NeuroImage 147 (2017) 952–959

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Is the statistic value all we should care about in neuroimaging?

Gang Chen*, Paul A. Taylor, Robert W. Cox

Scientific and Statistical Computing Core, National Institute of Mental Health, National Institutes of Health, Department of Health and Human Services, USA

ABSTRACT

Here we address an important issue that has been embedded within the neuroimaging community for a long time: the absence of effect estimates in results reporting in the literature. The statistic value itself, as a dimensionless measure, does not provide information on the biophysical interpretation of a study, and it certainly does not represent the whole picture of a study. Unfortunately, in contrast to standard practice in most scientific fields, effect (or amplitude) estimates are usually not provided in most results reporting in the current neuroimaging publications and presentations. Possible reasons underlying this general trend include (1) lack of general awareness, (2) software limitations, (3) inaccurate estimation of the BOLD response, and (4) poor modeling due to our relatively limited understanding of FMRI signal components. However, as we discuss here, such reporting damages the reliability and interpretability of the scientific findings themselves, and there is in fact no overwhelming reason for such a practice to persist. In order to promote meaningful interpretation, cross validation, reproducibility, meta and power analyses in neuroimaging, we strongly suggest that, as part of good scientific practice, effect estimates should be reported together with their corresponding statistic values. We provide several easily adaptable recommendations for facilitating this process.

1. Introduction

Just as cartography requires a balance to be struck between the loss of important detail and the exactitude of a map that has “the scale of a mile to the mile” (Carroll, 1889), so too science requires careful extraction and summarization following an experiment. In other words, to present concisely the important components of the data and analyses, an investigator reports the experiment and makes a generalized conclusion based on some supporting evidence: a small condensed set of numbers. The crucial question is: How much or to which extent should the investigator compress the information without sacrificing too much? There are arbitrary choices that have to be made, but there are some definite thresholds under which loss of information is too great for optimal utility.

For example, in a typical statistical analysis, two quantitative results are produced for each effect of interest: the estimation for the amplitude of the effect itself (e.g., a β value from regression analysis or GLM) and the associated statistic (e.g., t or z). The former provides the magnitude of a physical measurement, which is the essence of scientific investigation, while the latter offers statistical substantiation for the effect estimate in the form of a significance level (or confidence interval, the implied range that may contain the effect estimate with a certain likelihood). While the relationship between the two quantitates is tight, each conveys distinct information about the result of the experiment; in most scientific disciplines, it is considered unacceptable if only significance is reported (Sullivan and Feinn, 2012): the statistic value serves as auxiliary evidence for the existence of the targeted effect, and it is the effect estimate itself that is the center of investigation as the physical property of interest. For example, suppose that physicists would like to validate the predictions of the general relativity (Einstein, 1915) by investigating the gravitational waves from the merger of two black holes. It would be hard to imagine that they would only report a statistical value or the significance of their measurement (e.g., a chance of 1 event per 203,000 years, or a significance level of 3.4×10^{-7}), but that they would not reveal the strength of the signal they have detected (a peak gravitational-wave strain of 1.0×10^{-21} in the frequency range of 35 to 250 Hz) (Abbott et al., 2016).

However, within the field of neuroimaging, it has remained the predominantly common practice to report only statistical mapping tests in publications and presentations, a custom which has been largely (and perplexingly) immune to critical scrutiny. For instance, one typically sees brain results provided as blobs whose color spectrum corresponds to t- or z-values (or occasionally to p-values), and most of the time the underlying degrees of freedom are left out, rendering the statistics even harder to interpret. Similarly, in tabulated results for brain regions, standard reports usually contain the coordinates and statistic value at a single peak voxel (which is itself defined, again, as the maximum of the statistical values, not of the effect estimates, within

* Corresponding author.
E-mail address: gangchen@mail.nih.gov (G. Chen).

http://dx.doi.org/10.1016/j.neuroimage.2016.09.066
Received 7 July 2016Accepted 29 September 2016
Available online 10 October 2016
1053-8119/ Published by Elsevier Inc.

the region), and the effect estimate at such a peak voxel is rarely reported. The same phenomenon commonly occurs in reporting results of seed-based correlation analyses for resting-state data, where the brain maps and tables usually show the statistic (often *z*) values instead of and without including inter-regional correlations.

Recently there have been a number of discussions about the use and misuse of *p*-values in the scientific community (e.g., Wasserstein and Lazar, 2016; Nuzzo, 2014), and others have been more critical of the "cult" or "obsession" of statistical significance (e.g., Ziliak and McCloskey, 2009). The editors of the journal, Basic and Applied Social Psychology, have gone so far as to take the seemingly extreme step as to no longer accept papers with *p*-values due to the concern of the statistics being used to support lower-quality research (Trafimow, 2014). In a sense, our concern here is related, and addressing it would also alleviate many of these other topical issues, but the concern is specifically focused on the need for including the effect estimate in neuroimaging studies. To frame the discussion here, we quote the six guiding principles on *p*-values in a recent statement released by The American Statistical Association (ASA) (Wasserstein and Lazar, 2016):

1. " *p*-values can indicate how incompatible the data are with a specified statistical model.
2. *p*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis."

