Reproducible univariate brain-wide association studies do not necessarily require thousands of individuals

Luca Cecchetti[1] and Giacomo Handjaras[1]

1. Social and Affective Neuroscience (SANe) Group, MoMiLab, IMT School for Advanced Studies Lucca, Lucca, Italy

In brain-wide association studies (BWAS), researchers correlate behavior with the inter-individual variability in functional or structural properties of multiple brain regions. Marek, Tervo-Clemmens and colleagues[1] (hereinafter, M&TC) estimate statistical power of BWAS on the three largest available neuroimaging datasets (i.e., ABCD[2], Human Connectome Project[3] - HCP -, and UK Biobank[4]) and demonstrate that reproducible results require sample sizes in the order of thousands. Here, by performing power calculations on HCP and synthetic data, we show that reproducible univariate BWAS does not necessarily require thousands of individuals and that their calculation of power overestimates the sample size needed to detect region-specific effects.

In their paper, M&TC collate data from three large studies and use random resamplings at increasing sample sizes (ranging from 25 to approximately 4,000 individuals) to assess the reproducibility of brain-behavior correlations. Statistical significance is assessed across a range of alpha levels (from $p < 0.05$ uncorrected to $p < 0.05$ Bonferroni corrected) in the full sample (e.g., 3,928), which they consider a reference for the estimation of power at increasing sample sizes.

To summarize their procedure, we propose an example in which brain features (e.g., cortical thickness) of 4 distinct regions of interest (ROIs) correlate with behavior (e.g., cognitive abilities) and pass the Bonferroni corrected $p < 0.05$ threshold in the full sample (Figure 1A). Importantly, each one of these ROIs has a specific effect size (i.e., Pearson's correlation r), ranging from 0.16 (Figure 1A, green ROI) to 0.04 (Figure 1A, yellow ROI). In accordance with M&TC, random resamplings are performed at multiple sample sizes (we show hypothetical $n = 375$ in Figure 1A) and power is computed for each full-sample significant ROI as the percentage of random resamplings passing the statistical threshold. Thus, by doing so, they obtain a non-parametric estimate of statistical power at each sample size and ROI.

To support the claim that reproducible BWAS require thousands of individuals, M&TC compute the average power across ROIs at each sample size (line 53 in the function abcd_statisticalerorrs.m made available by the authors; Figure 3 of the original article). However, such a procedure disregards that larger correlations require fewer participants to be detected. In fact, one should consider region specific effects, as multiple significant ROIs in the full sample may be associated with behavior to a different extent. Therefore, to test the reproducibility of BWAS, instead of averaging power across ROIs, it would be more appropriate to consider ROI-specific power calculations.

In this regard, we use HCP data ($n = 1,096$) and test the relationship between age-adjusted cognitive ability scores (i.e., CogTotalComp_AgeAdj, the same variable used in M&TC) and Freesurfer[5] cortical thickness, and surface area across 68 ROIs[6].

We apply the same non-parametric approach of M&TC, and estimate average power across significant ROIs, as well as the power of the ROI with the largest effect size (i.e., ROI-specific power). When considering the relationship between cortical thickness and cognitive abilities, the maximum average power is 80% at $n \cong 935$ for the uncorrected $p < 0.05$ and 70% at $n = 1{,}000$ for the Bonferroni corrected $p < 0.05$ (Figure 1B, left panel). Conversely, using the ROI-specific power calculation the region showing the largest effect size (i.e., $r = 0.1167$) reaches 80% power at $n \cong 480$ for uncorrected $p < 0.05$, at $n \cong 870$ for $q < 0.05$ (False discovery rate correction[7]; FDR) and at $n \cong 1{,}000$ for Bonferroni corrected $p < 0.05$.

In many cases, covariates exert an effect on the relationship between brain and behavior. Thus, to test the performance of the two power calculation methods under these circumstances, we measure partial correlation between cortical surface area and cognitive abilities, accounting for parenchymal volume.

