Received: 7 December 2020 | Revised: 26 January 2021 | Accepted: 4 February 2021

DOI: 10.1002/hbm.25374

RESEARCH ARTICLE

WILEY

Improving the replicability of neuroimaging findings by thresholding effect sizes instead of *p*-values

Simon N. Vandekar[1,2] | **Jeremy Stephens**[2]

[1]Department of Biostatistics, Vanderbilt University, Nashville, Tennessee

[2]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee

Correspondence
Simon N. Vandekar, Department of Biostatistics, Vanderbilt University, 2525 West End Ave., #1136,
Nashville, TN 37203.
Email: simon.vandekar@vanderbilt.edu

Funding information
National Cancer Institute, Grant/Award Number: P50CA236733; National Institute of Mental Health, Grant/Award Number: 1R01MH123563-01

Abstract

The classical approach for testing statistical images using spatial extent inference (SEI) thresholds the statistical image based on the *p*-value. This approach has an unfortunate consequence on the replicability of neuroimaging findings because the targeted brain regions are affected by the sample size—larger studies have more power to detect smaller effects. Here, we use simulations based on the preprocessed Autism Brain Imaging Data Exchange (ABIDE) to show that thresholding statistical images by effect sizes has more consistent estimates of activated regions across studies than thresholding by *p*-values. Using a constant effect size threshold means that the *p*-value threshold naturally scales with the sample size to ensure that the target set is similar across repetitions of the study that use different sample sizes. As a consequence of thresholding by the effect size, the type 1 and type 2 error rates go to zero as the sample size gets larger. We use a newly proposed robust effect size index that is defined for an arbitrary statistical image so that effect size thresholding can be used regardless of the test statistic or model.

KEYWORDS
ABIDE, cluster extent inference, robust effect size index

1 | INTRODUCTION

Spatial extent inference (SEI) is the most common inference procedure in neuroimaging, where a statistical image is first thresholded to form spatially contiguous clusters and then *p*-values are computed based on the size of the suprathreshold clusters (Friston, Worsley, Frackowiak, Mazziotta, & Evans, 1994; Woo, Krishnan, & Wager, 2014; Yeung, 2018). The statistical image is often derived from a group level analysis using a linear model, a linear mixed effects model, or estimating equation (Friston et al., 1994; Guillaume, Hua, Thompson, Waldorp, & Nichols, 2014). The cluster forming threshold (CFT) is chosen to satisfy an uncorrected *p*-value threshold (e.g., $p < .001$ or $p < .01$) at the voxel level.

While the method is widely applied, there are two important limitations of thresholding statistical images based on the voxel-level *p*-value: (a) when the true effect size varies across voxels, the power of detecting a given voxel using a *p*-value threshold is dependent on the sample size, so the set of suprathreshold voxels across studies attempting to replicate an experiment is a function of the sample size—larger studies are targeting smaller effect sizes than smaller studies. This means that for fMRI studies, the amount of brain activation that should be detected in a group level analysis is dependent on the sample size. (b) *p*-value thresholding (PVT) is sensitive to arbitrarily small effect sizes in large sample sizes that are not of clinical interest. This can occur as a result of the null hypothesis fallacy (Bowring, Telschow, Schwartzman, & Nichols, 2019), which posits that the classical null hypothesis of an effect size exactly equal to zero is not satisfied in real neuroimaging data. The null hypothesis fallacy is supported by empirical evidence (Gonzalez-Castillo et al., 2012) and by biological features and processing steps (e.g., spatial smoothing) that imply the

mean function of a statistical image is spatially continuous (see table 1 of Chumbley & Friston, 2009).

In this article, we argue for using an effect size-based CFT, instead of a threshold based on a *p*-value. Our suggestion is motivated by increasing criticism of PVT in the context of hypothesis testing and an increased interest in potential alternatives (Bowring et al., 2019; Bowring, Telschow, Schwartzman, & Nichols, 2021; Chen et al., 2019; Chen, Taylor, & Cox, 2017; Wasserstein & Lazar, 2016; Wasserstein, Schirm, & Lazar, 2019). This approach resolves the limitations of PVT stated above: (a) Using effect size thresholding (EST), the set of voxels that are identified as activated in an analysis does not depend on the sample size, so it improves the consistency of the target regions of activation across studies. (b) The probability of detecting voxels with true effect sizes below the threshold approaches zero and the probability of detecting voxels with true effect sizes above the threshold approaches one, with increasing sample size. In other words, the type 1 and type 2 error rates go to zero as the sample size increases.

