nature human behaviour

Article

https://doi.org/10.1038/s41562-024-01931-7

Power and reproducibility in the external validation of brain-phenotype predictions

Received: 30 October 2023

Accepted: 18 June 2024

Published online: 31 July 2024

Check for updates

Matthew Rosenblatt[1] ✉, Link Tejavibulya[2], Huili Sun[1], Chris C. Camp[2], Milana Khaitova[3], Brendan D. Adkinson[2], Rongtao Jiang[3], Margaret L. Westwater[3], Stephanie Noble[3,4,5] & Dustin Scheinost[1,2,3,6,7]

Brain-phenotype predictive models seek to identify reproducible and generalizable brain-phenotype associations. External validation, or the evaluation of a model in external datasets, is the gold standard in evaluating the generalizability of models in neuroimaging. Unlike typical studies, external validation involves two sample sizes: the training and the external sample sizes. Thus, traditional power calculations may not be appropriate. Here we ran over 900 million resampling-based simulations in functional and structural connectivity data to investigate the relationship between training sample size, external sample size, phenotype effect size, theoretical power and simulated power. Our analysis included a wide range of datasets: the Healthy Brain Network, the Adolescent Brain Cognitive Development Study, the Human Connectome Project (Development and Young Adult), the Philadelphia Neurodevelopmental Cohort, the Queensland Twin Adolescent Brain Project, and the Chinese Human Connectome Project; and phenotypes: age, body mass index, matrix reasoning, working memory, attention problems, anxiety/depression symptoms and relational processing. High effect size predictions achieved adequate power with training and external sample sizes of a few hundred individuals, whereas low and medium effect size predictions required hundreds to thousands of training and external samples. In addition, most previous external validation studies used sample sizes prone to low power, and theoretical power curves should be adjusted for the training sample size. Furthermore, model performance in internal validation often informed subsequent external validation performance (Pearson's *r* difference <0.2), particularly for well-harmonized datasets. These results could help decide how to power future external validation studies.

Neuroimaging studies increasingly leverage large datasets to understand brain-phenotype associations[1]. However, even traditionally 'large' datasets, which include hundreds of participants, are underpowered for many association studies[2]. Low statistical power presents numerous roadblocks to the reproducibility of neuroimaging research, including false negatives, inflated effect sizes and replication failures[2–7].

In contrast to association studies, prediction frameworks can alleviate the poor reproducibility seen in certain neuroimaging studies[8–12]. Unlike association, prediction entails the evaluation of a model on unseen data, which minimizes the risk of overfitting, or instances where models fail to generalize beyond the training data. Thus, prediction provides a more robust measure of brain-phenotype associations than in-sample associations. Typically, prediction is achieved by dividing a dataset into training and test sets, such as through *k*-fold cross-validation. Although an improvement over in-sample associations, splitting a dataset into training and test samples does

A full list of affiliations appears at the end of the paper. ✉e-mail: matthew.rosenblatt@yale.edu

https://doi.org/10.1038/s41562-024-01931-7

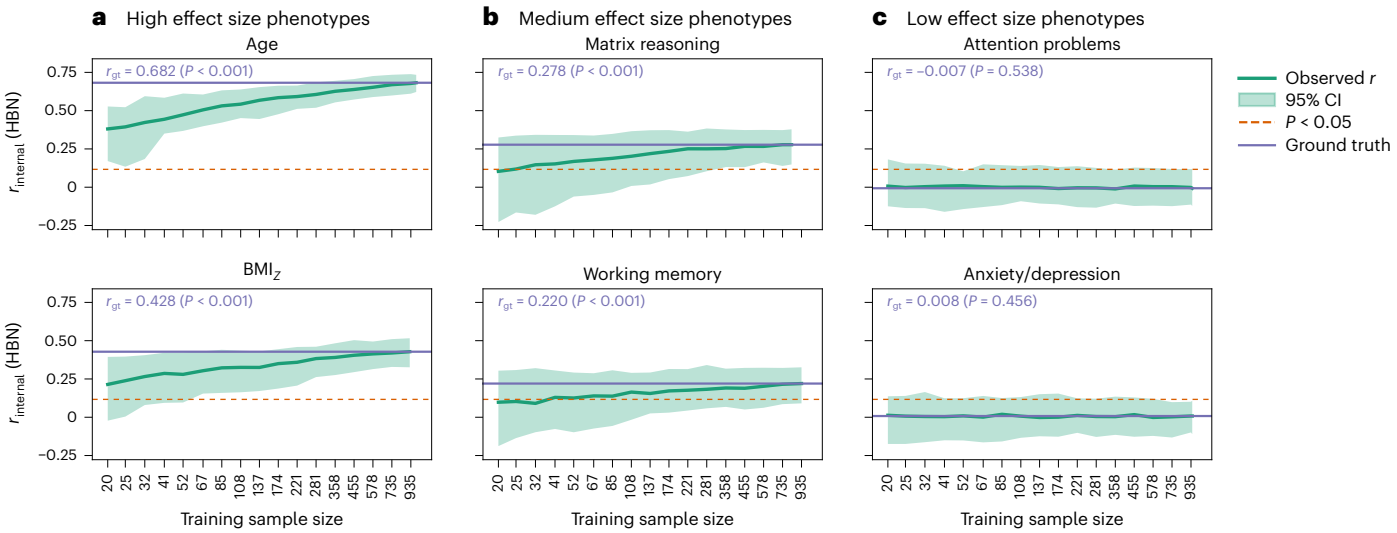

Fig. 1 | Internal validation performance in HBN. a–**c**, Prediction performance in HBN for (**a**) high (age, body mass index), (**b**) medium (matrix reasoning, working memory) and (**c**) low (attention problems, anxiety/depression symptoms) effect size phenotypes. Performance was evaluated in a randomly selected held-out sample of size $N_{held\ out}$ = 200, which was ~20% of the dataset. The centre line shows the median and the shaded area shows the 2.5th to 97.5th percentile among 100 repeats of resampling at each training sample size. The dotted line reflects the correlation value required for a significance level of $P < 0.05$. The solid line reflects the ground truth (gt) internal validation performance, as determined by a model made in the maximum training size and evaluated in the held-out sample. Similar results were observed for the ABCD, HCPD and PNC datasets (Supplementary Figs. 2–4) and for mean absolute error as an evaluation metric (Supplementary Figs. 5–8). *P* values were calculated with a one-sided test of significance of correlations. Multiple comparisons correction was not performed because each result simulates a separate study.

not fully capture the generalizability and utility of brain-phenotype associations. Even with cross-validation, a model can be overfit to the idiosyncrasies of a particular dataset[13,14].

External validation, or applying a model to an entirely different dataset, is the gold standard when evaluating the generalizability of predictive models. Generalizing a model to another dataset with different characteristics provides strong evidence of a robust and reproducible brain-phenotype association. As such, numerous works encourage generalization to external datasets[9,13–17]. Since few studies have the resources to collect two independent samples, external validation is usually performed using an existing publicly available dataset. As the availability of such datasets continues to increase, external validation will probably become more accessible and commonplace. Nevertheless, external datasets are rarely harmonized with the primary dataset, often differing in their phenotypic measures, neuroimaging data, or the relationship between phenotypic and neuroimaging data, which constitutes dataset shift[18,19]. Therefore, researchers typically resort to using the most similar dataset available. Statistical power is rarely a consideration for external validation studies.

Although power is computed for typical single-dataset studies as a function of the sample size, effect size and significance level α (ref. 20), power in external validation may depend on the effect size, both the training and external sample sizes, and potential dataset shift. Thus, traditional power calculations may be insufficient for external validation. A systematic evaluation of how the ground truth effect size, training sample size and external sample size contribute to external validation power will be critical to determining whether established theoretical power curves can be adapted for external validation.

In this work, we simulate the effects of the training and external dataset sizes on prediction power. Our analysis includes seven publicly available neuroimaging datasets, both functional and structural connectivity, and six main phenotypes, spanning various developmental, cognitive and psychiatric traits: age, age- and sex-adjusted body mass index (BMI_Z), matrix reasoning, working memory, attention problems and anxiety/depression symptoms. We first survey what training and external sample sizes have been used by existing external validation studies to contextualize our resampling simulations. Next, we resample the datasets across multiple sample sizes and evaluate internal (that is, within-dataset) and external (that is, across datasets) prediction performance across high, medium and low effect size phenotypes. Finally, we investigate the relationship between the internal and external prediction performance.

Results

External validation sample sizes in the literature

We performed a brief literature review of neuroimaging external validation papers published in 2022–2023 to contextualize typical training and external dataset sample sizes. We further combined our review with a previous analysis that included papers before 2022[14]. Among 54 qualifying articles, the median training sample size was $N = 129$ (IQR = 59.5–371.25) and the median external sample size was $N = 108$ (IQR = 50–281). These sample sizes informed our analysis of the simulated results presented in this work.

Resampling simulations of internal and external validation

We compared the typical external validation sample sizes in the field to resampling-based simulations of external validation in four developmental resting-state functional magnetic resonance imaging (fMRI) datasets: the Healthy Brain Network (HBN) Dataset[21], the Adolescent Brain Cognitive Development (ABCD) Study[22], the Human Connectome Project Development (HCPD) Dataset[23,24], and the Philadelphia Neurodevelopmental Cohort (PNC) Dataset[25,26]. Details about the datasets are presented in Supplementary Tables 1 and 2 and in Methods. Both internal (within-dataset) and external (across-datasets) validation were performed across six phenotypes. The phenotypes were broadly divided into high (Pearson's $r \geq 0.4$; age and BMI_Z), medium ($0.15 \leq r < 0.4$; matrix reasoning and working memory) or low effect size ($r < 0.15$; attention problems and anxiety/depression symptoms) phenotypes on the basis of internal validation performance (Supplementary Table 3). Phenotypes of differing effect sizes were selected because relationships between power and sample size also depend on the ground truth effect size, where stronger effect sizes have greater

https://doi.org/10.1038/s41562-024-01931-7

power[5]. In addition, our definitions of high, medium and low effect sizes differ from traditional classifications[27]. Our definitions were instead adjusted to align with typical results in the neuroimaging field.

In the remaining sections, we used simulations to understand the relationship between training dataset size, external dataset size and prediction power across effect sizes (Supplementary Fig. 1). Our primary results report findings from models that are trained in the HBN dataset and tested in all other datasets (see Supplementary Information for all possible combinations of training/external datasets).

Internal validation performance

For internal validation, a ridge regression[28] model was trained in a subset of one dataset and evaluated in a randomly held-out 20% of the same dataset. All models included covariate regression (self-reported sex, motion and age), feature selection of the top 1% of features most correlated with the outcome of interest[28] and 5-fold cross-validation within the training set to select the L2 regularization parameter α ($\alpha = 10^{\{-3,-2,-1,0,1,2,3\}}$). Family structure was accounted for to ensure that each set of family members was either in the training set or the test set.

As the training sample size increased, internal validation prediction performance, as measured by Pearson's r between the observed and predicted phenotypes, increased for high and medium effect size phenotypes (Fig. 1 and Supplementary Figs. 2–4). In low effect size phenotypes, the average performance increased in some cases (for example, PNC prediction of anxiety/depression symptoms in Supplementary Fig. 4) but otherwise remained near $r = 0$. Unsurprisingly, variability in performance was greater at small sample sizes across all datasets and phenotypes, but this effect was minor in low effect size phenotypes. Similar trends were observed when using mean absolute error (MAE) as an evaluation metric (Supplementary Figs. 5–8), although the effects of training sample size were more subtle in medium effect size phenotypes.

Furthermore, we compared between raw and scaled scores for working memory and matrix reasoning (Supplementary Table 3). Raw scores are used as the primary results throughout the remainder of this work because scaled scores were not available in PNC.

