MRI economics: Balancing sample size and scan duration in brain wide association studies

Leon Qi Rong Ooi[1-5], Csaba Orban[2-5], Thomas E Nichols[6], Shaoshi Zhang[1-5], Trevor Wei Kiat Tan[1-5], Ru Kong[2-5], Scott Marek[7], Nico U.F. Dosenbach[7-10], Timothy Laumann[9], Evan M Gordon[7], Juan Helen Zhou[1-4], Danilo Bzdok[11-13], Simon B Eickhoff[14,15], Avram J Holmes[16], B. T. Thomas Yeo[1-5*]

[1]Integrative Sciences and Engineering Programme (ISEP), National University of Singapore
[2]Centre for Sleep and Cognition & Centre for Translational MR Research, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[3]Department of Medicine, Human Potential Translational Research Programme & Institute for Digital Medicine (WisDM), Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[4]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[5]N.1 Institute for Health, National University of Singapore, Singapore
[6]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Oxford, UK
[7]Mallinckrodt Institute of Radiology, Washington University, School of Medicine, USA
[8]Department of Neurology, Washington University, School of Medicine, USA
[9]Department of Psychiatry, Washington University, School of Medicine, USA
[10]Deparments of Paediatrics, Biomedical Engineering, and Psychological and Brain Sciences, Washington University, School of Medicine, USA
[11]Department of Biomedical Engineering, McConnell Brain Imaging Centre, Montreal Neurological Institute, Canada
[12]Faculty of Medicine, School of Computer Science, McGill University, Montreal, QC, Canada
[13]Mila - Quebec Artificial Intelligence Institute, Montreal, QC, Canada
[14]Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Center Jülich, Jülich, Germany
[15]Institute for Systems Neuroscience, Medical Faculty, Heinrich-Heine University Düsseldorf, Düsseldorf, Germany
[16]Department of Psychiatry, Brain Health Institute, Rutgers University, Piscataway, NJ, USA

* Address correspondence to:
B.T. Thomas Yeo
CSC, TMR, ECE, N.1, WISDM
National University of Singapore
Email: thomas.yeo@nus.edu.sg

Abstract

A pervasive dilemma in neuroimaging is whether to prioritize sample size or scan duration given fixed resources. Here, we systematically investigate this trade-off in the context of brain-wide association studies (BWAS) using resting-state functional magnetic resonance imaging (fMRI). We find that total scan duration (sample size × scan duration per participant) robustly explains individual-level phenotypic prediction accuracy via a logarithmic model, suggesting that sample size and scan duration are broadly interchangeable. The returns of scan duration eventually diminish relative to sample size, which we explain with principled theoretical derivations. When accounting for fixed costs associated with each participant (e.g., recruitment, non-imaging measures), we find that prediction accuracy in small-scale BWAS might benefit from much longer scan durations (>50 min) than typically assumed. Most existing large-scale studies might also have benefited from smaller sample sizes with longer scan durations. Both logarithmic and theoretical models of the relationships among sample size, scan duration and prediction accuracy explain well-predicted phenotypes better than poorly-predicted phenotypes. The logarithmic and theoretical models are also undermined by individual differences in brain states. These results replicate across phenotypic domains (e.g., cognition and mental health) from two large-scale datasets with different algorithms and metrics. Overall, our study emphasizes the importance of scan time, which is ignored in standard power calculations. Standard power calculations inevitably maximize sample size at the expense of scan duration. The resulting prediction accuracies are likely lower than would be produced with alternate designs, thus impeding scientific discovery. Our empirically informed reference is available for future study design: WEB_APPLICATION_LINK

Introduction

A fundamental question in systems neuroscience is how individual differences in brain structure and function (as measured by MRI) are related to common variation in phenotypic traits, such as cognitive ability or physical health. Following our previous study (Marek et al., 2022), we define brain wide association studies (BWAS) as studies of the associations between common inter-individual variability in human brain structure/function and phenotypes. An important subclass of BWAS seeks to predict individual-level phenotypes using machine learning. Individual-level prediction is important for addressing basic neuroscience questions and critical for precision medicine (Finn et al., 2015; Gabrieli et al., 2015; Woo et al., 2017; Bzdok & Ioannidis, 2019; Eickhoff & Langner, 2019; Varoquaux & Poldrack, 2019).

