VIEWPOINT

Effect Size Estimation in Neuroimaging

Marianne C. Reddan, MA
Department of Psychology and Neuroscience, University of Colorado, Boulder.

Martin A. Lindquist, PhD
Department of Biostatistics, John Hopkins University, Baltimore, Maryland.

Tor D. Wager, PhD
Department of Psychology and Neuroscience, University of Colorado, Boulder; and Institute of Cognitive Science, University of Colorado, Boulder.

Corresponding Author: Tor D. Wager, PhD, Department of Psychology and Neuroscience, University of Colorado, Boulder, 345 UCB, Boulder, CO 80309 (tor.wager@colorado.edu).

A central goal of translational neuroimaging is to establish robust links between brain measures and clinical outcomes. Success hinges on the development of brain biomarkers with large effect sizes. With large enough effects, a measure may be diagnostic of outcomes at the individual patient level. Surprisingly, however, standard brain-mapping analyses are not designed to estimate or optimize the effect sizes of brain-outcome relationships, and estimates are often biased. Here, we review these issues and how to estimate effect sizes in neuroimaging research.

Effect size is a unit-free description of the strength of an effect, independent of sample size. Examples include Cohen *d*, Pearson *r*, and number needed to treat.[1,2] For a given sample size (N), these can be converted to a *t* or *z* score (eg, Cohen *d* is $t/[N]^{1/2}$). But *t*, *z*, *F*, and *P* values are sample size dependent and relate to the presence of an effect (statistical significance), not its magnitude. By contrast, effect size describes a finding's practical significance, which determines its clinical importance. This is an important distinction because small effects can reach statistical significance given a large enough sample, even if they are unlikely to be of practical importance or replicable across diverse samples.[3]

Traditional neuroimaging studies are not designed to estimate effect sizes. A typical analysis tests for effects at each of 50 000 to 350 000 brain voxels. Post hoc effect sizes are selectively reported for a small subset of significant voxels. This practice creates bias, making effect size estimates larger than their true values.[4] It is like a mediocre golfer who plays 5000 holes over the course of his career but only reports his 10 best holes. Bias is introduced because the best performance, selected post hoc, is not representative of expected performance.

The **Figure** shows a simulation in which the true effect size in a set of voxels is $d = 0.5$. Once noise is added and a statistical test (*t* test) is conducted across 30 individuals, all significant voxels have an estimated effect size greater than the true effect. Why does this occur? Voxels tend to be significant if they show a true effect and have noise that favors the hypothesis. Correcting for multiple comparisons reduces false positives but actually increases this optimistic bias.[6] As statistical thresholds become more stringent, an increasingly small subset of tests with favorable noise will reach significance, making the estimated post hoc effect size grow. In sum, conducting a large number of tests inherently induces selection bias, which invalidates effect size estimates.

To overcome selection bias, we must reduce the number of statistical tests performed. One solution is to test a single, predefined region of interest. However, it is rare to consider only 1 region and discard valuable data. In addition, many symptoms and outcomes of interest are increasingly thought to be distributed across brain networks.[5] It can also be tempting to redefine the boundaries of regions of interest post hoc after looking at the results—a form of *P* hacking that invalidates both hypothesis tests and effect size estimates.

An alternative approach is to integrate effects across multiple voxels into 1 model of the outcome, which is then tested on new observations (ie, new patients). Instead of testing each voxel separately, associations with clinical outcomes are combined into a single model, and a single prediction is made for each patient. This approach is common in clinical research; for example, multiple factors, like diet, exercise, and hormone levels, are combined into models of disease risk. Neuroimaging models are based on voxels or network measures rather than risk factors, but the principle is the same. As long as (1) the model makes a single prediction for each patient and (2) predictions are tested on patient samples independent of those used to derive the model, then effect size estimates are unbiased.

A growing number of studies use machine learning and multivoxel pattern analysis to integrate brain information into predictive models. Effect sizes are assessed via prospective application of the model to new, "out-of-training-sample" patients, often using an iterative strategy of training and testing on different subsets of patients, known as cross-validation[7] (see Chang et al,[5] for example). There are ways that cross-validation can fail, and it is possible to overfit a cross-validated data set by training many models and picking the best. However, if a model is tested prospectively on new, independent data sets without changing its parameters, then unbiased estimates of effect sizes can be obtained. Bias, or lack thereof, can also be assessed with permutation tests.

Because integrated models combine information distributed across the brain in an optimized way, these models can substantially outperform single regions in predicting outcomes (Figure, C [adapted from data in Chang et al[5]]). Thus, such models provide a promising way to establish meaningful associations between brain measures and clinically relevant outcomes.

