An Open Publishing Platform of the Organization for Human Brain Mapping
APERTURE NEURO

ORIGINAL RESEARCH ARTICLE

On the statistics of brain/behavior associations

Bertrand Thirion

Inria, CEA, Université Paris Saclay, France

The paper by Marek et al. published in *Nature*, 2022, addresses a question central to brain–behavior association studies, or more precisely, brain-wide association studies: namely, how much data are necessary to carry out statistically meaningful inference and the importance of replicating effects on independent cohorts. These questions are of wide interest, as they represent the first one a practitioner has to solve when running such a study.

As a second thought, one may wonder why such questions are addressed in a high-impact journal 30 years after the beginning of brain imaging, which has a particularly rich history of statistical contributions. The straightforward reason is opportunistic, that is, that the study has been driven by the availability of at least three large-scale cohorts (HCP, UK Biobank, and ABCD).

Still, one has to clarify whether the paper merely describes known statistical facts that had yet not been made fully explicit in the literature and thus had to be stated explicitly as they are? Did not the authors simply miss contributions that have been made in the field?

1. STATISTICS 101 FOR NEUROIMAGERS

It is obviously hard to disagree with the main message of the paper, summarized in the title, namely that neuroimaging group studies indeed require large samples to yield valid conclusions. Sample size effects are trivial, in the sense that they are the straightforward implication of the law of large numbers – studying them in the framework of brain imaging can only count as an illustration of the most basic statistical results. Large sample sizes bring power gains, that imply gains in reproducibility, in a framework where type-I error control has remained the standard to declare that associations are significant. More importantly, it will bring better estimates of effect sizes, alleviating the *winner's curse* (1). There is nothing to object to that – or merely the observation that all this had already been discussed in several classical works of the field, which Marek et al. fail to acknowledge: see for example (1–3) for univariate aspects and (4, 5) for multivariate aspects.

As a corollary, a finding that is not reproduced in a second cohort may not be meaningful. Or, put differently, drawing statistical conclusions from effect size on a discovery cohort entails some peculiar corrections, as this is a kind of post hoc inference (6). Note that lenient statistical thresholds ($p < 0.05$ uncorrected) should not be used, as they lead to unacceptable false discovery proportion, see below).

There is at least one aspect of the problem that is uniquely relevant to neuroimaging, namely that effect sizes (association with fluid intelligence, see Fig. 2 from Marek et al., 2022) are reproduced across studies. Unfortunately, this is asserted by Marek et al. on the bulk of brain-wide associations, most of which likely represent null effects. Effect size analysis should be carried out on non-null effects.

From this first view, one can conclude that the main merit of the paper is to make explicit and display graphically facts that had been identified previously, but with the pedagogical virtue of a smoking gun.

2. SOME NECESSARY UPDATES TO STATISTICAL ANALYSIS TOOLBOX

Nevertheless, I have some objections about the contribution. Namely that the paper is stuck in an old-fashioned correlation framework that needs to be overcome if one wants to build a science of brain–behavior associations.

First, the authors lazily report effect sizes as correlations. I dare to call that "lazy," as this setting forgoes the determination of directed links (A explains B rather than the converse), that is, a causal structure underlying the data. Population neuroscience is not going to yield any valid scientific insights if the community does not pay the effort to make such links explicit.

This will require more detailed conceptual and statistical settings, following the treatment effect theory for observational studies (7): certain variables can be considered as treatment, others cannot, some of them are confounders, some of them mediate the effect of others, some can be taken as instruments for unobserved

conditions, and some of them are proxy for unmeasured variables. Neuroimaging studies have to pay attention to these relationships. A major conceptual shift is thus needed, which is exactly the step that goes from data-driven analysis to science.

As a corollary of this comment, effect size should be given as treatment units (1 mm of cortical thickness yields gains of XX fluid intelligence units) or nondimensional Cohen's *d*, if meaningful units are not easily identified, but not as correlations. A noticeable benefit of this change is that effect sizes will have a more standard statistical behavior than correlations that require cumbersome *z*-transforms. Another obvious benefit will be to open the possibility to reuse actual effect sizes across studies, for instance, to build priors.

Second, the reliance on arbitrary type-I error control thresholds is not satisfactory. Researchers should instead keep an eye on the false discovery proportion – which becomes obviously easier with large same sizes. Doing so would quickly make it clear that some thresholds yield an unacceptable proportion of false detections. Use of adequate post doc corrections is necessary when this inference is carried out on the discovery data set, see for example (8).

Third, brain–behavior associations should target external validity explicitly. Indeed, the potential added value of imaging on top of cheaper behavioral assessments can only be grounded in the additional information it yields about the participants (mental or physical) health or on their cognitive abilities. This is done by issuing (and assessing the accuracy of) prediction. In that respect, the Marek et al. paper is quite confusing, since it calls Canonical Correlation Analysis (CCA) a predictive method, which it is not, as it only measures an in-sample correlation between multiple variables. Besides, for lack of proper explanation in the manuscript, it is unclear what the authors call "in-sample association" obtained by Support Vector Regression or CCA. It could well be what people call "overfit" in the machine learning literature. Overall, the paper seems to miss the point regarding predictive modeling.

A related, yet more technical difficulty concerns multivariate feature weights. The authors do not use state-of-the-art estimators to get these weights, which certainly leads to pessimistic reproducibility estimates. Indeed, much stronger estimation schemes have been proposed in the framework of neuroimaging that improve significantly classifier weights (local regularization, bagging, see (9), and that furthermore come with statistical guarantees and more power than nonparametric analysis of classifiers weights (10). The authors' statistical toolbox is thus outdated.

To conclude this comment, reliance on large-scale data is an obvious need for the community, which mandates the generalization of open data. In particular, the benefits of having exploration and inference data sets can only hold if some public data are available. But this is not enough: a methodological update is required to go from a collection of inconsistent results (what neuroimaging presents today) toward a science of brain–behavior relationship. For this, much more elaborate reasoning on the status of the measurements is necessary as well as proper predictive frameworks.

REFERENCES

1. Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., et al. Power failure: Why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 2013;14(5):365–376.
2. Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., Poline, J.B. Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. NeuroImage 2007;35(1):105–120.
3. Thyreau, B., Schwartz, Y., Thirion, B., Frouin, V., Loth, E., Vollstädt-Klein, S., et al. IMAGEN Consortium. Very large fMRI study using the IMAGEN database: Sensitivity-specificity and population effect modeling in relation to the underlying anatomy. NeuroImage 2012;61(1):295–303.
4. Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. NeuroImage 2017;145(Pt B):166–179.
5. Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. NeuroImage 2018;180(Pt A):68–77.
6. Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., Baker, C.I. Circular analysis in systems neuroscience: The dangers of double dipping. Nat Neurosci 2009;12(5):535–540.
7. Rubin, D.B. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 1974;66:688–701.
8. Rosenblatt, J.D., Finos, L., Weeda, W.D., Solari, A., Goeman, J.J. All-resolutions inference for brain imaging. NeuroImage 2018;181:786–796.
9. Hoyos-Idrobo, A., Varoquaux, G., Schwartz, Y., Thirion, B. FReM – Scalable and stable decoding with fast regularized ensemble of models. NeuroImage 2018;180(Pt A):160–172.
10. Chevalier, J.A., Nguyen, T.B., Salmon, J., Varoquaux, G., Thirion, B. Decoding with confidence: Statistical control on decoder maps. NeuroImage 2021;234:117921.

