*Psychoradiology*, 2022, **2**, 67–68
DOI: 10.1093/psyrad/kkac010
**Commentary**

# Scanning reproducible brain-wide associations: sample size is all you need?

Xiang-Zhen Kong[1,2,*], Chenghui Zhang[1], Yinuo Liu[1] and Yi Pu[3,*]

[1]Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou 310028, China
[2]Department of Psychiatry of Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 310016, China
[3]Department of Neuroscience, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main 60322, Germany
*Correspondence: Xiang-Zhen Kong, xiangzhen.kong@zju.edu.cn; Yi Pu, yi.pu@ae.mpg.de

Back to 1874, John Hughlings Jackson, a pioneer in neurology who had profound knowledge about the structure of the brain, posited that "difference of [brain] structure of necessity implies difference in function". Since then, understanding the relationship between the brain structure and function in humans has become an important endeavor in neuroscience research. It is important, because it is regarded to be helpful for not only guiding clinical practice in neuropsychiatric and other brain disorders, but also providing a deeper understanding of what makes us what we are. In the past decades, with the development of new techniques, such as magnetic resonance imaging, significant achievements have been made in understanding the relationship between brain structure and function. For instance, researchers take advantage of individual variations observed in both brain imaging and behavioral measures, and examine their correlations across people. This line of research has yielded very fruitful findings (e.g. Genon *et al.*, 2022). In recent years, however, accumulating evidence has started to challenge the reproducibility of these findings with current research practice. One of the most discussed problems regards the typical sample sizes (e.g. 20–30) (e.g. Button *et al.*, 2013), which limits the statistical power and causes reproducibility issues (e.g. Boekel *et al.*, 2015; Kong *et al.*, 2022).

A recently published article (Marek *et al.*, 2022) has put the sample size issue under the spotlight again. Taking advantage of cohort-level brain imaging datasets including the Adolescent Brain Cognitive Development (ABCD, $N = 11\,874$; 9–10 years), the Human Connectome Project (HCP, $N = 1200$; 22–35 years), and the UK Biobank (UKB, $N = 35\,735$; 40–69 years), the authors aimed to provide an accurate estimation of the effect size of the correlation between brain features (e.g. regional cortical thickness and functional connectivity) and behavioral phenotypes (e.g. cognitive ability and mental health). Since such large-scale datasets allowed randomly subsampling different numbers of individuals (from 25 to thousands) for smaller brain–behavioral association analyses, the authors also charted the observed effects sizes and reproducibility as a function of sample size.

The authors demonstrated that, in the full rigorously denoised ABCD samples ($N = 3928$), the brain–behavior correlations were very small, with the top 1% largest effects (of all ~11 million associations) being just >0.06 (median = 0.01). This effect size was far smaller than previously thought at ~0.20–0.80. Given the multi-site nature of the ABCD data (multiple scanner types), the authors further validated the results in single-site, single-scanner-type data in the HCP and the UKB. Size-matched samples selected from the three datasets ($N = 900$) showed similar effect size distributions (top 1% at ABCD >0.11; HCP >0.12; UKB >0.09). The subsampling of different numbers of individuals (from 25 to >3000) showed varying statistical errors (e.g. false negative rate and statistical power) and reproducibility rate of brain–behavior correlations. A standard power analyses given the effect size ($r > 0.06$) and typically used significance testing threshold ($P = 0.05$) suggested that a sample size of 2200 is required for achieving a power of 80% (Marek *et al.*, 2022).

This study provides a large-scale estimation of correlations between brain imaging and behavioral phenotypes, and calls for a rethink of common practice for more reproducible research. However, it is worrying that this article has been taken out of context and mistakenly interpreted by some news reports and readers. For instance, a commentary article published along with Marek *et al.*'s paper called this study a "bombshell study," and suggested that existing studies linking brain imaging features to traits such as cognitive ability and disease symptoms could be "junk" (Callaway, 2022). Another example suggested that "small neuroimaging association studies just generate noise" (see https://www.nature.com/articles/s41586-022-04492-9/metrics for more discussions). It was also suggested that funding agencies and journals "should be wary of" funding such projects and papers with sample sizes fewer than several hundred in a related commentary (Gratton *et al.*, 2022).

Such statements from news reports are eye-catching, and manuscripts submitted to journals are already being rejected simply because of "small sample size" compared to the proposed "thousands." While we recognize the many benefits of larger sample studies (e.g. increased statistical power and reduced false positives), we emphasize that caution is warranted when applying the results from Marek *et al.* (2022) to individual studies. This is because it is unknown to what extent the sample size estimations could apply to one particular study. In Marek *et al.* (2022), the implications about thousands of samples were based on the estimated distribution of effect sizes, which was based on two brain imaging features (thickness and functional connectivity) and 41 phenotype variables. This is far from representative of all possible measures used in human neuroscience.