ARTICLE INFORMATION

Published Online: January 11, 2017. doi:10.1001/jamapsychiatry.2016.3356

Conflict of Interest Disclosures: None reported.

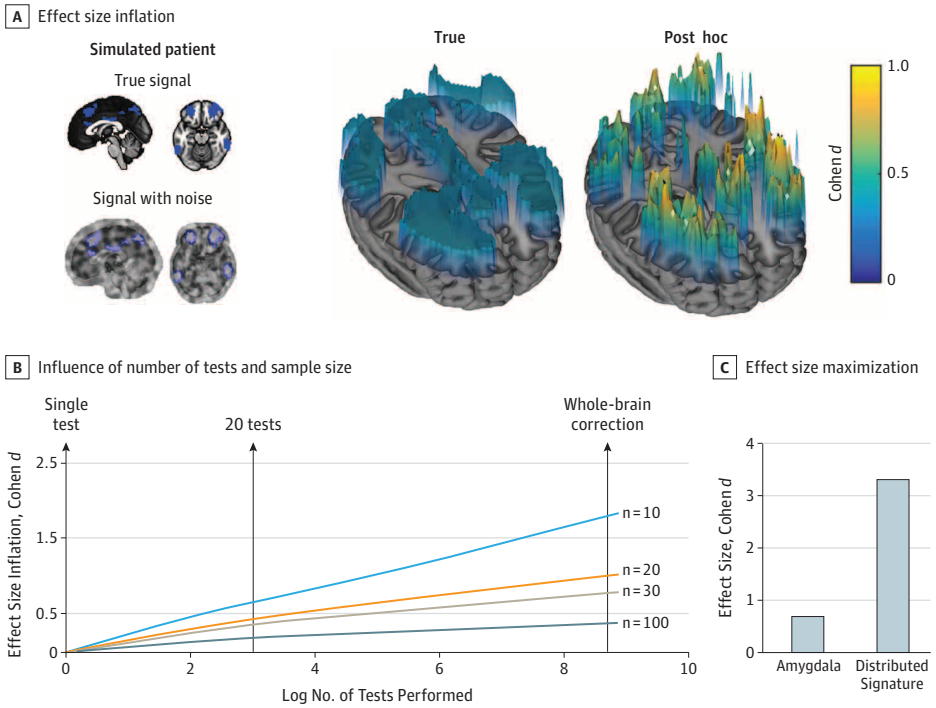
REFERENCES

1. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front Psychol*. 2013;4:863.

2. Grissom RJ, Kim JJ. *Effect Sizes for Research: Univariate and Multivariate Applications*. New York, NY: Routledge; 2012.

3. Button KS, Ioannidis JP, Mokrysz C, et al. Power failure: why small sample size undermines the

Figure. Inflation of Post Hoc Effect Sizes and Strategies for Maximizing Valid Estimates



A, Left panel: Top, Brain regions with true signal in one patient. Voxels in blue were assigned a true, moderate effect size of Cohen *d* = 0.5. Below, signal plus simulated noise; independent noise was added for each patient. Right panel: Results from a group *t* test (N = 30). Left, the true effect size. Right, post hoc effect sizes from significant voxels ($P < .001$ uncorrected). As expected, the estimated effect size of every significant voxel was higher than the true effect size. B, Expected effect size inflation for the maximal effect size across a family of tests. This bias, shown here using Monte Carlo simulation (Gaussian noise, 10 000 samples), increases as a function of the log number of tests performed and is approximated by the extreme value distribution (EV1). Effect size inflation increases as both the number of tests increases and the sample size decreases. C, Machine learning can maximize valid effect size estimates. The effect size for the difference between viewing negative and neutral images for an amygdala region of interest from the SPM Anatomy Toolbox version 2.2c (*d* = 0.66) and for a whole-brain multivariate signature optimized to predict negative emotion (*d* = 3.31). The test participants (N = 61) were an independent sample not used to train the signature, so the effect size estimates are unbiased. Adapted from data in Chang et al.[5]

reliability of neuroscience. *Nat Rev Neurosci*. 2013;14(5):365-376.

4. Lindquist MA, Mejia A. Zen and the art of multiple comparisons. *Psychosom Med*. 2015;77(2):114-125.

5. Chang LJ, Gianaros PJ, Manuck SB, Krishnan A, Wager TD. A sensitive and specific neural signature for picture-induced negative affect. *PLoS Biol*. 2015;13(6):e1002180.

6. Ioannidis JPA. Why most discovered true associations are inflated. *Epidemiology*. 2008;19(5):640-648.

7. Hastie T, Tibshirano R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Stanford, CA: Springer; 2008.