In the following analyses, we used bootstrap-based simulations to illustrate the advantages of EST over PVT using data from the Autism Brain Imaging Data Exchange (ABIDE; Di Martino et al., 2014, 2017). To demonstrate the two advantages above, we (a) compared the target region for PVT and EST across varying sample sizes and (b) compared the probability of a supra-threshold finding for both methods across varying sample sizes.

2 | METHODS

2.1 | ABIDE data set and statistical software

We downloaded fully processed amplitude of low frequency fluctuations (ALFF; Zang et al., 2007) from resting state functional magnetic resonance imaging scans for the ABIDE data set from the preprocessed connectomes project (http://preprocessed-connectomes-project.org/). The ABIDE is a collaboration of 16 international sites that have aggregated and openly share neuroimaging data from 1,112 scanning sessions, including 539 individuals with autism spectrum disorder (Di Martino et al., 2014). A total of 1,027 ALFF images were available for download from the preprocessed ABIDE data set, representing the subset of subjects who completed the resting state fMRI sequence. The data were preprocessed using the Configurable Pipeline for the Analysis of Connectomes (CPAC) pipeline and analyzed in MNI space at 3 mm isotropic resolution using the pbj package in R (currently, available for download at https://github.com/simonvandekar/pbj using the "ftest" branch). Simulations were executed using the NIsim R package (https://github.com/statimagcoll/NIsim). To improve coverage of the brain, we sequentially removed subjects from the sample until over 30,000 voxels were included in the study mask, which was defined as the intersection of all subjects ALFF images. At each step, the subject whose removal improved the study mask coverage the most was excluded from the study. A total of nine subjects were excluded yielding a mask with 30,272 voxels. This minimal data quality screening was performed in order to maximize the number of subjects available for the simulation analyses. Further details on the processing pipeline are provided on the preprocessed connectomes website (http://preprocessed-connectomes-project.org/abide/index.html) and code, from data download to figure production, is provided with this paper and available at https://github.com/statimagcoll/NIsim/blob/master/pbjESthresholding.Rmd. While our analyses use ALFF resting state images, the methods and theory discussed here apply to arbitrary statistical images obtained from analyses of functional or anatomical images.

2.2 | Statistical methods

2.2.1 | Effect size definition and image thresholds

We recently proposed a robust effect size index (RESI) that can be computed using the sample size, chi-squared statistic, and degrees of freedom (Vandekar, Tao, & Blume, 2020). The RESI is not model dependent and can be used for most statistical models including generalized linear models and mixed models. When a sandwich covariance estimator is used, the RESI estimator is asymptotically unbiased. The estimator for the effect size index is

$$\hat{S}(v) = [\max\{(T_n(v) - m_1)/(n-m), 0\}]^{1/2},$$

where $T_n(v)$ is the chi-squared statistic image for the test of the parameter of interest, m_1 is the degrees of freedom of the test statistic, m is the model degrees of freedom, and n is the number of independent samples. Note, the original formula (Vandekar et al., 2020) contains a typographical error because it subtracts by m instead of m_1 in the formula. Here, we use the factor n instead of $n - m$, because it has smaller positive bias.

To set the image thresholds, we converted *p*-value and *S* thresholds to chi-squared statistical values using the formulas

$$\chi^2(p) = F^{-1}_{\chi^2_{m_1}}(1-p) \quad (1)$$

$$\chi^2(S) = n \times s^2 + m_1, \quad (2)$$

where $F^{-1}_{\chi^2_{m_1}}$ is the inverse cumulative distribution function of a chi-squared distribution with m_1 degrees of freedom. We chose to convert the thresholds to the chi-squared scale for computational convenience, so that a single set of bootstrap images could be used for visualizing the effects of both thresholding methods.

