Commentary

Biological Psychiatry: CNNI

Thresholds, Power, and Sample Sizes in Clinical Neuroimaging

Cameron S. Carter, Tyler A. Lesh, and Deanna M. Barch

Replicability has become a major issue in the behavioral and biological sciences, and one of the goals of *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* is to publish articles that are designed and analyzed in a manner to ensure that they provide both novel and reliable insights into the cognitive and neural mechanisms underlying mental disorders and the mechanisms of treatment response. Essential to this process are methods for type I error control, given the massive univariate testing that is often used in whole-brain imaging studies. The most widely used approach involves cluster-based correction (1), in which a height threshold for significance is established at the voxel level and a minimum number of contiguous voxels is required to exceed the likelihood that a given cluster would occur by chance alone. This approach was introduced by Forman *et al*. (1) and has remained popular for several reasons. First, correcting for multiple comparisons using Bonferroni and related methods at the voxel level requires very large effect sizes that are beyond the sizes typically seen in neuroimaging studies. Second, correction methods focused on a single voxel fail to capitalize on the fact that in blood oxygen level–dependent imaging studies, most "true-positive" activations extend beyond a single voxel.

Concerns have been raised as to what the appropriate height threshold should be when cluster correction is applied. There are at least three competing concerns when making such a choice: false positives, false negatives, and the precision of spatial localization. Different researchers may prioritize different subsets of these concerns, depending on the type of study being conducted. For example, Woo *et al.* (2) used simulations to demonstrate that when lower height thresholds are used, clusters are larger, making it harder to detect activation in very small structures and limiting the ability to make precise anatomic inferences regarding the location of significant voxels. This constraint is important in studies in which a primary goal is testing hypotheses regarding functional specialization in subregions of the brain that are in close proximity. Based on these results, the authors suggested that height thresholds of at least $p < .001$ should be used in functional imaging studies that use cluster-based correction for multiple comparisons. In addition, Woo *et al.* noted that at low height thresholds (e.g., $p = .01$), the false-positive rate at the whole-brain level may be higher than expected. At much higher height thresholds (e.g., $p = .001$), the whole-brain false-positive rate is accurate or even conservative under some conditions (e.g., whole-brain false-positive rate $< .05$ with higher signal-to-noise levels).

Studies such as the one by Woo *et al.* (2) are very important as the field of clinical neuroimaging matures and addresses the complex statistical challenges that face us and the need to maximize the informativeness and replicability of results. However, there are additional complexities to the issue of using cluster-based correction that were not highlighted in the study by Woo *et al.*, especially as they apply to clinical neuroimaging studies. Woo *et al.* focused on single-group studies and prioritized concerns related to false-positive results and precise spatial localization. Clinical neuroimaging studies often involve two or more groups being contrasted, which leads to additional sample size and power challenges. Furthermore, clinical neuroimaging studies are also concerned with false-negative results and in some cases may be testing strong a priori hypotheses that may not always necessitate the same narrow spatial localization (e.g., hypotheses involving differential activation in segregated, distributed functional neural circuits). There may be informative studies that have aims or goals that might prioritize reducing false-negative results and enhancing power, at the expense of the narrowness of spatial localization.

The original motivation for introducing cluster-based correction was the reality that unrealistically large effect sizes were needed to detect significance at the voxel level with Bonferroni correction in a well-powered study. This raises the question as to what is a well-powered study. To assess this question, we generated power curve plots for effect sizes of .5 and .7 (moderate effects that would generally be considered meaningful in most areas of psychology and neuroscience) with α set at $p < .01$, $p < .005$, $p < .001$, and $p < .0001$. Power curves were calculated using G*Power 3.1 and indicate the total combined sample size needed at a range of power levels given two independent groups and a two-tailed *t* test (3). These power curves reflect the sample size necessary for a given voxel to reach a threshold of significance before subsequent cluster-level correction. Inspection of the power curves in Figure 1 suggests that for the moderate effect sizes tested, sample sizes needed to obtain .8 power vary greatly according to the voxel height threshold. For reference, the effect size averaged across nine regions of interest in the working memory task versus rest described by Desmond and Glover (4) was approximately .6, although power curves were displayed for a wide range of effect sizes in that article. It can be seen in Figure 1 that given a voxelwise threshold of $p < .0001$, the required combined sample sizes for each effect size (.7 and .5) are 192 and 366. For $p < .001$, a combined sample size of 146 and 280 is needed. The required combined sample sizes are 114 and 218 for $p < .005$ and 100 and 192 for $p < .01$.

