

# Replicable brain–phenotype associations require large-scale neuroimaging data

Received: 12 August 2022

Accepted: 25 May 2023

Published online: 26 June 2023


 Check for updates

Shu Liu <sup>1,2</sup> , Abdel Abdellaoui <sup>1,2,3</sup>, Karin J. H. Verweij <sup>1,2,3</sup>  
& Guido A. van Wingen <sup>1,2,3</sup> 

Numerous neuroimaging studies have investigated the neural basis of interindividual differences but the replicability of brain–phenotype associations remains largely unknown. We used the UK Biobank neuroimaging dataset ( $N = 37,447$ ) to examine associations with six variables related to physical and mental health: age, body mass index, intelligence, memory, neuroticism and alcohol consumption, and assessed the improvement of replicability for brain–phenotype associations with increasing sampling sizes. Age may require only 300 individuals to provide highly replicable associations but other phenotypes required 1,500 to 3,900 individuals. The required sample size showed a negative power law relation with the estimated effect size. When only comparing the upper and lower quarters, the minimally required sample sizes for imaging decreased by 15–75%. Our findings demonstrate that large-scale neuroimaging data are required for replicable brain–phenotype associations, that this can be mitigated by preselection of individuals and that small-scale studies may have reported false positive findings.

Neuroimaging studies have aimed to identify the neural basis of individual differences in various variables such as physical conditions, cognition, mental health and lifestyle behaviours, and have revealed structural and functional brain correlates of these phenotypes<sup>1–6</sup>. However, the replicability of the reported brain–phenotype associations remains unknown. There is growing concern about the replication of scientific results<sup>7–9</sup>. In particular, inadequate statistical power in neuroimaging research is thought to lead to high levels of false positive results<sup>10,11</sup> and assessing the replicability has always been a challenge because of the limited sample sizes<sup>12,13</sup>. The heterogeneity of brain–phenotype associations in the population can lead to inconsistent findings, particularly when small sample sizes are used<sup>14–17</sup>. Over the past years, large-scale and high-quality imaging datasets have been collected, which make the examination of replicability of brain–phenotype associations feasible. For example, UK Biobank started the world’s largest multimodal imaging study in 2014 and has released data of ~37,000 individuals with both T1 and resting-state fMRI scanning<sup>18,19</sup>.

A recent study has used data from the Adolescent Brain Cognitive Development (ABCD) study ( $N = 3,928$ ) and UK Biobank ( $N = 32,572$ ) to evaluate the variability of effect sizes of brain-wide associations with varying sampling sizes<sup>20</sup>. The study examined effect size stability, false positives, inflation, sign errors, statistical power, false negative rates and replication probability at different sample sizes. The results showed surprisingly small effect sizes that only reached reasonable statistical power of ~50% with <40% replication probability at half the full sample. However, replicability was assessed across structural and functional imaging modalities, across 41 psychological and demographic phenotypes and across the entire brain. The results therefore apply to a certain type of neuroimaging study where multiple neuroimaging modalities are examined in relation to multiple phenotypes simultaneously. In contrast, the conventional neuroimaging study typically focuses on exploring the relationship between a single neuroimaging modality and a single phenotype, with the goal of identifying brain regions or connections that are associated with that phenotype. As a

<sup>1</sup>Department of Psychiatry, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands. <sup>2</sup>Amsterdam Neuroscience, Amsterdam, the Netherlands. <sup>3</sup>These authors jointly supervised this work: Abdel Abdellaoui, Karin J.H. Verweij, Guido A. van Wingen.  e-mail: [s.liu@amsterdamumc.nl](mailto:s.liu@amsterdamumc.nl); [g.a.vanwingen@amsterdamumc.nl](mailto:g.a.vanwingen@amsterdamumc.nl)

result, it is unclear which modalities and phenotypes are more or less reliable in terms of their replicability.

To better examine the replicability of brain–phenotype associations, we used the UK Biobank dataset ( $N = 37,447$ ) to assess relationships of structural and functional brain measures with six representative variables for different domains<sup>21</sup>. For our main analysis, we selected physical variables that were expected to have strong associations with brain measures (age and body mass index (BMI)), variables assessing cognitive performance (fluid intelligence and numeric memory) and variables associated to mental health and lifestyle (neuroticism and alcohol consumption). We additionally included a variable that was expected to provide a lower bound on brain–phenotype associations (birth month)<sup>22</sup>. To evaluate whether the results would generalize to other variables from these phenotype domains, we tested brain–phenotype associations for another 23 variables. The structural brain measures included cortical surface area (CSA) and thickness (CT) based on 66 regions of the Desikan–Killiany (DK) atlas and the functional measures were derived from resting-state fMRI consisting of 210 functional connectivities (FCs) between different networks identified with independent component analysis (ICA). For each variable, we assessed the replicability of the pattern of brain–phenotype associations across the whole brain (global replicability) and of single brain–phenotype association separately (regional replicability). We also examined the improvement of replicability with sampling sizes from 100 to half the full sample.

## Results

### Research design

The UK Biobank participants included in our study are predominantly of European ancestry and range in age from 44 to 82 years old, with an approximately equal number of males and females. Supplementary Fig. 1 and Supplementary Table 1 show the distribution of six representative variables. The overview of the analytical steps of this study is presented in Fig. 1. In short, we randomly selected two non-overlapping subsamples from the full sample (total sample size  $N$  ranges from 25,231 to 37,447). We then conducted univariate Spearman's rank correlation analysis to examine the associations between brain measures and various variables in two independent subsamples. Next, we estimated the global replicability including intraclass correlation coefficients (ICC) and Jaccard index based on brain–phenotype associations from two subsamples<sup>23</sup>. We calculated the ICC for absolute agreement of brain–phenotype associations between subsamples for the strongest associated brain measures (10%, 25%, 50% or 100%). To calculate the Jaccard index, we used different thresholds ( $P < 0.05$ ,  $P < 0.01$ ,  $P_{\text{FDR}} < 0.05$  or  $P_{\text{Bonferroni}} < 0.05$ ) to select significant brain–phenotype associations and estimated the overlap between the subsamples. Here,  $P_{\text{FDR}}$  and  $P_{\text{Bonferroni}}$  represent the  $P$  values adjusted for false discovery rate (FDR) and Bonferroni correction, respectively. We also proposed a new regional replication index to estimate the replication probability for single brain–phenotype association across 100 sampling repetitions at a specific sample size. Similarly, four different significance thresholds were used to select significant associations (Methods).

### Brain–phenotype associations

Before assessing replicability, we first estimated the association strength between different brain measures and variables in the full sample (Supplementary Tables 2–4). Widespread significant CSA–phenotype associations were observed for age, BMI, fluid intelligence and numeric memory at the thresholds of  $P < 0.05$ ,  $P < 0.01$ ,  $P_{\text{FDR}} < 0.05$  or  $P_{\text{Bonferroni}} < 0.05$  (Supplementary Fig. 2a). For CT, age, BMI and alcohol consumption showed more significant brain–phenotype associations than the other variables (Supplementary Fig. 2b). For FC, all variables except for birth month showed widespread FC–phenotype associations at different thresholds (Supplementary Fig. 2c). Birth month, which was included to estimate the lower bound, did not show any

significant associations with brain measures after correction for multiple comparisons. As shown in Extended Data Fig. 1, except for age, the largest absolute effect sizes for brain–phenotype associations were small ( $\max |r_s| < 0.1$ ) and correlations were larger for FC than for CSA and CT. These results show that, in general, correlations between brain measures and phenotypes are small.

### Estimating global replicability

We then estimated the replicability of brain–phenotype associations across the brain as measured with ICC at increasing sampling sizes. In general, ICC values  $> 0.75$  were considered good, while ICC values between 0.5 and 0.75 were considered to indicate moderate replicability<sup>24</sup>. As shown in Fig. 2a, using the largest 25% brain–phenotype associations, ICC values for CSA reached 0.75 for age, BMI, numeric memory and fluid intelligence, at sample sizes of 700, 3,400, 5,500 and 10,500 individuals, respectively. Considerably fewer individuals were required for these four variables to achieve moderate replicability, with required sample size of 300, 1,200, 2,200 and 4,400, respectively. However, neuroticism and alcohol consumption did not exhibit good replicability, even with the largest sampling sizes of 15,200 and 17,100 individuals (half the full sample), respectively. Additionally, it required almost 13,700 and 14,500 individuals to attain moderate replicability of 0.5 for these two variables, respectively.

For CT, at least 300, 1,500, 6,100 and 14,900 individuals were needed to reach good ICCs for age, BMI, alcohol consumption and fluid intelligence, whereas ICCs for other variables did not reach 0.75 even including half the full sample (Fig. 2b). Except for birth month, all variables achieved moderate ICCs (age, 100; BMI, 600; alcohol consumption, 2,400; fluid intelligence, 6,700; neuroticism, 8,600; numeric memory, 10,500).

FC generally showed better ICCs than CSA and CT (Fig. 2c). For FC, associations with age, BMI, fluid intelligence, numeric memory, neuroticism and alcohol consumption required 800, 2,800, 4,500, 5,200, 6,300 and 9,100 individuals to achieve good ICCs and 300, 1,100, 1,900, 2,000, 2,900 and 3,500 individuals to reach moderate ICCs.

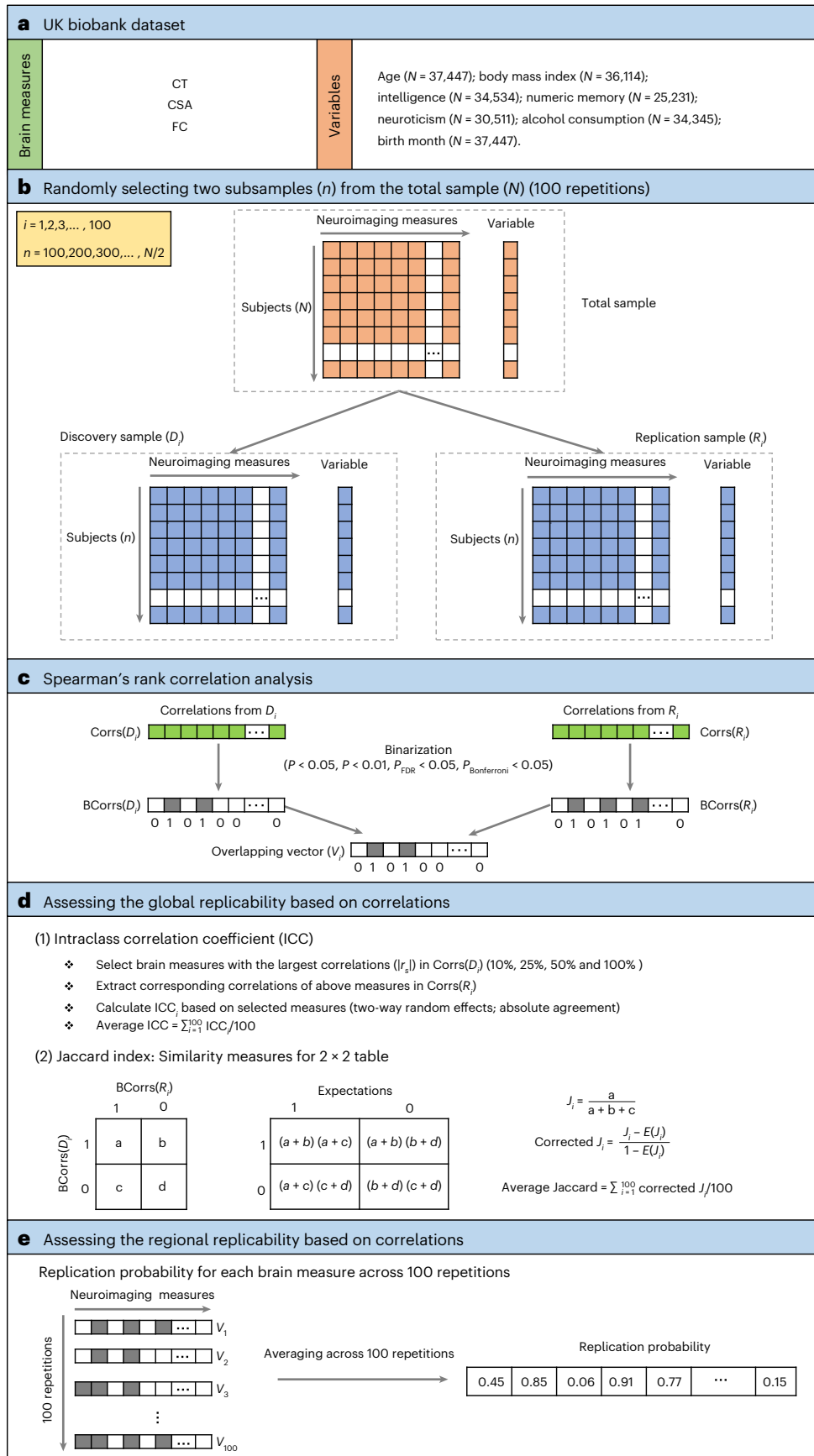
To assess whether ICCs were dependent on the number of included brain measures, we also used thresholds of 10%, 50% and 100%. This analysis showed that ICCs were lower when more brain measures were included (Supplementary Fig. 3).

In addition to ICC which directly measures the consistency of observed effect sizes, the Jaccard index was estimated to measure the overlap of significant brain–phenotype associations between two subsamples. At the significance threshold of  $P < 0.05$  (uncorrected for multiple testing), brain–phenotype associations for age and BMI in all three modalities (CSA, CT and FC) generally obtained Jaccard indices from 0.4 to 0.6 at half total sample size, indicating that  $> 40\%$  of the observed significant brain–phenotype associations were shared between the two independent subsamples (Fig. 2d–f). Moreover, the significant associations between CSA and numeric memory showed  $\sim 30\%$  overlap at half total sample size (Fig. 2d), whereas the significant associations between CT and alcohol consumption maximally reached a Jaccard index of 0.28 (Fig. 2e). For all six variables, significant association with FC generally showed higher overlap between the independent subsamples than the associations with CSA and CT (Fig. 2f).

Comparable results were obtained when estimating the Jaccard index at other significance thresholds ( $P < 0.01$ ,  $P_{\text{FDR}} < 0.05$  or  $P_{\text{Bonferroni}} < 0.05$ ) (Supplementary Fig. 4).

