Data-driven estimates of the reproducibility of univariate BWAS are biased.

Charles D.G. Burns[1*†], Alessio Fracasso[1] & Guillaume A. Rousselet[1*††]

[1] *School of Psychology and Neuroscience, University of Glasgow, Glasgow, G12 8QB*

ARISING FROM: Marek, Tervo-Clemmens *et al.* Reproducible brain wide association studies require thousands of individuals. https://doi.org/10.1038/s41586-022-04492-9 (2022)

Abstract

Recently, Marek, Tervo-Clemmens *et al.*[1] leveraged consortium neuroimaging data to answer a question on most researchers' minds: how many subjects are required for reproducible brain-wide association studies (BWAS)? Their approach could be considered a framework for testing the reproducibility of several neuroimaging models and measures. Here we test part of this framework, namely estimates of statistical errors of univariate brain-behaviour associations obtained from resampling large datasets with replacement. We suggest that reported estimates of statistical errors are largely a consequence of bias introduced by random effects when sampling with replacement close to the full sample size. We show that these biases can be largely avoided by only resampling up to 10% of the full sample size. Using this unbiased approach, sample size requirements for reproducible univariate BWAS tested by Marek, Tervo-Clemmens *et al.* are even worse.

[*] Corresponding authors.
[†] E-mail: charlesdgburns@gmail.com
[††] E-mail: guillaume.rousselet@glasgow.ac.uk

Main

Among numerous analyses in their study, Marek, Tervo-Clemmens *et al.*[1] estimated statistical errors of univariate BWAS as a function of sample size. Such univariate BWAS often involve tens of thousands of correlations between a brain measure and a behavioural measure, the vast majority of which fail to replicate even with thousands of participants. These replication failures can be explained by statistical errors of a study design such as statistical power[2–4]. To estimate statistical errors in univariate BWAS, Marek, Tervo-Clemmens *et al.* used a data-driven approach in which they treated a large discovery dataset as a population; they then drew replication samples by resampling with replacement (henceforth resampling) from that population. Here, we tested the validity of this data-driven approach by using ground truth simulated data.

First, we simulated a discovery null-sample with n = 1,000 subjects each with 1,225 brain connectivity measures (random Pearson correlations) and a single behavioural measure (normally distributed across participants). We correlated each brain connectivity measure with the behaviour across all subjects to obtain 1,225 brain-behaviour correlations. Since brain connectivity estimates and behavioural factors were simulated independently from each other, any resulting brain-behaviour correlations were entirely random. Data dimensions were chosen to be computationally feasible for reproducibility, however we invite readers to adjust these and re-run analyses using the openly available code (analyses recoded in R with supporting packages[5–7] for open-source accessibility https://github.com/charlesdgburns/rwr/). We then resampled our null-sample for 100 iterations across logarithmically spaced sample size bins (n = 25, to 1,000) and estimated statistical errors, following the methods described in Marek, Tervo-Clemmens *et al.*[1]. Surprisingly, we saw the same trends of statistical errors and reproducibility as those reported by Marek, Tervo-Clemmens *et al.*[1] but with random data (see Fig. 1), with strongly biased statistical power estimates.