External validation performance

Using the aforementioned pipeline, we performed external validation, where a ridge regression model was trained in one dataset and evaluated in another. Ground truth performances for each dataset and phenotype, evaluated using the full training and external dataset sizes, varied from non-predictive to strong (Supplementary Table 4; also labelled in all figures). We observed significant predictions in all 12 external validation models for age and BMI_Z, 11 models for matrix reasoning, 9 models for working memory, 4 models for attention problems and 5 models for anxiety/depression (Supplementary Table 4). Along with categorizing each phenotype as high, medium or low effect size, we also classified each specific combination of training dataset, external dataset and phenotype as high (Pearson's $r_{\text{ground truth}} > 0.4$), medium

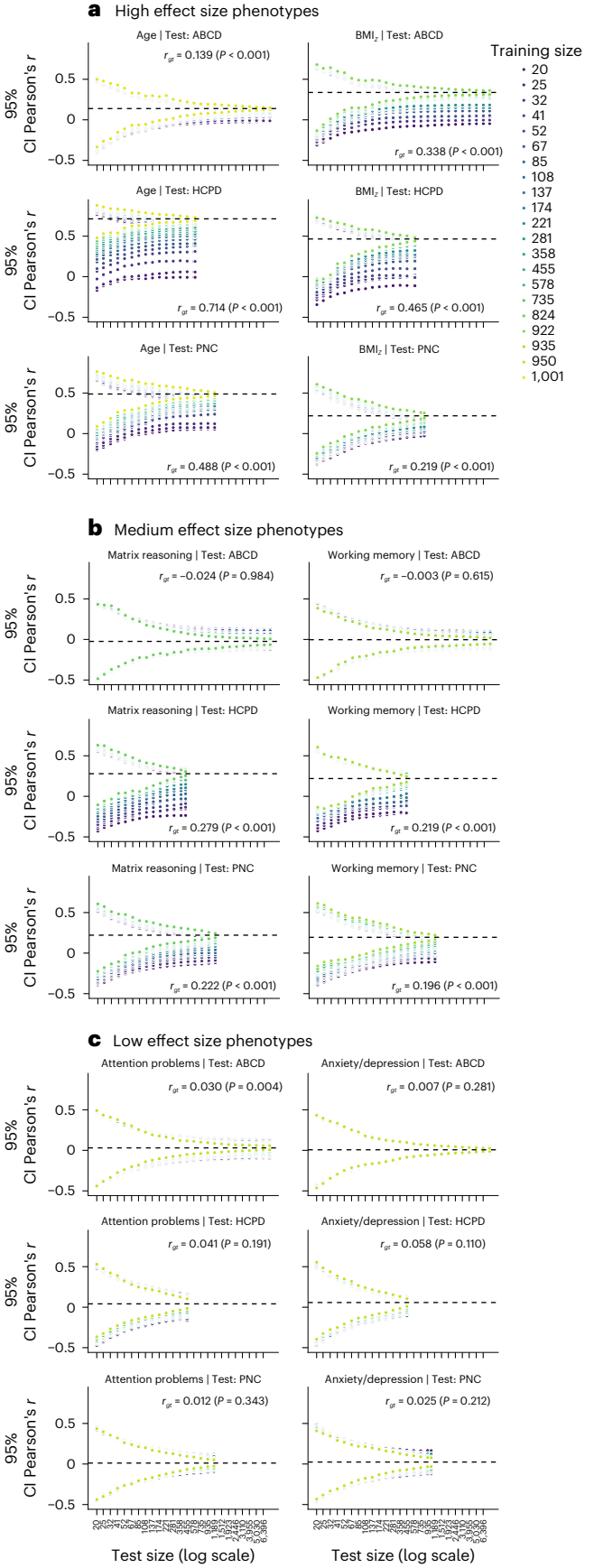

Fig. 2 | 95% confidence intervals of external validation performance, training in HBN. a–c, 95% confidence intervals of external validation performance for (**a**) high (age, body mass index), (**b**) medium (matrix reasoning, working memory) and (**c**) low (attention problems, anxiety/depression symptoms) effect size phenotypes are shown across training sample sizes (colour) and external sample sizes (x axis) when training in HBN and externally validating in ABCD, HCPD and PNC. The dashed line represents the ground truth r, as determined by using full training and external samples. Similar results were observed when training in ABCD, HCPD and PNC datasets (Supplementary Figs. 11–13), as well as for mean absolute error as an evaluation metric (Supplementary Figs. 14–17). In certain cases, points for lower training sample sizes (darker colours) are not visible due to overlap with points for higher training sample sizes (lighter colours). P values were calculated with a one-sided test of significance of correlations. Multiple comparisons correction was not performed because each result simulates a separate study.

https://doi.org/10.1038/s41562-024-01931-7

($0.15 < r_{\text{ground truth}} < 0.4$) or low effect size ($r_{\text{ground truth}} < 0.15$) on the basis of their ground truth prediction performance (Supplementary Table 4).

Notably, external validation was sometimes successful in one direction but not in the other. For instance, matrix reasoning was predicted at $r(417) = 0.329$ ($P < 0.001$) when training in ABCD and testing in HCPD, but $r(7{,}844) = 0.053$ ($P < 0.001$) when training in HCPD and testing in ABCD. This could be due to dataset shift in the distribution of phenotypes, as many phenotypes had statistically significant differences across datasets (Supplementary Table 6 and Fig. 9). In addition, a greater statistical difference between the phenotypes of two datasets was associated with greater differences in model performance when reversing the training and external datasets (Supplementary Fig. 10). External validation often occurs across datasets with different characteristics, which adds to the complexity of statistical power calculations.

Furthermore, the variability in prediction performance was greater as external sample size increased (see 95% CI in Fig. 2 and Supplementary Figs. 11–13). Small training sizes predicting high and medium effect size phenotypes exhibited higher variability in performance (Fig. 2 and Supplementary Figs. 11–13) and underestimation of the ground truth (Supplementary Figs. 18–21). When using MAE as an evaluation metric rather than Pearson's r, variability in performance was still greater with small external sample sizes, and small training sample sizes exhibited higher variability and worse performance (that is, higher MAE). However, the effects were less pronounced (Supplementary Figs. 14–17 for 95% CI and Figs. 22–25 for median performance), particularly in ABCD. Overall, these results highlighted the nuances of external validation. Factors such as effect size, similarity between datasets, training sample size and external sample size all affect prediction performance.

Power and false positive rate for external validation

The fraction of significant simulation results ($P < 0.05$; one-tailed significance test for correlations $r > 0$) was calculated for each combination of training dataset, external dataset, training sample size, external sample size and phenotype. For predictions with a significant ground truth, the proportion was labelled 'Power'. For predictions with a non-significant ground truth, this proportion was labelled 'False positive rate', which is the same as Type 1 error rate (Fig. 3).

In all datasets, external validation power was affected by both the training and external sample sizes (Fig. 3 and Supplementary Figs. 26–28). Higher power was achieved with larger training sample sizes and larger external sample sizes. Results were similar when using MAE as an evaluation metric (Supplementary Figs. 29–32), although some inconsistencies may be due to limiting significance testing to 500 permutations for computational feasibility.

Furthermore, we compared the simulated power to theoretical power curves[20] for one-tailed correlations ($P < 0.05$; see further details in 'Power calculation').

$$\text{power}(r_{\text{ground truth}}, N_{\text{external}}) = 1 - F\left(\tanh^{-1}(r_{\text{ground truth}}) \times \sqrt{N_{\text{external}} - 3}\right) \quad (1)$$

where F is the standard normal cumulative distribution function, $r_{\text{ground truth}}$ is the ground truth external validation prediction performance from the full training and test datasets, and N_{external} is the external

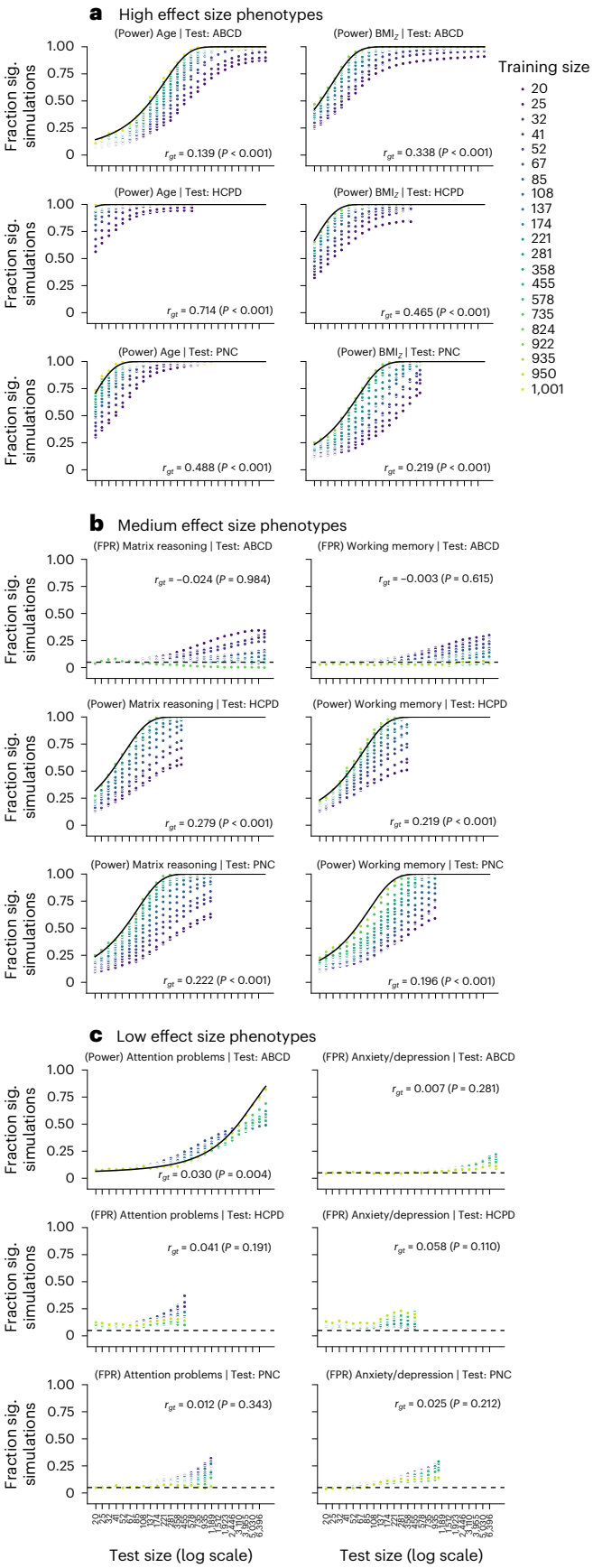

Fig. 3 | Power and false positive rate of external validation, training in HBN. a–c, The fraction of significant simulations as a function of training sample size (colour) and external sample size (x axis) for external validation is shown with training in HBN and testing in ABCD, HCPD or PNC for prediction of (**a**) high (age, body mass index), (**b**) medium (matrix reasoning, working memory) and (**c**) low (attention problems, anxiety/depression symptoms) effect size phenotypes. Non-significant ground truth effects are labelled as false positive rate (FPR). Significant ground truth effects are labelled as power and the black lines represent theoretical power assuming a known ground truth performance. Similar results were observed for the ABCD, HCPD and PNC datasets (Supplementary Figs. 26–28) and when using mean absolute error as an evaluation metric (Supplementary Figs. 29–32). P values were calculated with a one-sided test of significance of correlations. Multiple comparisons correction was not performed because each result simulates a separate study.

https://doi.org/10.1038/s41562-024-01931-7

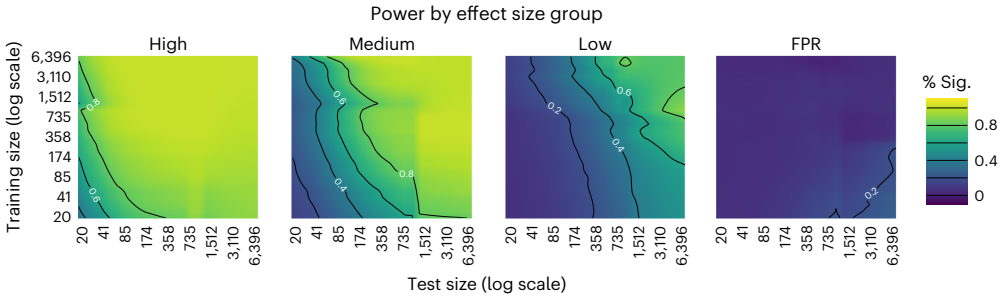

Fig. 4 | Power contour maps as a function of training and external sample sizes across HBN, ABCD, HCPD and PNC. Each prediction—or combination of training dataset, external dataset and phenotype—was categorized as a high effect size (Pearson's $r_{\text{ground truth}} > 0.4$), medium effect size ($0.15 < r_{\text{ground truth}} < 0.4$) or low effect size ($r_{\text{ground truth}} < 0.15$) prediction on the basis of its ground truth prediction performance (Supplementary Table 4). For non-significant ground truth effects, a contour map of the false positive rate is shown. These plots were generated with Plotly and portions of the contours are not covered by data but were filled in with smoothing. % Sig., percentage of simulations that were significant.

sample size. The external validation power closely followed the theoretical curve for power of correlations based on external sample size (Fig. 3 and Supplementary Figs. 26–28; see black lines). Unlike traditional power calculations, which assume that only one sample is used, external validation of predictive models relies on both the training and external sample size. Whereas power is by definition a direct function of the external sample size (equation 1), we also observed a strong, although indirect, contribution of training sample size to statistical power (Fig. 3 and Supplementary Figs. 26–28). Lower training sample sizes decreased the power for a given external sample size.

False positive rate, which we analysed in cases where the ground truth effect was non-significant, was highest for large external samples and small training samples (Fig. 3, and Supplementary Figs. 26–28 for additional datasets and Figs. 29–32 for MAE). At large external sample sizes, effects were significant despite very small effect sizes. Thus, with the high variability of models at a small training sample size, there was a risk of fitting a 'lucky' model, leading to false positives.

After categorizing each combination of training dataset, external dataset and phenotype as a high, medium or low effect size prediction (Supplementary Table 4), we found that high effect size predictions achieved adequate power (>80%) with training and external sample sizes of a few hundred individuals (Fig. 4). Medium effect size predictions required training and external sample sizes in the several hundreds to achieve >80% power, and low effect size predictions achieved >80% power only with hundreds to thousands of training and external samples. These findings suggest that strong effects, such as age, can be robustly detected in small samples.

We also investigated the simulated power for the median sample sizes on the basis of our literature review. Among the sample sizes we evaluated, the training sample size closest to the median was $N = 137$ and the external sample size closest to the median was $N = 108$. For these sample sizes, the median power was 99.67% (range 58.62–100.00%) for high effect size predictions, 49.04% (range 13.31–93.36%) for medium effect size predictions and 9.38% (range 5.74–33.88%) for low effect size predictions. Additional results at sample sizes comparable to the 25th and 75th percentiles in the field are available in Supplementary Table 7. Particularly for low and medium effect sizes, many external validation studies in the field appear to fall short of the typical 80% power standard.

Effect size inflation for external validation

Underpowered studies often lead to false negatives. For example, a study with 40% power has a false negative rate of 60%. Relatedly, effect size inflation, which occurs when the reported performance exceeds the ground truth, is another consequence of underpowered studies. In short, when a test sample is very small, a large effect is required to achieve significance. Thus, reported results that are significant in small, underpowered samples are often inflated relative to the ground truth[3,5,29].