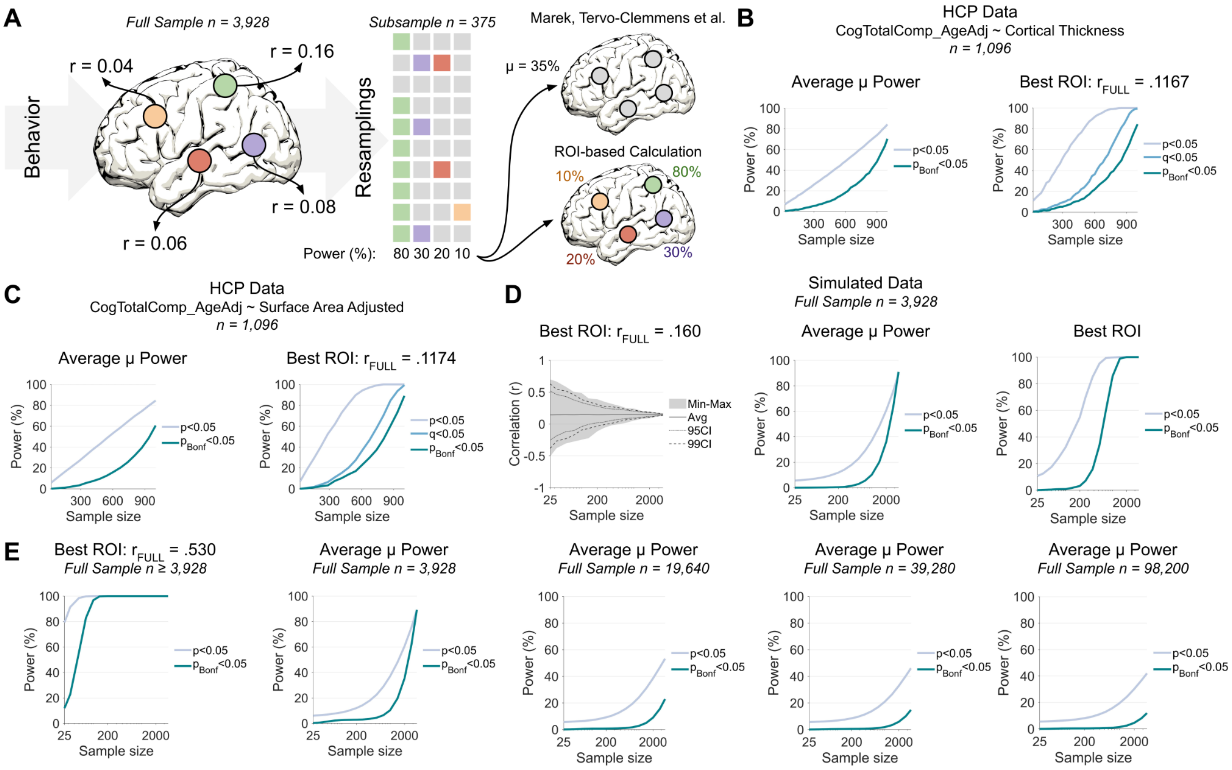

Figure 1. In **A**, we illustrate the analytical pipeline to assess reproducibility of univariate brain-wide association studies in Marek, Tervo-Clemmens and colleagues. At multiple sample sizes (e.g., $n = 375$), statistical power is computed for each ROI (i.e., green, purple, red and orange) of the full sample ($n = 3{,}928$) significantly associated with behavior. To represent statistical power as a function of sample size, the authors compute the average power across significant ROIs (i.e., average power method). We suggest an alternative ROI-based power calculation and test the reliability of the two methods across simulated and real data of the Human Connectome Project. In **B**, using average and ROI-based methods, we compute power as a function of sample size and alpha level for the relationship between cortical thickness and cognitive abilities in HCP data. In **C**, we compare the two methods in assessing power of the relationship between cortical surface area and behavior, also controlling for parenchymal volume. The left panel of **D** shows average correlation values (solid dark gray line) and their confidence intervals (CI95: dotted

lines; CI99: dashed lines; max-min: gray shaded area) at different sample sizes for r = 0.16 (i.e., the largest value that replicated out-of-sample in the original article). Central and right panels of **D** report the assessment of power at multiple sample sizes and alpha levels using the average and the ROI-specific methods, respectively. In **E**, we show that power estimates for the average method, but not for the ROI-specific one, depend on the size of the full sample.

Results show that 80% average power is reached at n ≅ 935 for $p < 0.05$ uncorrected, whereas the maximum average power for Bonferroni corrected $p < 0.05$ is ~61% at n = 1,000. Instead, the ROI showing the largest effect size (i.e., r = 0.1174) reaches 80% power at n ≅ 480 for the uncorrected $p < 0.05$, at n ≅ 870 for $q < 0.05$ and at n ≅ 935 for Bonferroni corrected $p < 0.05$. Of note, if one does not account for parenchymal volume, the correlation between the same ROI and behavior is r = 0.2657. This time the 80% ROI-specific power is reached at n ≅ 220 for Bonferroni corrected $p < 0.05$, while the average 80% power is reached at n ≅ 545 using the same alpha level (results not shown but reproducible using the code we made publicly available).

