a complex process that involves biophysics, neuroananatomy, neurophysiology, and statistics (experimental design, statistical modeling, and data analysis). fMRI data have a low signal-to-noise ratio (14, 15). As a consequence, all of the biophysics, neurophysiology, and neuroanatomy that underlie fMRI should be used to design experiments, formulate statistical models, and analyze the data to increase the signal-to-noise ratio and information extraction. Achieving more accurate fMRI data analyses is a challenging interdisciplinary task that requires concerted collaborations among physicists, statisticians, and neuroscientists who, together, can question the current approaches more deeply and construct more accurate analysis methods.

In an ideal fMRI statistical analysis, the relationships among the voxels would take account of the spatial and temporal properties of the experiment and the scanner

[a]Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114; [b]Department of Brain and Cognitive Sciences, Institute for Medical Engineering and Science and Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, MA 02139; and [c]Department of Psychology and Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA 15213

Author contributions: E.N.B. and M.B. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. Email: enb@neurostat.mit.edu.

www.pnas.org/cgi/doi/10.1073/pnas.1705513114

thermal noise (16). The experiment's spatial and temporal properties are dictated by the physiological changes (neural activity, blood flow, and blood oxygenation levels) induced by the particular behavioral task and background physiological activity and anatomy (white matter, gray matter, the ventricles, and blood vessels) of the relevant brain regions. The ideal fMRI acquisition scheme would be accompanied by a characterization of these spatial and temporal processes so that the subsequent data analysis can correctly take them into account (16). Improving fMRI statistical methods must combine research to decipher the meaning/origins of the blood oxygen level-dependent signal with characterizations of the spatio-temporal properties of task-related activity, background physiological activity, and scanner properties. Sharing data and methods would greatly expedite validation (9).

BRAIN 2025, the report of the NIH Brain Initiative, recommends fostering interdisciplinary collaborations among neuroscientists, physicists, engineers, statisticians, and mathematicians to properly collect, analyze, and interpret the data that result from the development of new neuroscience tools (https://www.braininitiative.nih.gov/2025/). The current exchange identifies fMRI as an existing tool that is perfect for pursuing such a collaboration. A possible goal could be to increase fMRI signal-to-noise ratios so that the technique can be used reliably to make inferences about an individual subject in a given paradigm.

Developing statistical methods based on detailed modeling of the fMRI process opens the door to using more direct, informative inference paradigms based on estimated effect sizes, confidence intervals, and Bayesian posterior assessments rather than more indirect approaches based on significance tests and *P* values. Linking statistical methodology development and fundamental fMRI research is crucial for developing more accurate analysis methods, attributing accurate scientific interpretations to results, and ensuring the reliability and reproducibility of fMRI studies. These points have been made before. However, their significance has perhaps not been considered to the extent required.

Acknowledgments

We thank Bruce Rosen, Michael Tarr, and Larry Wald for helpful comments.

1 Eklund A, Nichols TE, Knutsson H (2016) Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci USA* 113:7900–7905. Erratum in *Proc Natl Acad Sci USA* 113:E4929.
2 Crew B (2016) A bug in fMRI software could invalidate 15 years of brain research. Available at www.sciencealert.com/a-bug-in-fmri-software-could-invalidate-decades-of-brain-research-scientists-discover. Accessed April 10, 2017.
3 Murphy K (2016) Do you believe in god or is that a software glitch? Available at https://www.nytimes.com/2016/08/28/opinion/sunday/do-you-believe-in-god-or-is-that-a-software-glitch.html?_r=0. Accessed April 10, 2017.
4 Biello D (2016) Much of what we know about the brain may be wrong: The problem with fMRI. Available at ideas.ted.com/much-of-what-we-know-about-the-brain-may-be-wrong-the-problem-with-fmri/. Accessed April 10, 2017.
5 Eklund A, Nichols TE, Knutsson H (2016) Correction for Eklund et al. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci USA* 113:E4929.
6 Nichols T (2016) Bibliometrics of cluster inference. Available at blogs.warwick.ac.uk/nichols/entry/bibliometrics_of_cluster/. Accessed April 10, 2017.
7 Cox RW, Chen G, Glen DR, Reynolds RC, Taylor PA (2017) FMRI clustering and false positive rates. *Proc Natl Acad Sci USA* 114:E3370–E3371.
8 Kessler D, Angstadt M, Sripada C (2017) Reevaluating "cluster failure" using nonparametric control of false discovery rate. *Proc Natl Acad Sci USA* 114:E3372–E3373.
9 Eklund A, Nichols TE, Knutsson H (2017) Reply to Cox et al. and Kessler et al.: Data and code sharing is the way forward for fMRI. *Proc Natl Acad Sci USA* 114:E3374–E3375.
10 Poldrack RA, et al. (2017) Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nat Rev Neurosci* 18:115–126.
11 Munafo MR, et al. (2017) A manifesto for reproducible science. *Nature Human Behaviour* 1:0021.
12 Kass RE, et al. (2016) Ten simple rules for effective statistical practice. *PLOS Comput Biol* 12:e1004961.
13 Kass RE, Ventura V, Brown EN (2005) Statistical issues in the analysis of neuronal data. *J Neurophysiol* 94:8–25.
14 Parrish TB, Gitelman DR, LaBar KS, Mesulam MM (2000) Impact of signal-to-noise on functional MRI. *Magn Reson Med* 44:925–932.
15 Welvaert M, Rosseel Y (2013) On the definition of signal-to-noise ratio and contrast-to-noise ratio for FMRI data. *PLoS One* 8:e77089.
16 Wald LL, Rosseel Y (2017) Impacting the effect of fMRI noise through hardware and acquisition choices-impliations for controlling false positive rates. *Neuroimage*, 10.1016/j.neuroimage.2016.12.057.