Among significant results, we computed the median effect size inflation (or deflation) relative to the ground truth (Fig. 5 and Supplementary Figs. 33–35). Effect size inflation was most prevalent for low effect size phenotypes and small external sample sizes. For medium and high effect size phenotypes, effect size inflation was prevalent at small external sample sizes. Notably, small training sample sizes demonstrated effect size deflation for medium and high effect size phenotypes, where the performance was lower than the ground truth. Higher MAE (worse performance) was similarly observed with small training samples and decreased MAE (better performance) was observed at small external sample sizes, particularly for low effect size phenotypes (Supplementary Figs. 36–39).

We next calculated effect size inflation across all high, medium and low effect size predictions (Fig. 6). We found that effect size inflation in medium and low effect size predictions primarily depended on the external sample size (Fig. 6). Inflation is a consequence of low power, requiring moderate sample sizes in low-to-medium effect sizes (Fig. 4). For high effect size predictions, deflation occurred at low training sample sizes, and inflation was rare (Fig. 6).

Using the sample sizes closest to the median in the field ($N_{\text{training}} = 137$, $N_{\text{external}} = 108$), the inflation rates ranged from median Δr of −0.29 to −0.04 for high effect size predictions, −0.14 to 0.06 for medium effects and 0.05 to 0.28 for low effects, where negative inflation represents deflation. Additional results at sample sizes comparable to the 25th and 75th percentiles in the field are available in Supplementary Table 7. For high effect size predictions, typical sample sizes in the field could underestimate effect sizes, while effect sizes may be overestimated for low-to-medium effect size predictions.

Relating internal to external validation performance

A key remaining question is how internal and external validation performances are related, and whether a possible relationship can inform future external validation studies. When determining the desired statistical power for an external validation study, the most concerning case would be when r_{internal} is much greater than r_{external}. In this case, one might choose their external sample size for adequate power by using r_{internal}. However, dataset shift may diminish r_{external} and consequently external validation power, possibly leading to false negatives or effect size inflation. Thus, we investigated the proportion of cases where $r_{\text{internal}} - r_{\text{external}}$ was below various thresholds.

The difference between internal and external performances was highly variable for any given subsample, especially at smaller sample sizes (Fig. 7 and Supplementary Figs. 40–42). The internal and external performances were not always closely related. For example, BMI_z predictions had relatively poor generalizability to PNC, so $r_{\text{internal}} - r_{\text{external}}$

https://doi.org/10.1038/s41562-024-01931-7

was consistently greater than zero. Inversely, BMI_Z models from PNC performed better externally than internally, so $r_{internal} - r_{external}$ was negative.

At the training size closest to the existing median in the field ($N_{training} = 137$), 85.04% of evaluations across all datasets and phenotypes met the requirement of ($r_{internal} - r_{external} < 0.2$) and 67.57% met the criteria when restricting to ($r_{internal} - r_{external} < 0.1$). Similar effects were observed for the 25th and 75th percentiles of sample size and when considering only predictions with significant internal validation (Supplementary Table 8).

Moreover, models with significant internal performance generally had better external validation power than models with non-significant ground truth performance. The overall power of models with significant internal validation results was 99.06%, 89.40% and 65.02% for high, medium and low effect size ground truth predictions, respectively. For models with non-significant internal validation the power was 54.98%, 56.30% and 36.51%, respectively. When instead using MAE as a metric (Supplementary Figs. 43–46), the external validation power was 81.67% for models with significant internal validation performance and 44.11% for models with non-significant internal validation performance. Therefore, models that significantly predicted in internal validation also tended to significantly predict in external datasets.

Analysis in additional datasets

To determine whether our findings generalized beyond developmental resting-state functional connectivity data with participants based in the USA, we performed sensitivity analyses with (1) structural connectomes that were generated from diffusion tensor imaging (DTI) data, (2) datasets outside of the USA and (3) adult datasets.

For structural connectivity, we predicted age, working memory and anxiety/depression in three developmental datasets: HBN, HCPD and the Queensland Twin Adolescent Brain (QTAB) Project[30,31]. In addition, we predicted age and accuracy across relational and match conditions during a relational processing task[32,33] using structural connectivity data from two adult datasets, the Chinese Human Connectome Project (CHCP)[34,35] and the Human Connectome Project (HCP)[36]. We predicted task performance in CHCP and HCP because additional behavioural data were not yet available in CHCP. Notably, QTAB and CHCP are Australian and Chinese datasets, respectively, allowing for the assessment of cross-country external validation.

Across all structural connectivity datasets, we observed similar trends in prediction accuracy, effect size inflation and deflation, and power (Supplementary Figs. 47–51). We once again categorized all predictions as high, medium and low effect sizes and recreated the contour plots from Figs. 4 and 6 (Fig. 8). The results were consistent with our resting-state datasets, suggesting that our conclusions pertaining to power and effect size inflation extend to structural connectivity data.

Furthermore, extending the analyses to resting-state functional connectivity in two adult datasets (CHCP and HCP), we predicted age and accuracy across relational and shape conditions during a relational processing task. The results corroborated our results in developmental datasets (Supplementary Figs. 52 and 53). Crucially, effect size inflation was most frequent with small external samples, and effect size deflation

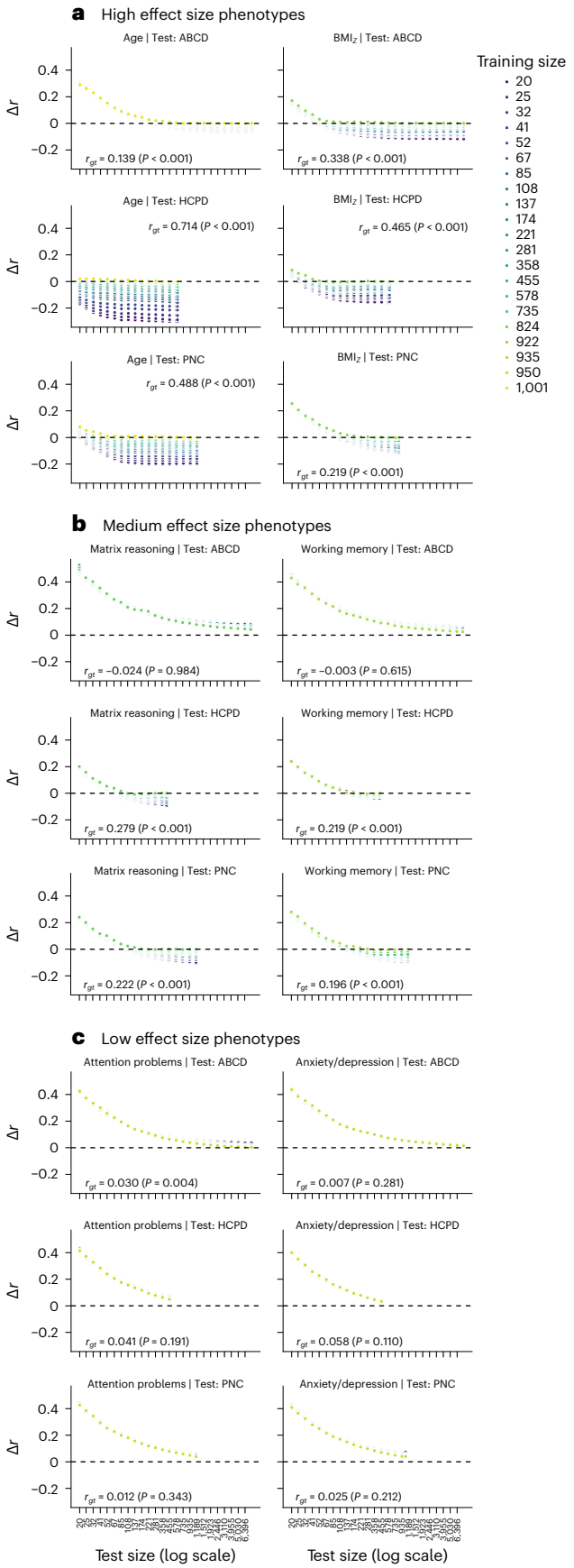

Fig. 5 | Effect size inflation or deflation of external validation predictions, training in HBN. a–c, Median effect size inflation is shown for external validation predictions as a function of training sample size (colour) and external sample size (*x* axis), training in HBN and testing in ABCD, HCPD or PNC for prediction of (**a**) high (age, body mass index), (**b**) medium (matrix reasoning, working memory) and (**c**) low (attention problems, anxiety/depression symptoms) effect size phenotypes. Δ*r* (*y* axis) represents the median difference between the observed, significant external validation performance and the ground truth performance. The dashed horizontal line shows Δ*r* = 0, which represents the line between inflation (Δ*r* > 0) and deflation (Δ*r* < 0). Similar results were observed for the ABCD, HCPD and PNC datasets (Supplementary Figs. 33–35) and when using mean absolute error as an evaluation metric (Supplementary Figs. 36–39). *P* values were calculated with a one-sided test of significance of correlations. Multiple comparisons correction was not performed because each result simulates a separate study.

https://doi.org/10.1038/s41562-024-01931-7

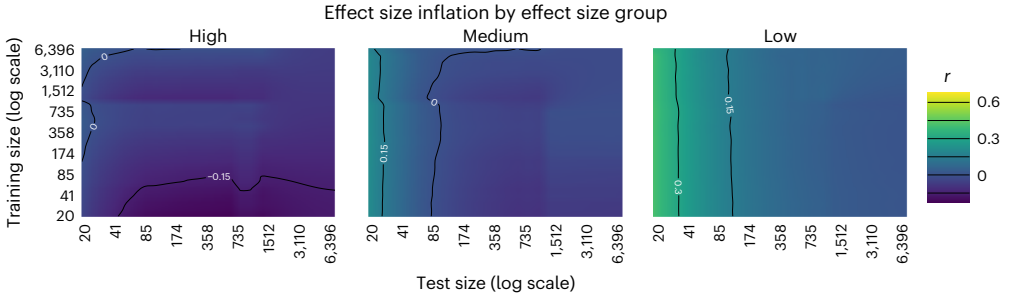

Fig. 6 | Effect size inflation contour maps as a function of training and external sample size across HBN, ABCD, HCPD and PNC. Each prediction—or combination of training dataset, external dataset and phenotype—was categorized as a high effect size (Pearson's $r_{\text{ground truth}} > 0.4$), medium effect size ($0.15 < r_{\text{ground truth}} < 0.4$) or low effect size ($r_{\text{ground truth}} < 0.15$) prediction on the basis of its ground truth prediction performance (Supplementary Table 4). The colour bar shows the difference in r between significant results and the ground truth as a function of training (y axis) and external (x axis) sample size. These plots were generated with Plotly and portions of the contours are not covered by data but were filled in with smoothing.

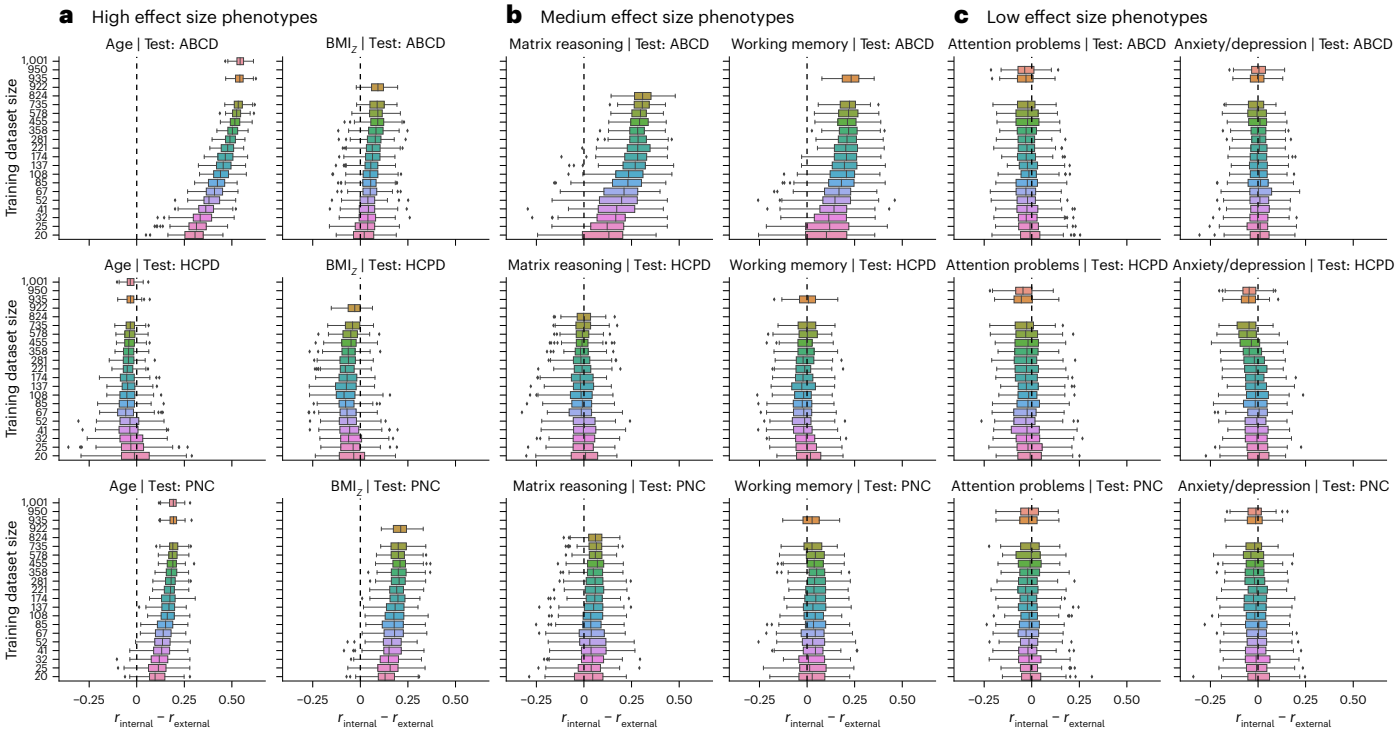

Fig. 7 | Difference between internal and external performance for each subsample of the training data, training in HBN. a–c, Boxplots of the difference between internal and external performance for (**a**) high (age, body mass index), (**b**) medium (matrix reasoning, working memory) and (**c**) low (attention problems, anxiety/depression symptoms) effect size phenotypes. For each training data size, 100 random subsamples were taken. The model was evaluated for internal performance in a held-out sample of size $N_{\text{held out}} = 200$, which was ~20% of the dataset. For external performance, the model formed in the training subsample was applied to the full external dataset. Boxplot elements: centre line, the median of the 100 random subsamples; box limits, the upper and lower quartiles; whiskers, 1.5× the interquartile range; points, outliers. The dashed vertical line shows $r_{\text{internal}} - r_{\text{external}} = 0$. Similar results were observed for the ABCD, HCPD and PNC datasets (Supplementary Figs. 40–42). In addition, we repeated the analyses using mean absolute error as an evaluation metric (Supplementary Figs. 43–46).

was most frequent with small training samples. Power was dependent on both training and external dataset sizes.