2.2.2 | Distinction between PVT and EST

If the null is false, the test statistic at a given location v is approximately chi-squared with noncentrality parameter $n \times S^2$, where S is the robust effect size index (Vandekar et al., 2020). When the null is true, $S = 0$. Because the PVT (1) chooses a chi-squared statistical threshold that is fixed across sample sizes, as n gets larger, regardless

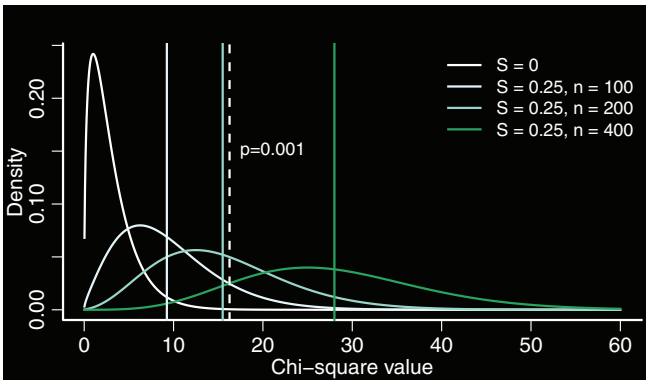

FIGURE 1 An illustration of the difference between *p*-value and effect size thresholding for a chi-squared statistic on 3 *df*. The white curve is the distribution of the test statistic under the null, which implies $S = 0$. The white dashed vertical line indicates the *p*-value threshold, $p < .001$, which is the same for all sample sizes. The green colored densities are the distribution of the chi-squared statistic under the alternative $S = 0.25$ for different sample sizes, *n*, indicated by the legend. As the sample size increases the distribution of the chi-squared statistic shifts right. The effect size threshold indicated by the vertical colored lines, $S^2 \times n + 3$, appropriately adjusts with the sample size

of how small a nonzero effect size is, the test statistic will eventually be large enough to reject the null (Figure 1). If the null hypothesis fallacy is true, then the effect size is zero almost everywhere, and the target region gets larger with increasing sample size (Chumbley & Friston, 2009; Gonzalez-Castillo et al., 2012).

In contrast, the EST (2) is a function of the sample size and increases at a linear rate with respect to *n*. Thresholding based on the formula (2) is equivalent to thresholding $(T_n(v) - m_1)/n > s^2$. Because $(T_n(v) - m_1)/n$ is an estimator for $S^2(v)$ (the square of the true, unknown, effect size image) and converges to the true value $(T_n(v) - m_1)/n \to S^2(v)$, as $n \to \infty$ the type 1 and type 2 error rates go to zero,

$$\mathbb{P}\left((T_n(v) - m_1)/n \geq s^2 | S^2(v) \leq s^2\right) \to 0$$
$$\mathbb{P}\left((T_n(v) - m_1)/n \geq s^2 | S^2(v) \geq s^2\right) \to 1.$$

The notational distinction between $S^2(v)$ and s^2 is that the former is the square of the true, unknown, effect size image, and the latter is the square of the chosen effect size threshold. To contrast the two approaches: PVT is always trying to target the subset of the image where the effect size is greater than zero, whereas EST only targets a range of interesting effect sizes.

2.2.3 | Simulation methods

We use simulations to illustrate the differences between PVT and EST. In order to simulate realistic imaging data, we sampled from the 1,027 ALFF images in the ABIDE data set with replacement for sample sizes $n \in \{25,50,100,200,400,800\}$. In each data set, we fit the model

$$Y_i(v) = \alpha_0(v) + \alpha_1(v) \times \text{mot}_i + \alpha_2(v) \times \text{sex}_i + \alpha_3(v) \times \text{dx}_i + \beta(v) \times \text{age}_i + \varepsilon_i(v),$$

where $Y_i(v)$ is the ALFF image for subject *i* at location *v*, mot_i is the mean frame displacement for subject *i*, sex_i, dx_i, age_i are the sex, diagnosis, and age of subject *i*, the $\alpha_j(v)$ and $\beta(v)$ terms are unknown parameter images, and the error term $\varepsilon_i(v)$ is assumed to have zero mean and finite variance. These covariates were chosen based on common demographic and developmental findings in the literature on resting fMRI activity and to control for the linear effect of motion (Dumontheil, 2016; Stevens, 2009, 2016). For each of 10,000 simulations, we estimated the chi-squared statistical image for the test of

$$H_0 : \beta(v) = 0, \quad (3)$$

distributed on 1 *df*, which we denote by $T_n(v)$. This null hypothesis implies that the effect size at a given voxel is equal to zero, $S(v) = 0$.