Many BWAS are underpowered, leading to low reproducibility and inflated prediction performance (Button et al., 2013; Arbabshirani et al., 2017; Bzdok & Meyer-Lindenberg, 2018; Kharabian Masouleh et al., 2019; Elliott et al., 2020; Poldrack et al., 2020). Larger sample sizes increase reliability of brain-behavior associations (Tian & Zalesky, 2021; Chen et al., 2023) and individual-level prediction accuracy (He et al., 2020; Schulz et al., 2020). Indeed, a recent study suggested that reliable BWAS require thousands of participants (Marek et al., 2022), although certain multivariate approaches might reduce sample size requirements (Chen et al., 2023).

In parallel, other studies have emphasized the importance of long functional MRI (fMRI) scan duration per participant during both resting and task states, which leads to improved data quality and reliability (Nee, 2019; Noble et al., 2019; Elliott et al., 2020; Lynch et al., 2020), as well as new insights into the brain (Laumann et al., 2015; Newbold et al., 2020; Gordon et al., 2023). When sample size is fixed, increasing resting-state fMRI scan duration per participant improves individual-level prediction accuracy of some cognitive measures (Feng et al., 2023).

Therefore, in a world with infinite resources, fMRI-based BWAS should maximize both sample size and scan duration for each participant. However, in reality, BWAS investigators have to decide between scanning more participants (for a shorter duration), or fewer participants (for a longer duration) within a fixed scan budget. Furthermore, there is a fundamental asymmetry between sample size and scan duration per participant because of inherent overhead cost associated with each participant, which can be quite substantial, e.g., when recruiting from a rare population. Surprisingly, the exact trade-off between sample size and scan duration per participant has never been studied. We emphasize that this trade-off is an issue for the design of small studies, as well as large-scale collection efforts with thousands of participants, given competing interests among multiple investigators and limited participant availability.

Here, we used the Adolescent Brain and Cognitive Development (ABCD) study and the Human Connectome Project (HCP) to systematically characterize the effects of sample size and scan duration of resting-state fMRI on BWAS prediction accuracy and reliability. We considered 37 phenotypes from the ABCD study and 59 phenotypes from the HCP dataset, spanning cognition, personality, emotion, physicality, well-being and mental health. We also explored how overhead cost per participant can impact the trade-off between sample size and scan duration in maximizing prediction accuracy within a fixed scan budget, thus providing an empirical reference for future study design.

Results

Larger sample size can compensate for shorter scan duration & vice versa

For each participant, we calculated a 419 × 419 resting-state functional connectivity (RSFC) matrix using the first *T* minutes of fMRI data (Schaefer et al., 2018). *T* was varied from 2 minutes to the maximum scan time in each dataset in intervals of 2 minutes. The RSFC matrices (from the first *T* minutes) served as input features to predict a range of phenotypes in each dataset using kernel ridge regression (KRR) via a nested inner-loop cross-validation procedure. Details of the phenotypes are found in Methods. The analyses were repeated with different numbers of training participants (i.e., different training sample size *N*). Within each cross-validation loop, test participants were fixed across different training set sizes, so that prediction accuracy was comparable across different training set sizes (Figure 1A). The whole procedure was repeated 50 times and averaged to yield stable results (Figure 1A).