Differences between internal and external performance were greater across the new datasets (Supplementary Figs. 54–60). Whereas 85.04% of simulations at the median sample size in the field ($N_{\text{training}} = 137$) in the developmental functional connectivity data met the requirement ($r_{\text{internal}} - r_{\text{external}} < 0.2$), the requirement was only achieved by 66.06% of simulations in developmental structural connectivity data, 46.00% in adult structural connectivity data and 50.00% in adult functional connectivity data. Notably, in the developmental structural connectivity sample, cross-country predictions (between QTAB and either HBN or HCPD) satisfied this requirement for 64.92% of simulations, although only for 37.83% when QTAB was the external dataset. In addition, internal and external performance differences were large (~$r = 0.5$) in adult datasets for age predictions. The discrepancy could be explained by the relatively small dataset size of CHCP or the difference in age ranges between the datasets (Supplementary Table 2). Still, models that demonstrated significant internal validation more frequently exhibited significant external validation. External validation power was 84.99% compared with 71.67% for models that had significant versus non-significant internal validation, respectively.

Effect of scan length

All resting-state scans in our primary analyses were of comparable length (Supplementary Table 11). Given recent evidence that scan length can improve prediction performance[37], we varied the scan length of

https://doi.org/10.1038/s41562-024-01931-7

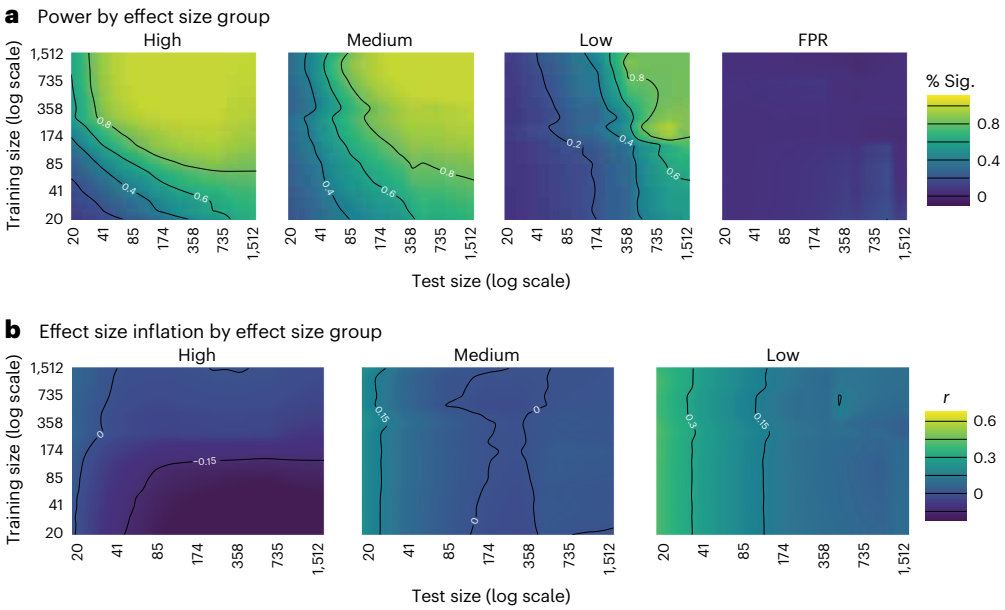

Fig. 8 | Power and effect size inflation contour maps in structural connectivity data. Each prediction—or combination of training dataset, external dataset and phenotype—was first categorized as a high effect size (Pearson's $r_{\text{ground truth}} > 0.4$), medium effect size ($0.15 < r_{\text{ground truth}} < 0.4$) or low effect size ($r_{\text{ground truth}} < 0.15$) prediction on the basis of its ground truth prediction performance. **a**, Power contour maps as a function of training and external sample size across HBN, HCPD and QTAB structural connectivity data. For non-significant ground truth effects, a contour map of the false positive rate is shown. **b**, Effect size inflation contour maps as a function of training and external sample size across HBN, HCPD and QTAB structural connectivity data. The colour bar shows the difference in *r* between significant results and the ground truth as a function of training (*y* axis) and external (*x* axis) sample size. These plots were generated with Plotly and portions of the contours are not covered by data but were filled in with smoothing.

ABCD functional connectivity data (5, 10, 15 or 20 min). Increasing scan length improved internal validation (Supplementary Fig. 61), consistent with ref. 37. External validation had only minor improvements by using longer training dataset scan lengths (Supplementary Fig. 62) or longer external dataset scan lengths (Supplementary Fig. 63). External power curves were similar when varying the scan lengths of either training (Supplementary Fig. 64) or external data (Supplementary Fig. 65). The difference between r_{internal} and r_{external} was greater when training with longer scans (Supplementary Fig. 66) because increasing scan lengths differentially benefited internal compared with external validation. Yet, little to no difference was observed when using longer scans in the external dataset (Supplementary Fig. 67). Overall, although increasing scan lengths improved internal validation performance, our results do not suggest that scan duration will alter power considerations for external validation.

Discussion

This work investigated the effects of training and external sample sizes on the external validation of brain-phenotype predictive models. Resampling-based simulations spanning seven datasets, six phenotypes, and both functional and structural connectivity suggested that previous external validation studies have relied on sample sizes prone to low power, potentially leading to false negatives or, in the case of significant results, effect size inflation. While traditional power calculations depend on the effect size and a single sample size, external validation presents a unique challenge given the presence of two sample sizes. Our simulation results demonstrated that power depended on both the training and external sample sizes and showed varying relationships by effect size. They highlight that traditional power calculations may be unfit for external validation.

Although external validation only occurs in a minority of neuroimaging prediction studies[14], it will probably become increasingly prevalent as the field addresses ongoing reproducibility challenges. Also, external validation may ameliorate certain ethical issues in machine learning[38,39], including bias[40–42] and (lack of) trustworthiness[43]. For bias, evaluating models in external datasets ascertains the robustness and generalizability of brain-phenotype associations in populations with different characteristics[44,45]. For trustworthiness, external validation ensures that data manipulations are not driving the results[43,46]. Given the promise of external validation for improving reproducibility, mitigating bias and increasing trustworthiness, neuroimaging may follow a similar trajectory as genome-wide association studies (GWAS). External replication is now a standard practice for GWAS[47,48]. Thus, a deeper understanding of statistical power in neuroimaging external validation is necessary.

Well-powered studies minimize the possibility of false negatives or misestimation of effect sizes, which, in turn, promotes the reproducibility and utility of scientific insights[2–7]. Adequate power has three components: external sample size, training sample size and effect size.

Statistical power is a direct function of external sample size, as verified by our simulation results. Small external sample sizes were generally underpowered and exhibited false negatives. In addition, low-powered simulations often showed effect size inflation. Intuitively, smaller external dataset sizes require larger effect sizes to achieve significance. Combined with the reporting bias toward significant effects[5,49–51], published effects with small test or external datasets may be inflated. Encouraging researchers to publish the results of external validation attempts, regardless of statistical significance, would ameliorate effect size inflation. A more realistic solution is to promote the use of large external dataset sizes, in which effect sizes are unlikely to be inflated. A general guideline is that if a sample is small enough that you would not train a model in it, you probably should not use it as an external dataset. At a minimum, the results should be interpreted with caution due to the potential for false negatives or, in the case of a significant result, effect size inflation. One caveat is that when using large datasets, statistical significance can be achieved with trivial effect sizes. For instance, an effect of $r = 0.03$ with an $N = 5{,}000$ may not be very meaningful, but it has a P value less than 0.05. However,

small effects can still be meaningful and affect policy[52,53] or inform our understanding of a more complex characteristic. Instead, reporting and interpreting both the effect size and significance are crucial in understanding brain-phenotype associations in large datasets[54,55].

Along with the external sample size, the training sample size plays a crucial role in adequately powering external validation. Large training datasets are needed to avoid overfitting or poor generalizability. Since the training sample size affects the quality of the model and, consequently, the external validation performance, it could be thought of as directly influencing the ground truth performance term in the theoretical power equation. Our simulations indicated that small training sample sizes may lead to false negatives. Moreover, small training samples underestimate the ground truth of large effects, resulting in unreliable effect estimates that decrease external validation performance. Therefore, we recommend using as large of a training sample size as possible to avoid false negatives or underestimation of the ground truth effect size. Furthermore, the combination of small training samples (<100) and large external samples (>500) increased the likelihood of false positives. However, focusing on effect sizes instead of *P* values could mitigate this.

In addition to large sample sizes improving power, we showed that higher effect size phenotypes are better-powered and reduce false negatives or effect size inflation. Several complementary approaches may increase the external validation power by increasing the effect size, including increasing the amount of fMRI data per participant[37] or designing behavioural measures to be better suited for prediction. Recent work has suggested that acquiring longer resting-state scans improves prediction performance[37], and our results support the conclusion that greater scan duration may increase effect sizes in internal validation. Nevertheless, the benefits of longer scan lengths were diminished in external validation. One possible explanation is that differences between datasets outweighed the reliability benefits provided by longer scans. In addition, models that are more reliable or perform better in a particular dataset do not necessarily generalize better. Future work should comprehensively investigate this matter. Regardless of scan length, the relationships between training sample size, external sample size, effect size and external validation power remained unchanged in a preliminary analysis in ABCD. Longitudinal or repeated-measures designs have greater power even with smaller sample sizes[52], and these may also increase effect sizes by acquiring more data for each participant. In addition, incorporating dimensionality reduction approaches, such as principal component analysis or factor analysis, may increase the effect size by creating a more informative composite score of the construct of interest, thus improving power.

Although we explained possible ways to improve the effect size, there is still no obvious way to accurately estimate the ground truth effect size. If the ground truth effect size for a given external validation brain-phenotype association were known, the required external sample size could be calculated directly using power curves. Unfortunately, perfect knowledge of the ground truth effect size is not feasible for two primary reasons. First, as we demonstrated, the observed effect size depends on the training sample size, so the ground truth for a given phenotype cannot be determined a priori. Second, this would require evaluation of external validation performance before performing the study, which is not feasible. Instead, one must rely on either internal validation prediction performance (if the main dataset has already been collected) or published effect sizes, which typically represent internal validation prediction rather than external validation. A decrease in external validation prediction performance compared with internal validation prediction is generally expected due to dataset shift, which is when the training and test populations are mismatched in a way that may degrade performance[18,56,57]. Based on our results, subtracting Pearson's $r = 0.2$ from the internal validation or literature correlation values may be a useful heuristic to account for the observed reductions in external prediction performance.

Additional effect size correction may be needed for external datasets that are poorly harmonized with the training set. A mismatch between datasets may come from differences in population characteristics, image acquisition, phenotypic measurements or brain-phenotype relationships[18,19]. The results in this work were presented in the presence of several additional dataset shifts, most notably significant differences in the distributions of phenotypes, or differences in the assessments used to obtain the measures. The relationships between training sample size, external sample size, ground truth effect size, power and effect size inflation/deflation held across all datasets in this work. However, the difference between internal and external performance tended to be larger in the presence of dataset shift. For example, predictive models using structural connectomes from the HBN or HCPD had poor generalization to QTAB, which may reflect cultural (that is, Australia versus USA) or ascertainment (that is, twins versus singletons) differences. As another example, predictive models of age had poor generalization to ABCD, probably due to the restricted age range of ABCD (9–11 years) compared with the other datasets. Only two non-USA datasets were used (CHCP, QTAB). The applicability of these results across various populations should be further investigated in future work. Nevertheless, harmonization between datasets and similarity between participants should be heavily considered when powering external validation studies because models may perform differently in participants with different characteristics[19,41,58]. Reducing dataset shift is another possible avenue to improve effect size. Some instances of dataset shift are unavoidable, but ongoing efforts to improve MR sequence and measurement harmonization, such as the Common Measures for Mental Health Science[59], may reduce dataset shift. Still, unharmonized, or even negatively correlated, measures can be used in external validation[60,61], such as predicting a clinical measure of attention-deficit/hyperactivity disorder from a sustained attention network despite a negative correlation between the clinical and sustained attention measures[60]. Finally, dataset shift can be mitigated with large, representative samples[57,62]. Future efforts should strive to collect data in diverse populations to better assess the generalizability of brain-phenotype models[63].