Thus, when power is computed at ROI level, reproducible BWAS require hundreds, not thousands, of individuals, even when effect sizes are relatively small and correction for multiple comparisons is applied (e.g., FDR).

Because of the differences in results obtained for the average and the ROI-specific power calculations, we further investigate the reliability of the two methods in detecting significant associations using synthetic data. Thus, we simulate a distribution of correlation values comparable to the one reported in Figure 1A of M&TC (Source Data Fig.1; sigma = 0.0336). We then generate synthetic behavioral scores correlated with data for 394 brain ROIs (i.e., the same number used in the original article) sampling from a multivariate normal distribution (mu = 0, sigma = 0.1). For one selected ROI, we impose a correlation of r ≅ 0.1600, which is the largest correlation that replicated out-of-sample in the M&TC article. The similarity between simulated and real M&TC data (see Figure 1F in the original article) can be appreciated by inspecting the average correlations and confidence intervals across sample sizes (Figure 1D, left panel).

In synthetic data, average (Figure 1D, central panel) and ROI-specific (Figure 1D, right panel) power calculations diverge. In particular, 80% average power is reached at n ≅ 2,800 and n ≅ 3,200 at uncorrected and Bonferroni corrected significance levels, respectively. Instead, using the ROI-specific method, 80% power is reached at n ≅ 375 and n ≅ 1,000 at uncorrected and Bonferroni thresholds, respectively.

Lastly, we estimate differences between the two methods in detecting medium/large effect sizes using synthetic data. Specifically, we impose a correlation of r ≅ 0.5300 between behavioral simulated data and a single ROI (out of 394). We select this particular value as it is detectable with 80% power in a sample of 25 participants (i.e., the smallest one tested by M&TC) using $p < 0.05$ uncorrected. All other correlations (i.e., between behavior and the remaining 393 ROIs) are comparable to those of Figure 1A in the M&TC article. Also, synthetic data allows for the simulation of a very large full sample size. M&TC use the full sample size to determine the ground truth brain-behavior correlations, as well as to select reference ROIs (i.e., those passing the significance threshold). Importantly, the same ROIs are used in average power calculation at multiple sample sizes using the resampling technique.

To assess whether the size of the full sample influences average and ROI-based power calculations, we generate synthetic brain and behavioral data for full samples of n = 3,928, 19,640 (5-fold increase), 39,280 (10-fold increase) and 98,200 (25-fold increase) individuals.

As expected, for $r \cong 0.5300$, ROI-based power calculation is not affected by the number of individuals in the full sample and 80% power is reached at $n \cong 25$ for $p < 0.05$ uncorrected and at $n \cong 70$ after adjusting for multiple comparisons using the Bonferroni method (Figure 1E, left panel).

Paradoxically, the average power decreases substantially as the full sample size increases: considering the subsample n = 3,604 (i.e., the largest explored in M&TC) and the threshold $p < 0.05$ uncorrected, average power drops from ~90% to ~45% with 25-fold increase in the full sample size (Figure 1E).

One possible explanation for this paradox is that any non-zero correlation becomes significant with a large enough full sample size. Thus, in addition to the ROI with medium/large correlation, the number of reference ROIs increases as a function of the full sample. However, the presence of multiple small, yet significant, effects decreases the estimate of the average power.

To support these claims we made publicly available our code at https://osf.io/2szkv/.

In summary, we show that computing the average power across ROIs may not be the optimal method to establish the minimum sample size needed to detect reproducible mass-univariate brain-behavior associations. ROI-based assessment of power reveals that BWAS do not necessarily require thousands of participants, even when correction for multiple comparisons is applied.

References

1. Marek, S., Tervo-Clemmens, B., Calabro, F.J. et al. Reproducible brain-wide association studies require thousands of individuals. Nature (2022). https://doi.org/10.1038/s41586-022-04492-9
2. Casey, B. J. et al. The Adolescent Brain Cognitive Development (ABCD) study: imaging acquisition across 21 sites. Dev. Cogn. Neurosci. 32, 43–54 (2018).
3. David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E.J. Behrens, Essa Yacoub, Kamil Ugurbil, for the WU-Minn HCP Consortium. (2013). The WU-Minn Human Connectome Project: An overview. NeuroImage 80(2013):62-79.
4. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12, e1001779 (2015).
5. Fischl B. (2012). FreeSurfer. NeuroImage, 62(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021
6. Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., ... & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage, 31(3), 968-980.
7. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1), 289-300.

Acknowledgement

Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.