fMRI clustering and false-positive rates

Robert W. Cox[a,1], Gang Chen[a], Daniel R. Glen[a], Richard C. Reynolds[a], and Paul A. Taylor[a]

Recently, Eklund et al. (1) analyzed clustering methods in standard fMRI packages: AFNI (which we maintain), FSL, and SPM. They claim that (*i*) false-positive rates (FPRs) in traditional approaches are greatly inflated, questioning the validity of "countless published fMRI studies"; (*ii*) nonparametric methods produce valid, but slightly conservative, FPRs; (*iii*) a common flawed assumption is that the spatial autocorrelation function (ACF) of fMRI noise is Gaussian-shaped; and (*iv*) a 15-y-old bug in AFNI's 3dClustSim significantly contributed to producing "particularly high" FPRs compared with other software. We repeated simulations from ref. 1 [Beijing_Zang data (2), cf. ref. 3] and comment on each point briefly.

AFNI and 3dClustSim

Fig. 1 *A*–*D* compares results of the "buggy" and "fixed" 3dClustSim. For each simulation, the typical difference was small: $\Delta\text{FPR} \lesssim 3-5\%$ at per-voxel $P=0.01$ and $\lesssim 1-2\%$ for $P=0.001$. The bug had only a minor impact.

Figures 1 and 2 of ref. 1 actually show similar FPRs for AFNI, FSL-OLS, and SPM: Most tests were in a range of $20-40\%$ FPR at $P=0.01$ and $5-15\%$ FPR at $P=0.001$ (nor did their famous 70% FPR come from AFNI). The data given in the Results section of ref. 1 simply do not support the statement in the Discussion section that AFNI had "particularly high" FPRs.

Smoothness

To test the effect of assuming a Gaussian ACF in fMRI noise, an empirical "mixed ACF" allowing for longer tails was computed from residuals (3). All FPRs (Fig. 1 *E* and *F*) decreased. Block designs remained >5%, likely reflecting dependence of the noise's spatial smoothness on temporal frequency. Heavy tails in spatial smoothness indeed have significant consequences for clustering.

Nonparametric Approach

A spatial model-free, nonparametric randomization approach was added to AFNI's group-level GLM program, 3dttest++ (3). All FPRs (Fig. 1 *G* and *H*) were within the nominal confidence interval. Although this approach shows promise (as in ref. 1), it may not be feasible to generalize nonparametric permutations to complicated covariate structures and models (e.g., complex ANOVA, analysis of covariance, or linear mixed effects) (4, 5).

Inflated FPRs

Several cases showed significant FPR inflation across existing fMRI software within the testing framework of ref. 1. However, deviations from nominal FPR were not uniformly large and depended strongly on several factors. Fig. 1 and figure 1 of ref. 1 show quite good cluster results for stricter per-voxel *P* values (which ref. 6 found to be predominantly used in fMRI analyses) and for event-related stimuli (emphasizing the importance of good experimental design): FPR inflation was often $\lesssim 10\%$ (Beijing) or $\lesssim 5\%$ (Cambridge), affecting only clusters with marginally significant volume.