As in the parent journal *Biological Psychiatry*, in *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* we expect that all articles submitted will apply a principled and

http://dx.doi.org/10.1016/j.bpsc.2016.01.005

ISSN: 2451-9022 Biological Psychiatry: Cognitive Neuroscience and Neuroimaging March 2016; 1:99–100 www.sobp.org/BPCNNI

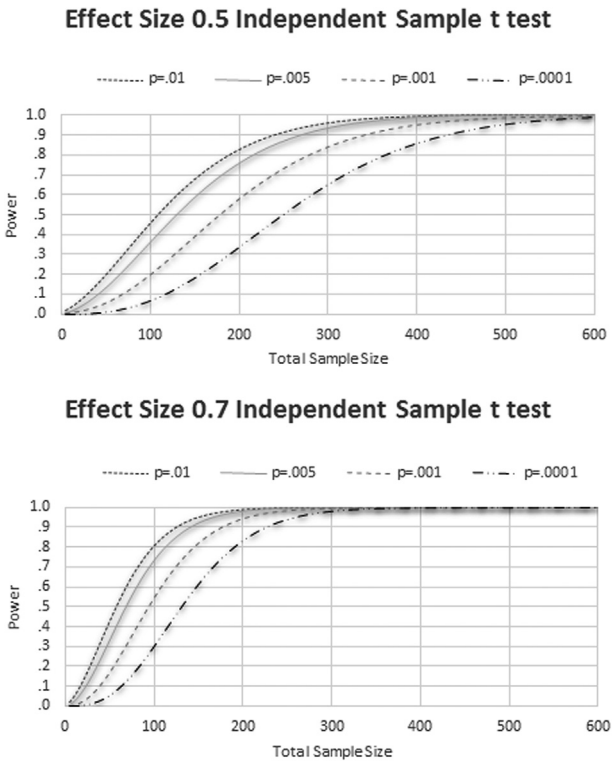

Figure 1. Plots indicating sample sizes required for .8 power to reject the null hypothesis at the voxel level at different statistical thresholds. The top panel shows these plots for data with an effect size of .5, and the bottom panel shows the data for an effect size of .7.

well-justified method for correction for multiple comparisons and protecting against type I error (5). Given the tradeoffs evident in the above-reviewed data as well as in studies such as those by Woo *et al.* (2), we believe there is no single approach that is appropriate for all studies or even within a method such as cluster-based correction that is optimal for all studies. For example, for studies seeking activation in very small regions or attempting to establish functional dissociations between adjacent regions, it would be most appropriate to use a high voxel threshold (e.g., $p < .001$ or less). For a study in which precise localization is not required for meaningful inference, the particular advantage afforded by spatial precision by a high threshold might not be required, and concerns related to false-negative results may have higher priority. When studies are testing for effects across the whole brain, the use of a high threshold comes at a cost regarding power and increases type II error. It can be seen in Figure 1 that at a given sample size, the cost in power when going from $p < .01$ to $p < .005$ is much smaller than when going from $p < .005$ to $p < .001$. Also, no matter what threshold is used, studies that have <100 subjects (i.e., 50 per group) have limited power unless large effect sizes are anticipated. Such studies may need to apply additional data reduction methods (i.e., a priori regions of interest) to reduce the risk of type II error.

In conclusion, in the rapidly developing field of neuroimaging, there is a need for ongoing analysis and method development to address concerns about replicability and provide appropriate type I and type II error protection. Cluster-based correction is likely to remain a popular method for type I error control for some time to come. At the present time, there appears to be no single voxel height threshold that is optimal for all studies. Rather, a given approach needs to be justified in light of the aims of the specific study and the experimental design and with the understanding that the limitation of enhanced spatial precision is decreased power and the need for larger sample sizes. Furthermore, there is a chance for error in any study, even the most rigorously designed and analyzed; therefore, in addition to such rigorous studies in *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, we will also be interested in seeing rigorous replications, as this might be considered the optimal test of reliability.

Acknowledgments and Disclosures

This work was supported by the National Institute of Mental Health (Grant No. 5 R01 MH059883 to CSC and Grant No. 5 R01 MH084840 to DMB) and the National Institute on Drug Abuse (Grant No. 1 U01 DA041120-01 to DMB).

CSC has served as a consultant for Sunovion. DMB has served as a consultant for Roche, Pfizer, Takeda, and Amgen. TAL reports no biomedical financial interests or potential conflicts of interest.

Article Information

From the Departments of Psychiatry (CSC, TAL) and Psychology (CSC) and Center for Neuroscience (CSC), University of California Davis, Davis, California; and Departments of Psychiatry (DMB), Psychology (DMB), and Radiology (DMB), Washington University in St. Louis, St. Louis, Missouri.

Address correspondence to Cameron S. Carter, M.D., University of California Davis, 4701 X Street, Sacramento, CA 95816; E-mail: cameron.carter@ucdmc.ucdavis.edu.

Received Jan 26, 2016; revised Jan 28, 2016; accepted Jan 28, 2016.

References

1. Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC (1995): Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. Magn Reson Med 33:636–647.
2. Woo CW, Krishnan A, Wager TD (2014): Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. Neuroimage 91: 412–419.
3. Faul F, Erdfelder E, Lang AG, Buchner A (2007): G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 39:175–191.
4. Desmond JE, Glover GH (2002): Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses. J Neurosci Methods 118:115–128.
5. Carter CS, Heckers S, Nichols T, Pine DS, Strother S (2008): Optimizing the design and analysis of clinical functional magnetic resonance imaging research studies. Biol Psychiatry 64:842–849.