There were several limitations to our study. First, we focused on external validation instead of replication in an independent sample, which entails repeating the entire analysis in an independent dataset. Both are valid strategies to improve reproducibility and replicability, but from a predictive sense, external validation is more common. Second, we only analysed multivariate brain-phenotype associations, as multivariate patterns are more reliable and are becoming more popular than univariate associations. Third, to evaluate internal validation performance, we used a small held-out sample (as small as $N_{\text{held-out}} = 75$). This limitation was due to the size of the datasets, but we repeated the evaluation for 100 different random subsamples to reduce the noise. Fourth, all datasets in the main analysis are composed of American youth; however, these cohorts differ in potentially important ways, including geographic region, psychopathology symptoms and behavioural measurements. Our sensitivity analyses found that the power and effect size inflation results generalize to developmental and adult cohorts in other countries, but the results regarding the difference between internal and external performance may not. The relatively small sample sizes of the QTAB and CHCP datasets may contribute to the lack of generalization. This warrants future investigation in larger, more diverse datasets. Furthermore, although twin structure was accounted for in the internal validation of QTAB, the twin structure of QTAB could cause the differences in external validation performance. Fifth, our initial analysis in ABCD suggested a minimal effect of scan length on external validation performance. However, future work should investigate how scan length and reliability could drive differences in external validation.

When selecting a dataset for external validation of a predictive model, one may have few options, depending on the phenotype of interest. If one must use a small training or external dataset in an external validation study, recognizing and explicitly acknowledging the sample size limitations will be crucial for promoting reproducibility. Despite the current reliance of the field on internal validation associations and predictions, external validation will become more widespread. This work provides a starting point for understanding what sample sizes are required to adequately power external validation studies.

Methods

Ethics approval

This study performed secondary data analysis on publicly available datasets. Data collection for these datasets was approved by the relevant ethics review boards for each cohort. Informed consent was obtained by the relevant data collection teams in these public datasets. For minors, a parent/legal guardian provided informed consent, and assent was obtained from the child. In addition, we have a Yale IRB exemption (HIC: 2000023326) to use public neuroimaging data. This study was not preregistered.

Literature review of external validation sample sizes

We performed a brief literature review to contextualize the power and external validation results. Using PubMed, we searched for articles with the following keywords to find functional connectivity prediction papers using external validation: ("functional connect*" OR ("fMRI" AND "connect*")) AND ("predict*") AND ("external" OR "cross-dataset" OR "across datasets" OR "generaliz*"). In cases where the articles used multiple training or external datasets, we recorded the sample size of the largest one. Articles were restricted to 2022 and 2023, which returned 117 articles as of July 2023. Articles were excluded for lacking external validation, not using fMRI connectivity data, or inadequate reporting details. Ultimately, 27 articles were included in our sample. The median sample size of the training dataset was $N = 161$ (IQR: 100–495), and the median sample size of the external dataset was $N = 94$ (IQR: 39.5–682). An additional analysis included papers before 2022[14], and they found 27 articles using external validation. In this sample, the median sample size of the training dataset was $N = 87$ (IQR: 25–343) and the median sample size of the external dataset was $N = 137$ (IQR: 60–197). In both our dataset and the ref. 14 dataset combined, the median training sample size was $N = 129$ (IQR: 59.5–371.25) and the median external sample size was $N = 108$ (IQR: 50–281).

Datasets

Resting-state fMRI data were obtained from four main datasets: the HBN Dataset[21], the ABCD Study[22], the HCPD Dataset[23,24] and the PNC Dataset[25,26]. Details about the datasets are presented in Supplementary Tables 1 and 2. In brief, the HBN dataset consists of participants aged 5–22 years recruited from four sites near the New York greater metropolitan area ($N = 1{,}024$–$1{,}201$). The ABCD dataset consists of 9–11-year-olds who underwent fMRI scanning across 21 sites in the United States ($N = 7{,}846$–$7{,}996$ across phenotypes). The HCPD dataset consists of participants aged 8–22 years who completed fMRI scanning across four sites in the United States (Harvard, UCLA, University of Minnesota, Washington University in St Louis) ($N = 419$–599). The PNC dataset consists of 8–21-year-olds in the Philadelphia area who received care at the Children's Hospital of Philadelphia ($N = 826$–$1{,}179$).

Throughout this work, we predicted age, age- and sex-adjusted body mass index (BMI_Z), matrix reasoning, working memory, attention problems and anxiety/depression symptoms in these four datasets. These measures span a wide range of effect sizes, making them particularly useful for investigating power and effect size inflation. Details about the measures are presented in Supplementary Table 1 and described below, and summary statistics of these measures in each dataset are presented in Supplementary Table 2.

Age in months was used across all datasets.

For BMI, we used the BMI z-score based on US Center for Disease Control (CDC) growth charts[64]. BMI z-scores were computed as follows[65]:

$$\mathrm{BMI}_Z = \frac{\left[\frac{\mathrm{BMI}}{M}\right]^L - 1}{LS} \quad (2)$$

where BMI was computed as the weight in kilograms divided by the square of the height in metres, and L (power in the Box-Cox transformation), M (median) and S (generalized coefficient of variation) were obtained from the CDC growth charts[64] (https://www.cdc.gov/growthcharts/percentile_data_files.htm). To obtain L, M and S values, ages were rounded to the nearest half month, and participants exceeding the age range of the CDC growth charts were considered to be of the maximum age available (240.5 months). While this approximation could slightly affect results, less than 2% of participants in this study exceeded 240.5 months in age and the oldest participant was only 277 months. Furthermore, participants were excluded if their modified z-scores for height, weight or BMI exceeded the CDC cut-offs for extreme values[66] (most recent available cut-offs: https://www.cdc.gov/nccdphp/dnpao/growthcharts/resources/sas.htm).

For the matrix reasoning measure, we used the WISC-V[67] Matrix Reasoning Total Raw Score in HBN, ABCD and HCPD. In PNC, we used the Penn Matrix Reasoning[68,69] Total Raw Score (PMAT_CR, accession code: phv00194834.v2.p2). While we used the Matrix Reasoning Raw Score in the main text, additional results using the Matrix Reasoning Scaled Score in HBN, ABCD and HCPD are presented in Supplementary Table 4.

For working memory, we used the NIH Toolbox List Sorting Working Memory Test[70] in HBN, ABCD and HCPD. In PNC, we used the Letter N-back task: Total Correct Responses to 0-Back, 1-Back and 2-Back Trials[71]. Results using a Working Memory Scaled Score in HBN, ABCD and HCPD are also presented in Supplementary Table 4.

For the attention problems measure, we used the Child Behavior Checklist (CBCL)[72] Attention Problems Raw Score in HBN, ABCD and HCPD. In PNC, we used the Structured Interview for Prodromal Symptoms[73]: Trouble with Focus and Attention Severity Scale (SIP001, accession code: phv00194672.v2.p2).

For anxiety/depression symptoms, we used the Anxiety/Depression CBCL Syndrome Scale Raw Score[72] in HBN, ABCD and HCPD. In PNC, we used the Anxious-misery factor obtained from exploratory factor analysis in ref. 74. This factor is the sum score of 39 questions in the PNC dataset, as detailed in Supplementary Table 1.

Notably, behavioural measures in HBN, ABCD, HCPD and PNC were adjusted to be on the same scale (that is, same possible range of values) across datasets to allow for the prediction and interpretation of mean absolute error (Supplementary Table 5).

Our primary results used HBN as the training dataset and all other datasets as the external datasets. ABCD was not selected because it has a limited age range (9–11 years) for age predictions. In addition, ABCD is the largest dataset in this study, so using it as an external dataset allows for evaluation over a wider range of external sample sizes. HCPD was not selected as the primary training dataset due to its smaller size. PNC was not selected as the primary training dataset because its phenotypes primarily used measures different from those of the other datasets (Supplementary Table 1). This rationale left HBN as the primary training dataset.

Preprocessing

Data in the four main developmental datasets were preprocessed using BioImage Suite (v.3.01)[75], as described in our previous work[76] (and duplicated as follows). This preprocessing included regression of covariates of no interest from the functional data, including linear and quadratic drifts, mean cerebrospinal fluid signal, mean white matter signal and mean global signal. Additional motion control was applied by

regressing a 24-parameter motion model, which included six rigid body motion parameters, six temporal derivatives and the square of these terms, from the data. Subsequently, we applied temporal smoothing with a Gaussian filter (approximate cut-off frequency = 0.12 Hz) and grey matter masking, as defined in common space[77]. Then, the Shen 268-node atlas[78] was applied to parcellate the denoised data into 268 nodes. Finally, we generated functional connectivity matrices by correlating each time series from pairs of nodes and applying the Fisher transform. In cases where multiple functional connectomes were available, the run with the lowest motion was used. Data were excluded for poor data quality (for example, artefacts in T1-weighted images, misaligned registrations), missing nodes due to lack of full brain coverage, high motion (>0.2 mm mean frame-wise displacement) or missing phenotypic data. After applying these exclusion criteria, 1,201, 7,996, 599 and 1,179 participants remained in HBN, ABCD, HCPD and PNC, respectively (Supplementary Table 2).

Data subsampling

For the internal validation, the main dataset was resampled without replacement and split into two subsets: a group to train predictive models (training group) and a group to evaluate the performance of the predictive models (held-out group). We chose to evaluate internal validation performance using a held-out group instead of *k*-fold cross-validation because of computational feasibility and consistency in the model evaluation. For consistency in the model evaluation, we wanted the internal performance to depend on the quality of the model and have similar test data across all models. For example, regardless of whether 20 or 1,000 participants were used to train a model in HBN, both were evaluated in 200 participants. If 10-fold cross-validation were instead used, then the model with 20 participants would be tested in only 2 participants in each fold, while the model with 1,000 participants would be tested in 100 participants in each fold. The held-out group size was selected to be ~20% of the dataset size across all phenotypes, which was $N_{\text{held-out}} = 200$ for HBN, 1,600 for ABCD, 100 for HCPD and 200 for PNC. For the additional datasets, the held-out sample size was 500 for HBN structural connectivity, 100 for HCPD structural connectivity, 100 for QTAB structural connectivity, 75 for CHCP functional and structural connectivity, and 200 for HCP functional and structural connectivity. In datasets with family structure, data were split to ensure that each set of family members was either in the training set or the test set. The training group was randomly subsampled at 25 sample sizes logarithmically spaced from $N = 20$ to $N = 6{,}396$ ($N = 20$, 25, 32, 41, 52, 67, 85, 108, 137, 174, 221, 281, 358, 455, 578, 735, 935, 1,189, 1,512, 1,923, 2,446, 3,110, 3,955, 5,030, 6,396). For external validation, we resampled both the training and external datasets. For each training sample, models were evaluated in random subsets of the external dataset at the sample sizes listed above. These sample sizes were selected to cover a wide range of training and external sample sizes. In addition, the above sample sizes allowed us to investigate external validation performance at sample sizes smaller than, equal to and larger than those common in existing literature.

The resampling procedure was repeated 100 times for the main dataset and the external dataset was resampled 100 times for each of these repeats. Thus, we performed 10,000 evaluations for each combination of the training dataset, external dataset, phenotype, training sample size and external sample size. In total, this paper included over 900 million model evaluations. A summary of the resampling procedure is presented in Supplementary Fig. 1.

Regression models

We referred to two types of results throughout this work: (1) internal validation and (2) external validation. For internal validation, we evaluated performance in a randomly selected held-out sample. Covariates (self-reported sex, motion and age, if applicable) were first regressed from the training data. Then, a ridge regression model was trained using the top 1% of features most correlated with the outcome of interest[28]. Fivefold cross-validation was performed within the training set to select the L2 regularization parameter α ($\alpha = 10^{\{-3,-2,-1,0,1,2,3\}}$). Afterwards, the entire pipeline was applied to the held-out test data. Crucially, the covariate regression parameters and features obtained from the training set were applied to the test set to avoid data leakage[76,79,80]. For external validation, we used the same models as above. However, the model was evaluated with the external dataset instead of the held-out test data. We used ridge regression instead of a more complex machine learning model because linear models are the most commonly used models in the field[14], and other complex, nonlinear models do not tend to show performance improvements in functional and structural connectivity[81].

Performance was evaluated with Pearson's correlation *r* as it is among the most common measures used in neuroimaging predictive studies. For instance, ref. 14 found that 97 of the 108 investigated studies used Pearson's correlation as the evaluation metric.

We additionally evaluated the performance using MAE as a metric, since Pearson's correlation *r* does not quantify the magnitude of errors[82]. Since different measures were on different scales, we divided MAE by the range of values. For example, the matrix reasoning scores ranged from 0–32, so we divided the MAE by 32. Because the anxiety/depression scores ranged from 0–26, we divided the MAE by 26.

We defined the 'ground truth' prediction performance as follows. For internal validation predictions, the ground truth referred to the performance in the total sample averaged over 100 random iterations of nested 5-fold cross-validation. The ground truth was operationalized for external predictions as the prediction performance when training in the whole primary dataset and testing with the entire external dataset.

Power calculation

We calculated power for all combinations of training dataset, test dataset and phenotype that had a significant ground truth effect. Since external validation involves testing a model in an independent dataset, directly converting *r* to *P* values using parametric statistics is appropriate, as opposed to cross-validation, where calculating *P* values requires permutation testing[83]. One-tailed significance testing was used since we only hypothesized that $r > 0$ to achieve significant prediction performance. To calculate power in external validation predictions, we computed the fraction of subsamples that achieved a significant prediction performance, as defined by the field-wide practice of $P < 0.05$.