We strongly disagree with Eklund et al.'s (1) summary statement: "Alarmingly, the parametric methods can give a very high degree of false positives (up to 70%, compared with the nominal 5%)." For comparison, their own nonparametric method's results actually showed up to 40% FPR. When characterizing results, medians or percentile ranges are generally more informative summary statistics than maxima. Looking backward, the typical ranges show much smaller FPR inflation than what had been highlighted, and looking forward they provide useful suggestions for experimental design and analyses (lower voxelwise *P*, event-related paradigms, etc.). By concentrating on the highest observed FPRs, the conclusions of Eklund et al. (1) are unnecessarily alarmist.

Acknowledgments

This work was supported by the National Institute of Mental Health and National Institute of Neurological Disorders and Stroke Intramural Research Programs ZICMH002888 of the NIH, US Department of Health and Human Services. This work used the computational resources of the NIH High-Performance Computing Biowulf cluster (https://hpc.nih.gov/).

[a]Scientific and Statistical Computing Core, National Institute of Mental Health, US Department of Health and Human Services, National Institutes of Health, Bethesda, MD 20892

Author contributions: R.W.C. designed research; R.W.C. performed research; R.W.C. contributed new reagents/analytic tools; R.W.C. and R.C.R. analyzed data; and R.W.C., G.C., D.R.G., R.C.R., and P.A.T. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. Email: robertcox@mail.nih.gov.