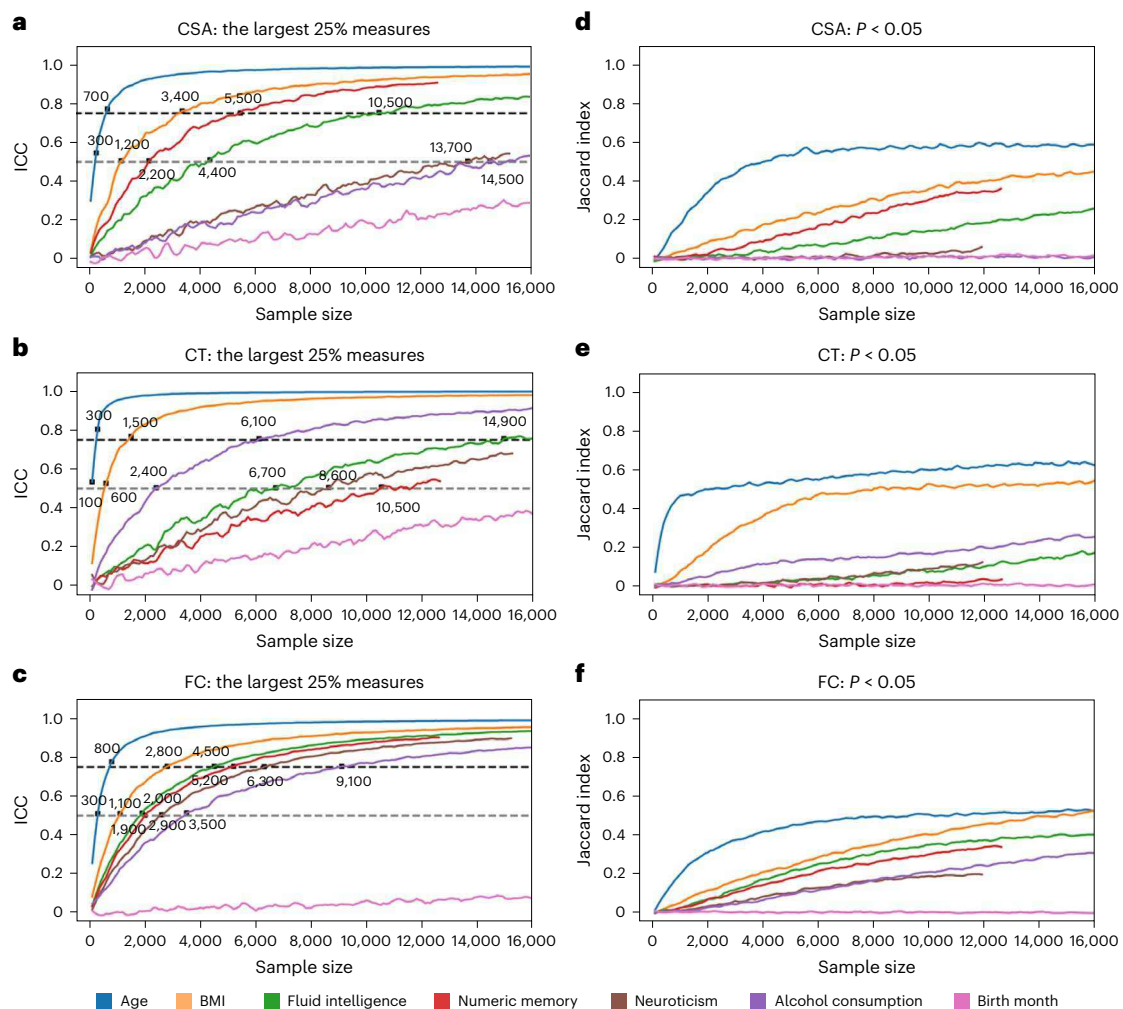
### Estimating regional replicability

The global replicability indices only measure the correspondence of brain-wide associations between two samples, which is not suitable for identifying the replicability for specific brain features. We developed a regional replicability index to estimate the replication probability between independent samples for single brain–phenotype association through repeated resampling. In line with the recommendation for



**Fig. 1 | Overview of the analytical steps in this study to estimate replicability.** **a**, The input data from UK Biobank. **b**, Randomly selecting two non-overlapping subsamples from the total sample. **c**, Spearman's rank correlation analysis in

two independent subsamples. **d**, Assessing the global replicability. **e**, Assessing the regional replicability. Corrs represents the correlation vector, while BCorrs represents the binary vector obtained through binarization of Corrs.



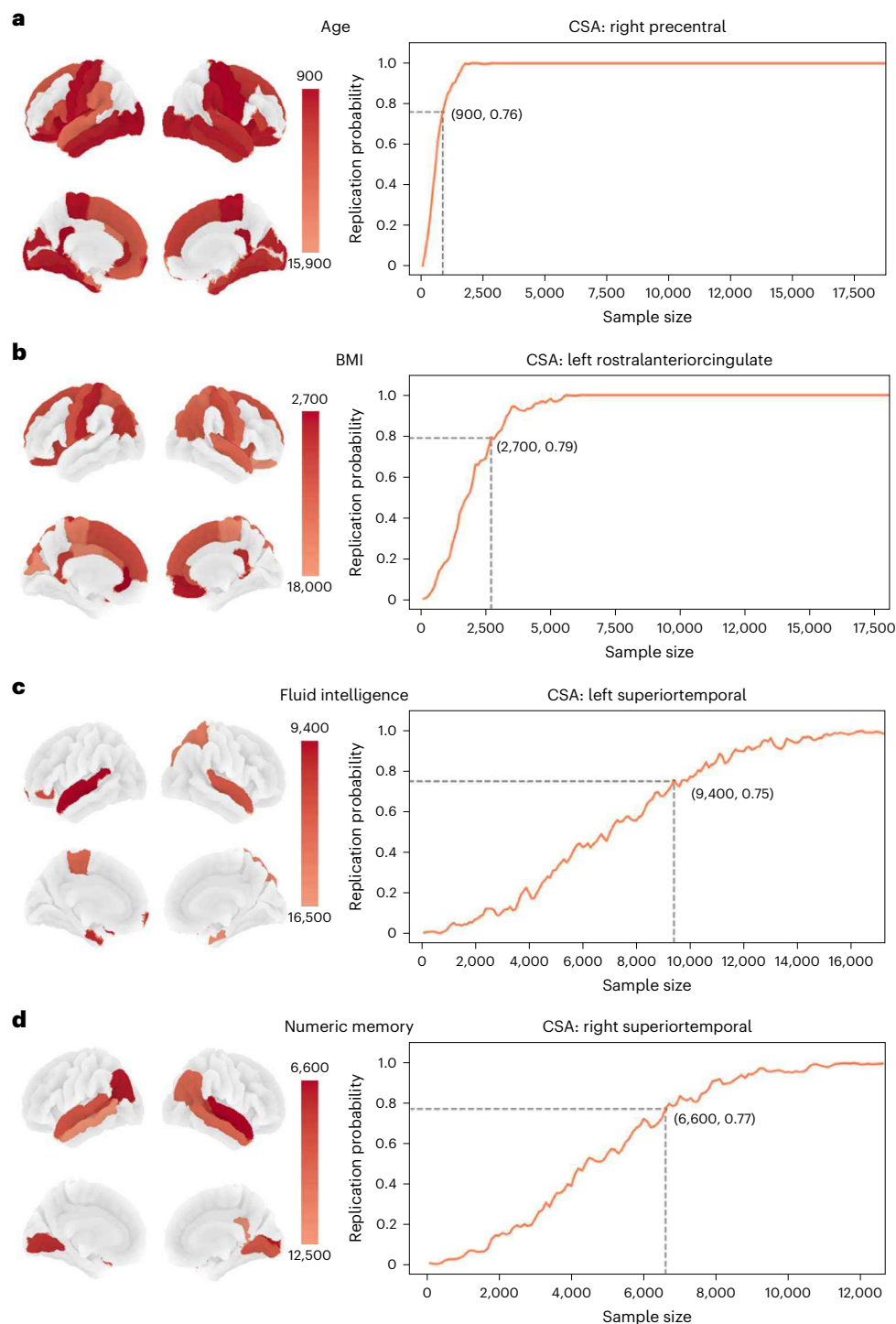
**Fig. 2 | Improvement of global replicability with increasing sample size.** **a, d**, The ICC (**a**) and Jaccard index (**d**) for CSA. **b, e**, The ICC (**b**) and Jaccard index (**e**) for CT. **c, f**, The ICC (**c**) and Jaccard index (**f**) for FC. The dotted lines indicate good and moderate replicability levels (0.75 and 0.5).

ICC<sup>24</sup>, values  $>0.75$  were considered to reflect good regional replicability, indicating that individual significant associations can be replicated independently with at least 75% probability. At different significance thresholds, widespread brain measures could achieve good regional replicability for age and BMI at smaller sample sizes than half the full sample (Extended Data Fig. 2).

When using the threshold of  $P < 0.05$ , significant associations between CSA and variables that achieved 75% replication probability are shown in Fig. 3 and Supplementary Table 5. For age, significant associations with 47 brain regions had regional replicability  $>0.75$ . Of them, the right precentral gyrus showed the highest regional replicability, requiring the smallest sample size (900 individuals) to achieve  $>75\%$  replication probability (Fig. 3a). For BMI, significant associations with 27 regions reached 75% replication probability, with the best replicability (left rostral anterior cingulate cortex) requiring 2,700 individuals to obtain 79% replication probability (Fig. 3b). Only 8 and 13 regions showed significant associations with fluid intelligence and numeric memory with good regional replicability, respectively. Of them, the same brain region, superior temporal cortex, was observed for intelligence and memory to have highest replicability, requiring 9,400 and 6,600 individuals, respectively (Fig. 3c,d). In contrast, for neuroticism and alcohol consumption, no regions were identified to have replicable associations with CSA, not even at the largest sampling sizes.

For CT, 48 and 40 regions had significant associations with age and BMI with regional replicability of  $>0.75$  at the significance threshold of  $P < 0.05$  (Supplementary Table 6), with the transverse temporal gyrus and superior parietal cortex having the highest replicability, requiring 300 and 1,800 individuals to achieve  $>75\%$  replication probability, respectively (Fig. 4a,b). CT of the transverse temporal gyrus and superior frontal cortex showed significant associations, with fluid intelligence achieving good replication at  $>13,600$  individuals (Fig. 4c). For neuroticism,  $>12,000$  individuals were required to obtain good replication for associations with CT in the superior parietal cortex and parahippocampal gyrus (Fig. 4d). In addition, 11 regions mainly located in frontal cortex and temporal cortex showed replicable significant associations with alcohol consumption and the best replicable region (left superior frontal cortex) required 3,500 individuals to obtain 75% replication probability (Fig. 4e). No regions were identified for numeric memory to have replicable associations with CT.

In general, FC required fewer individuals than CSA or CT to achieve good regional replicability (Extended Data Fig. 2 and Supplementary Table 7). Among 210 FCs between 21 ICA networks, 139 and 106 FCs had significant associations with age and BMI with good replication probability, requiring a sample of 500 and 1,500 individuals, respectively (Fig. 5a,b). For cognitive variables, 50 and 34 FCs had replicable significant associations with fluid intelligence and numeric memory.



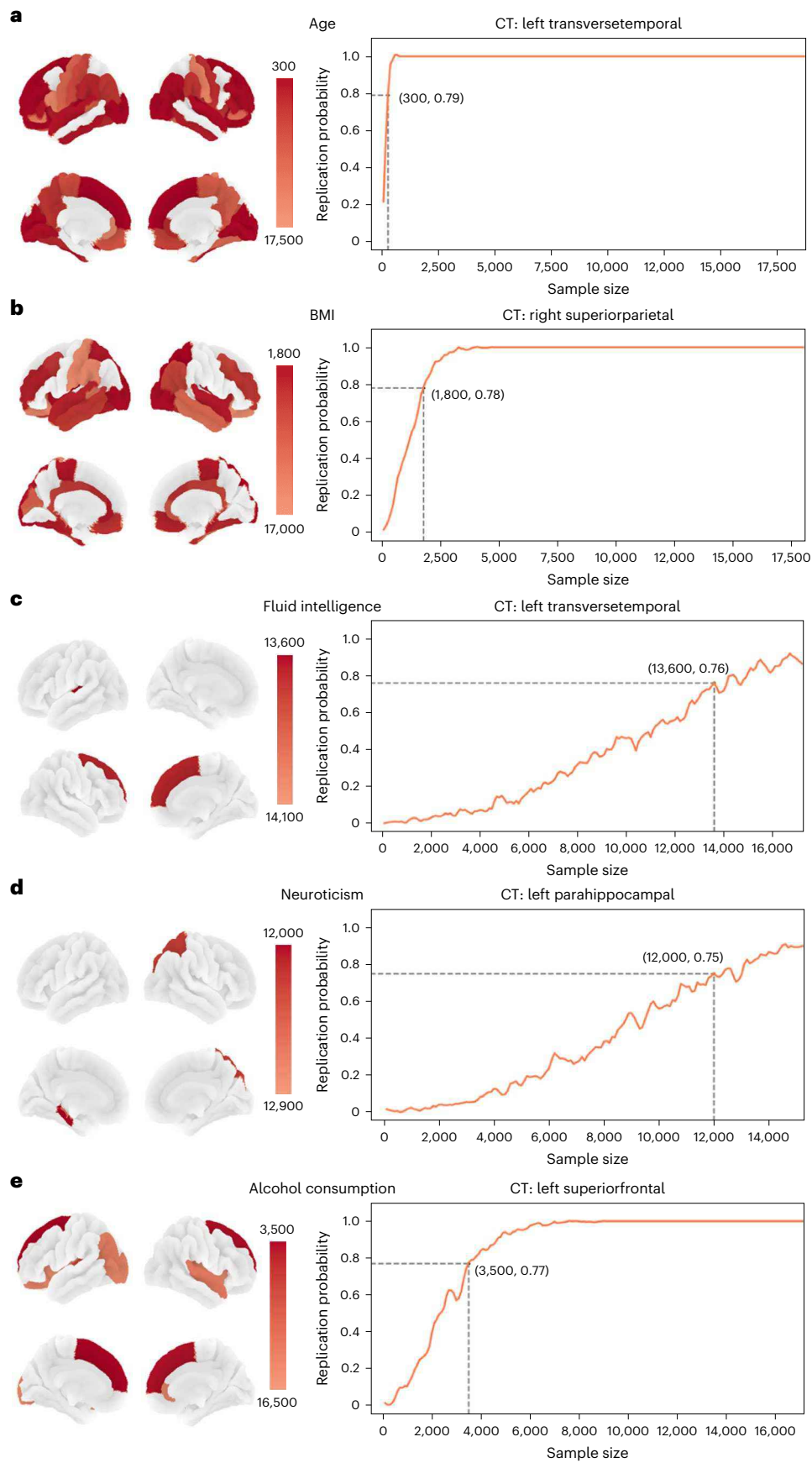
**Fig. 3 | Improvement of regional replicability with increasing sampling size for CSA.** Left panels show required sample sizes for brain regions to achieve good regional replicability of  $>0.75$ . Right panels show the improvement at increasing sample size for the best replicable brain region, requiring minimal sample sizes. **a–d**, Age (**a**); BMI (**b**); fluid intelligence (**c**); numeric memory (**d**).

