Correspondence

Two Methodologies in “Amygdala Activation and Connectivity to Emotional Processing Distinguishes Asymptomatic Patients With Bipolar Disorders and Unipolar Depression” That Can Produce False-Positive Results and Some Statistical Recommendations

To the Editor:

I was recently made aware of the article by Korgaonkar *et al.* (1) through a media article on ScienceDaily (2). Intrigued by the strong statements in the media article, I read the peer-reviewed article and was surprised to see the authors used two methodological procedures that have been demonstrated to produce high false-positive rates (3–5) and biased results (6,7). This correspondence is not to discredit the results of Korgaonkar *et al.* (1), but to emphasize the importance of using statistical tools that produce unbiased estimates and probabilities, so that readers can assess the amount of evidence provided by the results. The questionable methodological procedures I am referring to are Gaussian random field (GRF)–based spatial extent inference (SEI), which has been demonstrated to yield nominal type 1 error rates only at conservative cluster-forming thresholds (CFTs) ($p < .001$) and large smoothing kernels; and circularity analysis, where imaging variables are selected by testing associations in the full sample and subsequently used in prediction (6–8). Although the association of high false-positive rates with these methodological procedures has been highly publicized (3,6,9,10), use of these procedures is still common in neuroimaging studies (1,11,12). Both of these approaches produce apparently interesting associations with a high probability when, in fact, there is no association. I discuss these two analytical approaches and the importance of accurate procedures.

GRF-Based SEI False-Positive Rates

Type 1 error rates of GRF-based SEI have been demonstrated to be highly sensitive to the choice of CFT, image resolution, and image smoothness (3,5,13,14). At anticonservative CFTs ($p > .001$), type 1 error rates of GRF-based methods can be as high as 60%, and the actual error rates are highly dependent on the intrinsic and synthetic smoothness of the data and the software package used (3–5). Despite these results and recommendations about parameter settings from GRF proponents (13), some authors continue to use anticonservative thresholds.

The use of anticonservative testing procedures is caused, in part, by the overemphasis on an adjusted *p* value passing the .05 threshold. It is likely that some investigators select their *p* value adjustment software to pass this classical threshold. This approach affects the reader’s ability to interpret the strength of the results because the adjusted *p* values do not accurately describe the probability of a type 1 error. If correct study procedures are used, the *p* value contains important information; it describes the likelihood that the procedures used will produce a result as extreme or more extreme under the specified null hypothesis.

There are several practices that may ameliorate this problem:

1. Spatial extent *p* values should be reported without thresholding; when these are accurately computed, they represent the probability of a cluster equal or larger under the null that the image mean is unassociated with the covariate.
2. Corrected and uncorrected results should be reported and made available through an organized data sharing system so that evidence can be aggregated across studies (15,16).
3. Software tools that use GRF-based SEI should prohibit investigators from using thresholds where the theory is known to perform poorly.
4. Reviewers and journals should encourage the use of the latest statistical tools that robustly yield nominal type 1 error rates across any range of CFTs, voxel size, and smoothness (3–5). Many software packages (e.g., AFNI, FreeSurfer, FSL, SnPM) have introduced permutation SEI procedures (17) as well as cluster-free testing procedures (4,18). We recently proposed a robust semiparametric solution to this problem that is appropriate for multilevel models (19).

Circularity Errors

One form of a circularity error occurs when a scientist performs feature selection using the full sample and then assesses the accuracy of those features using cross-validation (1,8). In 2009, Kriegeskorte *et al.* (8) found that 56 of 134 sampled functional magnetic resonance imaging studies reported selectively obtained results. This error, combined with weak type 1 error control, can produce high prediction accuracy when no association exists. This is a critical problem because prediction accuracy is highly valued, and scientists unaware of the mistake may have overconfidence in the study findings. Moreover, prediction accuracy is more easily communicated to the public than other specialized concepts, so these artificially confident results are often strongly emphasized in media sources.

Circularity errors are difficult to identify without a detailed description of the methodological procedures. Two recent studies found that 14% and 11.6%, respectively, of the articles sampled did not describe methods clearly enough for the authors to determine whether circularity errors were made (8,12). Use of the general recommendations below may help to avoid this problem.

General Recommendations and Conclusions

Some of the following recommendations have been stated elsewhere:

1. Journals and reviewers should not place emphasis on a *p* value passing the .05 threshold. This is a reiteration of a statement made by the American Statistical Association (20). If a scientist has a well-founded hypothesis, a *p* value above the threshold may suggest that the study had low power or that the hypothesis should be revised; this can be determined from other statistics reported in the article.

https://doi.org/10.1016/j.bpsc.2018.12.007

ISSN: 2451-9022

2. Reproducible research practices should be required, such as code sharing, version tracking (e.g., with Git), reproducible software practices (e.g., R Markdown, Jupyter), and data sharing where possible (15,16).
3. Clear scientific reporting, as in the article by Korgaonkar *et al.* (1), is critical for reviewers to identify circular analyses.
4. Authors should be transparent about study procedures, such as data dredging, *p*-hacking, and use of multiple end points (21). An effort should be made to estimate the type 1 error of these procedures, e.g., with permutation testing.
5. Using or requiring data science experts as coauthors and reviewers allows scientists to more accurately implement advanced machine learning and statistical tools.

