Reproducible 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

Luca Cecchetti
IMT School for Advanced Studies Lucca
Piazza San Francesco, 19, 55100 Lucca - Italy
Email: luca.cecchetti@imtlucca.it
Ph: +39 0583 4326729

Giacomo Handjaras
IMT School for Advanced Studies Lucca
Piazza San Francesco, 19, 55100 Lucca - Italy
Email: giacomo.handjaras@imtlucca.it
Ph: +39 0583 4326729

In brain-wide association studies (BWAS), researchers correlate behavior with the inter-individual variability in functional or structural properties of distinct brain regions. Marek, Tervo-Clemmens, and colleagues[1] (hereinafter, M&TC) empirically assess statistical power, replication rate, type I error, sign error, and effect size inflation in BWAS using data from three large-scale neuroimaging initiatives (i.e., ABCD[2], Human Connectome Project[3] - HCP -, and UK Biobank[4]). Their results indicate that reproducible brain-behavior associations require thousands of observations. Here, leveraging synthetic and HCP data, we demonstrate that their calculations overestimate the sample size needed to detect reproducible effects by one order of magnitude.

In their article, M&TC assess the reproducibility of BWAS by randomly resampling (1,000 bootstraps with replacement) participants at increasing sample sizes (ranging from 25 to ~4,000). The correlation between brain features (e.g., cortical thickness) and behavior (e.g., cognitive ability) is computed at each sample size, allowing the empirical evaluation of power, type I error, sign error (i.e., type S error), inflation rate (i.e., type M error), and probability of replication as a function of the number of observations.
Their findings indicate that thousands of individuals are required for BWAS to detect true effects reliably. In addition, M&TC analyses reveal that studies on hundreds of participants - the typical size of BWAS - fail to replicate, produce inflated effect sizes, or even opposite sign associations (e.g., negative association when the actual correlation is positive).
We have read with great interest the M&TC article and value the authors' effort to perform billions of analyses aimed at testing BWAS reproducibility. However, we believe that their conclusions are based on a statistical procedure that biases sample size estimates and disregards region-specific effects: the averaging of metrics (e.g., power, replication rate) across brain parcels. Here, using synthetic and HCP data, we show that small effect sizes can be reliably detected with less than a thousand individuals when metrics are computed at a single region level.

To illustrate the M&TC procedure and the issue with sample size calculation, we use the estimate of statistical power as an example (Figure 1A). Features of 4 distinct brain regions (ROIs) significantly correlate with behavior in the full sample. Importantly, each ROI demonstrates a specific effect size (i.e., Pearson's correlation), ranging from r = 0.05 (Figure 1A, blue ROI) to r

= 0.20 (orange ROI). Following the M&TC procedure, random resamplings are performed at multiple sample sizes (we show hypothetical n = 375 in Figure 1A). For each significant ROI in the full sample, power is computed as the percentage of random resamplings passing the same statistical threshold and then averaged across regions (line 53 in the function *abcd_statisticalerrors.m* made available by the authors; Figure 3D of the original article). Yet, the introduction of this last step produces biased estimates of statistical power and sample size, which refer to a hypothetical effect not even measured across the four significant regions. The bias introduced by the averaging procedure penalizes larger effects (i.e., they are deemed to require larger samples to be detected; orange and purple ROIs in the example), whereas smaller effects (i.e., blue and green ROIs) are - paradoxically - favored. As detailed below, this consideration holds for other metrics (e.g., type M and type S errors), brain features (e.g., thickness, connectivity), and levels of anatomical resolution (e.g., components, networks, ROIs, edges/vertices) reported in the original article.

We argue that, rather than average metrics, region-specific estimates should be preferred when assessing BWAS reproducibility and substantiate our claim by simulating and analyzing brain-behavior associations like those reported in the original M&TC article (source data Figure 1).