**Received:** 7 September 2022; **Accepted:** 17 September 2022

Larger sample sizes can indeed result in stronger statistical power. In the meantime, however, a sample size of thousands incurs an extremely high financial cost, which is unrealistic for most researchers. Sometimes this could also result in increased ethical risk. In this commentary, we would like direct our readers' attention back to another fundamental contributing factor to statistical power, that is, the effect size per se. One might argue that the effect size (or true effect size) is constant. This is only true when the variables of interests had a perfect reliability, which is far from reality. It is important to note that the expected effect size (or practical effect size) largely depends on the measurement reliability of the variables. Mathematically, higher reliability of one variable is linked with smaller within-subject variability (measurement stability) and larger inter-subject variability (individual differences) (Zuo *et al.*, 2019). To reduce the within-subject variability, scientists should focus on developing more reliable measures of the brain and/or behaviors. In addition, researchers should also try to maximize the inter-subject variability. For instance, it would not be a good idea to focus on cross-sectional samples at the age of 9–10 years or on young adults of 22–35 years (although sometimes we may be interested in such ages) to investigate the association between age and brain structure, as was done in the simulations by Marek *et al.* (2022). The results would not reflect the truth regarding the brain development and at the same time generalization of conclusions will be limited. This rule also applies to studies of other associations.

To summarize, large-sample-size studies are appealing from multiple aspects, however, increasing the sample size is not the only solution to increase the reproducibility of research. Effect size is also an important factor to be considered. Scientists should make good use of their knowledge in a certain field to increase the expected effect sizes, e.g. via developing more sensitive experimental measurements and adopting better participant sampling strategies.

Besides, attempts at methodological innovation should be encouraged. One way to achieve this is to learn from other relevant fields. In Marek *et al.* (2022), the authors coined the term "brain-wide association study" for brain–behavioral relationship studies, which was named after the genome-wide association study. There are many methodological aspects the brain-wide association study could gain from the genome-wide association study besides the sample sizes of tens of thousands. These include unified and reliable genotypes, multi-site collaboration, multi-level analysis of univariate association (e.g. gene-based analysis, gene-set analysis, and polygenic risk scores), and, more importantly, the publicly available and continuously updated databases (e.g. gene ontology). In addition, we highlight advanced synthesis approaches such as meta- and mega-analysis that could make good use of existing public resources (e.g. UK Biobank and OpenNeuro) and multi-site collaborative datasets (e.g. via ENIGMA) (Thompson *et al.*, 2020). With all these improvements in the future, there is promise to open a new horizon for understanding the brain–behavior relationship in both health and disease.

## Author Contributions

X.K. & Y.P.: conception, preparing the first draft and editing; C.Z. & Y.L.: editing.

## Conflict of Interest

The authors declare no competing interests.

## Acknowledgements

Xiang-Zhen Kong was supported by the Fundamental Research Funds for the Central Universities (2021XZZX006), the National Natural Science Foundation of China (32171031), and the Information Technology Center of Zhejiang University.

## Reference

Boekel W, Wagenmakers E-J, Belay L, *et al.* (2015) A purely confirmatory replication study of structural brain-behavior correlations. *Cortex* **66**:115–33.

Button KS, Ioannidis JPA, Mokrysz C, *et al.* (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* **14**:365–76.

Callaway E (2022) Can brain scans reveal behaviour? Bombshell study says not yet. *Nature* 777–8.

Genon S, Eickhoff SB, Kharabian S (2022) Linking interindividual variability in brain structure to behaviour. *Nat Rev Neurosci* **23**:307–18.

Gratton C, Nelson SM, Gordon EM (2022) Brain-behavior correlations: two paths toward reliability. *Neuron* **110**:1446–9.

Kong X-Z, ENIGMA Laterality Working Group, Francks C (2022) Reproducibility in the absence of selective reporting: an illustration from large-scale brain asymmetry research. *Hum Brain Mapp* **43**:244–54.

Marek S, Tervo-Clemmens B, Calabro FJ, *et al.* (2022) Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**:654–60.

Thompson PM, Jahanshad N, Ching CRK, *et al.* (2020) ENIGMA and global neuroscience: a decade of large-scale studies of the brain in health and disease across more than 40 countries. *Transl Psychiatry* **10**:1–28.

Zuo X-N, Ting X, Michael PM (2019) Harnessing reliability for neuroscience research. *Nat Hum Behav* **3**:768–71.