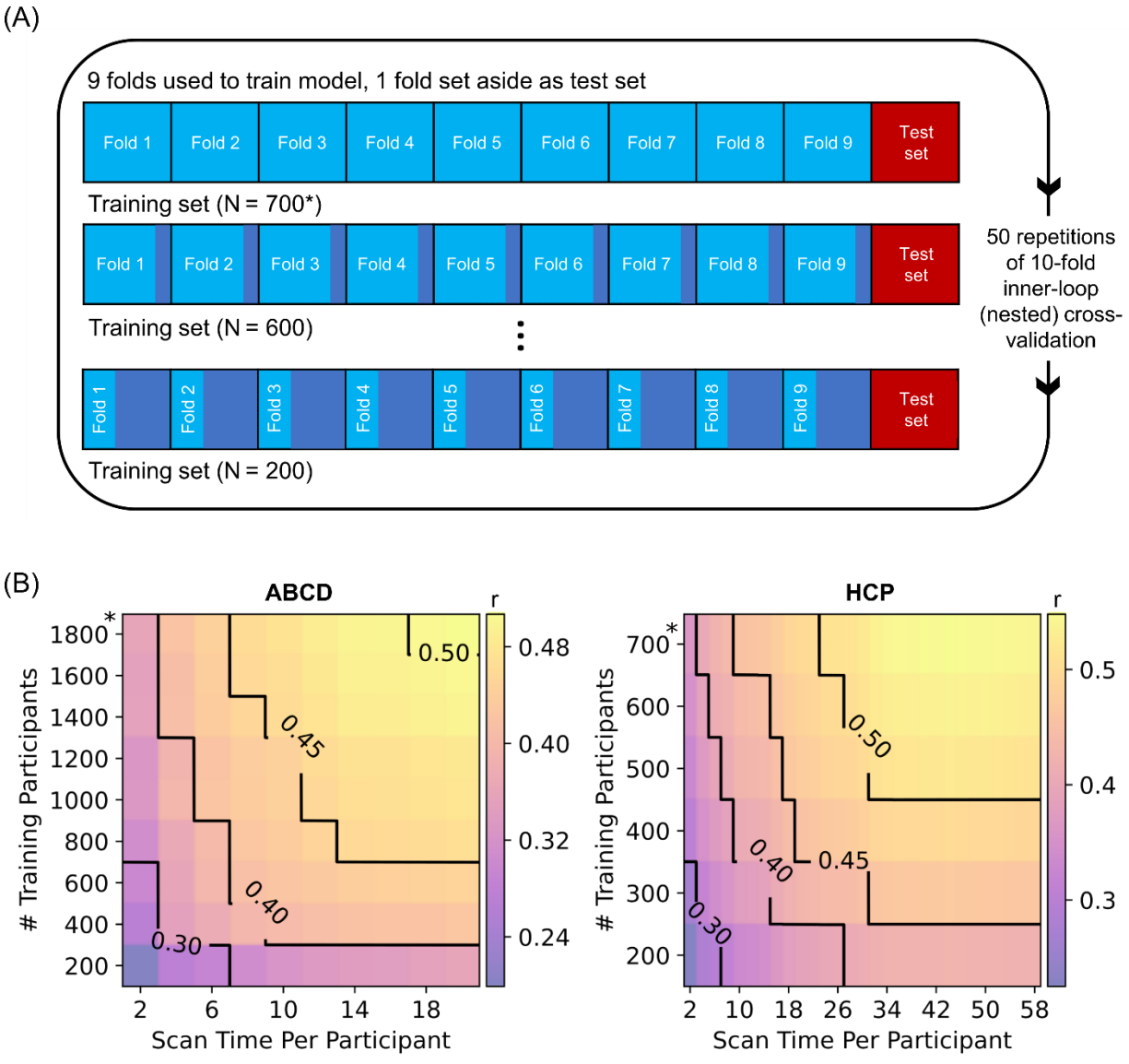


Figure 1. Increasing training participants and scan duration per participant lead to higher prediction accuracy of phenotypes. (A) Prediction workflow for the HCP dataset. The participants were split into 10 folds. One fold was set aside to be the test set. The remaining folds comprised the training set. Cross-validation was performed on the training set to select the best hyperparameter(s). The best hyperparameter(s) were then used to fit a final model from the full training set, which was then used to predict phenotypes in the test set. To vary training set size, each training fold was subsampled and the whole inner-loop nested cross-validation procedure was repeated with the resulting smaller training set. As shown in the panel, the test set remained the same across different training set sizes, so that prediction accuracy was comparable across different sample sizes. Each fold took a turn to be the test set (i.e., 10-fold inner-loop nested cross-validation) and the procedure was repeated with different amounts of fMRI data per participant *T* (not shown in panel). For stability, the entire procedure was repeated 50 times and averaged. A similar workflow was used in the ABCD dataset. We note that in the case of HCP, care was taken so siblings were not split across folds, while in the case of ABCD, participants from the same site were not split across folds. (B) Contour plot of prediction accuracy (Pearson's correlation) of the cognitive factor score as a function of the scan time *T* used to generate the functional connectivity matrix, and the number of training participants *N* used to train the predictive model in the Adolescent Brain and Cognitive Development (ABCD) and Human Connectome Project (HCP) datasets. Increasing training participants and scan time both improved prediction performance. The * in both figures indicates that all available participants were used, therefore the sample size will be close to, but not exactly the number shown. Multiple additional control analyses are found in Figures S1 to S5.