www.pnas.org/cgi/doi/10.1073/pnas.1614961114

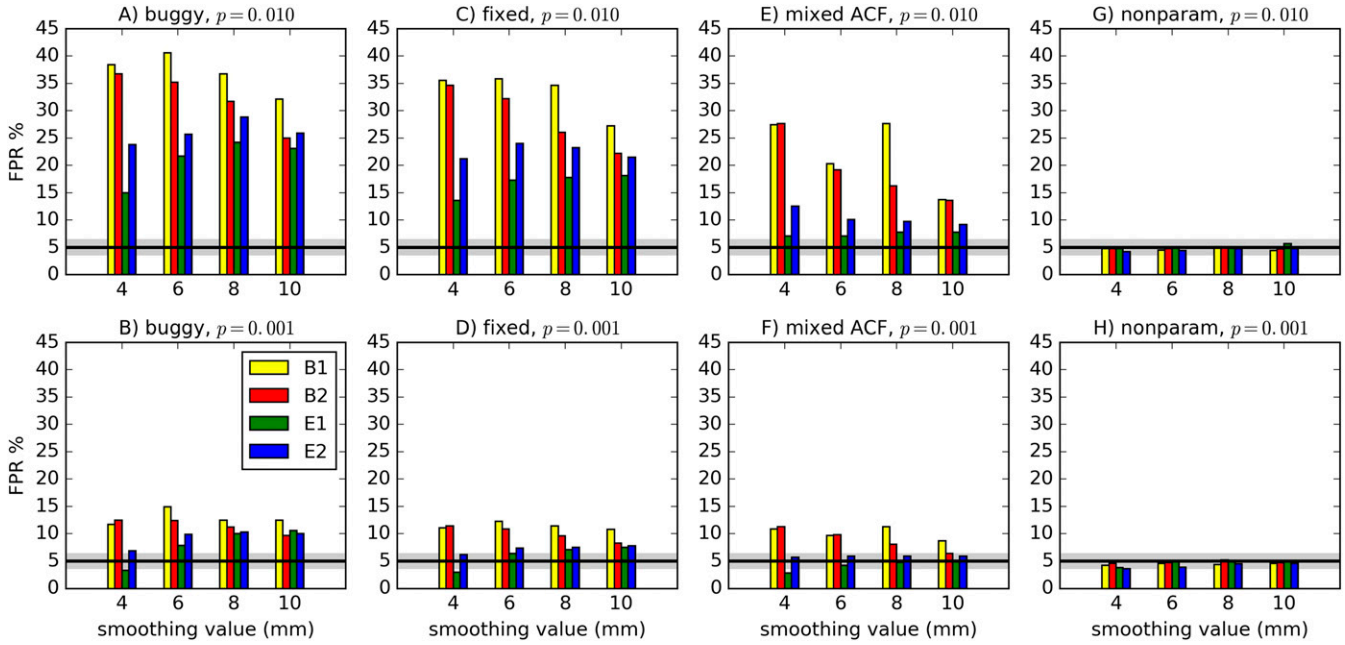

Fig. 1. FPRs for various software scenarios, with 1,000 two-sample one-sided *t*-tests (as in ref. 1; cf. ref. 3 for more details) using 20 subjects' data in each sample. For "buggy" (*A* and *B*) and "fixed" (*C* and *D*), cluster-size thresholds were selected using the Gaussian shape model with the FWHM being the median of the 40 individual subjects' values: "buggy" via 3dClustSim before the bug fix, "fixed" via 3dClustSim after the bug fix. For "mixed ACF" (*E* and *F*), the cluster-size threshold was selected using a non-Gaussian ACF model allowing for heavy tails (3). For "nonparam" (*G* and *H*), 3dttest++ was used to perform spatial model-free, nonparametric permutation testing (3); paired, two-sided, and tests with covariates gave similar results. Two different per-voxel *P*-value thresholds are shown. The black line shows the nominal 5% FPR (out of 1,000 trials), and the gray band shows its binomial 95% confidence interval, 3.65–6.35%. As in ref. 1, different smoothing values were tested (4–10 mm), and four test designs were used: B1 = 10-s block; B2 = 30-s block; E1 = regular event-related; E2 = randomized event-related.

1 Eklund A, Nichols TE, Knutsson H (2016) Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci USA* 113:7900–7905. Erratum in *Proc Natl Acad Sci USA* 113:E4929.

2 Biswal BB, et al. (2010) Toward discovery science of human brain function. *Proc Natl Acad Sci USA* 107:4734–4739.

3 Cox RW, Reynolds RC, Taylor PA (2016) AFNI and clustering: False positive rates redux. *bioRxiv* 065862, 10.1101/065862.

4 Chen G, Adleman NE, Saad ZS, Leibenluft E, Cox RW (2014) Applications of multivariate modeling to neuroimaging group analysis: A comprehensive alternative to univariate general linear model. *Neuroimage* 99:571–588.

5 Chen G, Saad ZS, Britton JC, Pine DS, Cox RW (2013) Linear mixed-effects modeling approach to FMRI group analysis. *Neuroimage* 73:176–190.

6 Carp J (2012) The secret lives of experiments: Methods reporting in the fMRI literature. *Neuroimage* 63:289–300.

Reevaluating "cluster failure" in fMRI using nonparametric control of the false discovery rate

Daniel Kessler[a,1,2], Mike Angstadt[a,1], and Chandra S. Sripada[a,1]

In a substantial contribution to the fMRI field, Eklund et al. (1) use nonparametric methods to demonstrate that random field theory (RFT)-based familywise error (FWE) correction for cluster inference does not control errors appropriately, 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, how should we interpret existing fMRI literature that used RFT-based, FWE-corrected *P* values ($p_{RFT\text{-}FWE}$)? To suggest caution is reasonable but incomplete; we require concrete, quantitative guidelines to enable appropriate calibration of skepticism.

Here, we undertake an initial attempt at such guidance. We heed Eklund et al.'s (1) warning and prefer nonparametric null distributions to 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 and coworkers 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). Thus, a reader considering a table of clusters significant under RFT–FWE might ask which of these results would have survived had the study instead used a nonparametric FDR-based method.

We address this question using the same task fMRI data (4, 5) analyzed by Eklund et al. (1) (available from openfMRI, ref. 6).

For each contrast, we generate 5,000 realizations of the data through sign flipping (code, data, and extended methods: https://github.com/mangstad/FDR_permutations). To obtain a null distribution of cluster extents (for an arbitrary 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 $\alpha_{FDR} = .05$ (cf. ref. 7 for a parametric implementation of clusterwise FDR).

We compare $p_{RFT\text{-}FWE}$ values to q_{FDR} values and note whether they survive FDR correction under $\alpha_{FDR} = .05$. We generate separate plots for this analysis conducted at CDT = {0.001, 0.01}.

Based on our results (Fig. 1), we suggest nearly all clusters identified as significant when using CDT = 0.001 and RFT–FWE correction are trustworthy by the nonparametric FDR benchmark. For clusters identified as significant with CDT = 0.01 and RFT–FWE correction, the guidance depends on the corrected *P* value: Clusters with $p_{RFT\text{-}FWE} < .00001$ seem consistently trustworthy by the nonparametric FDR benchmark, whereas clusters above this value are not reliably trustworthy.