To this aim, firstly, we simulate cognitive ability data and 100 resting-state functional connectivity (rsFC) components for 3,604 participants (i.e., the full sample in the rsFC dataset of M&TC; *generate_simulated_data.m*; MATLAB® code available at https://github.com/giacomohandjaras/BWAS/). The distribution of correlation values obtained for the 100 simulated rsFC components and synthetic cognitive ability scores (Figure 1B, left panel; correlations are obtained from source data of Figure 1 in the original article) well approximates (tolerance: r ± 1e-4) values reported in M&TC (Figure 1B, left panel of the original article). Using a bootstrapping procedure (*bootstrap_correlation.m*), we then obtain 1,000 random resamplings of brain data and behavioral scores at increasing sample sizes (from n = 25 to n = 3,000), as in the original report. We characterize sampling variability for the strongest association among the 100 correlations (r = 0.1652; Figure 1B, right panel; see Figure 1F of the original M&TC article for comparison) and assess power, replication rate, and statistical errors after controlling for the number of tests ($q < 0.05$ FDR corrected[5]).

For each reference rsFC component identified in the full sample ($q < 0.05$), we compute statistical power at multiple sample sizes as the percentage of random resamplings passing the same statistical threshold (*compute_power.m*). To assess type M error, we compute the ratio between the correlation value obtained for each significant resampling and the one obtained in the full sample. The inflation percentage is then estimated at each sample size as the average across resamplings (*compute_type_m_alternative.m*). Also, type S error is calculated as the percentage of resamplings in which the brain-behavior correlation has opposite sign compared to the one measured in the full sample. The calculation of type S error is repeated for unthresholded and FDR-corrected correlations (*compute_type_s.m*). For the probability of replication, instead,

we split the simulated dataset into discovery and replication sets of equal size (n = 1,802). The percentage of successful replication is computed for each simulated rsFC component and sample size as the proportion of resamplings yielding a significant brain-behavior correlation in both the discovery and replication sets (*compute_replication.m*). Crucially, to demonstrate how aggregated metrics produce biased estimates of BWAS reproducibility, we average power, type M error, type S error, and replication rates across rsFC components, as in M&TC. We report and visualize averaged as well as region-specific metrics.