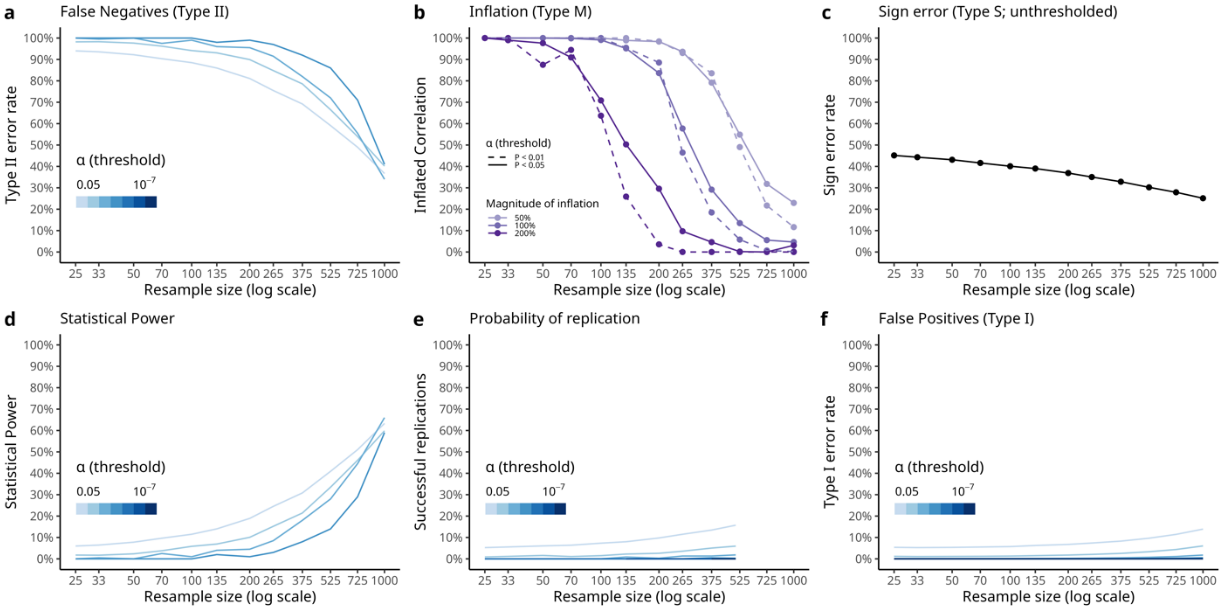


Figure 1: **Estimated statistical errors and reproducibility of random noise (ρ = 0).** We reproduce statistical errors estimated by Marek, Tervo-Clemmens *et al.*[1] after resampling from a discovery null-sample where observed significant effects are random ($\rho = 0$). We refer to Fig. 3 in Marek, Tervo-Clemmens *et al.*[1] for comparison and detailed descriptions of each plot. Here we further note that in **a**,**d**, only four lines are drawn because 1,225 random brain-behaviour correlations should only pass significance thresholds α = .05, .01, .001, and .0001 (1/8 chance). Similarly, for **e**,**f**, we consider correlations which are significant in resamples and notice that these pass thresholds as small as $\alpha = 10^{-7}$. The fact that false positive rates are greater than the significance threshold should be alarming, as it shows that simulating a replication by resampling will, on average, reject the null more frequently than we would expect even under the null.

These trends in statistical errors do not depend on absolute sample size, but the resample size relative to the full sample size. By repeatedly generating new null-samples, rather than resampling from a single null-sample, we verified that these statistical error estimates are indeed biased under the null as the resample size approaches the full sample size (Extended Data Fig. 1). For example, uncorrected (α = .05), statistical power was estimated to be 63% when resampling at the full sample size (*n* = 1,000, Fig 1. **d**), rather than the expected 5% obtained when generating new null-samples (*n* = 1,000, Extended Data Fig 1. **d**). One concern is that power is the most inflated while also being the most relevant for failed replications[3,4], which could potentially result in misleading meta-science.

To explain why this bias arises under the null, we investigated the underlying brain-behaviour correlations used in the calculation of statistical errors. Here we focused on resampling at the full sample size (*n* = 1,000) where these biases are most dramatic. As indicated by the false positive rate (Fig. 1 f.), the null distribution of brain-behaviour correlations is not preserved when resampling at the full sample size (Fig. 2). Instead, resampling subjects and computing correlations again results in a distribution wider than expected (comparing Fig. 2 **a**. and **c**.). This is because resampling involves two sources of sampling variability, first at the level of the discovery sample and again for the resampled replication sample. For instance, if a correlation in the discovery sample is randomly observed to be r = 0.11, then resampling participants and computing the same correlation again results in a correlation which varies around r = 0.11 (Fig. 2 **e**.). The influence on statistical error estimates such as power is two-fold. First, random correlations in the tail of a discovery sample are more likely to be in the tail of correlations in a resampled replication sample. This inflates power when estimated as the proportion of significant effects in the discovery sample which are significant again in the resampled replication sample (1 – false negative rates). Secondly, increased sampling variability alone leads to a wider-than-expected distribution of correlations with more extreme tails. These more extreme tails lead to an inflation of *P* values close to 0 in our resample (compare Fig. 2 **b.** and **d**.) when calculated using a standard correlation function (e.g., ‘corr’ in MATLAB).

In short, when resampling close to the full sample size, random correlations are more likely to be significant again in each resample due to the resampling process alone, biasing the statistical power estimates described above (see supplementary information for further details).