The best replicable FC for fluid intelligence and numeric memory required at least 3,700 and 2,600 individuals (Fig. 5c,d). In addition, 31 and 35 FCs showed significant associations with neuroticism and alcohol consumption with good replicability, requiring at least 3,900 and 6,400 individuals, respectively (Fig. 5e,f).

These findings were generally confirmed when using other significance thresholds of  $P < 0.01$ ,  $P_{\text{FDR}} < 0.05$  or  $P_{\text{Bonferroni}} < 0.05$  (Supplementary Tables 8–10). Paradoxically, stricter significance

thresholds required larger sample sizes to achieve good replicability (Supplementary Fig. 5).

We further examined the relationships between effect sizes and replicability. As shown in Extended Data Fig. 1, the largest absolute effect sizes for all variables have been derived from the full sample. We then calculated the correlations between the largest effect sizes and the minimally required sample sizes to achieve good regional replicability (Fig. 6). We found a significant negative power law relation between



**Fig. 4 | Improvement of regional replicability with increasing sampling size for CT.** Left panels show required sample sizes for brain regions to achieve good regional replicability of >0.75. Right panels show the improvement at

increasing sample size for the best replicable brain region, requiring minimal sample sizes. **a–e**, Age (**a**); BMI (**b**); fluid intelligence (**c**); neuroticism (**d**); alcohol consumption (**e**).

them (the fit line:  $y = 15.29x^{-1.82}$ ,  $R^2 = 0.993$ ,  $P = 2.42 \times 10^{-15}$ ), confirming that stronger brain–phenotype associations can be replicated better. Thus, the sample size required for good replicability increases drastically for weak associations.

To evaluate whether the results for these six representative variables would generalize to other phenotypes from the same domains, we investigated the associations between the brain and 23 additional phenotypes derived from UK Biobank (Supplementary Table 11 and Supplementary Fig. 6). We again estimated replicability at varying sample sizes with a threshold of  $P < 0.05$ . In general, these results were comparable to those for the six representative variables (Supplementary Figs. 7–13), indicating that the results in the main analysis generally represent the corresponding categories of physical, cognitive, mental health and lifestyle measures. Furthermore, we used the estimated power law relationship described above to predict the required sample sizes for new phenotypes based on effect sizes of brain–phenotype associations. We found that the predicted sample sizes closely matched the observed required sample sizes for these phenotypes, providing further evidence for the power law relation between effect size and the required sample size for replicable associations (Extended Data Fig. 3). For further information, please refer to the Supplementary Results.

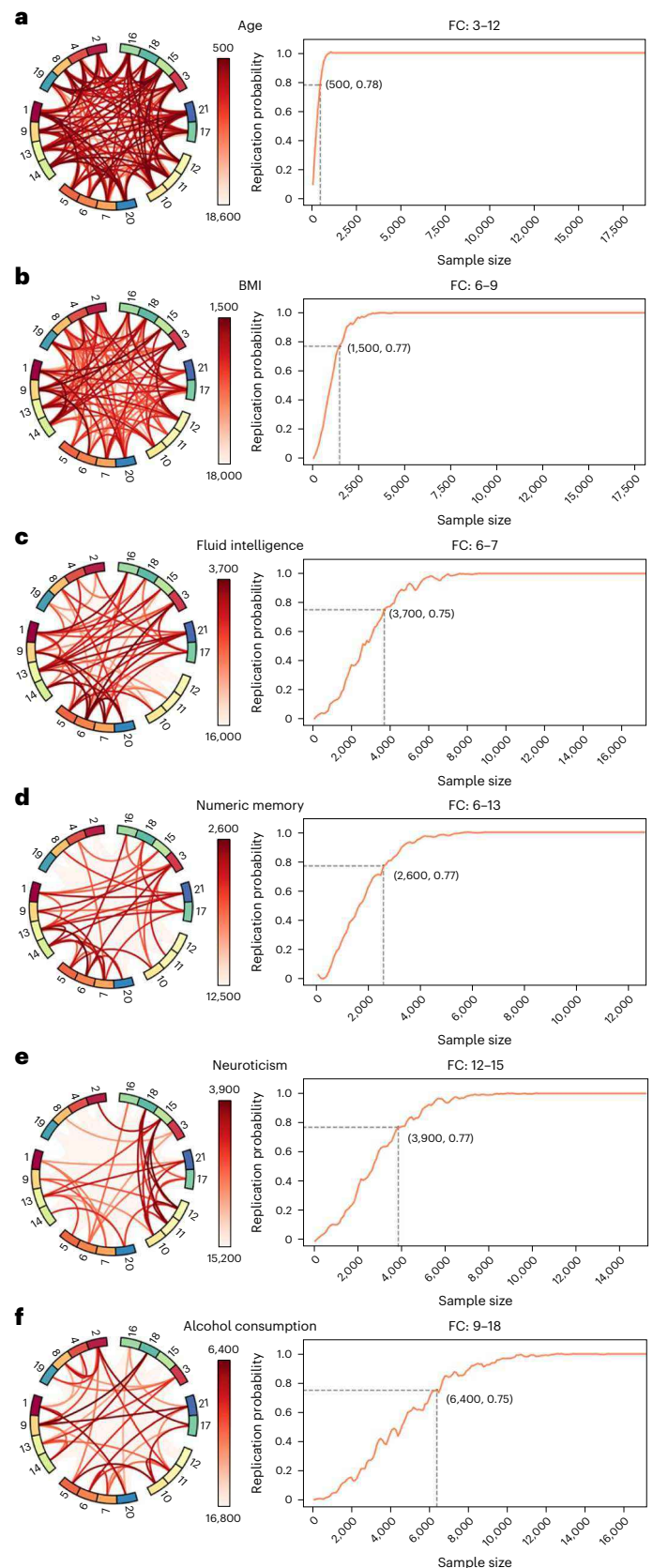
### Estimating replicability for group comparisons

The above correlation analyses showed that large samples are required to obtain replicable brain–phenotype associations for cognitive, mental health and lifestyle variables. We further investigated whether we could decrease the required sample sizes by preselection strategies.

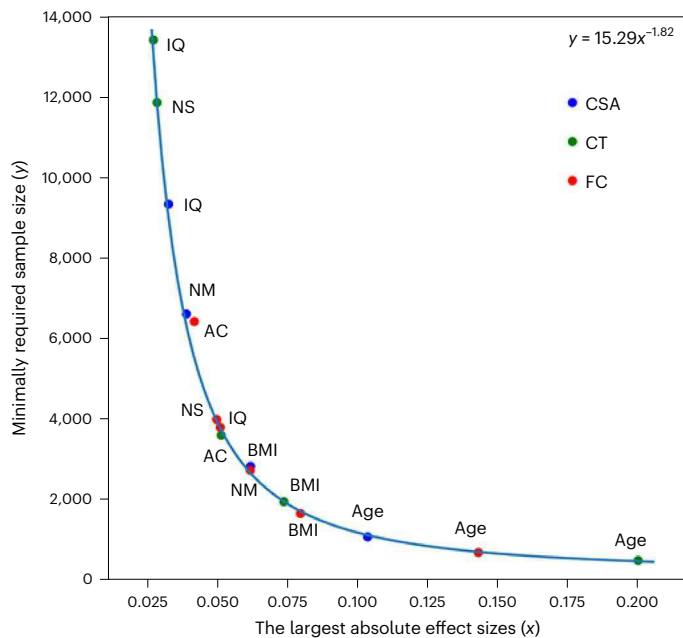
First, we used the median phenotype values to split the full sample into two groups while matching the sample sizes of the two groups as closely as possible (Supplementary Fig. 14). Then, we used two-sample  $t$ -tests to compare the neuroimaging measures between the two groups and estimated the regional replicability for each brain measure at different sampling sizes. When using a significance threshold of  $P < 0.05$  uncorrected, CSA–phenotype associations showed comparable replicability with the correlation analysis (Supplementary Fig. 15). For CT, the minimally required sample size for fluid intelligence to achieve good regional replicability dropped from 13,600 to 6,800 but this did not change much for the other variables (Supplementary Fig. 16). Conversely, for FC, replicable associations with BMI, numeric memory, neuroticism and alcohol consumption required more individuals (Supplementary Fig. 17). These results indicate that, in general, comparing small- and large-scoring individuals through median split with two-sample  $t$ -tests had little influence on the replicability (Extended Data Fig. 4). This was also the case when using other significance thresholds (Supplementary Tables 12–14).

Second, we calculated the quartiles and used the lower and upper quarters to select two groups with the smallest and largest 25% scores (Supplementary Fig. 18). In this situation, to achieve 75% replication probability, the minimally required sample sizes decreased 15–75% compared with correlation analysis (Fig. 7 and Supplementary Fig. 19). Age, BMI and alcohol consumption showed the best replicability with CT, requiring 200, 500 and 2,600 individuals of the pericalcarine cortex, superior parietal cortex and superior frontal cortex to reach 75% replication probability of the  $t$ -statistics, respectively, whereas fluid intelligence, numeric memory and neuroticism showed better replicability with FC, requiring at least 2,400, 1,900 and 2,800 individuals to achieve good regional replicability (Supplementary Figs. 20–22). Similarly, the minimally required sample sizes were also found at the other thresholds of  $P < 0.01$ ,  $P_{\text{FDR}} < 0.05$  or  $P_{\text{Bonferroni}} < 0.05$  (Supplementary Tables 15–17).

To further evaluate the influence of preselection on replicability, we conducted a more extensive analysis for the six representative variables and selected two groups on the basis of the smallest and largest scores at 10%, 20%, 30% and 40%. We then compared the minimum required sample size for each preselection criterion (Supplementary



**Fig. 5 | Improvement of regional replicability with increasing sampling size for FC.** Left panels show required sample sizes for connections to achieve good regional replicability of  $>0.75$ . Right panels show the improvement at increasing sample size for the best replicable functional connection, requiring minimal sample sizes. **a–f**, Age (**a**); BMI (**b**); fluid intelligence (**c**); numeric memory (**d**); neuroticism (**e**); alcohol consumption (**f**).



**Fig. 6 | The relationships of the largest absolute effect sizes derived from the full sample with the minimally required sample sizes needed to achieve regional replicability of 0.75.** A power trend line is fit for the points, in which the minimally required sample size is a function of the largest absolute effect sizes ( $y = 15.29x^{-1.82}$ ). BMI, body mass index; IQ, fluid intelligence; NM, numeric memory; NS, neuroticism score; AC, alcohol consumption.

Table 18 and Extended Data Fig. 5). Overall, stricter preselection led to lower minimum sample sizes. Specifically, we found that selecting two groups with scores falling below the lowest and above the highest 30% generally resulted in considerably lower sample size requirements for ensuring good replicability as compared to correlation analysis. Moreover, we also compared the minimum sample size necessary for additional phenotypes that were derived from different preselection methods (Supplementary Table 19 and Extended Data Figs. 6 and 7). All these findings provide support for the conclusion that preselection can effectively reduce the minimum required sample size for replicable brain associations. For further information, please refer to the Supplementary Results.

### Estimating replicability for multivariate methods

To investigate whether we could further reduce the required sample sizes for various variables, we used multivariate random forest regression to select the most important brain measures and estimated the replicability of feature selection (Supplementary Fig. 23). Extended Data Fig. 8 showed the improvement of replicability with increasing sampling size after selecting the top important brain features. Only age obtained Jaccard indices  $>0.6$  at its largest sampling size of  $N = 18,000$  (half the full sample size for age), which indicated that at best 60% of the selected structural and functional features were shared between the two independent samples. Other variables showed poor replicability after feature selection. Overall, the replicability of brain measures after multivariate feature selection was worse than the replicability of selected brain measures after univariate correlation analysis.

We additionally used partial least squares (PLS) regression to investigate whether a multivariate brain pattern could improve the replicability of brain–phenotype associations (Supplementary Fig. 24). We assessed the replicability of multivariate brain–phenotype associations by comparing the correlations of the first PLS component (PLS1) with variables between the discovery and replication sample (Extended Data Fig. 9a–c). Overall, the difference in the

correlations of PLS1 with variables decreased with increasing sample size, indicating that better replicability was achieved at larger sample sizes. Additionally, we assessed the replicability of PLS weights derived from two independent samples by calculating ICCs at increasing sample size (Extended Data Fig. 9d–f and Supplementary Fig. 25). These ICCs were comparable to those obtained from the univariate correlation analyses, indicating that simple multivariate statistical analysis alone cannot substantially improve the replicability of brain–phenotype associations.