We used two methods to evaluate the replicability of PVT and EST methods. First, across the 10,000 simulations, we computed the mean chi-squared image and then converted the *p*-value and *S* thresholds to chi-squared values using formulas (1) and (2). We used the converted chi-squared thresholds to visualize suprathreshold values in the images (Figure 2; Target). These images show target sets, where the true mean of the statistic surpasses the given threshold for each sample size, indicating that they should be rejected for that sample size. Here, we present the results for a medium effect size $S = 0.2$ and the standard CFT $p = .001$. Second, we estimated the probability of each voxel being identified in the target set by computing the proportion of simulated data sets where each voxel passed the given chi-squared threshold (Figure 2; Probability). Here, we present the results for a probability of .05 for a medium effect size $S = 0.2$ and the standard CFT $p = .001$. These images represent the probability that a given voxel is estimated to be in the target set for a random sample. The online interactive figures (https://statimagcoll.github.io/EST-2020/) allow the user to choose from a range of values for the robust and parametric statistical images.

3 | RESULTS

We used simulations to estimate the target region, defined as the region where the true effect size is larger than the targeted effect size determined by the *p*-value or S threshold. We also used the simulations to compute the probability that a random sample identifies each voxel as belonging to the target region. Ideally, voxels within the target region should have high probability of being identified as such, whereas voxels outside of the target region should have low probability of being identified as target voxels.

The target image identifies regions of the image where the true expected value of the test statistic surpasses the given chi-squared