We first considered the cognitive factor score from each dataset because the cognitive factor scores were previously found to exhibit the highest prediction accuracy across all phenotypes (Ooi et al., 2022). Figure 1B shows the prediction accuracy (Pearson's correlation) of the cognitive factors in the HCP and ABCD datasets as a function of both scan time per participant and number of training participants. Along a black iso-contour line, the prediction accuracy is (almost) constant even though scan duration and sample size are changing. Consistent with previous literature (He et al., 2020; Schulz et al., 2023), increasing the number of training participants (when scan time per participant is fixed) improved prediction performance. Similarly, increasing scan time per participant (when number of training participants is fixed) also improved prediction performance (Feng et al., 2023).

Similar conclusions were obtained when we measured prediction accuracy using coefficient of determination (COD) instead of Pearson's correlation (Figure S1), computed RSFC using the first *T* minutes of uncensored data (Figure S2), did not perform censoring of high motion frames (Figure S3), or utilized linear ridge regression (LRR) instead of KRR (Figures S4 & S5).

Sample size & scan time per participant are interchangeable

Next, we characterised the relative contributions of sample size and scan duration per participant to the prediction of different phenotypes. Figure 2A shows that the prediction accuracy of the cognitive factors increases with total scan duration (# training participants × scan time per participant), suggesting that sample size and scan time per participant were broadly interchangeable.

In the HCP dataset, we observed diminishing returns of scan time with respect to sample size for scan time beyond 30 minutes. For example, scanning 700 participants for 14 minutes per participant (with a total scan time of 9800 minutes) and scanning 300 participants for 58 minutes (with a total scan time of 17400 minutes) produced similar prediction accuracy (arrows in Figure 2A). The diminishing returns of scan time was not present in the ABCD study, which had a maximum scan time of 20 minutes.

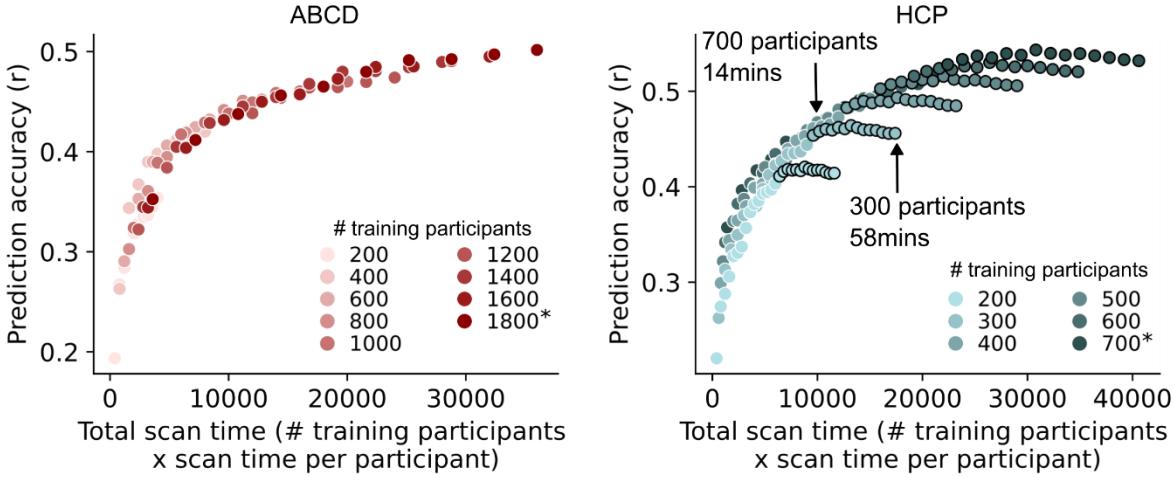
Looking beyond the cognitive factor scores, we focused on 28 (out of 59) HCP phenotypes and 23 (out of 37) ABCD phenotypes that were reasonably well-predicted with maximum prediction accuracies of $r > 0.1$ (Table S1A). Upon visual inspection, we found that 89% (i.e., 25 out of 28) HCP phenotypes exhibited diminishing returns of scan time beyond 20-30 minutes. Diminishing returns were not observed for all 23 ABCD phenotypes.