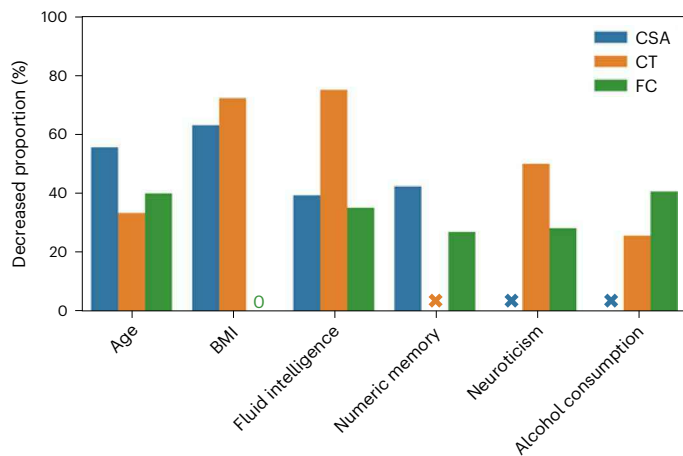
### Discussion

We systematically examined global and regional replicability for brain–phenotype associations for six phenotypes within different domains: physical, cognitive, mental health and lifestyle. Through correlation analysis, age started to show good global replicability for CT at a sample of 300 individuals, highlighting that replicable associations with brain measures do not necessarily require very large sample sizes. Brain–phenotype associations for the other variables required larger sample sizes than age and showed better regional than global replicabilities. Specifically, good regional replicability of associations with cognitive variables required at least 2,600 individuals and mental health and lifestyle variables at least 3,500 individuals. Corresponding numbers for global replicability were 4,500 and 6,100 individuals, respectively. The required sample size showed a negative power law relation with the estimated effect size and can be reduced considerably with 15–75% when only the lowest and highest scoring quarters are compared, although still requiring about 2,000 individuals. These findings are in line with the multimodal and multiphenotype results from ref. 20 and demonstrate that thousands of individuals are also required for good replicability of associations between one brain measure and one phenotype, suggesting that significant associations from previous studies with small samples may not be replicable and hence may represent false positive findings.

Physical variables showed better global replicability than the variables related to cognition and mental health. Additionally, functional brain measures generally showed better global replicability than structural brain measures. Through estimating regional replicability, we identified the brain measure with the best replicability for each variable. CT of the transverse temporal gyrus and superior parietal cortex were identified to have replicable significant associations with age and BMI, respectively, which required at least 300 and 1,800 individuals to achieve  $>75\%$  replication probability. Associations with FC for these two variables showed comparable replicability, with minimally required sample sizes of 500 and 1,500. In contrast, numeric memory, fluid intelligence, alcohol consumption and neuroticism required more than 2,600, 3,700, 2,600 and 3,900 individuals for replicable associations with CT or FC, respectively. Preselection of the first and fourth quarters of individuals reduced the required sample size for good replicability to generally  $\sim 2,000$  individuals. However, it should be noted that, even though the preselection procedure significantly reduces the minimally required sample sizes for imaging, it still relies on information from the entire sample to identify individuals in the upper and lower quarters.

Small sample sizes have been consistently considered as the critical factor for non-reproducibility in neuroimaging research<sup>25–28</sup>. In population-based research, results from smaller sample sizes are usually more contingent with a reduced likelihood that a statistically significant result reflects a true effect<sup>11,29</sup>. For example, previous studies demonstrated that correlation patterns between cognitive performance and structural measures had poor replication in two samples of  $\sim 100$  healthy adults<sup>13,30</sup>. A recent study also reported that large sample sizes are required to fully extract predictive information with machine learning models based on neuroimaging data<sup>31</sup>. Two main proposals explaining the poor reproducibility at small sample sizes have been put forward<sup>32</sup>. First, studies in small samples may only capture a partial and





**Fig. 7 | The decreased proportion of minimally required sample size to reach good replicability between a two-sample  $t$ -test (by the first and fourth quarters) and correlation analysis.** Bars indicate decreased minimally required sample sizes comparing the two-sample  $t$ -test to Spearman's correlation analysis. Cross symbol 'x' indicates that the minimally required sample sizes are missing for the two-sample  $t$ -test or Spearman's correlation analysis because the regional replicability of all brain measures did not reach 0.75 at any sample size. Zero '0' indicates that the minimally required sample sizes are the same between the two methods.

minor aspect of the true association pattern. Second, the measurement of both the phenotypes and brain measures are susceptible to noise and thus smaller samples are more likely to produce spurious significant results than are large samples<sup>33</sup>. This is similar to previous candidate-gene studies where many significant associations between common genetic variants and various phenotypes could not be replicated due to small sample sizes<sup>34,35</sup>. Since then, genomic datasets have grown to hundreds of thousands and even millions in genome-wide association studies (GWAS) to generate robust associations<sup>36</sup>. Therefore, to ensure replicable brain-phenotype associations, we will need to increase brain imaging sample sizes to minimize the effects of population sampling variability.

Our study found that the strength of the associations heavily influenced the replicability. We found a negative power law relation between the largest absolute effect sizes derived from the full sample and the minimally required sample sizes, indicating that stronger brain-phenotype associations can be better replicated. According to previous machine learning studies, relatively high prediction accuracy could be achieved for age<sup>31,37</sup>, whereas, until now, predictive accuracies for cognitive and behavioural variables and psychiatric diagnoses are still low<sup>38,39</sup>. Besides the reason that physical variables can be measured more objectively, the weak associations with cognitive and behavioural variables may lower the statistical power, which makes it difficult to replicate the significant observations<sup>40</sup>. Strict multiple comparison corrections are commonly used in neuroimaging research to identify statistically significant associations, which can reduce false positive results. However, stricter thresholds in correlation analysis will lower statistical power as well and result in high rates of false negatives<sup>41</sup>. In our study, at stricter thresholds, larger sample sizes were required to achieve 75% replication probability while estimating the regional replicability of brain-phenotype associations. Therefore, conservative multiple testing correction with strict thresholds hampered the replication of associations.

In addition to correlation analyses, we conducted group comparisons with two-sample  $t$ -tests. We divided the sample into four phenotype quarters and used the first and fourth quarters to select two groups with the largest and smallest 25% values and found better

replicability which generally required about 2,000 individuals. This improvement was not observed when comparing the lower with the upper half. Because of the negative correlation between effect sizes and replicability, this suggests that the effect sizes for comparing extreme cases were larger. This may translate to case-control neuroimaging studies comparing patients with controls. For example, ref. 42 examined the replicability for differences between major depression and healthy controls in functional brain measures based on rsfMRI dataset with 1,434 participants and found that at least 400 subjects in each group were required to replicate significant brain differences between groups. In addition, an abnormal gradient map in schizophrenia identified in a discovery dataset with 400 patients and 336 controls was found to be reproducible in the replication dataset with 279 patients and 262 controls<sup>43</sup>. Our study warrants further research to determine more specific sample size requirements for good replicability of case-control differences.

A recent study reported that brain-phenotype associations have small effect sizes and reported that thousands of individuals are required to ensure reproducibility<sup>20</sup>. However, the study merged different neuroimaging measures, phenotypes and brain regions to estimate the overall replication probability. The study therefore does not provide estimates for typical neuroimaging studies that aim to identify which brain regions are associated with a particular phenotype. By developing a measure for regional replicability and selecting particular phenotypes, our results demonstrate that 300 individuals may be sufficient for small ( $|r| = 0.15$ ) effect sizes and that >2,000 individuals are only required for very small ( $|r| < 0.05$ ) effect sizes. Furthermore, the results showed that good global replicability typically required considerably larger samples than regional replicability, suggesting that the required sample sizes in ref. 20 were overestimated. A commentary to the findings of ref. 20 emphasized that more data alone do not necessarily lead to better science<sup>44</sup>. The authors proposed that the key to building more robust brain-based predictive models is to make thoughtful choices about brain data acquisition, behavioural targets of prediction and approaches to model building. Authors from another commentary proposed that maximizing differences in the measures under investigation can improve the reliability of brain-behaviour associations<sup>45</sup>. This can be achieved through various means, such as carefully selecting subjects, conducting experimental or causal manipulations or focusing on measures that are theoretically motivated. Our study is consistent with the suggestions by previous research, which has shown that preselecting the sample can reduce the required sample size.

We showed that structural brain measures generally showed poorer replicability than functional brain measures. Paradoxically, test-retest reliability for functional MRI is lower than for structural MRI<sup>46-48</sup>. This indicates that the additional noise from lower test-retest reliability of functional MRI is less important than the higher association strength for the replication of brain-phenotype associations. This is in line with a previous study which reported that test-retest reliability of resting-state FC was not meaningfully correlated with the contribution of FC to behavioural prediction<sup>49</sup>.

In machine learning models, feature selection removes redundant input features to improve the generalization capability and lowers the computation costs. Therefore, robust feature selection is important to conduct classification or prediction based on neuroimaging features. Here, we used feature selection to determine whether this could improve replicability but we only found good replicability for age. The multivariate bootstrap-based feature selection showed poorer replicability than univariate correlation analysis, which indicated that the weak brain-phenotype associations may produce unstable feature importance. Therefore, even larger datasets may be required to produce robust feature selection in machine learning<sup>31</sup>. It should be noted that the comparisons we conducted only pertained to the replicability of initial feature selection. For final predictions, additional techniques

such as cross-validation can be used to estimate the generalization performance of machine learning models.

Although our results demonstrate the importance of large samples to identify replicable brain correlates with individual differences, small sample neuroimaging-only studies are adequately powered in some situations. For example, the neuroimaging data of 40 healthy adults were sufficient to produce replicable brain parcellation, the brainnetome atlas<sup>50</sup>. Small samples with ~20 individuals could accurately represent the central tendencies of human functional brain organization<sup>20</sup> and produce predictive network masks that successfully generalize across datasets and populations<sup>51</sup>. Moreover, small samples have been widely used to produce robust patterns of brain activation to tasks<sup>52,53</sup>. Thus, small sample neuroimaging studies can provide critical information for the development of brain research.

Several limitations of our study should be noted. First, we used the UK Biobank to investigate replicability because it is by far the largest neuroimaging dataset available. It is a population study and the results are therefore likely to generalize to other individuals. However, all participants were middle-aged and elderly individuals from the UK and there were some selection biases in sociodemographic, physical, lifestyle and health-related characteristics<sup>54</sup>. It therefore remains unclear whether the results may also generalize to younger individuals and those from other continents and backgrounds. On the basis of our results, we are not aware of the existence of any other public datasets available that included sufficient individuals to replicate most results. Second, we only considered six variables from the UK Biobank for our extensive experiments, which were selected to be representative for physical, cognitive, mental health and lifestyle domains. However, our results from more than 20 additional phenotypes suggest that the required sample sizes are comparable within domain and may therefore also translate to other variables from these domains but other sample sizes may be required for other variables and particularly other domains. Third, our study used 210 FC measures, which is coarse compared to tens of thousands of connections usually used in FC-based predictive modelling work.

In conclusion, we found that the replicability of associations between brain measures and variables assessing individual phenotype differences depend on sample size as well as the association strength. Through correlation analysis, we identified that the required sample size to obtain good replicability of brain–phenotype associations is 2,700 individuals for cognition and 3,900 for mental health and lifestyle variables. Moreover, the corresponding required sample sizes decreased to ~2,000 individuals when individuals are preselected for extreme scores. This study thus demonstrates that thousands of individuals are required for good replicability of brain–phenotype associations. This suggests that results from previous studies investigating interindividual differences may not be replicable and represent false positive findings.

## Methods

### Participants

This study used participants from UK Biobank, a large-scale biomedical database containing in-depth genetic and health information from about half-a-million individuals from the United Kingdom<sup>55</sup>. We included data from participants that had both structural and functional neuroimaging data (released early 2020,  $N = 37,447$ , 19,981 females and 17,466 males, age attending brain scanning ranges: 44–82 years old). Aside from brain measures, we used data on the following phenotypes: age at MRI scanning time ( $N = 37,447$ ), body mass index ( $N = 36,114$ ), fluid intelligence ( $N = 34,534$ ), numeric memory ( $N = 25,231$ ), neuroticism ( $N = 30,511$ ), alcohol consumption ( $N = 34,345$ ) and birth month ( $N = 37,447$ ) (Supplementary Table 1). All participants gave written informed consent and the UK Biobank project received ethical approval from the National Health Service North West Centre for Research Ethics Committee (reference 11/NW/0382).

### Imaging processing

We used the structural and functional brain measures derived from T1 and resting-state fMRI, respectively, as processed by WIN FMRIB on behalf of UK Biobank<sup>56</sup>. The detailed preprocessing information can be found online ([https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain\\_mri.pdf](https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf)). In short, a standard Skyra 3 T scanner was used for image acquisition. For raw T1 imaging data, the field-of-view (FoV) was reduced to minimize the amount of non-brain tissue and gradient distortion correction was applied. The reduced-FoV T1 imaging data were nonlinearly warped to MNI152 space using FNIRT<sup>57</sup>. Next, the T1 images were processed with FreeSurfer to achieve accurate cortical modelling. A total of 132 surface structural measures including CSA and CT were estimated on the basis of the 66 regions of the DK atlas. For resting-state fMRI data, the following preprocessing was applied: motion correction, intensity normalization, highpass temporal filtering, EPI unwarping, GDC unwarping and structured artefacts removal. In addition, group independent component analysis (group-ICA) was applied to decompose the resting-state fMRI data into a specified number of networks based on >4,000 participants and the temporal correlations of averaged time series between different networks were estimated as the FC. We used the partial correlation matrix on 25 dimensions which produced 210 FCs.

### Statistical analyses

The SciPy Stats package<sup>58</sup> in Python was used to calculate the Spearman's rank correlation coefficients between brain features and seven variables. We controlled for covariates such as sex and scanning sites to investigate the brain associations with age. For the other variables, we included age as an additional covariate in the analysis. In addition, for CT and CSA analysis, we also adjusted for average CT and total CSA, respectively. Notably, head motion can significantly impact brain functional measures. Therefore, we regressed out head motion to effectively remove its effects on our FC measures (Supplementary Fig. 26). According to the distribution of the variables (Supplementary Fig. 1), we also split the sample into four quarters: Q1, Q2, Q3 and Q4. We first used the median values to split the full sample in half and applied the SciPy Stats package to conduct two-sample *t*-test to compare the neuroimaging differences between the two groups. Second, we removed the individuals in the middle part and only included Q1 and Q4 as the two groups. The same covariates as the main correlation analysis were included. Then, we compared the replicability of correlation analysis and two-sample *t*-test. Moreover, we applied bootstrap-based random forest regression in the sklearn.ensemble module<sup>59</sup> of Python to compute impurity-based feature importance, which in turn were used to select features with highest importance. We also compared the replicability of feature selection with correlation analysis.