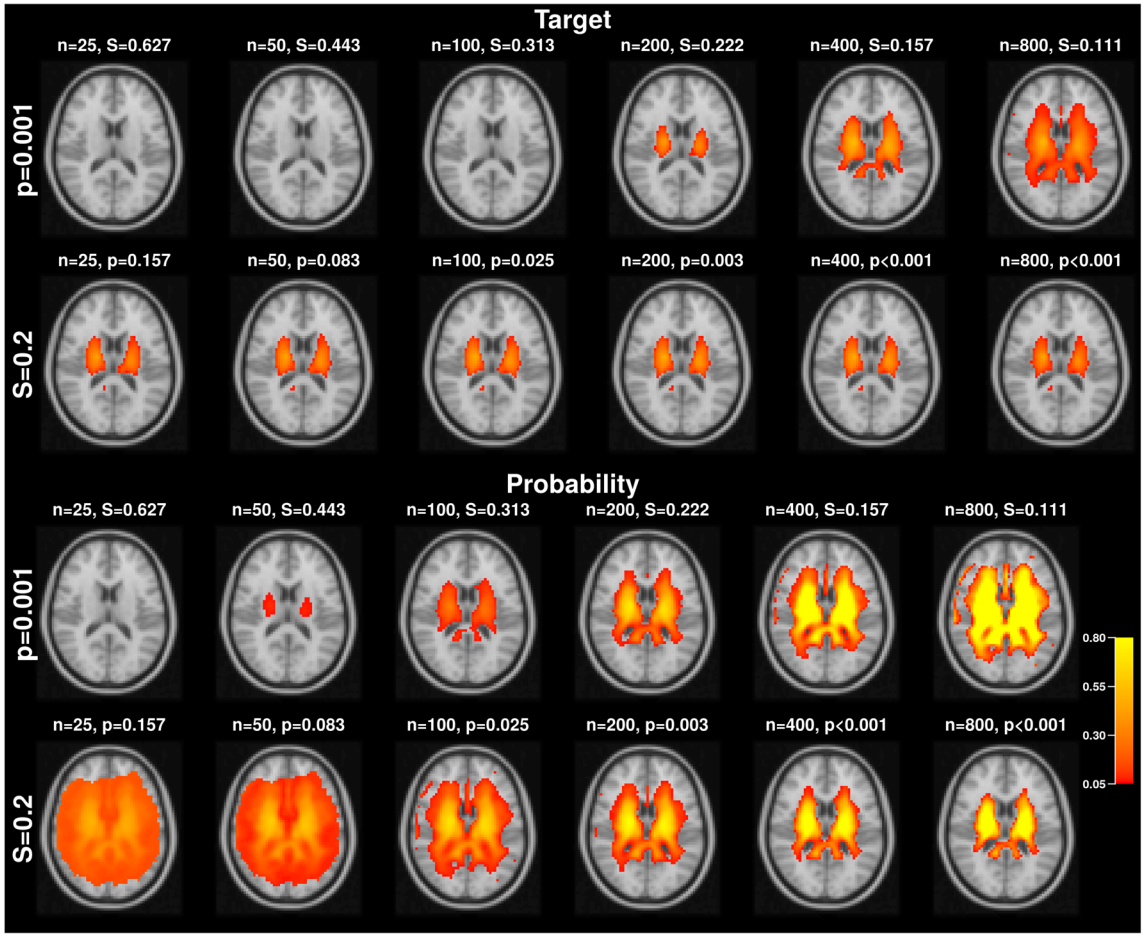

FIGURE 2 The target image and probability image for a *p*-value and *S* threshold using a parametric test statistic image. The target images show regions of the image where the true expected value of the statistic surpasses the given *p* or *S* threshold. The probability images show the probability that a given voxel is found to be in the target set across random samples. Sample sizes (*n*) are shown in each column with the corresponding threshold using the other thresholding method (that is not held constant across sample sizes)

threshold. For *p*-value thresholding the target set changes due to the fact that the effect size threshold is dependent on the sample size, because the type 1 error rate is fixed instead; smaller sample sizes are only targeting regions of the image with larger effect sizes (Figure 2; Target $p = .001$). Because we are sampling from a real data set, the mean function is continuous, so the null hypothesis is likely false almost everywhere (Chumbley & Friston, 2009; Gonzalez-Castillo et al., 2012). For this reason, the target set continues to increase across samples. In contrast, for the EST (Figure 2; Target $S = 0.2$) the target set remains consistent across sample sizes and the *p*-value threshold gets smaller with larger sample sizes.

The probability image is thresholded above .05 and shows the probability that a given voxel is included in the estimated target set across replications of the experiment (Figure 2; Probability). For the *p*-value threshold, this image can be thought of as the voxelwise power and type 1 error depending on whether the null hypothesis (3) is true at that voxel (Figure 2; Probability $p = .001$). If the null hypothesis is false almost everywhere, then the implication is that the images for PVT represent power. For the EST, the target set is determined by a threshold on the true effect size being larger than $S = 0.2$ instead of the null hypothesis (3). Smaller samples ($n = 25$, $n = 50$, $n = 100$) have greater uncertainty about which voxels belong in the target set, but the certainty increases with the sample size (Figure 2; Probability $S = 0.2$). Interactive versions of these figures are available online (https://statimagcoll.github.io/EST-2020/) that allow the reader to vary the thresholds and type of test statistic.

4 | DISCUSSION

We used bootstrap-based simulations to argue for thresholding statistical images by their effect size in SEI, over the classical approach of thresholding images by their *p*-value. EST has the clear advantage that the target set is consistent across different sample sizes and that the type 1 error goes to zero as the sample size increases. We hope that this approach may encourage more rigor in reporting neuroimaging by emphasizing a distinction of strength of evidence (effect size) and the probability of misleading evidence (Blume, 2002; Kang, Blume, Ombao, & Badre, 2015): The effect size summarizes the strength of statistical evidence conditional on the given data, and the *p*-value is

(often) an estimate of the type 1 error rate; the probability of a result more extreme under the null, considering variability across repeated samplings of the data. As a consequence, EST maintains a constant target set and guarantees that the type 1 error rate goes to zero as the sample size increases, because of the increasingly stringent threshold on the probability scale.

Choosing an effect size threshold in a small sample appears to be advantageous because the corresponding *p*-value threshold is less conservative, but this leads to greater uncertainty about the results (Figure 2; Probability). In this framework, small studies may identify regions where the effect size estimate is suprathreshold, but will likely not have to power to obtain small SEI *p*-values for that finding. The fact that a study cannot make a strong probabilistic statement is a result of it being underpowered for the target effect size. For a fixed CFT and sample size, the adjusted *p*-value assigned to an observed cluster using SEI is a function of the size of the study mask and the covariance structure of the image within the mask. A larger study mask will have reduced power because there is more likely to be larger clusters by chance. For this reason, if there is a specific hypothesis for a region of interest (ROI) based on prior literature, we recommend restricting inference to the ROI for the test, especially in small samples.