Overall, this suggests that for almost all phenotypic measures (that were reasonably well-predicted), sample size and scan duration per participant were broadly interchangeable for the ABCD study and up to 30 minutes in the HCP dataset. As will be seen in a later section, the diminishing returns of scan time in the HCP dataset might potentially be the result of inter-individual differences in brain states captured by the HCP study design.

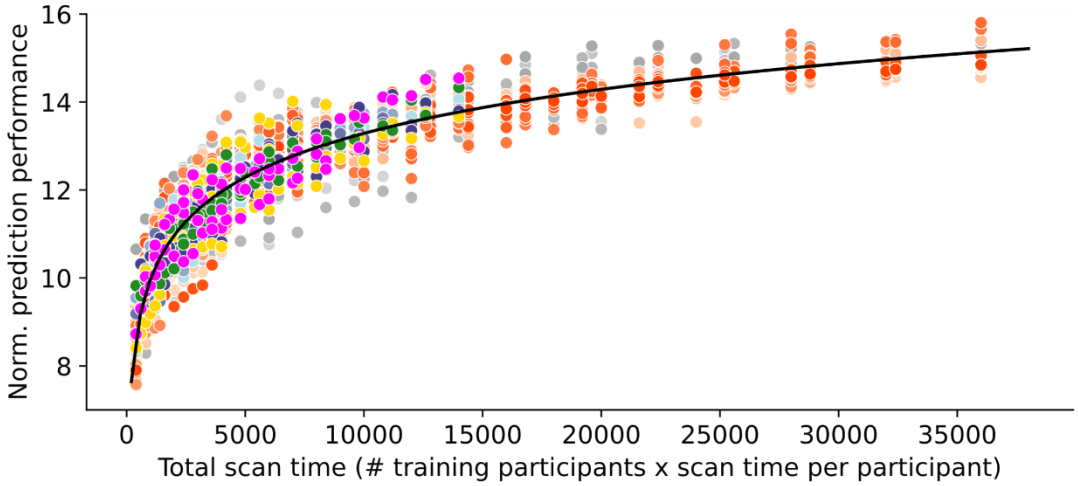
Total scan duration explains prediction accuracy via a logarithmic trend

Among the 25 HCP and all 23 ABCD phenotypes exhibiting broad interchangeability of sample size and scan duration, a logarithmic pattern was evident in 76% (19 out of 25) HCP and 74% (17 out of 23) ABCD phenotypes (Table S1A; Figures S6 and S7). To assess the universality of a logarithmic relationship between total scan time and prediction accuracy, for each of the 19 HCP and 17 ABCD phenotypes, we fitted a logarithm curve (with two free parameters) between prediction accuracy and total scan time (ignoring data beyond 20 minutes per participant). The logarithm fit allowed phenotypic measures from both datasets to be plotted on the same normalized prediction performance scale (Figures 2B). See Methods for details.

(A) **Accuracy of cognition factor scores plotted against total scan time**



(B1) **Fit of logarithmic relationship across phenotypes**



(B2) **Fit of logarithmic relationship across phenotypes (plotted against a log scale)**