We believe that the neuroimaging field needs to move forward to promote the reportage of the effect estimates along with the corresponding statistics. We first discuss the statistical terms in the context of FMRI analyses, highlighting specific features related to that field. We then argue that full reporting in FMRI is necessary and promotes good scientific practice, clarity, increased reproducibility, cross-study comparability and allows for proper meta and power analyses. Finally, we provide several recommendations for researchers and software designers to facilitate these "best practices" actions.

2. What is the effect estimate in neuroimaging?

In neuroimaging, the ultimate focus is on the physical evidence for the brain's neuronal response, which evidence is typically embodied in the strength of the FMRI BOLD signal. For task-related experiments, the response strength is reflected in the effect estimate (or β value) associated with a task/condition or with a linear combination of β's from multiple tasks, such as the contrast between two tasks. For seed-based correlation analyses with resting-state data, time series correlation captures the relationship between a seed and the rest of the brain. Similarly, for naturalistic scanning, one measure is the "inter-subject correlation" (ISC) at a region that features the synchronization or similarity among subjects (Hasson et al., 2004). It is worth noting that, in typical multivariate pattern analysis (Haxby et al., 2001), the sensitivity measure showing the percentage of cases in which a classifier makes correct predictions is not an effect estimate, but it is a metric that combines the size of the effect (i.e., how discriminable the experimental conditions are) with the statistical reliability with which the effect is estimated (i.e. the noise level on the activity patterns). Similarly, some model-based methods have been adopted to account for rich sets of FMRI measurements in fields such as vision studies. Even though an effect estimate in the conventional sense cannot be defined under such scenarios, the proportion of variance in the data that could be accounted for by a model (Kay et al., 2013) or by a representational similarity matrix (Khaligh-Razavi and Kriegeskorte, 2014) can effectively serve as a physical metric that characterizes the model performance.

Here, we use the term "effect estimate" to refer generally to any of these or analogous cases: the estimated response magnitude (e.g., β value) of a regression model or GLM, the estimated correlation coefficient in the context of correlation analyses, etc. We note that in the statistical literature, the phrase "effect size" can typically encompass two distinct scenarios: one for describing absolute (or unstandardized) effect size (the estimated magnitude of an effect under investigation, e.g., sample mean or the estimated β in a regression model), and the other for describing standardized effect magnitude (e.g., Cohen's *d*), which is typically used when the measurement units have no intrinsic meaning (e.g., Likert-type scale adopted in survey research), when a comparison is performed between two different scales (e.g., relative effect sizes among different confounders such as age and sex), or when data variability is the focus of study (Sullivan and Feinn, 2012). While it is well known that the acquired BOLD signal has only arbitrary units, it might seem that the second usage of effect size is a good candidate. However, FMRI data are commonly scaled to a more meaningful evaluation in terms of percent signal change (as discussed further below). As such, here we use the term "effect estimate" in FMRI to refer to the unit-bearing case of "effect size" in the context of percent signal change.

3. What does a *t*-statistic value reveal in neuroimaging?

A *t*-statistic value for an effect estimate is calculated as the latter divided by its standard error, which represents the reliability or accuracy of the effect estimate. Thus, the *t*-statistic is a mixture of two components, the effect estimate and the noise estimate. However, both components vary across the brain. For example, the variability of BOLD response may partially result from the inhomogeneity of vascularization, and to some extent the variability of the noise level may be caused by the heterogeneous sensitivity profiles of RF coils across the brain. The combined impact from the two components makes the *t*-statistic unsuitable for comparing effects across regions, subjects, groups, or studies, which is one of the reasons that FMRI group analysis is typically performed on the effect estimate, not the *t*-statistic. In addition, as a dimensionless measure, the *t*-statistic is more susceptible to sample size (number of trials or subjects), signal-to-noise ratio (SNR), preprocessing steps/methods, experimental designs, unexplained confounds, and scanner parameters than the effect estimate itself. Therefore, statistic values only serve the purpose of a binary inference of null (e.g., there is no difference between the two conditions) versus alternative (e.g., there is difference between the two conditions) hypotheses, and it does not provide any information about the specific response magnitude. For example, two voxels (or regions) with the same *t*-statistic value in the brain do not mean the same response amplitude, and *vice versa* (Fig. 1). That is to say, the *t*-statistic does not carry enough interpretation information for the effect of interest.

4. Practical realities/difficulties of FMRI

There are several features inherent to FMRI acquisition and analysis that present challenges to an investigator interpreting and reporting results. At first glance, some of these may seem to explain the present practices of reporting only statistic values as results. We describe them briefly here, and then discuss how they actually necessitate, rather than discourage, the inclusion of effect estimates in the end.