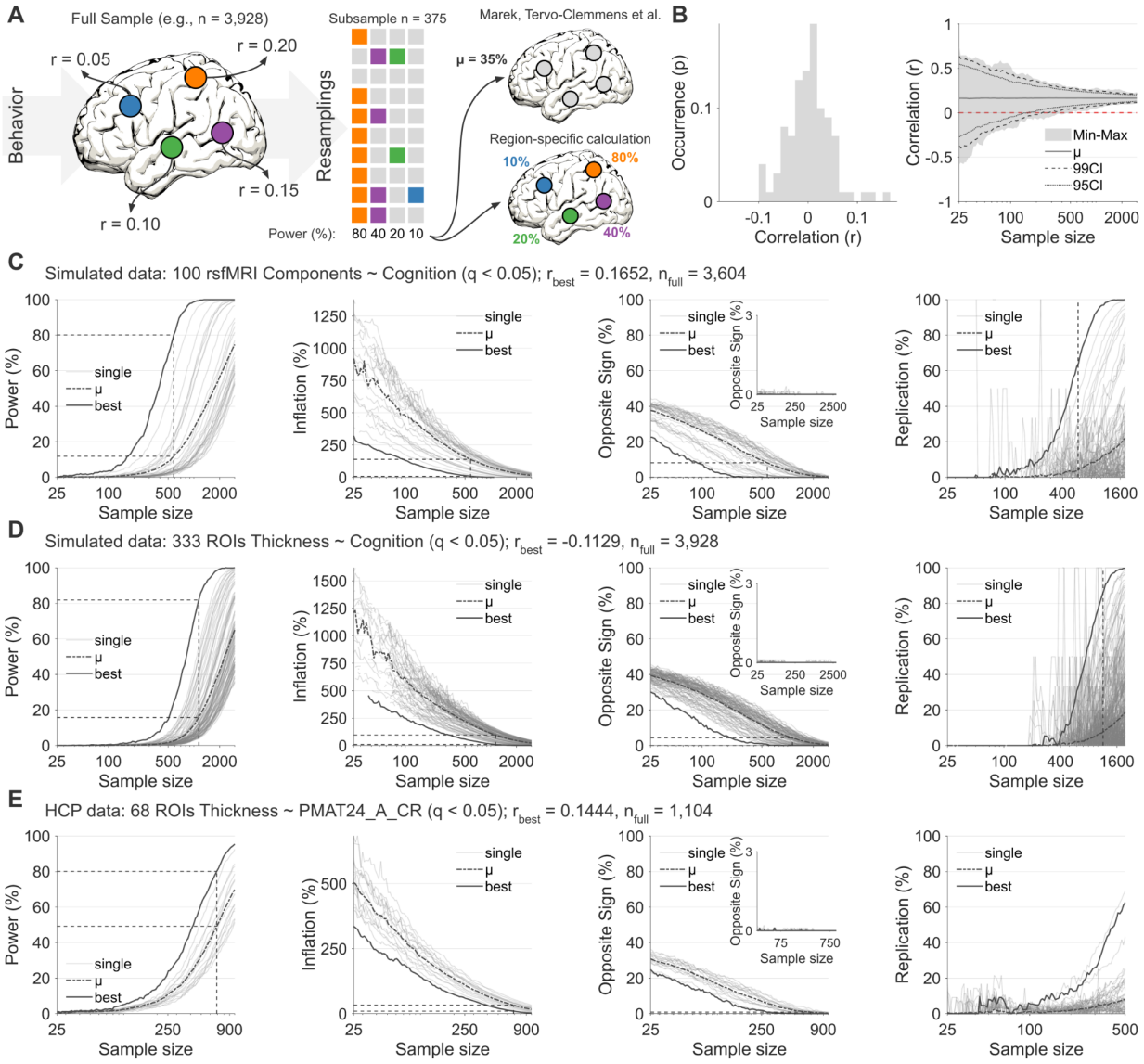

Figure 1. Panel **A** illustrates the analytical pipeline adopted by M&TC to assess BWAS reproducibility, as well as the region-specific calculation we propose. Statistical power is computed at multiple sample sizes for ROIs passing the statistical significance in the full sample (orange, purple, green, and blue regions; n = 3,928). Afterward, the authors compute the average power across significant ROIs. We suggest an alternative region-specific power calculation and test the reliability of the two methods across simulated and HCP data. In **B**, we show the distribution of synthetic correlations for the rsFC components vs. cognitive ability (left panel). Correlations are generated

starting from source data in Figure 1 of the original M&TC article. To ease the comparison between our simulated and original M&TC data, we also report the sampling variability for the largest simulated correlation (r = 0.1652, right panel; Figure 1F in M&TC). In **C**, **D,** and **E,** we report statistical power, effect size inflation (type M error), percentage of opposite sign correlations (type S error; unthresholded: larger plot; thresholded: insert plot), and replication rate as a function of sample size. Metrics are computed on simulated (panel **C**: rsFC components vs. cognitive ability; panel **D**: ROI thickness vs. cognitive ability) and real data (panel **E**: HCP ROI thickness vs. fluid intelligence) after applying the false discovery rate correction for multiple comparisons ($q < 0.05$). In each plot, the dotted line represents the average across regions (as in M&TC), the dark solid line refers to the brain parcel showing the best correlation, whereas the solid gray lines represent all other regions passing the statistical threshold in the full sample. Where present, dashed lines indicate the sample size required for 80% power based on the region-specific calculation. We draw dashed projection lines crossing the average metric at the same sample size to highlight the difference between the M&TC method and the one proposed here.

As in the original article, the strongest association between simulated rsFC components and cognition in the full sample (n = 3,604) is r = 0.1652 (*simulation_parent_script.m*). For this correlation, 80% power is reached at n = 580 after controlling for multiple comparisons ($q < 0.05$; Figure 1C, solid black line). Also, at n = 580, the percentage of inflation is 8%, there is a 0% probability of observing opposite sign correlations for both unthresholded, and FDR corrected data, and the replication rate is 64%. Importantly, in addition to the best simulated rsFC component, other associations attain 80% power with $n \leq 1{,}000$ ($q < 0.05$; Figure 1C, solid grey lines), and most of these correlations reach a 90% replication rate with $n \cong 1{,}500$. Conversely, at n = 580, the average statistical power computed across rsFC components is 12% ($q < 0.05$; Figure 1C, dotted line). Also, for the same sample size, the average inflation is 140%, there is an 8% probability of observing opposite sign correlations across rsFC components, and the average replication rate is 4%. These biased estimates obtained from simulated data are in line with those documented in the original M&TC article, where - using real data - the authors indicate a 5% replication rate for BWAS with $n < 500$ (pg. 657) and only 68% power at $n \cong 4{,}000$.

To ensure that the discrepancies between region-specific and average metrics are not due to the distribution of correlations observed for rsFC components, we repeat all analyses on simulated ROI-based cortical thickness values (333 ROIs) and cognitive ability scores (source data of Figure 1A in M&TC).