### Intraclass correlation coefficient

An ICC<sup>60</sup> is generally used to measure the accordance of statistic values from two independent samples. In this study, ICC has been used to measure the replicability of Spearman's rank correlation coefficients. The ICC value ranges from 0 to 1, with large value indicating correlations are more replicable between two samples. The pingouin statistical package<sup>61</sup> in Python was used to calculate ICC. We selected two-way random effects model, in which both the sampling process and subjects are considered sources of random effects. The absolute differences was used to measure the agreement between the correlations of two samples and the correlations of a single sample was considered as the basis for measurement. Detailed description of ICC can be found in a previous paper<sup>24</sup>.

### Jaccard index

We used the Jaccard index to quantify the level of overlap of significantly identified brain measures between two samples. The index

ranged from 0 to 1, with a higher index indicating more overlap between two subsamples. We can calculate the Jaccard index by a  $2 \times 2$  table (Fig. 1d):

$$\text{Jaccard} = \frac{a}{a + b + c}.$$

Moreover, we have corrected for the agreement chance<sup>62</sup> using a formula

$$\text{Corrected Jaccard} = \frac{\text{Jaccard} - E(\text{Jaccard})}{1 - E(\text{Jaccard})},$$

where the value 1 is the maximum Jaccard index and  $E(\text{Jaccard})$  is the expectation values of the Jaccard index calculated by conditional upon fixed marginal totals of the  $2 \times 2$  table. As ref. 62 showed, different forms of expectation are applied in univariate correlation analysis and multivariate feature selection depending on whether the frequency distributions are different between two samples in the  $2 \times 2$  table (Fig. 1d and Supplementary Fig. 23d).

### Sampling and estimating replicability

As shown in Fig. 1, we estimated the global and regional replicability of brain–phenotype associations and examined the improvement of replicability with increasing sampling size. We randomly selected two non-overlapping subsamples ( $n$ ) from the total sample ( $M$ ) and then put the two subsamples back (Fig. 1b). The sampling procedure will be repeated 100 times. The discovery sample ( $D_i$ ) and replication sample ( $R_i$ ) were produced by the  $i$ th selection. The size of  $n$  ranged from 100 individuals to half the full sample size. Univariate Spearman's correlation analysis was then conducted to examine the associations between brain measures and each variable in two independent subsamples  $D_i$  and  $R_i$  (Fig. 1c). The correlations across brain measures were then transferred to binary vectors by the significance threshold of  $P < 0.05$ ,  $P < 0.01$ ,  $P_{\text{FDR}} < 0.05$  or  $P_{\text{Bonferroni}} < 0.05$ . Next, we obtained the overlapping vector ( $V_i$ ) between binary vectors of  $D_i$  and  $R_i$ . The global replicability was estimated on the basis of brain-wide correlations (Fig. 1d). We selected brain measures with the top 10%, 25%, 50% or 100% largest correlations ( $|r_{i,j}|$ ) in  $D_i$  and extracted the corresponded correlations of these brain measures in  $R_i$ . For selected brain measures, the similarity of the correlation coefficients between  $D_i$  and  $R_i$  was then calculated by ICC. We also calculated the Jaccard index according to binary vectors of  $D_i$  and  $R_i$ . Notably, the ICC model we used measures the degree of replicability of true correlation coefficients, rather than absolute ones. In the estimation of Jaccard index, correlations that reach significance levels in both the discovery and replication samples and have the same signs are transferred to a value of one. Finally, the average ICC and Jaccard index of 100 random selections were used to measure the global replicability at the subsample size  $n$ .

To measure the replication probability for each specific brain measure across 100 random selections, we proposed a new regional replicability index (Fig. 1e). We averaged the overlapping vectors ( $V_1, V_2, \dots, V_{100}$ ), which indicated the probabilities that significant associations identified in one subsample can be replicated in another independent subsample.

Similar to what we did above for brain–phenotype associations, here we also estimated replicability by sampling using two-sample  $t$ -test (Supplementary Figs. 14 and 18). In addition, we conducted multivariate statistical methods including random forest and partial least square regression in two independent subsamples (Supplementary Figs. 23 and 24).

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

This research used data from the UK Biobank resource (<https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=100>). Access to UK Biobank data requires the submission and approval of a research project by the UK Biobank committee. The DK atlas was used to parcellate the human cortex into 66 regions for structural brain measures (<https://surfer.nmr.mgh.harvard.edu/fswiki/CorticalParcellation>). Twenty-one functional networks were used to estimate functional brain measures ([https://www.fmrib.ox.ac.uk/ukbiobank/group\\_means/rfMRI\\_ICA\\_d25\\_good\\_nodes.html](https://www.fmrib.ox.ac.uk/ukbiobank/group_means/rfMRI_ICA_d25_good_nodes.html)).

### Code availability

Python code on Jupyter notebook used for statistical analyses is available in GitHub: <https://github.com/deeppsyche/Replicability-ukbb>.

### References

- Niu, X., Zhang, F., Kounios, J. & Liang, H. Improved prediction of brain age using multimodal neuroimaging data. *Hum. Brain Mapp.* **41**, 1626–1643 (2020).
- Kaczurkin, A. N., Raznahan, A. & Satterthwaite, T. D. Sex differences in the developing brain: insights from multimodal neuroimaging. *Neuropsychopharmacology* **44**, 71–85 (2019).
- Steegers, C. et al. The association between body mass index and brain morphology in children: a population-based study. *Brain Struct. Funct.* **226**, 787–800 (2021).
- Radonjić, N. V. et al. Structural brain imaging studies offer clues about the effects of the shared genetic etiology among neuropsychiatric disorders. *Mol. Psychiatry* **26**, 2101–2110 (2021).
- Spear, L. P. Effects of adolescent alcohol consumption on the brain and behaviour. *Nat. Rev. Neurosci.* **19**, 197–214 (2018).
- Hilger, K. et al. Predicting intelligence from brain gray matter volume. *Brain Struct. Funct.* **225**, 2111–2129 (2020).
- Aarts, A. A. et al. Estimating the reproducibility of psychological science. *Science* <https://doi.org/10.1126/science.aac4716> (2015).
- Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016).
- Replication studies offer much more than technical details. *Nature* **541**, 259–260 (2017).
- Poldrack, R. A. et al. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.* **18**, 115–126 (2017).
- Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
- Satzubal, C. L. et al. Genetic architecture of subcortical brain structures in 38,851 individuals. *Nat. Genet.* **51**, 1624–1636 (2019).
- Boekel, W. et al. A purely confirmatory replication study of structural brain–behavior correlations. *Cortex* **66**, 115–133 (2015).
- Munson, B. A. & Hernandez, A. E. Inconsistency of findings due to low power: a structural MRI study of bilingualism. *Brain Lang.* **195**, 104642 (2019).
- Zhou, Z. W. et al. Inconsistency in abnormal functional connectivity across datasets of ADHD-200 in children with attention deficit hyperactivity disorder. *Front. Psychiatry* **10**, 692 (2019).
- Schmaal, L. et al. ENIGMA MDD: seven years of global neuroimaging studies of major depression through worldwide data sharing. *Transl. Psychiatry* **10**, 172 (2020).
- Müller, V. I. et al. Altered brain activity in unipolar depression revisited: meta-analyses of neuroimaging studies. *JAMA Psychiatry* **74**, 47–55 (2017).
- Littlejohns, T. J. et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nat. Commun.* **11**, 2624 (2020).
- Smith, S. M. et al. An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nat. Neurosci.* **24**, 737–745 (2021).

20. Marek, S. et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
21. Miller, K. L. et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* **19**, 1523–1536 (2016).
22. Meyer, C. et al. Seasonality in human cognitive brain responses. *Proc. Natl Acad. Sci. USA* **113**, 3066–3071 (2016).
23. Kampa, M. et al. Replication of fMRI group activations in the neuroimaging battery for the Mainz Resilience Project (MARP). *Neuroimage* **204**, 116223 (2020).
24. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **15**, 155–163 (2016).
25. Ingre, M. Why small low-powered studies are worse than large high-powered studies and how to protect against ‘trivial’ findings in research: comment on Friston (2012). *Neuroimage* **81**, 496–498 (2013).
26. Button, K. S. et al. Confidence and precision increase with high statistical power. *Nat. Rev. Neurosci.* **14**, 585–586 (2013).
27. Schönbrodt, F. D. & Perugini, M. At what sample size do correlations stabilize? *J. Res. Pers.* **47**, 609–612 (2013).
28. Grady, C. L., Rieck, J. R., Nichol, D., Rodrigue, K. M. & Kennedy, K. M. Influence of sample size and analytic approach on stability and interpretation of brain–behavior correlations in task-related fMRI data. *Hum. Brain Mapp.* **42**, 204–219 (2021).
29. Varoquaux, G. Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage* **180**, 68–77 (2018).
30. Genon, S. et al. Searching for behavior relating to grey matter volume in a-priori defined right dorsal premotor regions: lessons learned. *Neuroimage* **157**, 144–156 (2017).
31. Schulz, M.-A., Bzdok, D., Haufe, S., Haynes, J.-D. & Ritter, K. Performance reserves in brain-imaging-based phenotype prediction. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.02.23.481601> (2022).
32. Masouleh, S. K., Eickhoff, S. B., Hoffstaedter, F. & Genon, S. Empirical examination of the replicability of associations between brain structure and psychological variables. *eLife* <https://doi.org/10.7554/eLife.43464> (2019).
33. Loken, E. & Gelman, A. Measurement error and the replication crisis. *Science* **355**, 584–585 (2017).
34. Border, R. et al. No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples. *Am. J. Psychiatry* **176**, 376–387 (2019).
35. Marigorta, U. M., Rodríguez, J. A., Gibson, G. & Navarro, A. Replicability and prediction: lessons and challenges from GWAS. *Trends Genet.* **34**, 504–517 (2018).
36. Tam, V. et al. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 67–484 (2019).
37. Cole, J. H. Multimodality neuroimaging brain-age in UK biobank: relationship to biomedical, lifestyle, and cognitive factors. *Neurobiol. Aging* **92**, 34–42 (2020).
38. Woo, C. W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365–377 (2017).
39. Abrol, A. et al. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nat. Commun.* **12**, 353 (2021).
40. Sullivan, G. M. & Feinn, R. Using effect size—or why the *P* value is not enough. *J. Grad. Med. Educ.* **4**, 279–282 (2012).
41. Albers, C. The problem with unadjusted multiple and sequential statistical testing. *Nat. Commun.* **10**, 1921 (2019).
42. Xia, M. et al. Reproducibility of functional brain alterations in major depressive disorder: evidence from a multisite resting-state functional MRI study with 1,434 individuals. *Neuroimage* **189**, 700–714 (2019).
43. Wang, M. et al. Reproducible abnormalities of functional gradient reliably predict clinical and cognitive symptoms in schizophrenia. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.11.24.395251> (2020).
44. Rosenberg, M. D. & Finn, E. S. How to establish robust brain–behavior relationships without thousands of individuals. *Nat. Neurosci.* **25**, 835–837 (2022).
45. Gratton, C., Nelson, S. M. & Gordon, E. M. Brain–behavior correlations: two paths toward reliability. *Neuron* **110**, 1446–1449 (2022).
46. Melzer, T. R. et al. Test–retest reliability and sample size estimates after MRI scanner relocation. *Neuroimage* **211**, 116608 (2020).
47. Noble, S., Scheinost, D. & Constable, R. T. A decade of test–retest reliability of functional connectivity: a systematic review and meta-analysis. *Neuroimage* **203**, 116157 (2019).
48. Tozzi, L., Fleming, S. L., Taylor, Z. D., Raterink, C. D. & Williams, L. M. Test–retest reliability of the human functional connectome over consecutive days: identifying highly reliable portions and assessing the impact of methodological choices. *Netw. Neurosci.* **4**, 925–945 (2020).
49. Noble, S. et al. Influences on the test–retest reliability of functional connectivity MRI and its relationship with behavioral utility. *Cereb. Cortex* **27**, 5415–5429 (2017).
50. Fan, L. et al. The human brainnetome atlas: a new brain atlas based on connectonal architecture. *Cereb. Cortex* **26**, 3508–3526 (2016).
51. Kardan, O. et al. Differences in the functional brain architecture of sustained attention and working memory in youth and adults. *PLoS Biol.* **20**, e3001938 (2022).
52. Harvey, J. L., Demetriou, L., McGonigle, J. & Wall, M. B. A short, robust brain activation control task optimised for pharmacological fMRI studies. *PeerJ* **6**, e5540 (2018).
53. Suda, A. et al. Functional organization for response inhibition in the right inferior frontal cortex of individual human brains. *Cereb. Cortex* **30**, 6325–6335 (2020).
54. Fry, A., Littlejohns, T., Sudlow, C., Doherty, N. & Allen, N. OP41 The representativeness of the UK Biobank cohort on a range of sociodemographic, physical, lifestyle and health-related characteristics. *J. Epidemiol. Community Health* **70**, A26 (2016).
55. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
56. Alfaro-Almagro, F. et al. Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage* **166**, 400 (2018).
57. Andersson, J. L. R., Jenkinson, M. & Smith, S. *Non-linear Registration aka Spatial Normalisation* FMRIB Technical Report TRO7JA2 (FMRIB Centre, 2007).
58. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
59. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
60. Shrout, P. E. & Fleiss, J. L. Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* **86**, 420–428 (1979).
61. Vallat, R. Pingouin: statistics in Python. *J. Open Source Softw.* <https://doi.org/10.21105/joss.01026> (2018).
62. Warrens, M. J. Similarity measures for 2×2 tables. *J. Intell. Fuzzy Syst.* **36**, 3005–3018 (2019).

## Acknowledgements

This research has been conducted using the UK Biobank Resource under application no. 30091. A.A. and K.J.H.V. are supported by the Foundation Volksbond Rotterdam. This work was supported by a China Scholarship Council (CSC) grant to S.L. G.v.W. has received research funding by Philips for an unrelated project. The funders have no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

### Author contributions

S.L., A.A., K.J.H.V. and G.v.W conceived and designed the study. S.L. analysed the data and wrote the manuscript. A.A., K.J.H.V. and G.v.W provided significant feedback on the data analysis and the revision of the manuscript. A.A., K.J.H.V. and G.v.W jointly supervised the work.

### Competing interests

The authors declare no competing interests.

### Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41562-023-01642-5>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41562-023-01642-5>.

**Correspondence and requests for materials** should be addressed to Shu Liu or Guido A. van Wingen.

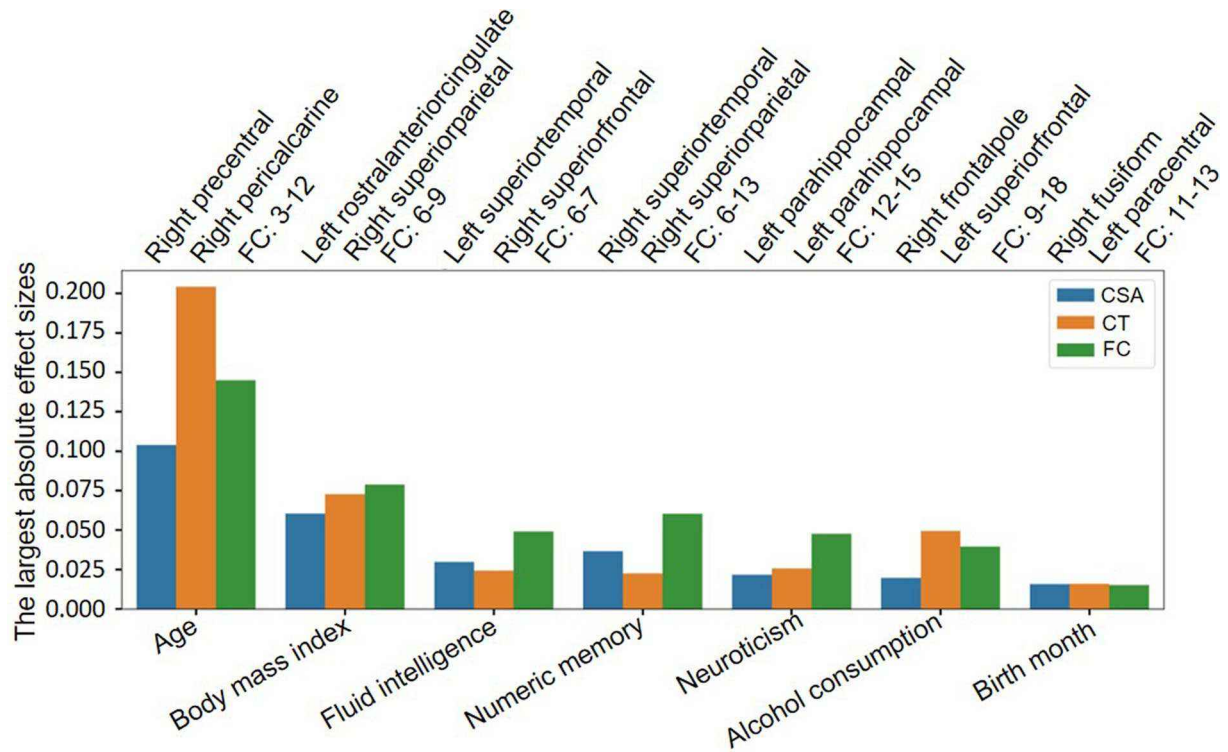
**Peer review information** *Nature Human Behaviour* thanks Deanna Barch, Omid Kardan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

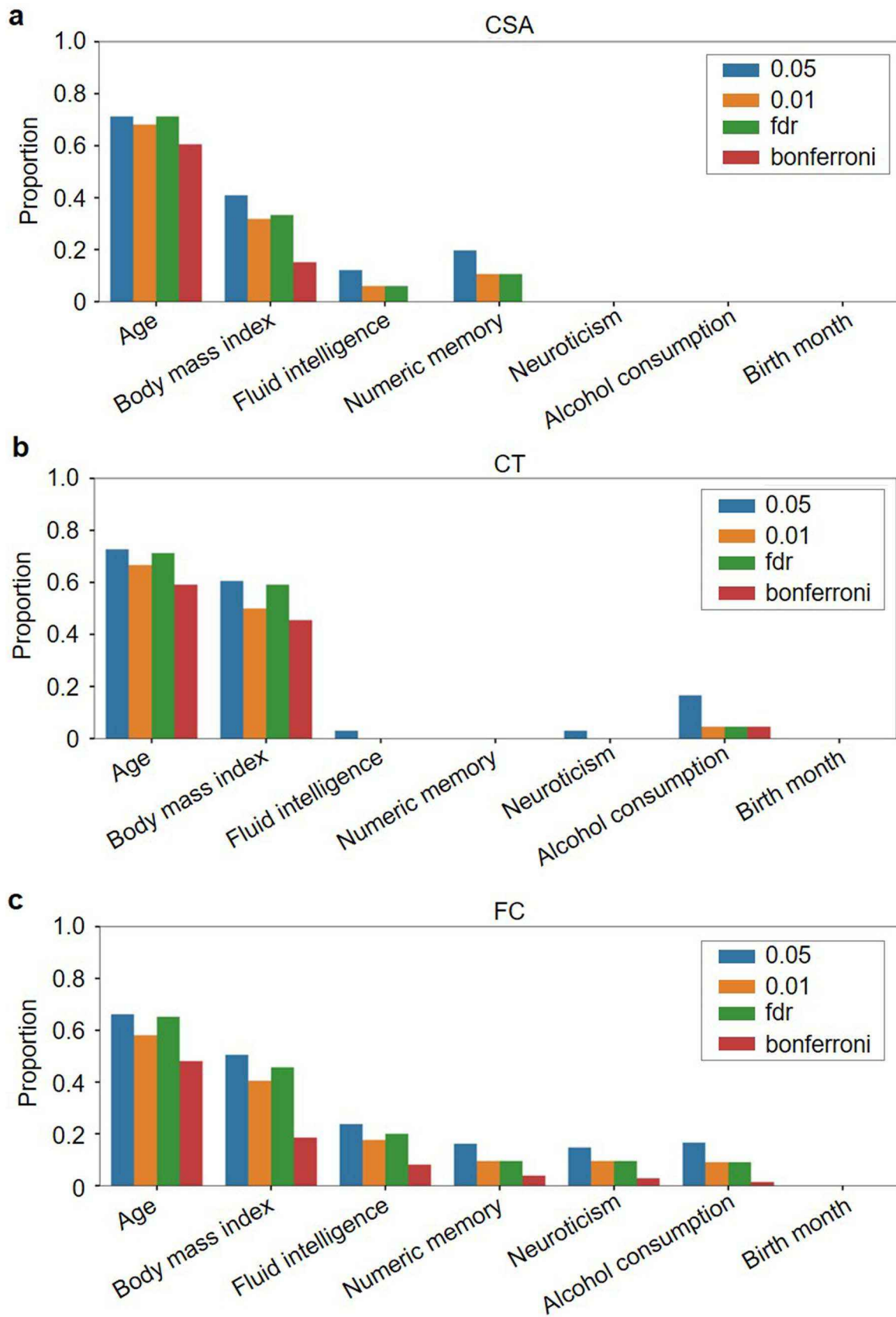
**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

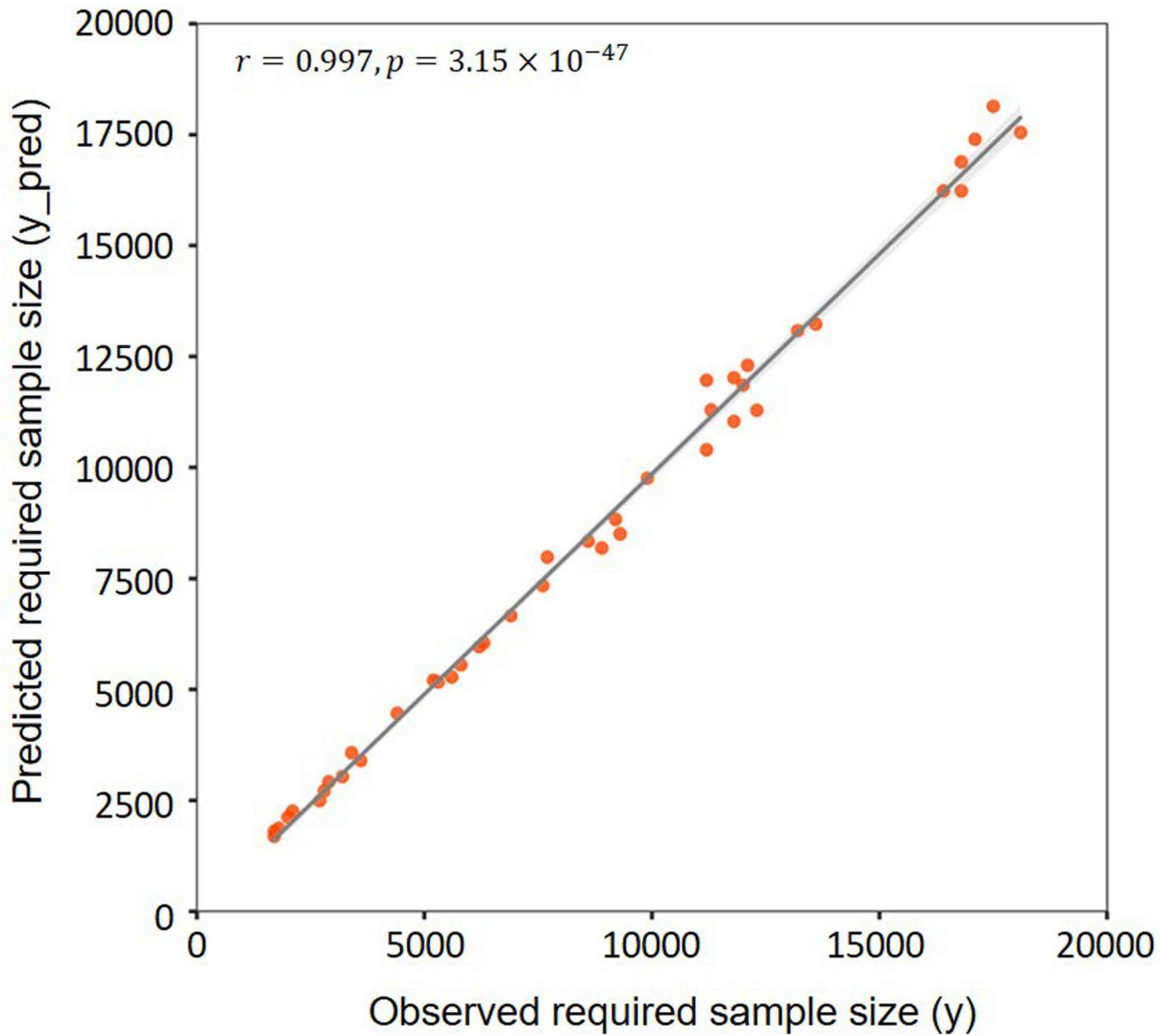
© The Author(s), under exclusive licence to Springer Nature Limited 2023



**Extended Data Fig. 1 | The largest absolute effect sizes of brain–phenotype associations for each imaging modality and phenotype.** The upper text shows the specific brain measures which have the largest absolutes of effect sizes across brain–phenotype associations. Cortical surface area, CSA; Cortical thickness, CT; Functional connectivity, FC.

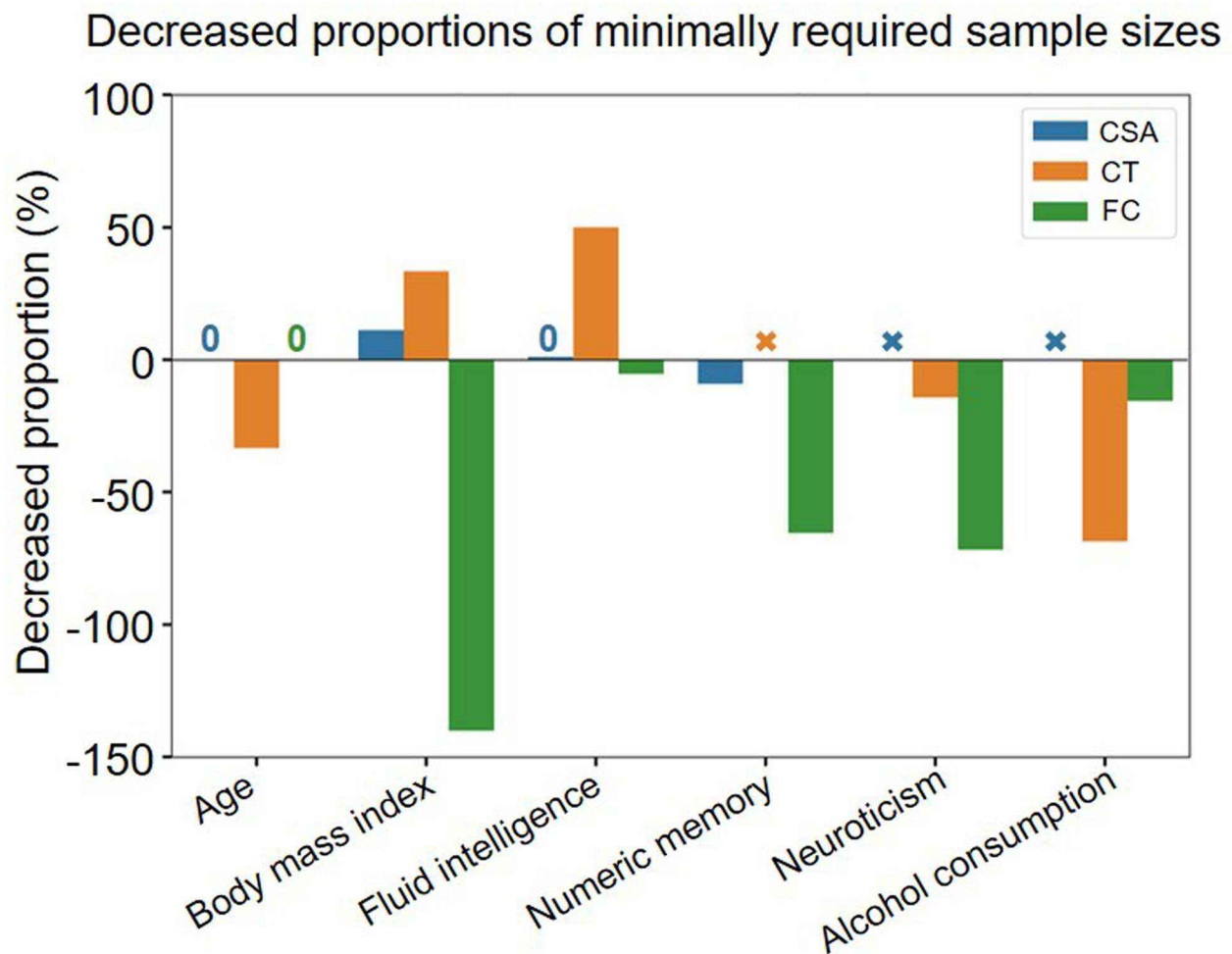


**Extended Data Fig. 2 | The proportion of brain features with regional replicability of larger than 0.75. (a) Cortical surface area (CSA); (b) Cortical thickness (CT); (c) Functional connectivity (FC).**