These findings have promising implications for past fMRI studies using RFT-based cluster-level inference that used CDT = 0.001, estimated to be upward of 8,500 reports (8, 9). Although the story is mixed for CDT = 0.01 (used in ~3,500 studies) (8, 9), our findings suggest that not all such previously reported clusters are unreliable. We identify 0.00001 as a potential cutoff for trustworthiness.

Our results provide more granular guidance on the relationship between $p_{RFT\text{-}FWE}$ and trustworthiness of results. A more comprehensive examination of fMRI task datasets that used RFT-based FWE can 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.

[a]Department of Psychiatry, University of Michigan, Ann Arbor, MI 48109

Author contributions: D.K., M.A., and C.S.S. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

[1]D.K., M.A., and C.S.S. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: kesslerd@umich.edu.

www.pnas.org/cgi/doi/10.1073/pnas.1614502114

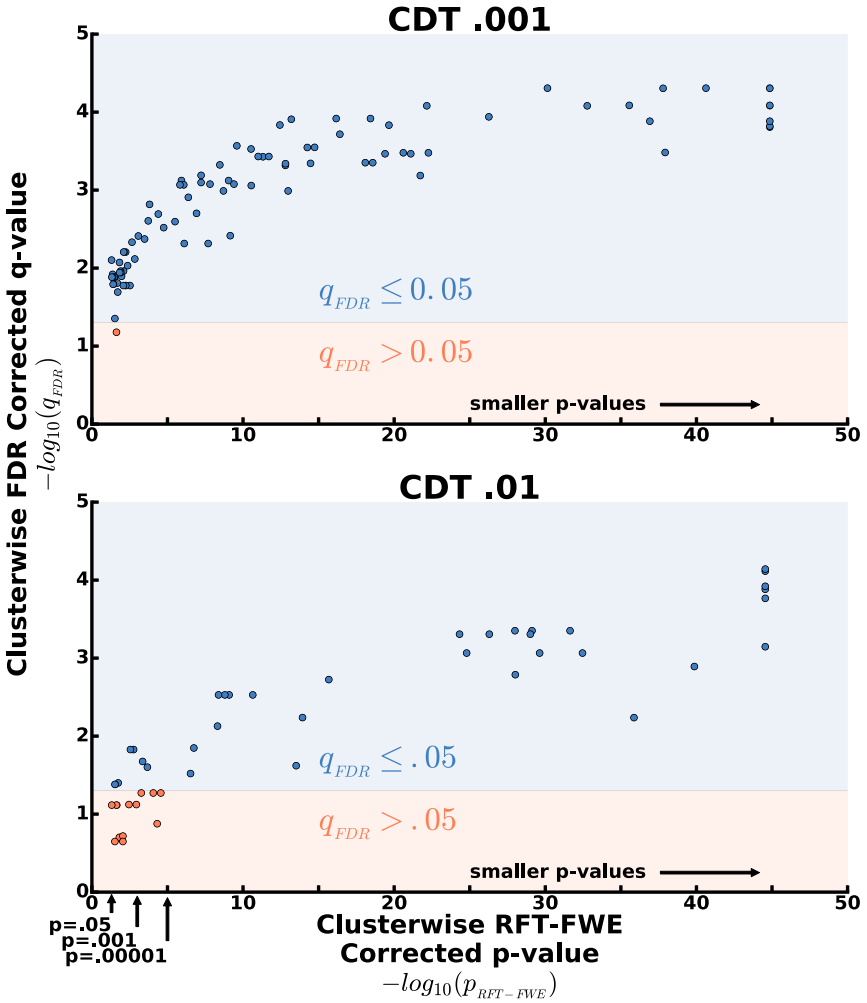

Fig. 1. Assessing RFT-based FWE using an FDR benchmark. We submitted the same task data analyzed by Eklund et al. (1, 5, 6) to nonparametric clusterwise FDR analysis. For CDT = .001 (*Top*), RFT-based FWE approximates effective FDR control with $\alpha_{FDR} = .05$. For CDT = .01 (*Bottom*), only clusters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

1 Eklund A, Nichols TE, Knutsson H (2016) Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci USA* 113:7900–7905. Erratum in *Proc Natl Acad Sci USA* 113:E4929.
2 Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc B* 57:289–300.
3 Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 15:870–878.
4 Duncan KJ, Pattamadilok C, Knierim I, Devlin JT (2009) Consistency and variability in functional localisers. *Neuroimage* 46:1018–1026.
5 Tom SM, Fox CR, Trepel C, Poldrack RA (2007) The neural basis of loss aversion in decision-making under risk. *Science* 315:515–518.
6 Poldrack RA, et al. (2013) Toward open sharing of task-based fMRI data: The OpenfMRI project. *Front Neuroinform* 7:12.
7 Chumbley JR, Friston KJ (2009) False discovery rate revisited: FDR and topological inference using Gaussian random fields. *Neuroimage* 44:62–70.
8 Nichols TE (2016) Bibliometrics of cluster inference. Available at 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.