In this case, the strongest association between simulated thickness and cognition in the full sample (n = 3,928) is r = -0.1129. For this correlation, 80% power is reached at n = 1,140 after controlling for multiple comparisons ($q < 0.05$; Figure 1D, solid black line). Also, with a sample size of n = 1,140, the percentage of inflation is 9%, there is a 0% probability of observing opposite sign correlations for both unthresholded, and FDR corrected data, and the replication rate is 87%. In line with the previous simulation, other ROIs reach 80% power with $n \leq 2{,}000$ ($q < 0.05$; Figure 1D, solid grey lines). Instead, at n = 1,140, the average statistical power is 16% ($q < 0.05$; Figure 1D, dotted line). For the same sample size, the average inflation is 95%, there is a 4% probability of observing opposite sign correlations across ROIs, and the average replication rate is 8%. As expected, smaller effect sizes ($|r| \cong 0.11$ as compared with $|r| \cong 0.16$) require larger

samples (n ≅ 1,200 as compared to n ≅ 600) to be detected. However, region-specific estimates of power, replication rate, and statistical errors show that small effect sizes ($0.11 < |r| < 0.16$) can be detected reliably with hundreds, not thousands, of participants.

To further prove that our findings do not depend on the synthetic nature of simulated correlations, we use HCP data (n = 1,104) and test the relationship between Freesurfer[6] estimates of cortical thickness for 68 ROIs[7] and a measure of fluid intelligence (PMAT24_A_CR; *hcp_parent_script.m*). Of note, among the numerous behavioral measures provided by the HCP consortium, we select this measure of fluid intelligence as there is at least one brain-behavior correlation in the full sample (r = 0.1444; Figure 1E) comparable to the largest reported in M&TC. In line with simulation results, the strongest relationship in HCP data reaches 80% power at n = 689 after controlling for multiple comparisons ($q < 0.05$; Figure 1E, solid black line). Using the same sample size and statistical threshold, we observe a 10% inflation rate and a 0% occurrence of opposite sign unthresholded and FDR-corrected correlations. Instead, at n = 689, the average statistical power is 49%, the average inflation rate is 33%, and there is a 1% probability of observing opposite sign results across ROIs. As far as replication rate is concerned, the strongest correlation has 63% reproducibility with n ≅ 500 (i.e., approximately half of the full HCP sample). In contrast, the average replication rate is 8% for the same sample size.
Hence, in line with simulations' findings, real data results indicate that small effect sizes can be detected reliably with hundreds of observations. Of note, estimates of statistical power obtained with the bootstrapping procedure in HCP data align with those computed using parametric statistics. Setting r = 0.1444 and $\alpha = 0.05$ (uncorrected), 80% power is achieved at n = 380 according to bootstrapping calculations and at n = 374 using a parametric method[8].

That average estimates of power, replication rate, and statistical errors are unreliable is also evident when comparing ABCD results with those obtained from the UK Biobank (Figure 3 and Supplementary Figure 9, in the M&TC article). As pointed out by the authors, correlations obtained from the two datasets are comparable in terms of magnitude (Figure 2 in the original M&TC article). However, with a sample size of n ≅ 3,000, the estimated average power in the UK Biobank is ~25% for the uncorrected threshold and ~0% for the Bonferroni corrected one (Supplementary Figure 9 of the original article). These estimates are substantially lower than those documented in the ABCD sample (i.e., ~65% at n ≅ 3,000). One possible explanation for this discrepancy is that any non-zero correlation becomes significant with a large enough full sample size. Thus, the number of reference regions (i.e., those for which power, replication rate, and statistical errors are computed) increases with the number of observations in the full sample, and the presence of several small - yet significant - effects decreases substantially the estimate of the average power at smaller sample sizes. To corroborate this claim, we simulate a distribution of brain-behavior correlations at two full sample sizes: n = 4,000 and n = 256,000. Despite the distribution of correlation values being the same across the two full sample sizes, the M&TC

estimate of average power decreases as the full sample increases (Figure 2A; *power_decrease_parent_script.m*). Paradoxically, at n = 3,000, average power ($q < 0.05$) drops from ~74% to ~33% with a 64-fold increase (from n = 4,000 to n = 256,000) of the full sample size. Importantly, single ROI estimates of power (as well as other metrics) do not depend on the size of the full sample. A single simulated correlation value of $r = 0.5300$ can be detected with 80% power at $n \cong 60$ ($q < 0.05$), regardless of the size of the full sample (Figure 2A). This further evidence indicates that average power estimates are inadequate to assess the reproducibility of effects distributed on the cortical mantle.