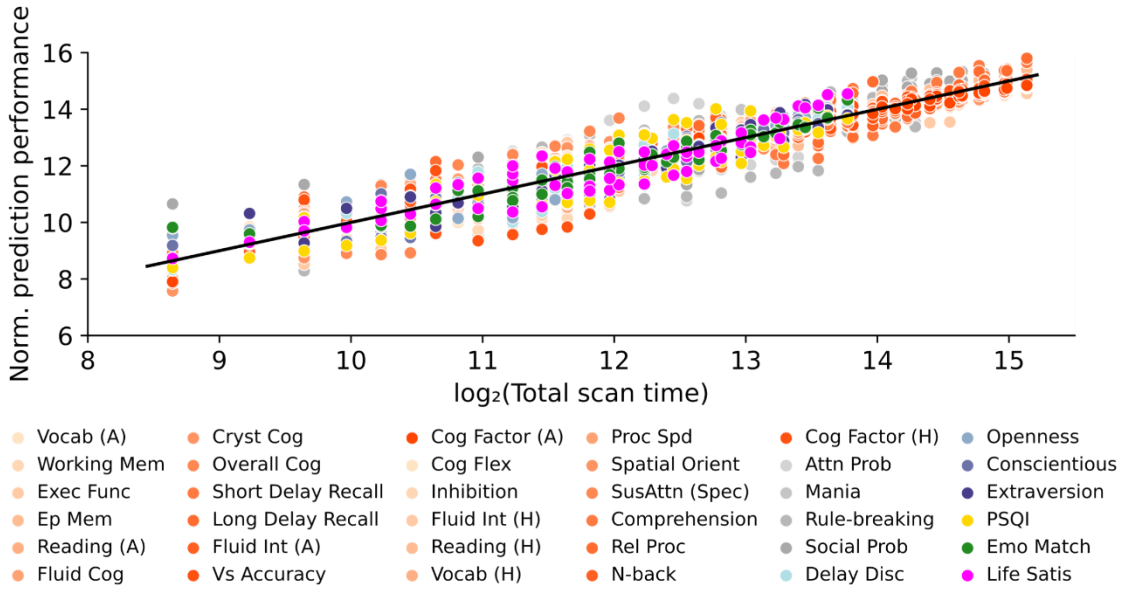


Figure 2. Sample size and scan time are broadly interchangeable for individual-level phenotypic prediction. (A) Scatter plot showing prediction accuracy (Pearson's correlation) of the cognitive factor as a function of total scan time (defined as # training participants x scan time per participant). Each color shade represents different number of total participants used to train the prediction algorithm. Plots are repeated for the Adolescent Brain and Cognitive Development (ABCD) study and Human Connectome Project (HCP). The * indicates that all available participants were used, therefore the sample size will be close to, but not exactly the number shown. There was a diminishing returns of scan time per participant beyond 30 minutes in the HCP dataset; data points with more than 30 minutes of scan time are shown with black outlines. As shown by the black arrows, scanning 700 participants for 14 minutes and 300 participants for 58 minutes yielded the same prediction accuracy, although the total scan duration of the former was almost 2 times lower: 700 × 14 = 9800 vs 300 × 58 = 17400. In the ABCD dataset, where maximum scan time per participant was 20 minutes, the diminishing returns of scan time was not observed. (B1) Scatter plot showing normalized prediction accuracy of the cognitive factor scores and 34 other phenotypes versus total scan duration ignoring data beyond 20 minutes of scan time. Cognitive, mental health, personality, physicality, emotional and well-being measures are shown in shades of red, grey, blue, yellow, green and pink, respectively. The logarithmic black curve suggests that total scan time explained prediction performance well across phenotypic domains and datasets. (B2) Same as Figure 2B1, except the horizontal axis (total scan duration) is plotted on a logarithm scale. The linear black line suggests that the logarithm of total scan duration explained prediction performance well across phenotypic domains and datasets.

The black curve (Figures 2B) indicated the quality of the logarithmic fit of the phenotypes (dots in Figure 2B). Overall, total scan duration explained prediction accuracy across HCP and ABCD phenotypes remarkably well: coefficient of determination (COD) or R^2 = 0.88 and 0.89 respectively. For example, scanning 300 participants for 28 minutes (total scan time = 300 × 28 = 8400 minutes) in the HCP dataset, or 600 participants for 14 minutes (total scan time = 600 × 14 = 8400 minutes) in the ABCD dataset yielded very similar normalized prediction accuracies for the cognitive factor scores (arrows in Figure S8). Quantitative goodness of fit measures are reported in Table S1B.

The logarithm curve was also able to explain prediction accuracy well across different prediction algorithms (KRR and LRR) and different performance metrics (COD and r), as illustrated for the cognitive factor scores in Figure S8. The logarithm fit was also excellent when we considered 30 minutes of scan time, instead of 20 minutes (Figure S9).

As scan time increases, sample size becomes more important than scan time

In the previous sections, we showed that sample size and scan time per participant were broadly interchangeable in the ABCD study and up to 20-30 minutes of scan time per participant in the HCP dataset. To examine this interchangeability more closely, we considered the prediction accuracy of the HCP factor score across six combinations of sample size and scan time totalling 6000 minutes of total scan duration (Figure 3A).

We observed that prediction accuracy decreased with increasing scan time per participant, despite maintaining 6000 minutes of total scan duration (Figure 3A). However, the accuracy reduction was modest below 30 minutes of scan time, and was not significant. Similar conclusions were obtained for all 19 HCP and 17 ABCD phenotypes that followed a logarithmic fit (Figure S10).