The null hypothesis was that no positive correlation exists (H_0: $r_{\text{ground truth}} \leq 0$, where $r_{\text{ground truth}}$ is the true correlation). We tested whether there was evidence that the prediction correlation was greater than zero (H_1: $r_{\text{ground truth}} > 0$). To perform parametric statistical testing, we converted correlations with the Fisher arctanh transform[20]:

$$C(r) = \frac{1}{2}\ln\left(\frac{(1+r)}{(1-r)}\right) \tag{3}$$

where for an observed correlation *r* found from *N* observations, $C(r)$ is centred around the true correlation $r_{\text{ground truth}}$ with variance $\sigma^2 = \frac{1}{N-3}$.

For a power analysis, *r* was the ground truth, $r_{\text{ground truth}}$. The relevant test statistic was[20]:

$$Z = C(r_{\text{ground truth}})\sqrt{N_{\text{external}} - 3} \tag{4}$$

The test statistic was then converted to theoretical power using the cumulative distribution function of the standard normal distribution, *F*. Power was calculated assuming a one-tailed $\alpha = 0.05$. This resulted in the final equation as presented in equation (1) in 'Power and false positive rate for external validation predictions', which is restated below:

$$\begin{aligned}&\text{power}(r_{\text{ground truth}}, N_{\text{external}})\\ &= 1 - F(\tanh^{-1}(r_{\text{ground truth}}) \times \sqrt{N_{\text{external}} - 3})\end{aligned} \tag{5}$$

Notably, only the external sample size, $N_{external}$, is a term in the theoretical power calculation. The training sample size, $N_{training}$, only indirectly affects power calculations. An alternative way to think of this is that $N_{training}$ modifies the quality of the model and thus the ground truth correlation $r_{ground\ truth}$, which in turn influences power.

For MAE as an evaluation metric, no theoretical power equation was available because MAE relies on permutation testing to determine significance. For internal validation, 10,000 permutations were used to determine one-tailed significance at $\alpha = 0.05$. For external validation, we used only 500 permutations for computational feasibility.

False positive rate

We computed the false positive rate for all external validation predictions that did not have a significant ground truth effect. The false positive rate is the proportion of simulated examples for which the observed effect is significant ($P < 0.05$) despite a ground truth effect that is not significant.

Performance effect size inflation

Another important consideration is the inflation of reported effect sizes, as documented by numerous previous studies[2,3,5,6]. Low power reduces the likelihood of detecting an actual effect and leads to the inflation of reported significant effects[3,5]. In other words, if significant results are reported in a low-powered sample, such as due to a small sample size, then the effect size is probably inflated.

We first examined all results that achieved significant prediction performance to approximate the inflation of effect sizes because this aligns with publication bias surrounding positive results. We agree with other works that non-significant results should still be published[5,84], but the current reality of the field is that most published results are significant predictions. Among the significant prediction results, we compared the effect size to the ground truth effect size and calculated the inflation relative to the ground truth ($\Delta r = r_{reported} - r_{ground\ truth}$).

Relating internal and external performance

After looking at internal validation performance and external validation performance separately, we compared the two to determine whether internal validation performance could inform how well a model would generalize. We calculated the difference between the internal validation held-out performance ($r_{internal}$) and the performance in the full external dataset ($r_{external}$) for each training sample. We then assessed the performance difference across 100 iterations of random subsampling for each training dataset size.

Structural connectivity datasets

Additional structural connectivity datasets included three developmental datasets: HBN, HCPD and the QTAB Project[30,31]; and two adult datasets: the CHCP[34,35] and the HCP[36]. Behavioural information for these datasets is available in Supplementary Tables 9 and 10, and diffusion imaging acquisition parameters are available in Supplementary Table 12. Certain text about the structural connectivity processing is copied from https://brain.labsolver.org/ to encourage consistency and reproducibility.

For the developmental datasets (HBN, HCPD and QTAB), FIB files were downloaded from https://brain.labsolver.org/, and several extra correction steps were taken. The susceptibility artefact was estimated using reversed phase-encoding b0 by TOPUP from the Tiny FSL package (http://github.com/frankyeh/TinyFSL), a recompiled version of FSL TOPUP (FMRIB, Oxford) with multithread support. The correction was conducted through the integrated interface in DSI-Studio (https://dsi-studio.labsolver.org/). The diffusion MRI data were rotated to align with the AC-PC line. The accuracy of b-table orientation was examined by comparing fibre orientations with those of a population-averaged template[85].

For the adult datasets (CHCP, HCP), diffusion data were downloaded after processing with the HCP minimal preprocessing pipeline (v.3.4.0)[35].

For all datasets, the diffusion data were reconstructed using generalized q-sampling imaging[86] with a diffusion sampling length ratio of 1.25. For the developmental datasets, the tensor metrics were calculated and analysed using the resource allocation (TG-CIS200026) at Extreme Science and Engineering Discovery Environment (XSEDE) resources[87].

As stated in previous works and duplicated here for consistency[88,89], whole-brain fibre tracking was conducted with DSI-Studio with quantitative anisotropy (QA) as the termination threshold. QA values were computed in each voxel in their native space for each subject and were then used to warp the brain to the template in Montreal Neurological Institute (MNI) space using the statistical parametric mapping nonlinear registration algorithm. Once in MNI space, spin density functions were again reconstructed with a mean diffusion distance of 1.25 mm using three fibre orientations per voxel. Fibre tracking was performed in DSI-Studio with an angular cut-off of 60 degrees, step size of 1.0 mm, minimum length of 30 mm, spin density function smoothing of 0.0, maximum length of 300 mm and a QA threshold determined by the diffusion-weighted imaging signal. Deterministic fibre tracking using a modified FACT algorithm[90] was performed until 10,000,000 streamlines were reconstructed for each individual. We used the Shen atlas[78] in MNI space with 268 nodes to construct individual structural connectomes: the pairwise connectivity strength was calculated as the average QA value of each fibre connecting the two end regions and thresholded at 0.001, which results in a 268 × 268 adjacency matrix for each participant.

For structural connectivity models in the developmental data, we predicted age, working memory and anxiety/depression symptoms. For HBN and HCPD, the measures were previously detailed. For QTAB, we used the same measures for age (age in months) and working memory (NIH Toolbox List Sorting Working Memory Test[70]), but used the Spence Children's Anxiety Scale[91] Score for anxiety/depression symptoms (Supplementary Table 1). Summary statistics of these phenotypes are presented in Supplementary Table 9.

For structural connectivity models in the adult data, we predicted age and accuracy across relational and match conditions during a relational processing task[32,33]. We predicted task performance in CHCP and HCP because additional behavioural data were not yet available in CHCP. Summary statistics of these phenotypes are presented in Supplementary Table 10.

Adult functional connectivity datasets

For the CHCP and HCP functional data, data were downloaded from the HCP minimal preprocessing pipeline (v.3.4.0)[35]. Using the minimally preprocessed data as a starting point, the same steps described in 'Preprocessing' were applied.

We predicted age and accuracy across relational and shape conditions during a relational processing task[32,33]. Summary statistics of these phenotypes are presented in Supplementary Table 10.

Effect of scan length

In the ABCD dataset, participants were first restricted to those with four resting-state fMRI scans meeting our inclusion criteria. The resulting sample size was $N = 3{,}946$. Among these participants, either one (5 min), two (10 min), three (15 min) or four (20 min) scans were used. Connectomes were formed by averaging across connectomes from each scan.

Reporting summary

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

Data availability

The following datasets are publicly available but require permission to access. Relevant instructions for data access are available at each individual link below.

The main datasets are available through the Healthy Brain Network Dataset[21] (International Neuroimaging Data-sharing Initiative, https://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/), the Adolescent Brain Cognitive Development Study[22] (NIMH Data Archive, https://nda.nih.gov/abcd), the Human Connectome Project Development Dataset[23,24] (NIMH Data Archive, https://www.humanconnectome.org/study/hcp-lifespan-development/data-releases) and the Philadelphia Neurodevelopmental Cohort Dataset[25,26] (dbGaP Study Accession: phs000607.v3.p2, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000607.v3.p2).
For the additional datasets, the Queensland Twin Adolescent Brain Project dataset[30] is available via OpenNeuro (https://openneuro.org/datasets/ds004146/versions/1.0.4) and the non-imaging phenotypes are available via Zenodo at https://zenodo.org/records/7765506 (ref. 92). Preprocessed structural connectivity data were downloaded from the developmental datasets from https://brain.labsolver.org/hbn.html (HBN), https://brain.labsolver.org/hcp_d.html (HCPD) and https://brain.labsolver.org/greesland_twin.html (QTAB). The Chinese Human Connectome Project dataset[36] is available via the Science Data Bank: https://www.scidb.cn/en/detail?dataSetId=f512d085f3d3452a9b14689e9997ca94. The Human Connectome Project[34] is available via the ConnectomeDB database (https://db.humanconnectome.org). Source data are provided with this paper.

Code availability

We used Python 3.11.3 to conduct the analyses. Code for the analyses is available on GitHub at https://github.com/mattrosenblatt7/external_validation_power (ref. 93) and on Zenodo at https://zenodo.org/records/10975870 (ref. 94). Preprocessing was carried out using Bioimage Suite v.3.01, which is freely available (https://medicine.yale.edu/bioimaging/suite/). Additional preprocessing was performed with the Human Connectome Project minimal preprocessing pipeline v.3.4.0 (https://github.com/Washington-University/HCPpipelines/releases).

References

1. Horien, C. et al. A hitchhiker's guide to working with large, open-source neuroimaging datasets. *Nat. Hum. Behav.* **5**, 185–193 (2021).
2. Marek, S. et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* **605**, E11 (2022).
3. 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).
4. Yarkoni, T. & Braver, T. S. in *Handbook of Individual Differences in Cognition: Attention, Memory, and Executive Control* (eds Gruszka, A. et al.) 87–107 (Springer, 2010).
5. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
6. Cremers, H. R., Wager, T. D. & Yarkoni, T. The relation between statistical power and inference in fMRI. *PLoS ONE* **12**, e0184923 (2017).
7. Liu, S., Abdellaoui, A., Verweij, K. J. H. & van Wingen, G. A. Replicable brain–phenotype associations require large-scale neuroimaging data. *Nat. Hum. Behav.* **7**, 1344–1356 (2023).
8. Klapwijk, E. T., van den Bos, W., Tamnes, C. K., Raschle, N. M. & Mills, K. L. Opportunities for increased reproducibility and replicability of developmental neuroimaging. *Dev. Cogn. Neurosci.* **47**, 100902 (2021).
9. Rosenberg, M. D. & Finn, E. S. How to establish robust brain–behavior relationships without thousands of individuals. *Nat. Neurosci.* **25**, 835–837 (2022).
10. Spisak, T., Bingel, U. & Wager, T. D. Multivariate BWAS can be replicable with moderate sample sizes. *Nature* **615**, E4–E7 (2023).
11. Goltermann, J. et al. Cross-validation for the estimation of effect size generalizability in mass-univariate brain-wide association studies. Preprint at *bioRxiv* https://doi.org/10.1101/2023.03.29.534696 (2023).
12. Makowski, C. et al. Leveraging the adolescent brain cognitive development study to improve behavioral prediction from neuroimaging in smaller replication samples. *Cereb. Cortex* **34**, bhae223 (2024).
13. Genon, S., Eickhoff, S. B. & Kharabian, S. Linking interindividual variability in brain structure to behaviour. *Nat. Rev. Neurosci.* **23**, 307–318 (2022).
14. Yeung, A. W. K., More, S., Wu, J. & Eickhoff, S. B. Reporting details of neuroimaging studies on individual traits prediction: a literature survey. *Neuroimage* **256**, 119275 (2022).
15. Rosenberg, M. D., Casey, B. J. & Holmes, A. J. Prediction complements explanation in understanding the developing brain. *Nat. Commun.* **9**, 589 (2018).
16. 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).
17. Wu, J. et al. Cross-cohort replicability and generalizability of connectivity-based psychometric prediction patterns. *Neuroimage* **262**, 119569 (2022).
18. Dockès, J., Varoquaux, G. & Poline, J.-B. Preventing dataset shift from breaking machine-learning biomarkers. *Gigascience* **10**, giab055 (2021).
19. Kopal, J., Uddin, L. Q. & Bzdok, D. The end game: respecting major sources of population diversity. *Nat. Methods* **20**, 1122–1128 (2023).
20. Lachin, J. M. Introduction to sample size determination and power analysis for clinical trials. *Control. Clin. Trials* **2**, 93–113 (1981).
21. Alexander, L. M. et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Sci. Data* **4**, 170181 (2017).
22. Casey, B. J. et al. The Adolescent Brain Cognitive Development (ABCD) study: imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* **32**, 43–54 (2018).
23. Somerville, L. H. et al. The lifespan human connectome project in development: a large-scale study of brain connectivity development in 5–21 year olds. *Neuroimage* **183**, 456–468 (2018).
24. Harms, M. P. et al. Extending the Human Connectome Project across ages: imaging protocols for the Lifespan Development and Aging projects. *Neuroimage* **183**, 972–984 (2018).
25. Satterthwaite, T. D. et al. Neuroimaging of the Philadelphia Neurodevelopmental Cohort. *Neuroimage* **86**, 544–553 (2014).
26. Satterthwaite, T. D. et al. The Philadelphia Neurodevelopmental Cohort: a publicly available resource for the study of normal and abnormal brain development in youth. *Neuroimage* **124**, 1115–1119 (2016).
27. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (Routledge, 1988).
28. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
29. Ioannidis, J. P. A. Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648 (2008).
30. Strike, L. T. et al. The Queensland Twin Adolescent Brain Project, a longitudinal study of adolescent brain development. *Sci. Data* **10**, 195 (2023).
31. Strike, L. T. et al. Queensland Twin Adolescent Brain (QTAB). *OpenNeuro* https://doi.org/10.18112/openneuro.ds004148.v1.0.1 (2022).
32. Barch, D. M. et al. Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage* **80**, 169–189 (2013).