In summary, our results indicate that (1) average estimates of power, replication rate, and statistical errors are biased and should not be used to draw inferences on the reproducibility of regional brain-behavior associations; (2) reproducible univariate BWAS do not necessarily require thousands of individuals; (3) when applying a rigorous correction for multiple comparisons the probability of detecting and, thus, publishing opposite sign associations is null, even when sample sizes are extremely small (tens of individuals; Figure 1C-E opposite sign plot insert); (4) therefore, results obtained in fairly conducted small BWAS (e.g., no p-hacking, correction for multiple comparisons - all variables that affect the study reliability irrespectively of sample size) are more likely to be true and inflated, than false positives (with FWE $\alpha = 0.05$, only five studies out of 100 are expected to be false by chance).

A possible counterargument to our criticism is that we selectively report results for the largest effect size and that a complex cognitive trait is unlikely to relate to the properties of a single brain region.

Firstly, using M&TC data, we show that multiple univariate brain-behavior associations can be detected with ≥ 80% statistical power in less than a thousand individuals (Figure 2B; *count_powered_correlations.m*). Of note, cognitive abilities significantly correlate ($q < 0.05$) with 32 (out of 100) rsFC components in the full sample (n = 3,604). Two of these associations are detected with ≥ 80% statistical power in less than 1,000 individuals ($q < 0.05$). Considering the same power and significance threshold, the recruitment of 1,500 individuals allows the detection of 6 (out of 32) significant correlations ($q < 0.05$; ≥ 80% power). As far as the relationships between rsFC networks and cognitive abilities are concerned, 6 out of 61 reach ≥ 80% statistical power ($q < 0.05$) with less than 1,000 individuals. Instead, results for the relationship between cortical thickness and cognitive abilities indicate that the detection of multiple associations requires between 1,500 and 2,000 individuals ($q < 0.05$; ≥ 80% power; Figure 2B).

Secondly, brain-behavior associations are often studied in a multivariate manner. In this regard, M&TC use canonical correlation to associate patterns of brain features (i.e., rsFC and cortical thickness) with behavioral data obtained from the NIH Toolbox (i.e., cognition) and the Child Behavior Checklist (CBCL; i.e., psychopathology). Using source data from Figure 4 of the original M&TC article, we assess the statistical power of this multivariate approach. For each

brain-behavior association, we select out-of-sample correlations (i.e., those measured in the replication dataset), estimate the statistical significance at multiple sample sizes (15 sample sizes; range: n = 25 : 1,475), and compute power across resamplings (100 bootstraps; *multivariate_cca_power.m*). Overall, effect sizes obtained from canonical correlation are higher than those observed in univariate analyses, even when considering out-of-sample values (see Figure 4 of M&TC). Indeed, at n = 350, the relationship between cognition and rsFC is detected with 89% power ($p < 0.05$; Figure 2C). Also, when considering the association between cognition and cortical thickness ($p < 0.05$), 82% power is achieved after recruiting 615 participants. Psychopathology correlates with brain features (i.e., rsFC and cortical thickness) to a lesser extent, thus requiring more than 1,500 individuals to achieve power ≥ 80% (Figure 2C).

These results indicate that multiple univariate effects are detected with sufficient statistical power in less than 1,000 individuals and that multivariate analyses further increase the probability to detect reliable brain-behavior associations when the sample size is in the order of a few hundred. Thus, when establishing general BWAS guidelines, it is essential to emphasize that some effects can be detected reliably even when the sample size is relatively small. Alternatively, researchers (as well as reviewers and editors) would be induced to believe that no effect has sufficient power to replicate, even when the sample size is 30,000 (as documented in Supplementary Figure 9 of M&TC).

Moreover, it could be argued that researchers should aim at recruiting the largest possible sample in their studies. Although it may seem controversial to disagree with such a suggestion, we would like to emphasize that one of the reasons for conducting power calculations is to establish when to stop data collection. This is of particular importance as (1) even negligible effects with limited practical applications reach statistical significance in very large cohorts[9], and (2) it helps to optimize the balance between study value (i.e., relevance for the society) and participants' burden (e.g., scanning costs, participants' time, and effort)[10].