LETTER

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 (nonparametric) 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

[a]Division of Medical Informatics, Department of Biomedical Engineering, Linköping University, S-581 85 Linköping, Sweden; [b]Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University, S-581 83 Linköping, Sweden; [c]Center for Medical Image Science and Visualization, Linköping University, S-581 83 Linköping, Sweden; [d]Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom; and [e]WMG, 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.

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

www.pnas.org/cgi/doi/10.1073/pnas.1620285114

improvements to the AFNI software. Furthermore, we never would have been able to identify the bug in 3dClustSim were AFNI not open-source software. Kessler et al. (4) also used the same task datasets from OpenfMRI (8) to find the empirical cluster FDR. Together, these examples show the importance of data sharing (11, 12), open-source software (13), code sharing (14, 15), and reproducibility (16).

Acknowledgments

This research was supported by the Neuroeconomic Research Initiative at Linköping University, by Swedish Research Council Grant 2013-5229 ("Statistical Analysis of fMRI Data"), the Information Technology for European Advancement 3 Project BENEFIT (better effectiveness and efficiency by measuring and modelling of interventional therapy), the Swedish Research Council Linnaeus Center CADICS (control, autonomy, and decision-making in complex systems), and the Wellcome Trust.

1 Eklund A, Nichols TE, Knutsson H (2016) Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci USA* 113:7900–7905. Erratum in *Proc Natl Acad Sci USA* 113:E4929.
2 Brown EN, Behrmann M (2017) Controversy in statistical analysis of functional magnetic resonance imaging data. *Proc Natl Acad Sci USA* 114:E3368–E3369.
3 Cox RW, Chen G, Glen RD, Reynolds RC, Taylor PA (2017) fMRI clustering and false positive rates. *Proc Natl Acad Sci USA* 114:E3370–E3371.
4 Kessler D, Angstadt M, Sripada CS (2017) Reevaluating "cluster failure" in fMRI using nonparametric control of the false discovery rate. *Proc Natl Acad Sci USA* 114:E3372–E3373.
5 Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE (2014) Permutation inference for the general linear model. *Neuroimage* 92:381–397.
6 Cox RW, Reynolds RC, Taylor PA (2016) AFNI and clustering: False positive rates redux. *bioRxiv*, 10.1101/065862.
7 Guillaume B, Hua X, Thompson PM, Waldorp L, Nichols TE; Alzheimer's Disease Neuroimaging Initiative (2014) Fast and accurate modelling of longitudinal and repeated measures neuroimaging data. *Neuroimage* 94:287–302.
8 Poldrack RA, et al. (2013) Toward open sharing of task-based fMRI data: the OpenfMRI project. *Front Neuroinform* 7:12.
9 Gorgolewski KJ, et al. (2016) NeuroVault.org: A repository for sharing unthresholded statistical maps, parcellations, and atlases of the human brain. *Neuroimage* 124:1242–1244.
10 Biswal BB, et al. (2010) Toward discovery science of human brain function. *Proc Natl Acad Sci USA* 107:4734–4739.
11 Poldrack RA, Gorgolewski KJ (2014) Making big data open: Data sharing in neuroimaging. *Nat Neurosci* 17:1510–1517.
12 Poline JB, et al. (2012) Data sharing in neuroimaging research. *Front Neuroinform* 6:9.
13 Ince DC, Hatton L, Graham-Cumming J (2012) The case for open computer programs. *Nature* 482:485–488.
14 Baker M (2016) Why scientists must share their research code. *Nature*, 10.1038/nature.2016.20504.
15 Eglen S, et al. (2016) Towards standard practices for sharing computer code and programs in neuroscience. *bioRxiv*, 10.1101/045104.
16 Gorgolewski KJ, et al. (2016) BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLoS Comput Biol* 13:e1005209.