The observation that increasing scan time per participant has diminishing returns relative to sample size suggests that a simple logarithmic model does not explain all the factors that contribute to prediction accuracy. In the next section, we derived a mathematical theory that better explains the relative contributions of scan time and sample size to prediction as empirically observed.

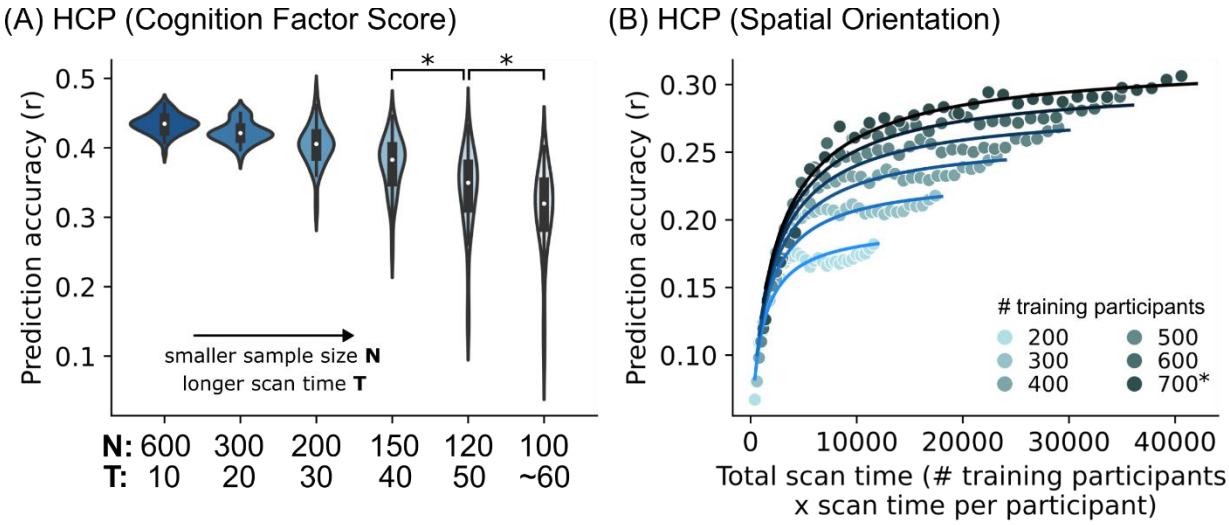


Figure 3. As scan time increases, sample size eventually becomes more important than scan time. (A) Prediction accuracy of the HCP cognition factor score when total scan duration is fixed at 6000 minutes, while varying scan time per participant from 10 to 60 minutes. Each violin plot shows the distribution of prediction accuracies across 50 random cross-validation splits. * indicates that the distributions of prediction accuracies were significantly different after false discovery rate (FDR) $q < 0.05$ correction. (B) Scatter plot of prediction accuracy against total scan duration for a representative phenotype (spatial orientation) in the HCP dataset. The curves were obtained by fitting a theoretical model to the prediction accuracies of the cognitive factor score. The theoretical model explains why sample size is more important than scan time (see main text).

Theoretical relationship of prediction accuracy with sample size & scan time explains why sample size is more important than scan time

Even though sample size and scan time are broadly interchangeable, there is a diminishing return of scan time per participant relative to sample size (Figure 3A). To gain insights into this phenomenon, we derived a closed-form mathematical relationship relating prediction accuracy (Pearson's correlation) with scan time per participant T and sample size N under certain mild assumptions (see Methods).

We found that prediction accuracy can be written as a function of sample size "N" and total scan duration "NT". The theoretical model with three free parameters was estimated by fitting to real data (Figure 1B), yielding an excellent fit with actual prediction accuracies for the 19 HCP and 17 ABCD phenotypes (Figures 3B, S11 & S12): $R^2 = 0.89$ for both datasets (Table S1B).

Based on the estimated model parameters, we find that when T is small, the NT term dominates the N term, which explains the almost 1-to-1 interchangeability between scan duration and sample size for shorter scan duration. The existence of the N term ensures that sample size is still slightly more important than scan time even for small T. As T increases, the N term becomes comparable and then dominates the NT term, so sample size becomes much more important than scan time.