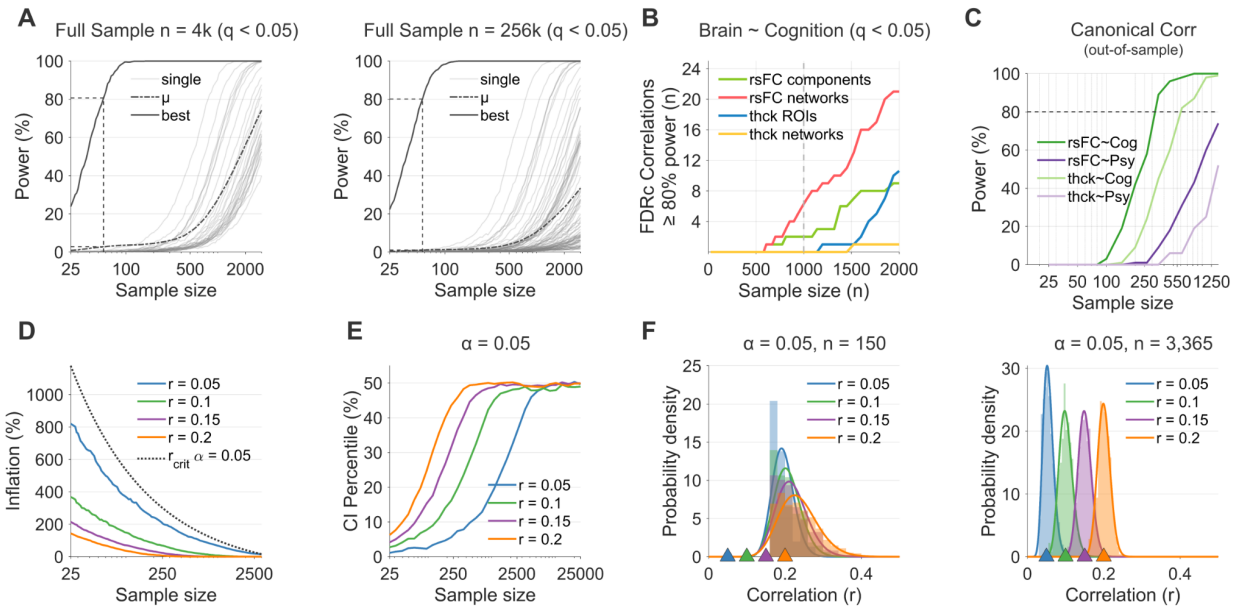

Figure 2. In **A**, we show that power estimates for the average method depend on the size of the full sample (n = 4,000 in the left panel and n = 256,000 in the right panel). Because any non-zero correlation becomes significant with a large enough full sample size, the number of reference regions (solid gray lines) increases with the number of observations in the full sample. The presence of several small - yet significant - effects (please see bottom right part of the two panels) decreases the average power estimate (dotted line) at smaller sample sizes. Instead, region-specific power estimates (dark solid line) do not depend on the size of the full sample. In **B**, we report the cumulative distribution of univariate significant associations ($q < 0.05$) with statistical power $\geq 80\%$ as a function of sample size. Phenotype is cognitive ability and brain features are rsFC components (green), rsFC networks (red), ROI cortical thickness (blue), and network cortical thickness (yellow). The grey dashed line marks n = 1,000. Panel **C** shows statistical power as a function of sample size for multivariate canonical correlation analysis in M&TC. Correlations between brain and behavior are obtained from the replication dataset (i.e., out-of-sample correlations; M&TC Figure 4, source data). Phenotypes are NIH Toolbox (cognition; green) and CBCL (psychopathology; purple). Brain features are rsFC (darker colors) and cortical thickness (lighter colors). The grey dashed line represents 80% power. Panel **D** shows that effect size inflation results from statistical thresholding (dotted line). However, because larger population effects (r = 0.20, solid orange line) are more easily detected and less inflated in smaller samples, typically sized BWAS are more likely to report less inflated estimates of relatively larger true effects (as measured in the population) than extremely inflated estimates of relatively smaller population effects (e.g., r = 0.05 solid blue line). In **E**, we show that - because of statistical thresholding - effect sizes reported in the literature (i.e., the mean of the confidence interval) align with true effects measured in the population only when the sample size is large enough ($n \geq 453$ for r = 0.20, solid orange line; $n \geq 1{,}104$ for r = 0.15, solid purple line; $n \geq 3{,}365$ for r = 0.10, solid green line; $n \geq 6{,}566$ for r = 0.05, solid blue line). Not accounting for this effect determines overly optimistic estimates of the sample size required for replication (**F**, left panel; n = 150). At larger sample sizes (**F**, right panel; n = 3,365), the impact of statistical thresholding diminishes, and effects reported in single studies align with the population ground truth (colored triangles; population size n = 100,000 in the current simulation).