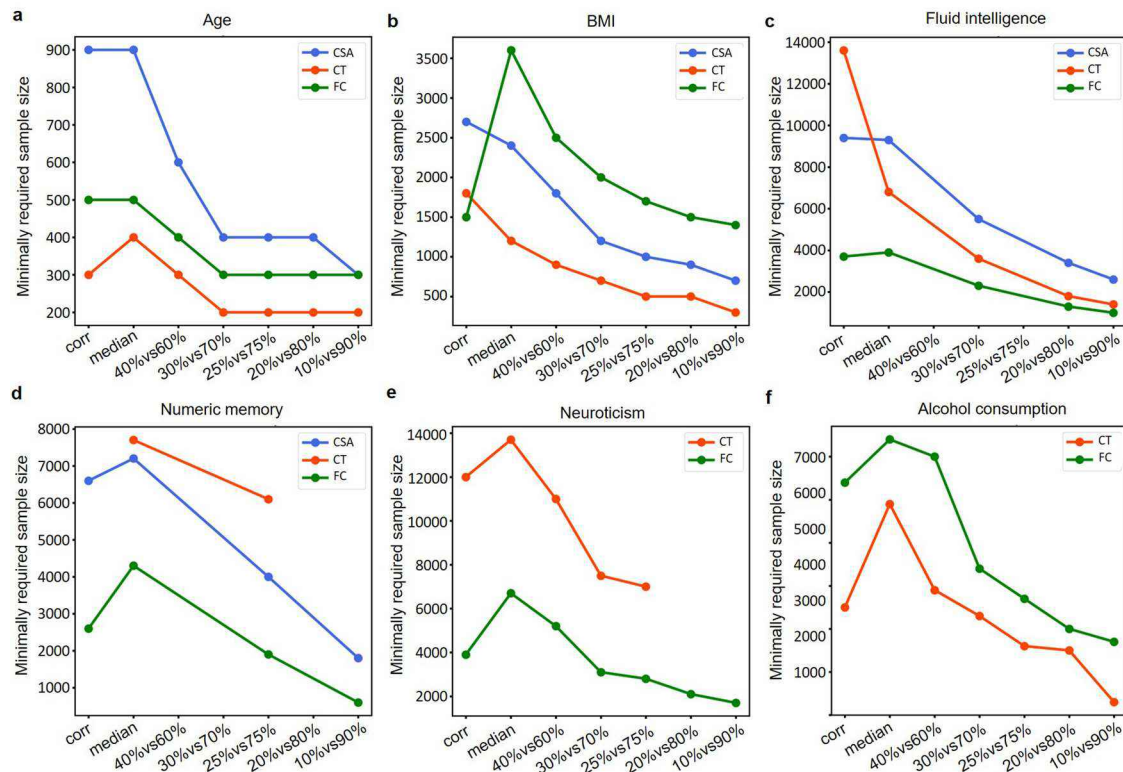
**Extended Data Fig. 3 | The relationship between observed and predicted sample sizes for added phenotypes.** x axis represents true minimally required sample sizes to obtain the replicable brain associations; y axis represents the predicted required sample sizes according to the effect sizes of brain associations with added phenotypes.





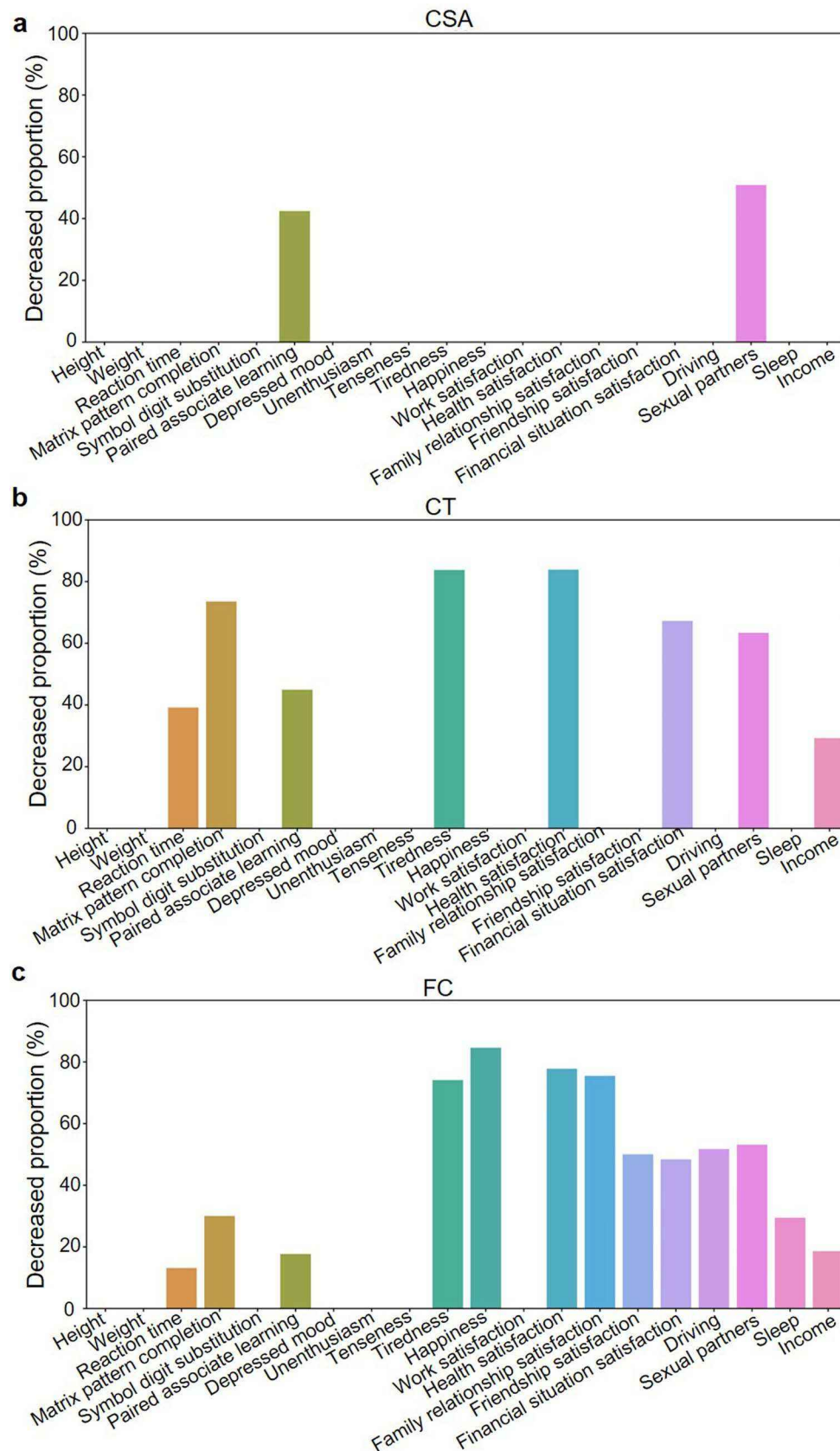
**Extended Data Fig. 4 | Differences of minimally required sample size to reach 75% replication probability between two-sample *t*-test using median splits and correlation analysis.** Positive values indicate lower minimally required sample sizes for the two-sample *t*-test, whereas negative values indicate higher minimally required sample sizes. Cross symbol 'x' indicates that the minimally

required sample sizes are missing for the two-sample *t*-test or Spearman's correlation analysis, because the regional replicability of all brain measures did not reach 0.75 at any sample size. Zero '0' indicates that the minimally required sample sizes are the same between two-sample *t*-test and Spearman's correlation analysis.



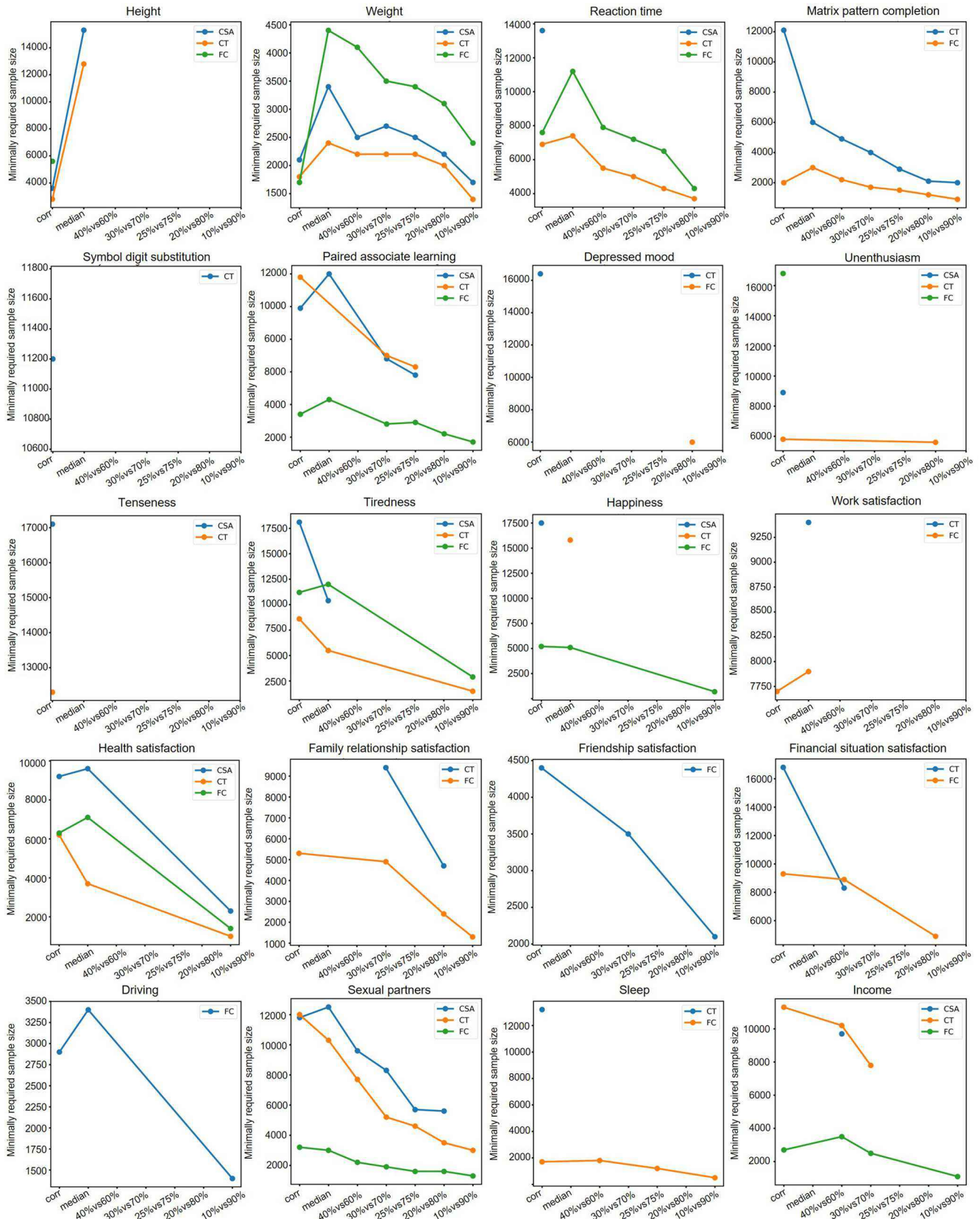
**Extended Data Fig. 5 | The minimally required sample sizes with preselection procedures at 10%, 20%, 25%, 30%, 40%, and 50% (median) for six representative phenotypes, when using a significance threshold of  $p < 0.05$  uncorrected. (a) Age; (b) Body mass index (BMI); (c) Fluid intelligence; (d) Numeric memory; (e) Neuroticism; (f) Alcohol consumption. Missing**

estimates (dots) could be related to two situations: first, good replicability is not obtained even at the largest sampling size; second, the specific preselection is not conducted because of the distribution of the variable. Cortical surface area (CSA), cortical thickness (CT), and functional connectivity (FC).