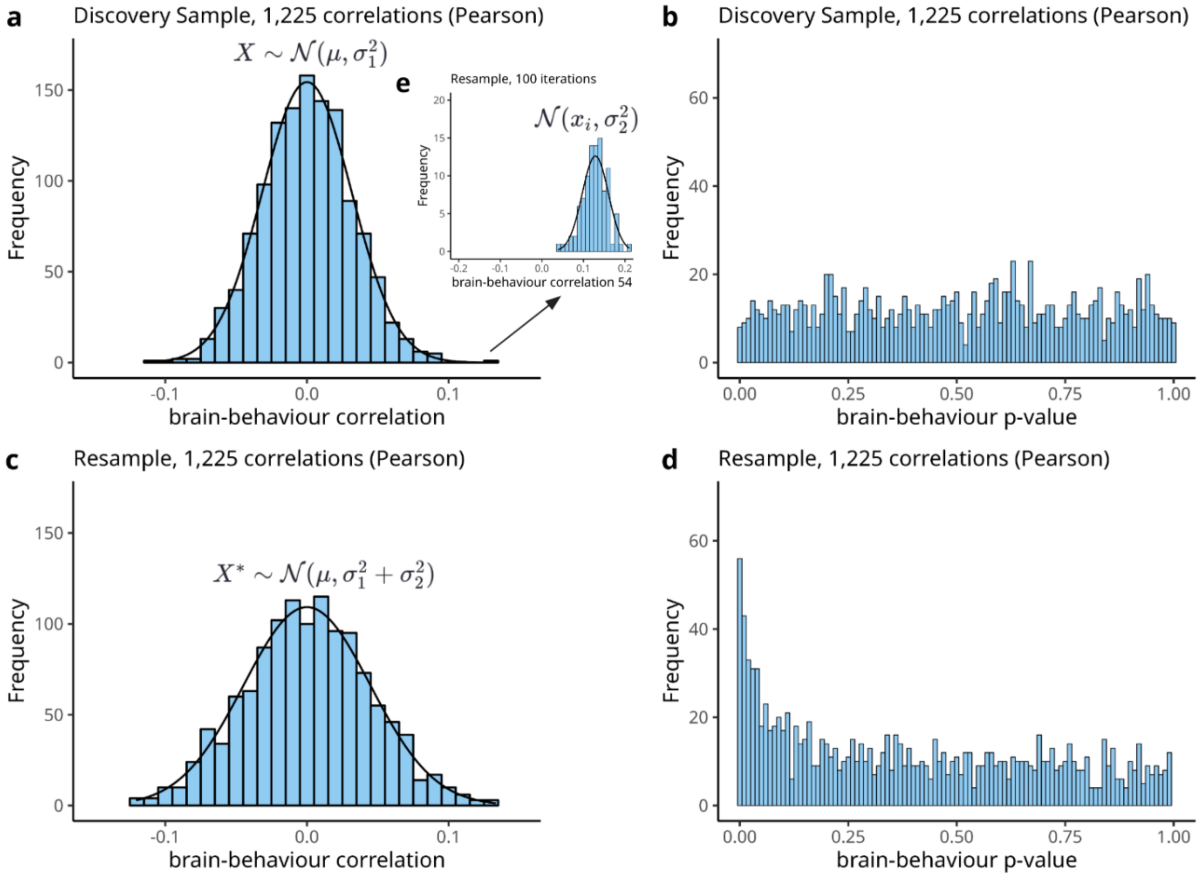


Figure 2: **Null distributions not preserved when resampling with replacement at the full sample size ($\rho = 0$).** **a**, Distribution of simulated random brain-behaviour correlations (1,225 total) treated as a discovery sample. There are n = 1,000 subjects for each random Pearson correlation, here compared to a Gaussian curve with mean $\mu = 0$ and variance σ_1^2 = .001 drawn in black. **b**, We verify that the two-tailed P values of our 1,225 random brain-behaviour correlations are uniformly distributed. **c**, Distribution of all 1,225 brain-behaviour correlations computed after resampling subjects at full sample size (n = 1,000). This distribution is clearly wider than our discovery sample null-distribution. The solid black line shows a parametric fit derived from an interaction of sampling variability (see **e.**), namely a Gaussian distribution with mean μ = 0 and variance $\sigma_1^2 + \sigma_2^2$ = .002. **d,** Distribution of two-tailed P values of all 1,225 brain-behaviour correlations computed after resampling at full sample size (n = 1,000). The distribution is inflated around 0 due to the wider tails in our null distribution. **e,** To help explain the widened distribution, we track the largest correlation observed in our original null-sample (r = 0.11), plotting the distribution of corresponding

brain-behaviour correlations across the 100 iterations of resampling at the full sample size (*n* = 1,000). The solid black line represents a Gaussian with mean $\mu = 0.11$ and variance $\sigma_2^2 = .001$. The interaction of variability across iterations (**e**) and variability in the discovery sample (**a**) results in the widened distribution (**c**) by additive variance[8].