An important point raised by M&TC is that, in small sample studies, effect size inflation is the consequence of statistical thresholding and is aggravated by the pressure to publish positive findings. We document such an association by estimating the theoretical inflation percentage for correlation values corresponding to the critical $\alpha = 0.05$ at multiple sample sizes (Figure 2D; dotted line; *inflation_effect_parent_script.m*). In the original article, the authors refer to this issue as the BWAS paradox. They suggest that the most inflated effects are most likely to be statistically significant at smaller sample sizes (pg. 657-8). However, by simulating the trajectory of inflation for four distinct effect sizes (r = 0.05, r = 0.10, r = 0.15, and r = 0.20; n = 4,000; $p < 0.05$ uncorrected), we show that the larger the effect size in the full sample, the smaller the inflation in studies with fewer participants (Figure 2D; e.g., orange vs blue curves). Importantly, larger effects (as measured in the full sample) are less inflated as well as more likely to pass the statistical significance threshold at smaller sample sizes. To demonstrate this, we use the M&TC resampling method and simulate 1,000 studies (sample size: n = 200) for each of the abovementioned correlations (4,000 studies in total). Results show that 1,410 out of 4,000 studies pass the significance threshold after correcting for multiple comparisons ($q < 0.05$). Of these 1,410 significant studies, the 49.79% are obtained from bootstrapping the full sample with a brain-behavior correlation of r = 0.20, whereas the 5.39% by resampling the one with a correlation of r = 0.05 (for completeness: the 29.57% comes from r = 0.15 and the 15.25% from r = 0.10). Therefore, typically sized BWAS are more likely to report less inflated estimates of

relatively larger true effects (as measured in the population) than extremely inflated estimates of relatively smaller correlations observed in the population.

If typically sized BWAS can reliably detect small effects ($0.10 < |r| < 0.20$), why do we have replication failures? Leaving aside factors that affect study reliability regardless of sample size, such as malpractices (e.g., p-hacking, file-drawer issue), the inflation documented in smaller samples, and the lack of consideration for the confidence interval of the effect size may play a crucial role. Let's assume our goal is to replicate a BWAS reporting a correlation of $r = 0.30$ measured in 150 individuals (FWEp < 0.05; 95CI: 0.12 - 0.48). Firstly, since the authors have applied a rigorous correction for multiple comparisons and there is no reason to suspect misconduct, we should trust this finding. As mentioned above, correction for multiple comparisons ensures that the published study is more likely to report a true (inflated) effect than a false positive association. Because the aim is to replicate the effect for one significant region, we set $\alpha = 0.05$ and estimate that 85 individuals are required to detect an effect of $r = 0.30$ with 80% power. However, due to the relatively small sample included in the original study ($n = 150$), $r = 0.30$ is likely to be an inflated estimate of the true effect size in the population. Also, it should be pointed out that the range of estimates for the unknown parameter (i.e., the confidence interval) goes from $r = 0.12$ to $r = 0.48$. On these premises, estimating the replication sample size based on the reported effect (i.e., the mean of the confidence interval) is an overly optimistic choice. To demonstrate this and explore if and when the mean of the confidence interval (i.e., the effect size typically reported in published articles) should be trusted for power calculation, we simulate four effects ($r = 0.05$, $r = 0.10$, $r = 0.15$, and $r = 0.20$) in a full sample of 100,000 individuals (*sample_adjustment_parent_script.m*). For each effect, we create random resamplings at multiple sample sizes (as prescribed in M&TC), identify significant correlations at $p < 0.05$, and calculate the percentile of the confidence interval that best approximates the true effect (as measured in 100,000 individuals). In case of no inflation, the effect measured in the full sample corresponds to the 50th percentile of the confidence interval of the effect reported in smaller studies. We show that the mean of the confidence interval is a proper estimate of the true effect $r = 0.20$ starting from $n = 453$ (Figure 2E). Smaller effects require even more observations before the mean of the confidence interval aligns with the effect size measured in the full sample: $n \geq 1{,}104$ for $r = 0.15$, $n \geq 3{,}365$ for $r = 0.10$, and $n \geq 6{,}566$ for $r = 0.05$ (Figure 2E). As expected, when the correction for multiple comparisons is applied (e.g., 100 ROIs, $p < 0.0005$), the alignment of study effect size with the population requires even larger samples: $n \geq 884$ for $r = 0.20$ and $n \geq 12{,}810$ for $r = 0.05$ (data not shown).