We only evaluated the thresholding part of SEI here, and our approach assumes that it is possible to accurately compute SEI *p*-values using any CFT. Arbitrary thresholds are known to fail with Gaussian random field approximations (Eklund, Nichols, & Knutsson, 2016; Friston, Worsley, et al., 1994; Kessler, Angstadt, & Sripada, 2017; Silver, Montana, Nichols, & Initiative, 2011) using less stringent *p*-value CFTs ($p > .001$), so our approach is more appropriate using modern resampling-based SEI methods that leverage permutation testing or bootstrapping, which are more likely to be robust to the CFT (Guillaume et al., 2014; Vandekar et al., 2018, 2019; Winkler, Ridgway, Webster, Smith, & Nichols, 2014). In future work it will be important to evaluate the procedures at a range of CFTs in simulations to ensure that the procedure maintains nominal error rates regardless of the chosen effect size threshold.

We do not suggest a particular effect size threshold in this paper and relied on thresholds published by Cohen (1988). There is not a single threshold appropriate for all research studies and we suggest that effect size images should be published along with the paper results in order to transparently present study findings. There are many ways for a researcher to choose effect size thresholds for static presentation in a publication. For example, an effect size threshold can be determined from well-known published effect sizes in the field. Alternatively, Kang et al. (2015) used an evidentialist approach for inference in task-base fMRI where the null hypothesis was determined by statistical image intensities in the cerebrospinal fluid. A similar approach could be chosen to determine the threshold for null effect sizes in neuroimaging. Another approach that circumvents thresholding the image, called Threshold-free Cluster Enhancement (TFCE), is a permutation inference method that does not require selection of a cluster forming threshold (Smith & Nichols, 2009). The TFCE statistic at given location, *w*, is a weighted sum of the size of the cluster that includes *w* across all possible thresholds, where the weight is a function of the threshold for a *t*- or Z-statistic image. The TFCE approach could be extended to use an effect size image instead of a raw statistical image, which could make the results more similar across studies with different sample sizes.

Our proposed methods still rely on hypothesis testing using SEI so are not completely free of the limitations of PVT. For imaging, modern approaches that construct confidence sets using effect size thresholding approaches or Bayesian inference procedures hold promise as true alternatives to PVT-based inference for neuroimaging (Bowring et al., 2019, 2021; Chen et al., 2019; Sommerfeld, Sain, & Schwartzman, 2018). Our suggestions here of EST demonstrates the advantage of considering alternatives to classical PVT.

ACKNOWLEGMENTS

Financial support from National Institutes of Health grants: (5P30CA068485-22; P50CA236733; 1R01MH123563-01).

DATA AVAILABILITY STATEMENT

We downloaded fully processed amplitude of low frequency fluctuations \citep[ALFF;][]{zang_altered_2007} from resting state functional magnetic resonance imaging scans for the ABIDE data set from the preprocessed connectomes project (\url{http://preprocessed-connectomes-project.org/}).

ORCID

Simon N. Vandekar https://orcid.org/0000-0002-7457-9073

REFERENCES

Blume, J. D. (2002). Likelihood methods for measuring statistical evidence. *Statistics in Medicine*, *21*(17), 2563–2599.

Bowring, A., Telschow, F., Schwartzman, A., & Nichols, T. E. (2019). Spatial confidence sets for raw effect size images. *NeuroImage*, *203*, 116187.

Bowring, A., Telschow, F. J. E., Schwartzman, A., & Nichols, T. E. (2021). Confidence sets for Cohen's d effect size images. *NeuroImage*, *226*, 117477.

Chen, G., Taylor, P. A., & Cox, R. W. (2017). Is the statistic value all we should care about in neuroimaging? *NeuroImage*, *147*, 952–959.

Chen, G., Xiao, Y., Taylor, P. A., Rajendra, J. K., Riggins, T., Geng, F., … Cox, R. W. (2019). Handling multiplicity in neuroimaging through Bayesian lenses with multilevel modeling. *Neuroinformatics*, *17*(4), 515–545.