While we have shown clear biases when there are no true effects, this does not directly imply biases when true effects are present. We note that Marek, Tervo-Clemmens *et al.*[1] already showed that the largest univariate effect is highly replicable even for moderate sample sizes, so there are at least some true BWAS effects in the real world. However, as the average true effect size remains unknown, we systematically simulated a range of discovery samples, each representing a study where the size of the underlying true effects corresponded with different levels of statistical power. While we refer to supplementary information for detailed methods, we note that since the bias under the null is driven by the false rejection of null hypotheses, we adopt a fixed significance threshold after Bonferroni correction which controls for at least one false positive. Focusing on statistical power estimates, we see that the bias near the full sample size depends on the true statistical power of the discovery sample (Fig. 3.). Power estimates are inflated if the discovery sample is underpowered, but on the other hand a highly powered discovery sample may give conservative power estimates.

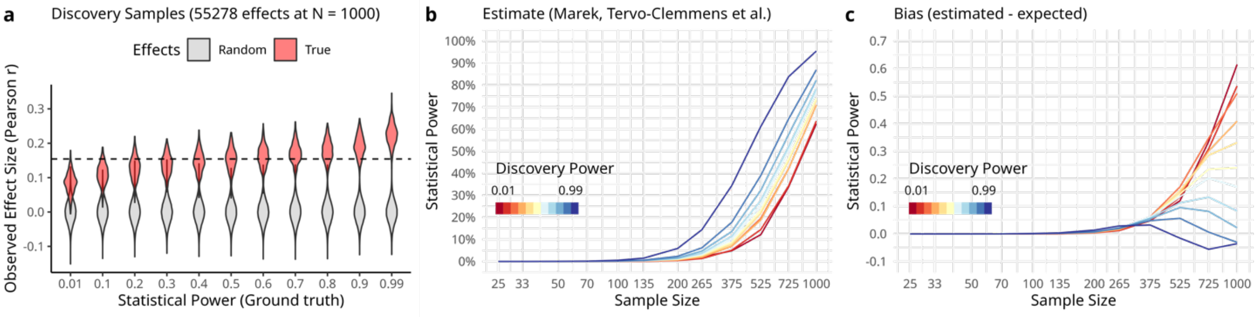


Figure 3: **Bias in estimated statistical power depends on true statistical power. a,** Simulated discovery samples are represented for each of the underlying power scenarios. For each scenario, grey violin plots show the distribution of 54,778 random effects and red violin plots represent the distribution of 500 true effects. The dashed line represents the critical Pearson r for a Bonferroni corrected significance level (α = 0.05/55278). **b**, We estimated power across sample size by simulating resampling methods in Marek, Tervo-

Clemmens *et al.*[1] using a Bonferroni corrected significance threshold. Line colour represents the ground truth statistical power of the study at full sample size (n = 1,000, α = 0.05/55278). **c**, To demonstrate bias across different sample sizes, we subtracted analytical power curves from the estimated power (panel **b**), with lines coloured as in panel **b**. Note that underpowered discovery samples inflate statistical power estimates.

Note that regardless of power at the full sample size, bias in statistical power is largely avoided when subsampling up to around 10% of the full sample size, which is consistent with the use of resampling techniques in a recent meta-scientific paper evaluating statistical power and false discovery rates for genome-wide association studies (GWAS)[9].

