

The cluster failure and the self-correction of neuroimaging research

Yuelu Liu*

Center for Mind and Brain, University of California, Davis, 267 Cousteau Pl, Davis, CA, USA

2016 is a turbulent year for neuroimaging research. A recent article by Anders Eklund, Thomas Nichols, and Hans Knutsson, published in the high-impact journal "Proceedings of the National Academy of Sciences" (PNAS), called into question the methods used in three highly popular software suites for analyzing functional magnetic resonance imaging (fMRI) data [1]. The specific issue discovered in the three fMRI processing suites (AFNI, FSL, and SPM) involves the way in which cluster-wise thresholding algorithms handle multiple comparison correction in fMRI activations, which can lead to inflated false positive rates (in some cases as high as 50% instead of the intended 5% nominal level). What seems even more alarming is that this issue lurked around and went undetected for many years since the introduction of cluster-wise multiple comparison correction methods and hence could potentially affect a great number of fMRI studies.

The article generated huge waves in the media. As can be readily seen from headlines such as "40,000 neuroscience papers on fMRI might be wrong" or "25 years of inflated fMRI false-positive rates", there is undoubtedly a great deal of disbelief in the general media toward fMRI research. Coupled with recent articles reporting generally low statistical power and replicability in psychological and neuroscience research [2,3], the work by Eklund *et al.* [4] surely looks like another devastating blow further shaking the tarnished credibility of fMRI research.

Despite the negativity in the public media, I would like to argue that the Eklund *et al.* [4] study instead represents a perfect example of the important self-correcting mechanism of science. It promotes constructive discussions within the neuroimaging community on the future of fMRI research. First, let us take a careful look at what really went wrong in the 3 software suites called into question. We know that task-based fMRI studies identify systematic differences between experimental conditions by statistically comparing the strength of the hemodynamic response measured in tens of thousands of voxels (a 3-D pixel) across the brain. Performing this many statistical tests requires multiple comparisons correction to properly control the family-wise error rate (FWE). The two most common approaches to control FWE are voxel-wise and cluster-wise thresholding techniques, which operate on the activation intensity at each specific voxel (hence voxel-wise) or on the extent of contiguous voxels (a cluster), respectively. The study by Eklund *et al.* [4] discovered that while voxel-wise thresholding is conservative, the different cluster-wise thresholding techniques implemented by the 3 fMRI software suites incorrectly model the spatial smoothness of fMRI data, therefore leading to invalid control of FWE. Specifically, it was found that the spatial autocorrelation function in fMRI data does not follow the squared exponential distribution assumed in random field theory. This discrepancy is more pronounced for larger clusters given that the empirical spatial autocorrelation

functions estimated from actual fMRI data had heavier tails compared to the theoretically assumed squared exponential function. According to the Eklund *et al.* [4] study this might explain why the cluster-wise thresholding works especially poor (false positive rate as high as 50%) when a low cluster-defining threshold of $p = 0.01$ is used (which typically generates larger clusters).

The Eklund *et al.* [4] paper indeed revealed a critical flaw in 3 of the mainstream fMRI processing software suites. As neuroscientists, we embrace the paper and reflect on the future of fMRI research. It is through the discovery of errors in existing methods scientists could continuously perfect research tools, which eventually leads to new discoveries that will withstand even more rigorous scientific scrutiny under future technology and standards. In fact, this self-correcting process has been constantly shaping the direction of fMRI research. The Eklund *et al.* [4] paper is just the latest one in a series of recent papers examining the validity of methods used in fMRI analyses. One of the findings in the Eklund *et al.* [4] paper has already prompted the developers of AFNI to fix a bug in their 3dClustSim code which partially addressed the issue of inflating FWE. Non-parametric methods, which rely on fewer assumptions as pointed out in the Eklund *et al.* [4] paper, is a very attractive direction in the near future.

Yet a disturbing fact is that the software error reported in the Eklund *et al.* [4] study was undetected despite all the research efforts in neuroimaging throughout the years. A multitude of factors are likely to have contributed to this. Objectively speaking, one important factor is that most previous research studies only used simulated rather than real fMRI dataset as used in the Eklund *et al.* [4] study. Simulated datasets, although making the ground truth (whether there is an activation or not) more easily accessible, generally fall short in sufficiently modelling the complex spatiotemporal properties of fMRI data. The mathematical models used to generate the data would quickly become too complex and intractable if one intends to capture all the nuances. In this regard, the use of sham task analyses in the Eklund *et al.* [4] study on large open-domain resting state fMRI datasets (499 healthy participants; where no systematic brain activation is expected given no task was actually implemented) is an innovative approach to validate statistical methods. This is made possible only as a result of recent data sharing initiatives in the neuroimaging community such as the 1000 Functional Connectomes Project, the Open fMRI Project, and the Human Connectome Project.