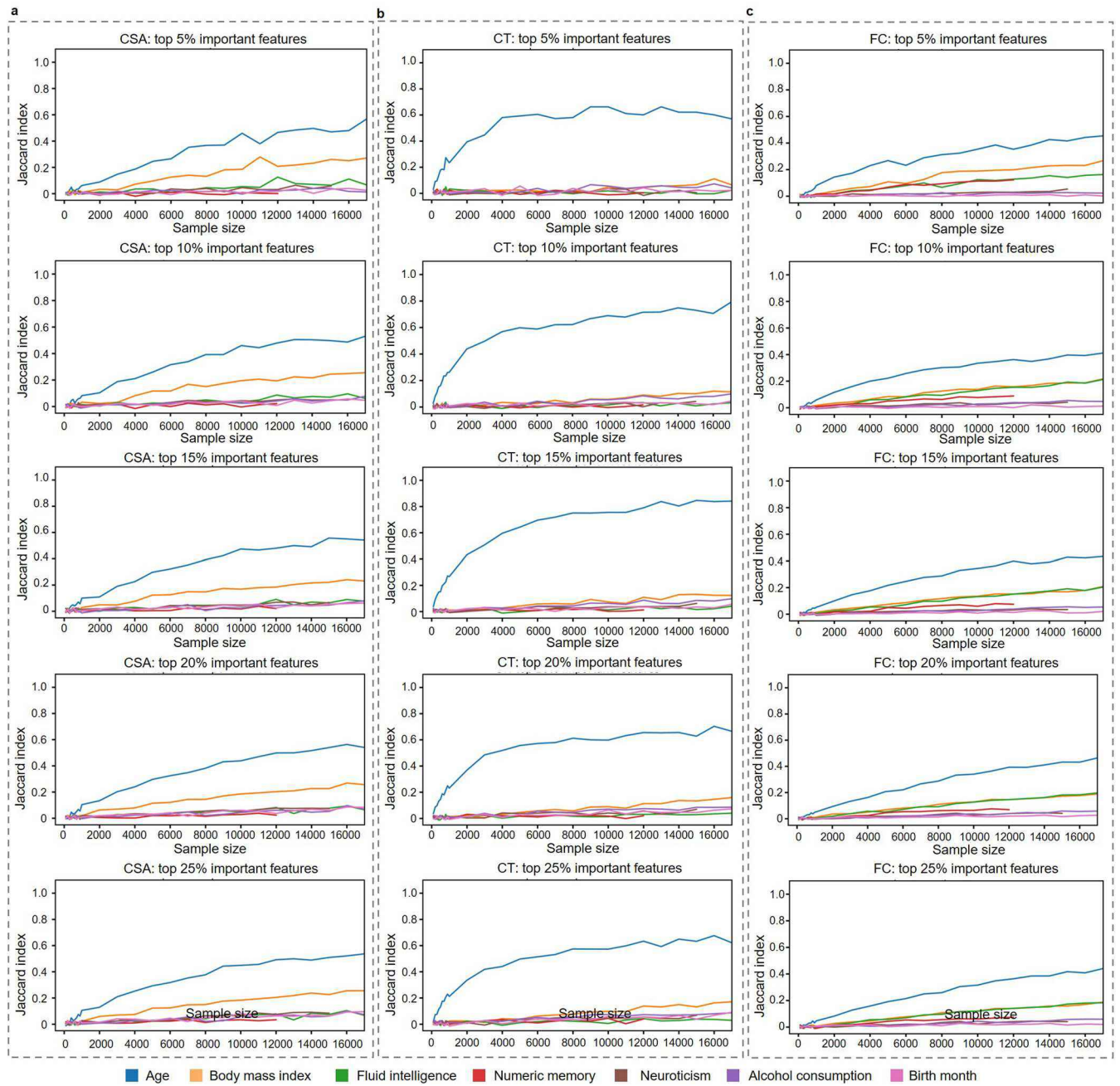
**Extended Data Fig. 6 | Decrease proportion of additional phenotypes in minimally required sample size to reach good replicability between a two-sample t-test (by sample selection) and correlation analysis for added phenotypes. (a)** Cortical surface area (CSA); **(b)** Cortical thickness (CT); **(c)** Functional connectivity (FC). Empty bars indicate that the minimal required

sample sizes are missing for the two-sample t-test or Spearman's correlation analysis, because the regional replicability of all brain measures did not reach 0.75 at any sample size. Coloured bars indicate that minimally required sample size decreased in a two-sample t-test (by sample selection) compared to correlation analysis.

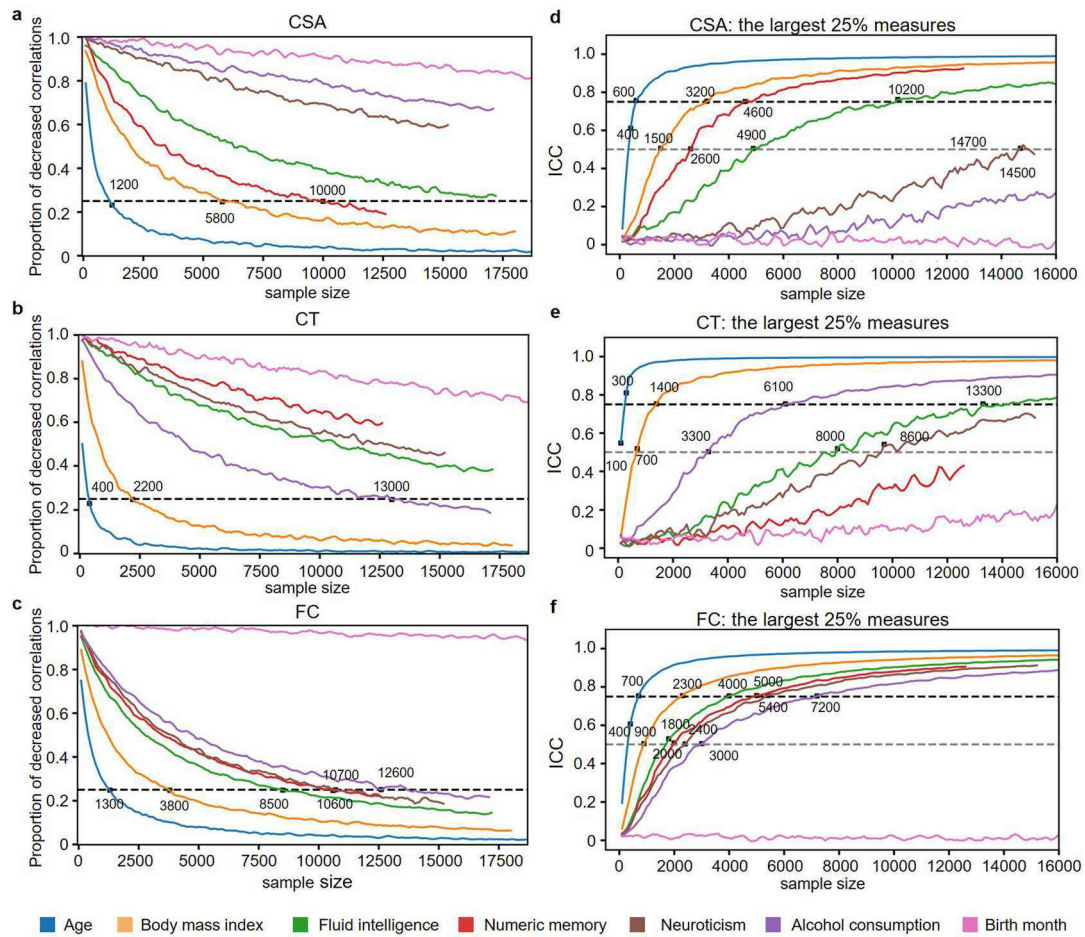


**Extended Data Fig. 7 | The minimally required sample sizes with preselection procedures at 10%, 20%, 25%, 30%, 40%, and 50% for additional phenotypes.** Missing estimates (dots) could be related to two factors: first, good replicability

is not obtained even at the largest sampling size; second, the specific preselection is not conducted because of the distribution of the variable. Cortical surface area (CSA), cortical thickness (CT), and functional connectivity (FC).



**Extended Data Fig. 8 | Improvement of the replicability of feature selection with increasing sample size at the thresholds ranging from 5% to 25%.** (a) shows the Jaccard index for cortical surface area (CSA) at different thresholds. (b) shows the Jaccard index for cortical thickness (CT) at different thresholds. (c) shows the Jaccard index for functional connectivity (FC) at different thresholds.



**Extended Data Fig. 9 | Improvement of replicability for partial least square (PLS) regression analysis with increasing sample size.** (a), (b), and (c) show the proportion of decreased correlations of PLS1 with the phenotypes for cortical

surface area (CSA), cortical thickness (CT), and functional connectivity (FC); (d), (e), and (f) show the intraclass correlation coefficient (ICCs) of PLS weights. The dotted lines indicate good and moderate replicability levels (0.75 and 0.5).

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | n/a                                 | Confirmed  |
|-------------------------------------|--|
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of all covariates tested   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated   |

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

- |                 |   |
|-----------------|---|
| Data collection | No software was employed for data collection.   |
| Data analysis   | Python 3.7 was utilized for conducting the data analyses. The study employed various packages, including pandas (1.1.3), numpy (1.21.6), Scipy (1.7.3), pingouin (0.3.2), sklearn (0.22.1), matplotlib (3.0.3), seaborn (0.11.0), Nilearn (0.8.1), and mne (0.19.2). The code for this study can be found on GitHub at: <a href="https://github.com/deeppsyche/Replicability-ukbb">https://github.com/deeppsyche/Replicability-ukbb</a> |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

This research utilized data from the UK Biobank resource (<https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=100>). Access to UK Biobank data requires the submission and approval of a research project by the UK Biobank committee. The Desikan-Killiany (DK) atlas was used to parcellate the human cortex into 66

regions for structural brain measures (<https://surfer.nmr.mgh.harvard.edu/fswiki/CorticalParcellation>). 21 functional networks was used to estimate functional brain measures ([https://www.fmrib.ox.ac.uk/ukbiobank/group\\_means/fMRI\\_ICA\\_d25\\_good\\_nodes.html](https://www.fmrib.ox.ac.uk/ukbiobank/group_means/fMRI_ICA_d25_good_nodes.html)).

## Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

### Reporting on sex and gender

Our findings don't apply to only one sex or gender. Our study includes 19,989 females and 17,473 males without bias. The sex is just used as a covariate. As described by UKBB, it is a mixture of the sex the United Kingdom National Health Service had recorded for the participant and self-reported sex.

### Reporting on race, ethnicity, or other socially relevant groupings

Our study do not use socially constructed or socially relevant categorization variable. Because the participants are recruited in UK, they are predominantly of European ancestry. For details, please see: Bycroft, C. et al (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203.

### Population characteristics

The participants range in age from 44 to 82 years old, with an approximately equal number of males and females.

### Recruitment

Half a million volunteer participants were recruited in Great Britain. The participation rate however was 5.45% and was biased towards older, more healthy, and female residents. The UK Biobank sample does reflect nationally representative data sources to a significant degree.

### Ethics oversight

The UK Biobank project has received ethical approval from the National Health Service North West Centre for Research Ethics Committee (reference: 11/NW/0382).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences  Behavioural & social sciences  Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

### Sample size

Based on the full UK Biobank dataset, participants who had missing important data for analysis were excluded from the study. The details of the data exclusions are provided below. Ultimately, we included data from participants that had both structural and functional neuroimaging data (released early 2020, N = 37447, 19989 females and 17473 males, age attending brain scanning ranges: 44-82 years old). Aside from brain measures, we incorporated data on several phenotypes in our analysis. They included age at MRI scanning time (N = 37,447), body mass index (N = 36,114), fluid intelligence (N = 34,534), numeric memory (N = 25,231), neuroticism (N = 30,511), alcohol consumption (N = 34,345), and birth month (N = 37,447). Except data missing, no other methods were used to predetermine sample size.

### Data exclusions

Participants who did not provide brain scans or only provided T1 or fMRI scans were excluded from the analysis. Furthermore, participants without corresponding phenotypes were also excluded. The exclusion of these participants was necessary due to the absence of crucial data required for the analysis.

### Replication

No other cohort with large scale brain imaging data is available for replication at this stage.

### Randomization

We controlled for covariates such as age, sex, scanning sites, average cortical thickness, total cortical surface area, and head motion.

### Blinding

Our study used all available subjects that have completed data for analysis from UK Biobank. There was no equivalent process of randomization that comes into this analysis, and there were no experimental groups. Therefore, there is no step equivalent to blinding involved.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.



## Materials &amp; experimental systems

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

## Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input type="checkbox"/>	<input checked="" type="checkbox"/> MRI-based neuroimaging

## Magnetic resonance imaging

## Experimental design

Design type	resting-state fMRI; structural MRI (cortical thickness and cortical surface area)
Design specifications	The experimental design has been done previously and is fully described online: <a href="https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf">https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf</a>
Behavioral performance measures	For our main analysis, we selected six representative variables including age, body mass index, fluid intelligence, numeric memory, neuroticism and alcohol consumption. In addition, birth month and 23 additional variables were also included in our analysis. Please see Supplementary Table 11 and Supplementary Figure 6.

## Acquisition

Imaging type(s)	T1-weighted MRI, resting-state fMRI
Field strength	3T
Sequence & imaging parameters	Open-source MRI data was used in this study. The details of MRI data acquisition for these 2 modalities can be seen in Methods or Elliott, L. T. et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank. Nature 562, 210–216 (2018).
Area of acquisition	Whole brain scans
Diffusion MRI	<input type="checkbox"/> Used <input checked="" type="checkbox"/> Not used

## Preprocessing

Preprocessing software	The full set of image analysis pipeline scripts are available from <a href="https://www.fmrib.ox.ac.uk/ukbiobank/">https://www.fmrib.ox.ac.uk/ukbiobank/</a> - at present the scripts primarily call tools from FSL and FreeSurfer.
Normalization	1. The external surface of the skull is estimated from the T1, and used to normalise brain tissue volumes for head size; 2. for resting-state processing, grand-mean intensity normalisation of the entire 4D dataset by a single multiplicative factor.
Normalization template	MNI152 template
Noise and artifact removal	1. Several Quality control (QC) measures quantitate the discrepancy between the T1 structural image and the standard (population average) template image as well as each of the other modalities (for that same subject). All of these “discrepancy” QC measures are the unitless “correlation ratio” cost function, that is used by FLIRT to optimise alignments, and which is used here to quantify the discrepancy between any two images. 2. Several other QC measures quantitate signal to noise ratio (SNR) in some of the modalities. For the T1, the tissue-type segmentation is used to estimate within-tissue-type noise level (standard deviation), as well as mean intensities for grey and white matter. These quantities are used to estimate overall image SNR and also CNR (contrast to noise - white-grey mean intensity difference normalised by noise level). From the preprocessed rfMRI (both before and after artefact removal) timeseries data, similar measures are calculated, but in this case the “noise” level is the temporal standard deviation. For detailed information, please refer to the online document: <a href="https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf">https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf</a>
Volume censoring	Please refer to Methods or Elliott, L. T. et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank. Nature 562, 210–216 (2018).

## Statistical modeling &amp; inference

Model type and settings	Univariate analysis: Spearman's rank correlation analysis, two-sample t-test Multivariate random forest feature selection Partial least regression analysis
-------------------------	---

Effect(s) tested

Specify type of analysis:  Whole brain  ROI-based  Both

Anatomical location(s)

Statistic type for inference

(See [Eklund et al. 2016](#))

Correction

## Models & analysis

n/a | Involved in the study

Functional and/or effective connectivity

Graph analysis

Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Multivariate modeling and predictive analysis