What are the implications for the results presented by Marek, Tervo-Clemmens *et al.*[1]? For the strictly denoised Adolescent Brain Cognitive Development (ABCD) sample (n = 3,928), they reported around 68% power at n = 3,928 after Bonferroni correction when resampling at the full sample size (Marek, Tervo-Clemmens *et al.*[1] Fig. 3 **d**.). Our true effect simulation results indicate that this could be inflated from a true average power anywhere between 1% and 40%. Furthermore, in their Supplementary Figure 9 **d**. Marek, Tervo-Clemmens *et al.*[1] report around 1% power for n = 4,000 and $\alpha = 10^{-7}$ when subsampling from the UK Biobank with a full sample size of n = 32,572. We therefore argue that the 68% power reported for the full ABCD sample (n = 3,928, $\alpha = 10^{-7}$) more likely reflects methodological bias, rather than a result of increased signal after strict denoising of brain data. While the largest BWAS effects may be highly reproducible with 4,000 participants, the average univariate BWAS effect is most likely not reproducible. On the other hand, our true effect simulations also indicate that the UK Biobank estimates at the full sample size are more reliable, with an underlying power likely between 70% and 90% at n = 32,572 after Bonferroni correction, suggesting that replicable univariate BWAS tested in Marek, Tervo-

Clemmens *et al.* require tens-of-thousands of individuals.

We strongly agree with Marek, Tervo-Clemmens *et al.*[1] that their results shouldn't be overinterpreted beyond BWAS[10]. Stressing this further, their results should be interpreted in the context of the specific study design they tested. For example, we may remind ourselves that inter-individual correlation studies offer "as little as 5%-10% of the power" of within-subject t-test studies with the same number of participants[4]. Other methodological choices, such as data modelling, should also be carefully considered. The lack of power in univariate BWAS considered by Marek, Tervo-Clemmens *et al.* could also be influenced by the choice of a group-averaged brain parcellation[11], which fails to account for individual level variations in resting state functional connectivity[11]. Brain models which do account for such individual variability generalise better, as demonstrated by stronger out-of-sample prediction[12,13], and could also lead to higher replication rates.

In summary, we showed that statistical error estimates of univariate correlations are methodologically biased, when obtained by resampling with replacement close to the full sample size of a large dataset. Notably, statistical power is inflated when the true power of the discovery sample is low and slightly deflated when true power is high. We further showed that this bias is largely avoided when subsampling only up to 10% of the full sample size after Bonferroni correction. When we avoided this bias in Marek, Tervo-Clemmens *et al.*'s[1] results, by subsampling from the UK Biobank results, we argued that the univariate BWAS tested are generally not reproducible even with thousands of individuals. We stress that this may not have wider implications outside of the specific study design tested by Marek, Tervo-Clemmens *et al.* and ultimately emphasize the importance of study design and model selection in neuroimaging approaches. To demonstrate this, further investigations of reproducibility of wider BWAS methods should be carried out. We hope such future studies may take into account the methodological considerations for obtaining statistical error estimates when resampling with replacement which we discussed here.

References

1. Marek, S. *et al.* Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
2. Ioannidis, J. P. A. Why Most Published Research Findings Are False. *PLOS Med.* **2**, e124 (2005).
3. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
4. Yarkoni, T. Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power—Commentary on Vul et al. (2009). *Perspect. Psychol. Sci.* **4**, 294–298 (2009).
5. Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. (2022).
6. Ripley, B. *et al.* MASS: Support Functions and Datasets for Venables and Ripley's MASS. (2023).
7. Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
8. El Otmani, S. & Maul, A. Probability distributions arising from nested Gaussians. *Comptes Rendus Math.* **347**, 201–204 (2009).
9. Chen, Z., Boehnke, M., Wen, X. & Mukherjee, B. Revisiting the genome-wide significance threshold for common variant GWAS. *G3 GenesGenomesGenetics* **11**, jkaa056 (2021).
10. Callaway, E. Can brain scans reveal behaviour? Bombshell study says not yet. *Nature* **603**,

777–778 (2022).

11. Gordon, E. M. *et al.* Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations. *Cereb. Cortex* **26**, 288–303 (2016).

12. Kong, R. *et al.* Individual-Specific Areal-Level Parcellations Improve Functional Connectivity Prediction of Behavior. *Cereb. Cortex* **31**, 4477–4500 (2021).

13. Farahibozorg, S.-R. *et al.* Hierarchical modelling of functional brain networks in population and individuals from big fMRI data. *NeuroImage* **243**, 118513 (2021).

Acknowledgments

A.F. is supported by a grant from the Biotechnology and Biology research council (BBSRC, grant number: BB/S006605/1) and the Fundação Bial, Fundação Bial Grants Programme 2020/21, A-29315, number 203/2020, grant edition: G-15516.