33. Smith, R., Keramatian, K. & Christoff, K. Localizing the rostrolateral prefrontal cortex at the individual level. *Neuroimage* **36**, 1387–1396 (2007).
34. Van Essen, D. C. et al. The WU-Minn Human Connectome Project: an overview. *Neuroimage* **80**, 62–79 (2013).
35. Glasser, M. F. et al. The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **80**, 105–124 (2013).
36. Ge, J. et al. Increasing diversity in connectomics with the Chinese Human Connectome Project. *Nat. Neurosci.* **26**, 163–172 (2023).
37. Ooi, L. Q. R. et al. MRI economics: balancing sample size and scan duration in brain wide association studies. Preprint at *bioRxiv* https://doi.org/10.1101/2024.02.16.580448 (2024).
38. Chandler, C., Foltz, P. W. & Elvevåg, B. Using machine learning in psychiatry: the need to establish a framework that nurtures trustworthiness. *Schizophr. Bull.* **46**, 11–14 (2020).
39. Mitchell, M. et al. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* 220–229 (Association for Computing Machinery, 2019).
40. Benkarim, O. et al. The cost of untracked diversity in brain-imaging prediction. Preprint at *bioRxiv* https://doi.org/10.1101/2021.06.16.448764 (2021) .
41. Greene, A. S. et al. Brain–phenotype models fail for individuals who defy sample stereotypes. *Nature* **609**, 109–118 (2022).
42. Li, J. et al. Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci. Adv.* **8**, eabj1812 (2022).
43. Rosenblatt, M. et al. Connectome-based machine learning models are vulnerable to subtle data manipulations. *Patterns* https://doi.org/10.1016/j.patter.2023.100756 (2023).
44. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **54**, 1–35 (2021).
45. Tejavibulya, L. et al. Predicting the future of neuroimaging predictive models in mental health. *Mol. Psychiatry* **27**, 3129–3137 (2022).
46. Finlayson, S. G. et al. Adversarial attacks on medical machine learning. *Science* **363**, 1287–1289 (2019).
47. Uffelmann, E. et al. Genome-wide association studies. *Nat. Rev. Methods Primers* **1**, 59 (2021).
48. Poldrack, R. A. et al. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.* **18**, 115–126 (2017).
49. Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
50. Munafò, M. R., Stothart, G. & Flint, J. Bias in genetic association studies and impact factor. *Mol. Psychiatry* **14**, 119–120 (2009).
51. Greenwald, A. G. Consequences of prejudice against the null hypothesis. *Psychol. Bull.* **82**, 1–20 (1975).
52. Gratton, C., Nelson, S. M. & Gordon, E. M. Brain-behavior correlations: two paths toward reliability. *Neuron* **110**, 1446–1449 (2022).
53. Searle, A. K. et al. Tracing the long-term legacy of childhood lead exposure: a review of three decades of the port Pirie cohort study. *Neurotoxicology* **43**, 46–56 (2014).
54. Cohen, J. The earth is round (p < .05). *Am. Psychol.* **49**, 997–1003 (1994).
55. Gigerenzer, G. Mindless statistics. *J. Socio Econ.* **33**, 587–606 (2004).
56. Subbaswamy, A. & Saria, S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* **21**, 345–352 (2020).
57. Finlayson, S. G. et al. The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.* **385**, 283–286 (2021).
58. Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. & Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc. Natl Acad. Sci. USA* **117**, 12592–12594 (2020).
59. Barch, D. M. et al. Common measures for National Institute of Mental Health funded research. *Biol. Psychiatry* **79**, e91–e96 (2016).
60. Rosenberg, M. D. et al. A neuromarker of sustained attention from whole-brain functional connectivity. *Nat. Neurosci.* **19**, 165–171 (2016).
61. Adkinson, B. D. et al. Brain-phenotype predictions can survive across diverse real-world data. Preprint at *bioRxiv* https://doi.org/10.1101/2024.01.23.576916 (2024).
62. Lea, A. S. & Jones, D. S. Mind the gap — machine learning, dataset shift, and history in the age of clinical algorithms. *N. Engl. J. Med.* **390**, 293–295 (2024).
63. Ricard, J. A. et al. Confronting racially exclusionary practices in the acquisition and analyses of neuroimaging data. *Nat. Neurosci.* **26**, 4–11 (2023).
64. Kuczmarski, R. J. et al. 2000 CDC Growth Charts for the United States: methods and development. *Vital Health Stat.* **11**, 1–190 (2002).
65. Cole, T. J., Bellizzi, M. C., Flegal, K. M. & Dietz, W. H. Establishing a standard definition for child overweight and obesity worldwide: international survey. *BMJ* **320**, 1240–1243 (2000).
66. Freedman, D. S. et al. Validity of the WHO cutoffs for biologically implausible values of weight, height, and BMI in children and adolescents in NHANES from 1999 through 2012. *Am. J. Clin. Nutr.* **102**, 1000–1006 (2015).
67. Wechsler, D. *WISC-V: Technical and Interpretive Manual* (Pearson, 2014).
68. Bilker, W. B. et al. Development of abbreviated nine-item forms of the Raven's Standard Progressive Matrices test. *Assessment* **19**, 354–369 (2012).
69. Moore, T. M., Reise, S. P., Gur, R. E., Hakonarson, H. & Gur, R. C. Psychometric properties of the Penn Computerized Neurocognitive Battery. *Neuropsychology* **29**, 235–246 (2015).
70. Tulsky, D. S. et al. NIH Toolbox Cognition Battery (NIHTB-CB): list sorting test to measure working memory. *J. Int. Neuropsychol. Soc.* **20**, 599–610 (2014).
71. Gur, R. C. et al. A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: standardization and initial construct validation. *J. Neurosci. Methods* **187**, 254–262 (2010).
72. Achenbach, T. M. & Ruffle, T. M. The Child Behavior Checklist and related forms for assessing behavioral/emotional problems and competencies. *Pediatr. Rev.* **21**, 265–271 (2000).
73. Miller, T. J. et al. Prodromal assessment with the structured interview for prodromal syndromes and the scale of prodromal symptoms: predictive validity, interrater reliability, and training to reliability. *Schizophr. Bull.* **29**, 703–715 (2003).
74. Moore, T. M. et al. Development of a computerized adaptive screening tool for overall psychopathology ('p'). *J. Psychiatr. Res.* **116**, 26–33 (2019).
75. Papademetris, X. et al. BioImage Suite: an integrated medical image analysis suite: an update. *Insight J.* **2006**, 209 (2006).
76. Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S. & Scheinost, D. Data leakage inflates prediction performance in connectome-based machine learning models. *Nat. Commun.* **15**, 1829 (2024).
77. Holmes, C. J. et al. Enhancement of MR images using registration for signal averaging. *J. Comput. Assist. Tomogr.* **22**, 324–333 (1998).
78. Shen, X., Tokoglu, F., Papademetris, X. & Constable, R. T. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *Neuroimage* **82**, 403–415 (2013).

79. Snoek, L., Miletić, S. & Scholte, H. S. How to control for confounds in decoding analyses of neuroimaging data. *Neuroimage* **184**, 741–760 (2019).
80. Chyzhyk, D., Varoquaux, G., Milham, M. & Thirion, B. How to remove or control confounds in predictive models, with applications to brain biomarkers. *Gigascience* **11**, giac014 (2022).
81. Schulz, M.-A. et al. Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nat. Commun.* **11**, 4238 (2020).
82. Wu, J., Li, J., Eickhoff, S. B., Scheinost, D. & Genon, S. The challenges and prospects of brain-based prediction of behaviour. *Nat. Hum. Behav.* **7**, 1255–1264 (2023).
83. Shen, X. et al. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nat. Protoc.* **12**, 506–518 (2017).
84. Dwan, K. et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE* **3**, e3081 (2008).
85. Yeh, F.-C. et al. Population-averaged atlas of the macroscale human structural connectome and its network topology. *Neuroimage* **178**, 57–68 (2018).
86. Yeh, F.-C., Wedeen, V. J. & Tseng, W.-Y. I. Generalized *q*-sampling imaging. *IEEE Trans. Med. Imaging* **29**, 1626–1635 (2010).
87. Towns, J. et al. XSEDE: accelerating scientific discovery. *Comput. Sci. Eng.* **16**, 62–74 (2014).
88. Gu, S. et al. The energy landscape of neurophysiological activity implicit in brain network structure. *Sci. Rep.* **8**, 2507 (2018).
89. Sun, H. et al. Network controllability of structural connectomes in the neonatal brain. *Nat. Commun.* **14**, 5820 (2023).
90. Yeh, F.-C., Verstynen, T. D., Wang, Y., Fernández-Miranda, J. C. & Tseng, W.-Y. I. Deterministic diffusion fiber tracking improved by quantitative anisotropy. *PLoS ONE* **8**, e80713 (2013).
91. Spence, S. H., Barrett, P. M. & Turner, C. M. Psychometric properties of the Spence Children's Anxiety Scale with young adolescents. *J. Anxiety Disord.* **17**, 605–625 (2003).
92. Strike, L. T. et al. Queensland Twin Adolescent Brain (QTAB) non-imaging phenotypes. *Zenodo* https://doi.org/10.5281/zenodo.7765506 (2022).
93. Rosenblatt, M. External_validation_power. *GitHub* https://github.com/mattrosenblatt7/external_validation_power (2024).
94. Rosenblatt, M. External_validation_power: v1.0.0a. *Zenodo* https://doi.org/10.5281/zenodo.10975870 (2024).

Acknowledgements

This study was supported by the National Institute of Mental Health grant R01MH121095 (obtained by D.S.). M.R. was supported by the National Science Foundation Graduate Research Fellowship under grant DGE2139841. L.T. was supported by the Gruber Science Fellowship. C.C.C. was supported by the Gruber Science Fellowship and the National Science Foundation Graduate Research Fellowship under grant DGE2139841. B.D.A. was supported by NIH Medical Scientist Training Program Training Grant T32GM136651. S.N. was supported by the National Institute of Mental Health under grant R00MH130894. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the funding agencies. The Healthy Brain Network (http://www.healthybrainnetwork.org) and its initiatives are supported by philanthropic contributions from the following individuals, foundations and organizations: Margaret Bilotti; Brooklyn Nets; Agapi and Bruce Burkard; James Chang; Phyllis Green and Randolph Cōwen; Grieve Family Fund; Susan Miller and Byron Grote; Sarah and Geoff Gund; George Hall; Jonathan M. Harris Family Foundation; Joseph P. Healey; The Hearst Foundations; Eve and Ross Jaffe; Howard & Irene Levine Family Foundation; Rachael and Marshall Levine; George and Nitzia Logothetis; Christine and Richard Mack; Julie Minskoff; Valerie Mnuchin; Morgan Stanley Foundation; Amy and John Phelan; Roberts Family Foundation; Jim and Linda Robinson Foundation, Inc.; The Schaps Family; Zibby Schwarzman; Abigail Pogrebin and David Shapiro; Stavros Niarchos Foundation; Preethi Krishna and Ram Sundaram; Amy and John Weinberg; Donors to the 2013 Child Advocacy Award Dinner Auction; Donors to the 2012 Brant Art Auction. Additional data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9–11 and follow them over 10 years into early adulthood. The ABCD Study® is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089, U24DA04112 and U24DA041147. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/consortium_members/. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. The Human Connectome Project Development data was supported by the National Institute Of Mental Health of the National Institutes of Health under Award Number U01MH109589 and by funds provided by the McDonnell Center for Systems Neuroscience at Washington University in St Louis. The HCP-Development 2.0 Release data used in this report came from https://doi.org/10.15154/1520708. Additional data were provided by the PNC (principal investigators H. Hakonarson and R. Gur; phs000607.v1.p1). Support for the collection of these datasets was provided by grant RC2MH089983 awarded to R. Gur and RC2MH089924 awarded to H. Hakonarson. This research has been conducted in part using the QTAB project resource, which was funded by the National Health and Medical Research Council (NHMRC), Australia (Project Grant ID: 1078756 to M.L.W.), the Queensland Brain Institute, University of Queensland, and with the assistance of resources from the Centre for Advanced Imaging and the Queensland Cyber Infrastructure Foundation, University of Queensland. Additional data were provided in part by the Chinese Human Connectome Project (CHCP, PI: J.-H. Gao) funded by the Beijing Municipal Science and Technology Commission, Chinese Institute for Brain Research (Beijing), National Natural Science Foundation of China, and the Ministry of Science and Technology of China. Data were provided in part by the HCP, WUMinn Consortium (principal investigators D. Van Essen and K. Ugurbil; 1U54MH091657) funded by the 16 National Institutes of Health institutes and centres that support the National Institutes of Health Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

Author contributions

M.R., S.N. and D.S. conceptualized the study. L.T., M.R., H.S., M.K., B.D.A. and D.S. curated the data. M.R. performed the formal analysis. M.R. and D.S. drafted the manuscript. L.T., H.S., C.C.C., M.K., B.D.A., R.J., M.L.W., S.N. and D.S. reviewed and edited the manuscript. M.R., R.J. and D.S. contributed to the visualizations. D.S. supervised the project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-024-01931-7.

Correspondence and requests for materials should be addressed to Matthew Rosenblatt.