Thus, in case we aim to replicate a brain-behavior correlation found in a relatively small sample (e.g., 150 individuals), power calculation should not be based on the effect size reported in the original article. When power is computed assuming no inflation and ignoring the confidence interval of the effect, one cannot exclude that replication failures are due to optimistic estimates of the required sample size.

Unfortunately, the size of the effect in the population is unknown; therefore, it is difficult to predict how inflated an effect of r = 0.30 measured in 150 individuals precisely is. Nonetheless, considering the probability density function of significant correlations at n = 150 and $\alpha = 0.05$, it is implausible that such a correlation reflects an effect in the population of r = 0.05 (Figure 2F, first panel). Yet, it would be problematic to conclude whether a power calculation should be conducted considering r = 0.20 rather than r = 0.15. We cannot even rule out completely the possibility of a 200% inflation of the population effect r = 0.10. As expected, uncertainty decreases with increasing sample size (Figure 2F, second panel).

Hence, for the effect reported in the original article of our example (r = 0.30, n = 150), a more cautious power calculation based on the estimated population effect of r = 0.15 (target power 80%, $\alpha = 0.05$) prescribes the recruitment of 347 individuals, rather than 85 (see above).

Importantly, it is the pressure to publish significant results the main responsible for the effect size inflation present in the BWAS literature. Although the sampling variability in smaller BWAS is high (Figure 1B, right panel), when no statistical threshold is applied, the average effect across studies (Figure 1B, right panel; solid dark line) represents an adequate estimator of the population effect, even when n = 25.

In conclusion, reproducible brain-behavior correlations do not necessarily require thousands of individuals. Average estimates of power, replication rate, and statistical errors are biased and should not be used to assess BWAS reproducibility. Indeed, fairly conducted small BWAS can detect small effects reliably and are unlikely to report false-positive results or opposite sign associations. Leaving malpractice aside, rather than the sample size per se, what biases the literature and exacerbates the replication crisis is the higher probability of publishing significant results. Publication bias determines the inflation of effect sizes, which is seldom considered when estimating the replication sample size.

Very large datasets are certainly needed to advance our knowledge of the association between brain and behavior. However, this should not translate into a hegemony of large consortia. We are convinced that studies on smaller samples are as important as those on large datasets. Typically sized BWAS carried out in individual laboratories increase sample diversity and representativeness in the published literature and promote the discovery of novel behavioral measures that, perhaps, relate more strongly to brain features[11].

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. 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.
6. Fischl B. (2012). FreeSurfer. NeuroImage, 62(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021
7. 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.
8. Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods, 39(2), 175-191.
9. Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12).
10. Bacchetti, P., Wolf, L. E., Segal, M. R., & McCulloch, C. E. (2005). Ethics and sample size. American journal of epidemiology, 161(2), 105-110.
11. Gratton, C., Nelson, S. M., & Gordon, E. M. (2022). Brain-behavior correlations: Two paths toward reliability. Neuron, 110(9), 1446-1449.

Acknowledgment

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. We thank Scott Marek and Brenden Tervo-Clemmens for sharing their code and for the constructive discussion. We thank Jack Van Horn for his insightful comments.

Competing interests

The authors declare no competing interests.

Author Contributions

Conception, data analysis and interpretation, manuscript writing and revising: LC and GH.

Corresponding authors

Correspondence to Luca Cecchetti (luca.cecchetti@imtlucca.it), and Giacomo Handjaras (giacomo.handjaras@imtlucca.it).