Chumbley, J. R., & Friston, K. J. (2009). False discovery rate revisited: FDR and topological inference using Gaussian random fields. *NeuroImage*, *44*(1), 62–70.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum Associates.

Di Martino, A., O'Connor, D., Chen, B., Alaerts, K., Anderson, J. S., Assaf, M., … Milham, M. P. (2017). Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Scientific Data*, *4*, 170010.

Di Martino, A., Yan, C., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., … Milham, M. P. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, *19*(6), 659–667.

Dumontheil, I. (2016). Adolescent brain development. *Current Opinion in Behavioral Sciences*, *10*, 39–44.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, *113*, 7900–7905.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*(4), 189–210.

Friston, K. J., Worsley, K. J., Frackowiak, R. S., Mazziotta, J. C., & Evans, A. C. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, *1*(3), 210–220.

Gonzalez-Castillo, J., Saad, Z. S., Handwerker, D. A., Inati, S. J., Brenowitz, N., & Bandettini, P. A. (2012). Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(14), 5487–5492.

Guillaume, B., Hua, X., Thompson, P. M., Waldorp, L., & Nichols, T. E. (2014). Fast and accurate modelling of longitudinal and repeated measures neuroimaging data. *NeuroImage*, *94*, 287–302.

Kang, H., Blume, J., Ombao, H., & Badre, D. (2015). Simultaneous control of error rates in fMRI data analysis. *NeuroImage*, *123*, 102–113.

Kessler, D., Angstadt, M., & Sripada, C. S. (2017). Reevaluating "cluster failure" in fMRI using nonparametric control of the false discovery rate. *Proceedings of the National Academy of Sciences*, *114*(17), E3372–E3373.

Silver, M., Montana, G., Nichols, T. E., & Initiative, A. D. N. (2011). False positives in neuroimaging genetics using voxel-based morphometry data. *NeuroImage*, *54*(2), 992–1000.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98.

Sommerfeld, M., Sain, S., & Schwartzman, A. (2018). Confidence regions for spatial excursion sets from repeated random field observations, with an application to climate. *Journal of the American Statistical Association*, *113*(523), 1327–1340.

Stevens, M. C. (2009). The developmental cognitive neuroscience of functional connectivity. *Brain and Cognition*, *70*(1), 1–12.

Stevens, M. C. (2016). The contributions of resting state and task-based functional connectivity studies to our understanding of adolescent brain network maturation. *Neuroscience & Biobehavioral Reviews*, *70*, 13–32.

Vandekar, S., Tao, R., & Blume, J. (2020). A robust effect size index. In *Psychometrika* (Vol. 85, p. 232). Publisher: Springer.

Vandekar, S. N., Satterthwaite, T. D., Rosen, A., Ciric, R., Roalf, D. R., Ruparel, K., … Shinohara, R. T. (2018). Faster family-wise error control for neuroimaging with a parametric bootstrap. *Biostatistics*, *19*(4), 497–513.

Vandekar, S. N., Satterthwaite, T. D., Xia, C. H., Adebimpe, A., Ruparel, K., Gur, R. C., … Shinohara, R. T. (2019). Robust spatial extent inference with a semiparametric bootstrap joint inference procedure. *Biometrics*, *75*, 1145–1155.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05". *The American Statistician*, *73*, 1–19.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, *92*, 381–397.

Woo, C.-W., Krishnan, A., & Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage*, *91*, 412–419.

Yeung, A. W. K. (2018). An updated survey on statistical Thresholding and sample size of fMRI studies. In *Frontiers in Human Neuroscience* (Vol. 12). Publisher: Frontiers.

Zang, Y.-F., Yong, H., Chao-Zhe, Z., Qing-Jiu, C., Man-Qiu, S., Meng, L., … Yu-Feng, W. (2007). Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI. *Brain and Development*, *29*(2), 83–91.

How to cite this article: Vandekar SN, Stephens J. Improving the replicability of neuroimaging findings by thresholding effect sizes instead of *p*-values. *Hum Brain Mapp*. 2021;42: 2393–2398. https://doi.org/10.1002/hbm.25374