The two methodologies in the article by Korgaonkar *et al.* (1) are well characterized, yet they persist in neuroimaging studies (12). These procedures produce bias that can deceive readers about the strength of the findings. This is particularly impactful when the results are publicized in the media where study findings are often enthusiastically overstated. Honest and clear scientific reporting along with the practices described above may help to improve the replicability and validity of study findings.

Simon N. Vandekar

Acknowledgments and Disclosures

I thank my friend Micah Rees and my wife Lillie Vandekar for making me aware of the news article describing the results of the referenced study.

The author reports no biomedical financial interests or potential conflicts of interest.

Article Information

From the Department of Biostatistics, Vanderbilt University, Vanderbilt University Medical Center, Nashville, Tennessee.

Address correspondence to Simon N. Vandekar, Ph.D., Department of Biostatistics, Vanderbilt University, Vanderbilt University Medical Center, 2525 West End Avenue, Suite 1100, Nashville, TN 37203; E-mail: simon.vandekar@vanderbilt.edu.

See also associated correspondence: https://doi.org/10.1016/j.bpsc.2018.12.006.

Received Sep 11, 2018; revised Dec 17, 2018; accepted Dec 18, 2018.

References

1. Korgaonkar MS, Erlinger M, Breukelaar IA, Boyce P, Hazell P, Antees C, *et al.* (2019): Amygdala activation and connectivity to emotional processing distinguishes asymptomatic patients with bipolar disorders and unipolar depression. Biol Psychiatry Cogn Neurosci Neuroimaging 4:361–370.
2. Westmead Institute for Medical Research: Brain scans could distinguish bipolar from depression: Looking inside the brain to distinguish bipolar from depression. ScienceDaily. Available at: https://www.sciencedaily.com/releases/2018/09/180904093747.htm. Accessed October 3, 2018.
3. 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.
4. Cox RW, Chen G, Glen DR, Reynolds RC, Taylor PA (2017): FMRI clustering in AFNI: False-positive rates redux. Brain Connect 7:152–171.
5. Greve DN, Fischl B (2018): False positive rates in surface-based anatomical analysis. Neuroimage 171:6–14.
6. Vul E, Harris C, Winkielman P, Pashler H (2009): Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect Psychol Sci 4:274–290.
7. Kriegeskorte N, Lindquist MA, Nichols TE, Poldrack RA, Vul E (2010): Everything you never wanted to know about circular analysis, but were afraid to ask. J Cereb Blood Flow Metab 30:1551–1557.
8. Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI (2009): Circular analysis in systems neuroscience: The dangers of double dipping. Nat Neurosci 12:535–540.
9. Linköping University: Softwares for fMRI yield erroneous results: Cluster failure: Why fMRI inferences for spatial extent have inflated false positive rates. ScienceDaily. Available at: https://www.sciencedaily.com/releases/2016/06/160627160927.htm. Accessed October 3, 2018.
10. Lehrer J: Voodoo correlations: Have the results of some brain scanning experiments been overstated? Scientific American. Available at: https://www.scientificamerican.com/article/brain-scan-results-overstated/. Accessed October 3, 2018.
11. Pang Y, Cui Q, Duan X, Chen H, Zeng L, Zhang Z, *et al.* (2017): Extraversion modulates functional connectivity hubs of resting-state brain networks. J Neuropsychol 11:347–361.
12. Pulini AA, Kerr WT, Loo SK, Lenartowicz A (2019): Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: Effects of sample size and circular analysis. Biol Psychiatry Cogn Neurosci Neuroimaging 4:108–120.
13. Flandin G, Friston KJ (2017): Analysis of family-wise error rates in statistical parametric mapping using random field theory [published online ahead of print Nov 1]. Hum Brain Mapp.
14. Woo CW, Krishnan A, Wager TD (2014): Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. Neuroimage 91:412–419.
15. Maumet C, Auer T, Bowring A, Chen G, Das S, Flandin G, *et al.* (2016): Sharing brain mapping statistical results with the neuroimaging data model. Sci Data 3:160102.
16. Gorgolewski KJ, Varoquaux G, Rivera G, Schwarz Y, Ghosh SS, Maumet C, *et al.* (2015): NeuroVault.org: A web-based repository for collecting and sharing unthresholded statistical maps of the human brain. Front Neuroinform 9:8.
17. Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE (2014): Permutation inference for the general linear model. Neuroimage 92:381–397.
18. Smith SM, Nichols TE (2009): Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44:83–98.
19. Vandekar SN, Satterthwaite TD, Xia CH, Ruparel K, Gur RC, Gur RE, Shinohara RT (2018): Robust spatial extent inference with a semiparametric bootstrap joint testing procedure. arXiv:180807449 [stat.ME]. Available at: http://arxiv.org/abs/1808.07449. Accessed September 6, 2018.
20. Wasserstein RL, Lazar NA (2016): The ASA's statement on p-values: context, process, and purpose. Am Stat 70:129–133.
21. Ioannidis JP (2005): Why most published research findings are false. PLoS Med 2:e124.