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

[1]Department of Biomedical Engineering, Yale University, New Haven, CT, USA. [2]Interdepartmental Neuroscience Program, Yale University, New Haven, CT, USA. [3]Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, CT, USA. [4]Department of Bioengineering, Northeastern University, Boston, MA, USA. [5]Department of Psychology, Northeastern University, Boston, MA, USA. [6]Child Study Center, Yale School of Medicine, New Haven, CT, USA. [7]Department of Statistics and Data Science, Yale University, New Haven, CT, USA. ✉e-mail: matthew.rosenblatt@yale.edu

nature portfolio

Corresponding author(s): Matthew Rosenblatt

Last updated by author(s): Jun 16, 2024

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
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ 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)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ 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 data collection was performed in this study. Preprocessing was carried out using Bioimage Suite v3.01, which is freely available (https://medicine.yale.edu/bioimaging/suite/). Additional preprocessing was performed with the Human Connectome Project minimal preprocessing pipeline v3.4.0 (https://github.com/Washington-University/HCPpipelines/releases).
Data analysis	We used Python 3.11.3 to conduct the analyses. Code for the analyses is available on GitHub (https://github.com/mattrosenblatt7/external_validation_power) and Zenodo (https://zenodo.org/records/10975870).

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

These following datasets are publicly available but require permission to access. Relevant instructions for data access are available at each individual link below.

The main datasets are available through the Healthy Brain Network Dataset21 (International Neuroimaging Data-sharing Initiative, https://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/), the Adolescent Brain Cognitive Development Study22 (NIMH Data Archive, https://nda.nih.gov/abcd), the Human Connectome Project Development Dataset23,24 (NIMH Data Archive, https://www.humanconnectome.org/study/hcp-lifespan-development/data-releases), and the Philadelphia Neurodevelopmental Cohort Dataset25,26 (dbGaP Study Accession: phs000607.v3.p2,, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000607.v3.p2).

For the additional datasets, the Queensland Twin Adolescent Brain Project dataset30 is available via OpenNeuro (https://openneuro.org/datasets/ds004146/versions/1.0.4) and the non-imaging phenotypes are available via Zenodo (https://zenodo.org/records/7765506). Preprocessed structural connectivity data were downloaded in the developmental datasets from https://brain.labsolver.org/hbn.html (HBN), https://brain.labsolver.org/hcp_d.html (HCPD), and https://brain.labsolver.org/greesland_twin.html (QTAB). The Chinese Human Connectome Project dataset36 is available via the Science Data Bank: https://www.scidb.cn/en/detail?dataSetId=f512d085f3d3452a9b14689e9997ca94. The Human Connectome Project34 is available via the ConnectomeDB database (https://db.humanconnectome.org).

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	Self-reported sex was included as a covariate when training the predictive models to account for potential sex differences.
Reporting on race, ethnicity, or other socially relevant groupings	No race or ethnicity data were reported in this study. However, the original datasets do describe race and ethnicity data.
Population characteristics	ABCD consists of 9-10 year olds in the United States (Casey et al., 2018). HBN consists of 5-22 year olds from the greater New York area (Alexander et al., 2017). HCPD includes healthy participants ages 8-22, with imaging data acquired at four sites across the United States (Harvard, UCLA, University of Minnesota, Washington University in St. Louis) (Harms et al., 2018; Somerville et al., 2018). The PNC dataset consists of 8-21 year-olds in the Philadelphia area who received care at the Children's Hospital of Philadelphia (Satterthwaite et al., 2014, 2016). QTAB is composed of adolescent twins in Australia. CHCP consists of adults of Han Chinese origin (Ge et al., 2023). HCP consists of healthy adults from the United States (Van Essen et al., 2013). Detailed population descriptions are in the relevant papers for each dataset.
Recruitment	Full recruitment procedures are available from the Adolescent Brain Cognitive Development Study (Casey et al., 2018), the Healthy Brain Network Dataset (Alexander et al., 2017), the Human Connectome Project Development Dataset (Harms et al., 2018; Somerville et al., 2018), the Philadelphia Neurodevelopmental Cohort Dataset (Satterthwaite et al., 2014, 2016), the Queensland Twin Adolescent Brain Project dataset (Strike et al., 2023), the Chinese Human Connectome Project dataset (Ge et al., 2023), and the Human Connectome Project (Van Essen et al., 2013).
Ethics oversight	The four datasets used in this study were each supervised by their relevant ethical review boards. We have a Yale IRB exemption (HIC: 2000023326) to use public neuroimaging data.

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

Life sciences study design

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

Sample size	We did not pre-determine sample size since we did not collect data. Instead, large, open-source neuroimaging datasets were selected for this study. The training group was randomly subsampled at 25 sample sizes logarithmically spaced from N=20 to N=6396 (N=[20, 25, 32, 41, 52, 67, 85, 108, 137, 174, 221, 281, 358, 455, 578, 735, 935, 1189, 1512, 1923, 2446, 3110, 3955, 5030, 6396]). For external validation, we resampled both the training and external datasets. For each training sample, models were evaluated in random subsets of the external dataset at the sample sizes listed above. These sample sizes were selected to cover a wide range of training and external sample sizes. In addition, the above sample sizes allow us to investigate external validation performance at sample sizes smaller than, equal to, and larger than those common in existing literature.
Data exclusions	Data were excluded for missing behavioral measurements, high motion (framewise displacement >0.2 mm), artifacts, misaligned registrations, or missing coverage in the fMRI scan.
Replication	We demonstrated our findings across several datasets. The methods include internal validation (within-dataset) and external validation (cross-dataset).
Randomization	For within-dataset predictions, a held-out subset was randomly selected, and the training data were randomly subsampled to explore the

Randomization	effect of sample size. For external validation, both the training and external datasets were randomly subsampled to determine statistical power at various sample sizes.
Blinding	There was no group allocation (continuous or ordinal measures were used)

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 & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Plants

Seed stocks	N/A (see above, plants are checked as "n/a")
Novel plant genotypes	N/A
Authentication	N/A

Magnetic resonance imaging

Experimental design

Design type	Resting-state fMRI; diffusion tensor imaging
Design specifications	Details are published in the relevant papers for the Adolescent Brain Cognitive Development Study (Casey et al., 2018), the Healthy Brain Network Dataset (Alexander et al., 2017), the Human Connectome Project Development Dataset (Harms et al., 2018; Somerville et al., 2018), the Philadelphia Neurodevelopmental Cohort Dataset (Satterthwaite et al., 2014, 2016), the Queensland Twin Adolescent Brain Project dataset (Strike et al., 2023), the Chinese Human Connectome Project dataset (Ge et al., 2023), and the Human Connectome Project (Van Essen et al., 2013).
Behavioral performance measures	Age in months was used across all datasets. For BMI, we used the BMI z-score based on United States Center for Disease Control (CDC) Growth Charts(Kuczmarski et al., 2002). To obtain L, M, and S values, ages were rounded to the nearest half month, and participants exceeding the age range of the CDC growth charts were considered the maximum age available (240.5 months). While this approximation could slightly affect results, less than 2% of participants in this study exceeded 240.5 months in age, and the oldest participant was only 277 months. Furthermore, participants were excluded if their modified z-scores for height, weight, or BMI exceeded the CDC cutoffs for extreme values(Freedman et al., 2015) (most recent available cutoffs: https://www.cdc.gov/nccdphp/dnpao/growthcharts/resources/sas.htm). For matrix reasoning, we used the WISC-V(Wechsler, 2014) Matrix Reasoning Total Raw Score in HBN, ABCD, and HCPD for the matrix reasoning measure. In PNC, we used the Penn Matrix Reasoning(Bilker et al., 2012; Moore et al., 2015) Total Raw Score (PMAT_CR, accession code: phv00194834.v2.p2). While we used the Matrix Reasoning Raw Score in the main text, additional results using the Matrix Reasoning Scaled Score in HBN, ABCD, and HCPD are presented in Table S4. For working memory, we used the NIH Toolbox List Sorting Working Memory Test(Tulsky et al., 2014) in HBN, ABCD, and HCPD. In PNC, we used the Letter N-back task: Total Correct Responses to 0-Back, 1-Back, and 2-Back Trials(Gur et al., 2010). Results using a Working Memory Scaled Score in HBN, ABCD, and HCPD were also presented in Table S4. For the attention problems measure, we used the Child Behavior Checklist (CBCL) (Achenbach and Ruffle, 2000) Attention Problems Raw Score in HBN, ABCD, and HCPD. In PNC, we used the Structured Interview for Prodromal Symptoms(Miller et al., 2003): Trouble with Focus and Attention Severity Scale (SIP001, accession code:

phv00194672.v2.p2).
For anxiety/depression symptoms, we used the Anxiety/Depression CBCL Syndrome Scale Raw Score(Achenbach & Ruffle, 2000) in HBN, ABCD, and HCPD. In PNC, we used the Anxious-misery factor obtained from exploratory factor analysis by Moore et al.(Moore et al., 2019). This factor is the sum score of 39 questions in the PNC dataset, as detailed in Table S1.
For structural connectivity models in the developmental data, we predicted age, working memory, and anxiety/depression symptoms. For HBN and HCPD, the measures were previously detailed. For QTAB, we used the same measures for age (age in months) and working memory (NIH Toolbox List Sorting Working Memory Test(Tulsky et al., 2014)), but used Spence Children's Anxiety Scale(Spence et al., 2003) Score for anxiety/depression symptoms (Table S1). Summary statistics of these phenotypes are presented in Table S9.
For structural connectivity models in the adult data, we predicted age and accuracy across relational and match conditions during a relational processing task(Barch et al., 2013; Smith et al., 2007). We predicted task performance in CHCP and HCP because additional behavioral data were not yet available in CHCP. Summary statistics of these phenotypes are presented in Table S10.
For functional connectivity models in the adult data, we predicted age and accuracy across relational and shape conditions during a relational processing task(Barch et al., 2013; Smith et al., 2007). Summary statistics of these phenotypes are presented in Table S10.

Acquisition

Imaging type(s): functional (primary analysis); diffusion (secondary analysis of structural connectivity)

Field strength: All datasets were collected at 3T, except a portion of the Healthy Brain Network dataset was collected at 1.5T

Sequence & imaging parameters: Details of the sequence and imaging parameters can be found in the relevant papers for the Adolescent Brain Cognitive Development Study (Casey et al., 2018), the Healthy Brain Network Dataset (Alexander et al., 2017), the Human Connectome Project Development Dataset (Harms et al., 2018; Somerville et al., 2018), the Philadelphia Neurodevelopmental Cohort Dataset (Satterthwaite et al., 2014, 2016), the Queensland Twin Adolescent Brain Project dataset (Strike et al., 2023), the Chinese Human Connectome Project dataset (Ge et al., 2023), and the Human Connectome Project (Van Essen et al., 2013).

Area of acquisition: Whole brain scan

Diffusion MRI ☒ Used ☐ Not used

Parameters:
See Table S12 in the supplementary information.
HBN: b=1000 and 2000 s/mm2; sampling direction 64,64; in-plane resolution 1.8 mm; slice thickness 1.8 mm
HCPD b=1500 and 3000 s/mm2; sampling direction 93,92 in-plane resolution 1.5 mm; slice thickness 1.5 mm
QTAB b=1000 and 3000 s/mm2; sampling direction 10,30; in-plane resolution 2 mm; slice thickness 2 mm
CHCP b=1000, 2000, and 3000 s/mm2; sampling direction 90,90; in-plane resolution 1.25 mm; slice thickness 1.25 mm
HCP b=1000, 2000, and 3000 s/mm2; sampling direction 90,90; in-plane resolution 1.25 mm; slice thickness 1.25 mm

Preprocessing

Preprocessing software: Preprocessing was carried out using BioImage Suite, which is freely available here: (https://medicine.yale.edu/bioimaging/suite/), and DSI Studio (https://dsi-studio.labsolver.org/).

Normalization: nonlinear normalization into MNI space

Normalization template: MNI304

Noise and artifact removal: Several covariates of no interest were regressed from participants' functional data including linear and quadratic drifts, mean cerebrospinal fluid signal, mean white matter signal, and mean global signal. For additional control of possible motion-related confounds, a 24-parameter motion model (including six rigid body motion parameters, six temporal derivatives, and these terms squared) was regressed from the data. The data were temporally smoothed with a Gaussian filter (approximate cutoff frequency=0.12 Hz).

Volume censoring: Subjects with >0.2 mm mean framewise displacement were excluded

Statistical modeling & inference

Model type and settings:
We used the following pipeline for prediction of phenotypes from functional and structural connectivity data:
1) Regression of covariates (self-reported sex, head motion, age except when predicting age)
2) Selection of the top 1% of features, as determined by their univariate correlations with the phenotype of interest
3) Ridge regression (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html) with 5-fold cross-validation within the training data to select the L2 regularization parameter via a grid search
4) Application of steps 1-3 to the test data

Effect(s) tested: We tested the prediction of age, age- and sex-adjusted body mass index, matrix reasoning, working memory, attention problems, and anxiety/depression symptoms from neuroimaging connectivity data.

Specify type of analysis: ☒ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference

(See Eklund et al. 2016)

We reported the correlation between the observed and predicted phenotype (Pearson r). We also reported mean absolute error.

Correction

We ran many simulations to estimate statistical power at various sample sizes of the training and external dataset. Since each simulation is meant to represent a single study, we did not correct for multiple comparisons.

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

Functional connectivity data was obtained by calculating the Pearson correlation between each pair of parcels and then taking the Fisher transform. The Shen 268 atlas was used (Shen et al., 2013).

Multivariate modeling and predictive analysis

Within the training data, covariates (self-reported sex, head motion, age) were regressed from the functional connectivity data, and then the top 1% of features most significantly correlated with the phenotypic variable of interest were selected. After this, a ridge regression model was fit to predict the phenotypic measure from the functional connectivity features. The training process included a 5-fold cross-validation grid search within the training data to select the L2 regularization parameter.