Correspondence to: Yuelu Liu, Centre for Mind and Brain, University of California, Davis, 267 Cousteau Pl, Davis, CA, USA, Email: vlliu@ucdavis.edu

Received: August 11, 2016; **Accepted:** September 09, 2016; **Published:** September 12, 2016

Aside from the above, one still wonders why it took so long for scientists to identify this error and let it to threaten the validity of so many studies using fMRI (about 3500 studies could have been affected according to the latest estimate by the authors). It seems that although in theory scientific research bears the self-correcting mechanism, in reality we need to do a much better job to ensure the efficient and timely correction of errors. Unfortunately, the hyper-competitiveness in today's scientific research system may have hindered the self-correcting process [5]. Slowdowns in government budgets for research in the past decade coupled by rising costs of scientific research have made obtaining research funding increasingly challenging for principal investigators. Further, stringent budgets in universities also created increased difficulty for young aspiring scientists to land a tenure-track position after their PhD training. Many of them now face prolonged postdoctoral trainings or need to seek alternative non-tenure track academic positions (e.g., lecturer, research scientist, temporary visiting positions etc.) after graduation. This generates a difficult situation for scientists to engage in deeply creative activity and perform meaningful research. In terms of grant application, investigators are often better off to submit grant proposals that have a high chance of generating immediate scientific results in order to increase the likelihood of securing funding continuously. Proposals and hypotheses that are well-based on existing theories/knowledge are hence often considered more competitive among grant reviewers than those seek to invalidate and correct previous findings because the latter could be considered too risky and more likely to fail. A similar situation exists when it comes to publishing research articles in scientific journals. When the careers of scientists (securing jobs, funding, tenure promotions) become so heavily dependent on their publication record, the increased competition entices scientists to publish eye-catching results that "advances" the field rather than those merely trying to determine whether a previous work is replicable or not. In addition, the hyper-competitiveness also tends to encourage people to publish more papers at a faster rate. This is likely to negatively impact the overall quality of the published studies, generating large amount of noise in the field. When so many published studies are not replicable [2,3], even attempting to correct them becomes a daunting task, let alone when scientists are preoccupied with writing and revising their grant proposals.

Many journals may also have the tendency toward publishing positive instead of negative results. It is natural that people are drawn toward exciting new discoveries. Commercial journals, even the most prestigious ones, may unintentionally develop such a bias since publishers rely on broad readership and subscriptions to be profitable. Exciting new discoveries typically generates more interests and citations compared to negative results, therefore more efficiently boosting the ranking of a journal. From the standpoint of researchers who publish

results questioning the validity of prior studies, doing so may inevitably put themselves in confrontation with the research group they are questioning. Moreover, researchers who published the original studies questioned by the new study may be chosen as anonymous reviewers, which might further increase the difficulty for the new study to get published. Even if the study manages to get published, the authors are still likely to continue facing resistance from defenders of the old norm. In these regards, publishing negative results questioning the validity of existing ones is difficult in itself as authors have to face an uphill battle against journal editors, anonymous reviewers, as well as various research groups defending the studies called into question.

Despite all these hurdles toward science's self-correction process, I remain positive and hopeful that we are making progress to eliminate such hurdles. The latest media attention on the validity of fMRI research generated by the Eklund *et al.* [4] paper is exactly what we need to get the process started. Yet significantly more efforts are still needed to restore the normal pace of self-correction in scientific research. These efforts include mitigating the hyper-competition in the scientific community, addressing journals' tendency to publish positive results, as well as fostering an open and positive attitude toward publishing negative results among scientists. While we hope federal research funding will steadily increase in the future to ease the hyper-competitiveness in the scientific research system, editors of scientific journals should welcome more submissions of high quality research studies reporting findings that are either incompatible with or seek to invalidate prior studies. We scientists as reviewers, when assigned with the task of reviewing such a manuscript, should also focus more on examining the validity of the study's design, soundness of the methods, and appropriateness of the conclusion rather than overly fixate on the inconsistency between the manuscript and prior studies. These efforts will help reduce the disadvantage and bias these studies face and might gradually encourage scientists to publish more of such studies.

References

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 U S A* 113: 7900-7905. [[Crossref](#)]
2. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, et al. (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 14: 365-376. [[Crossref](#)]
3. Open Science Collaboration (2015) PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* 349: aac4716. [[Crossref](#)]
4. Eklund A, Andersson M, Josephson C, Johansson M, Knutsson H (2012) Does parametric fMRI analysis with SPM yield valid results? An empirical study of 1484 rest datasets. *Neuroimage* 61: 565-578. [[Crossref](#)]
5. Alberts B, Kirschner MW, Tilghman S, Varmus H (2014) Rescuing US biomedical research from its systemic flaws. *Proc Natl Acad Sci U S A* 111: 5773-5777. [[Crossref](#)]