Authors' information

School of Psychology and Neuroscience, University of Glasgow, Glasgow, G12 8QB, UK

Charles D.G. Burns, Alessio Fracasso, Guillaume Rousselet

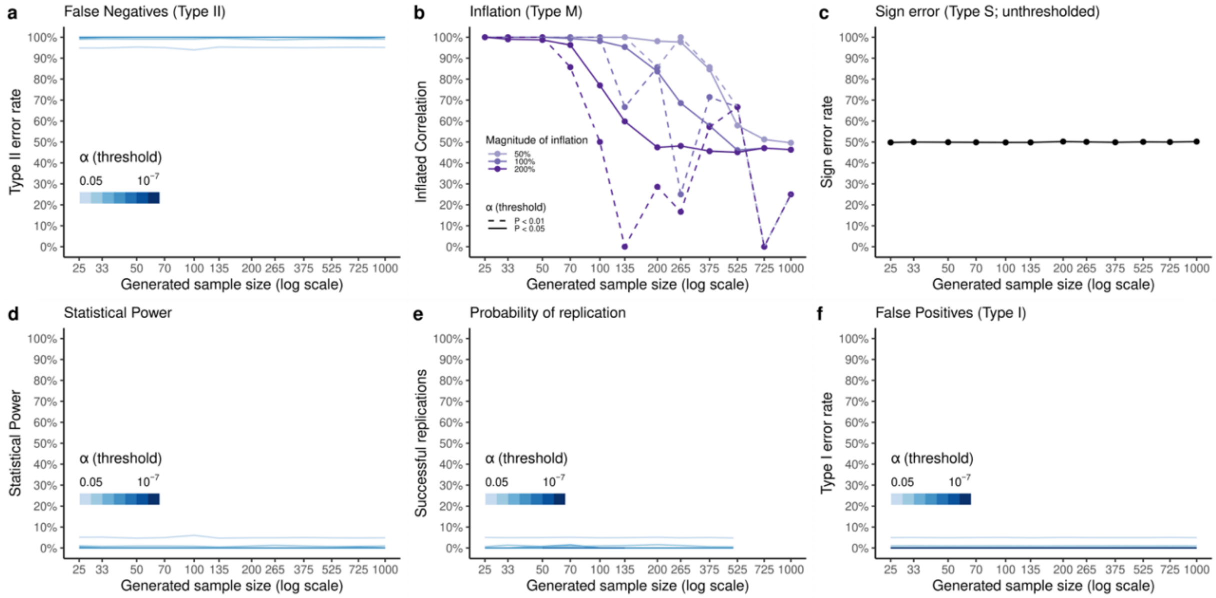
Authors' contributions

C.D.G.B.: Conceptualisation, design, implementation, analysis, interpretation, writing - original draft. A.F.: Interpretation of results, writing - review & editing. G.A.R.: Conceptualisation, design, interpretation, writing - review & editing, supervision.

Competing interests

None.

Extended Data



Extended Data Figure 1: **Expected statistical errors and reproducibility of random noise (ρ = 0)**. We obtain a ground truth of statistical errors under the null by iteratively generating null-samples at increasing sample sizes (*n* = 25, ... , 1,000) instead of resampling from a single null-sample. This corresponds to sampling from an infinite-size population. We average estimates for each sample size over 100 simulations. **a,** False negative rates under the null are constant across sample sizes and equivalent to 1 - α for a given significance threshold (α = .05, .01, .001 plotted). **b,** Since inflation rates are given as a proportion of replicated (same sign, and significant across discovery and replication sample) correlations, we can expect these to be high for small sample sizes as the critical r for significance is higher, and so the likelihood of being inflated decreases as sample sizes increase. Since we are averaging across correlations, few of these will be very inflated while many will be less inflated so we obtain a weighted average of 50% across inflation thresholds for large sample sizes. Noisy estimates are obtained for p<0.01 as only 1,225 edges were simulated, very few were significant across discovery and replication samples (12 expected) and even fewer of the same sign (6 expected). **c,** We expect 50% sign errors regardless of sample size as the sign of a given correlation in a replication null-sample will be random. **d, e, f,** estimates of statistical power, probability of replication, and false positives are based on proportions of significant correlations in replication null-samples, so in each case the probability of a correlation being significant in a newly generated null-sample is exactly determined by the significance threshold (α = .05, ..., 10^{-7}).