Taking a step back, we note that the theoretical model agreed with the intuition that larger sample size is necessary to capture inter-individual variability in both brain measures and phenotypes (Kharabian Masouleh et al., 2019), which cannot be achieved by just increasing scan duration per participant alone. Scan duration per participant is still important for accounting for within-individual variability. However, a larger sample size can still implicitly account for within-individual variation (Orban et al., 2020), which might explain why ultimately sample size is still slightly more important than scan duration per participant.

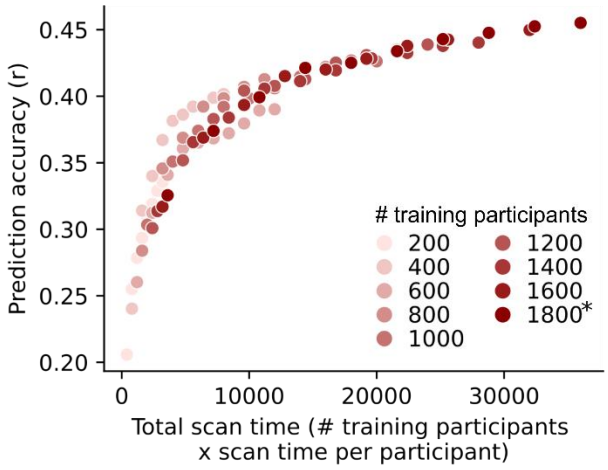
Models work better for well-predicted phenotypes

For phenotypes that were predicted with maximum prediction accuracies of Pearson's $r > 0.1$, the logarithmic and theoretical models were able to explain the prediction accuracies well with an average explained variance >75% (Table S1B). If we loosened the prediction threshold to include phenotypes whose prediction accuracies (Pearson's r) were positive in at least 90% of all combinations of sample size *N* and scan time *T* (Table S1A), the model fit was lower but still relatively high with average explained variance >67% (Table S1B).

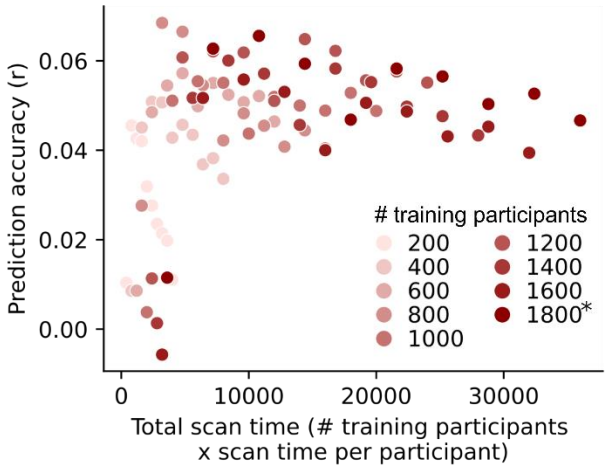
More generally, phenotypes with high overall prediction accuracies adhered to the logarithmic and theoretical models well (example in Figure 4A), while phenotypes with poor prediction accuracies resulted in poor adherence to both models (example in Figure 4B). Indeed, model fit for both models was strongly correlated with prediction accuracy across phenotypes in both datasets (Figures 4C to 4F). These findings suggest that the imperfect fit of the theoretical and logarithmic models for some phenotypes may be partially due to their poor intrinsic predictability, rather than due to true variation in their response patterns.

Examples of phenotypes with good and poor prediction accuracies

(A) ABCD (Vocabulary)

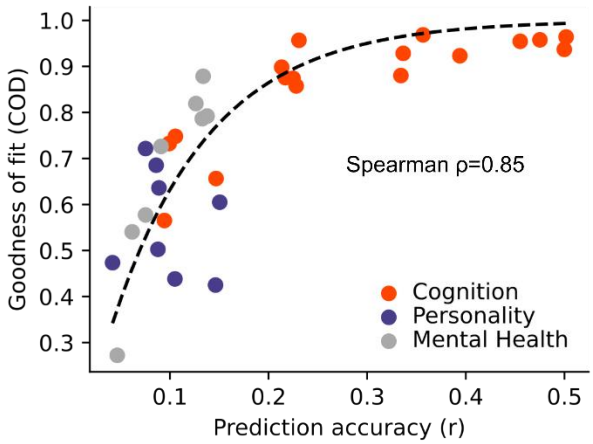


(B) ABCD (Anxious Depresssed)

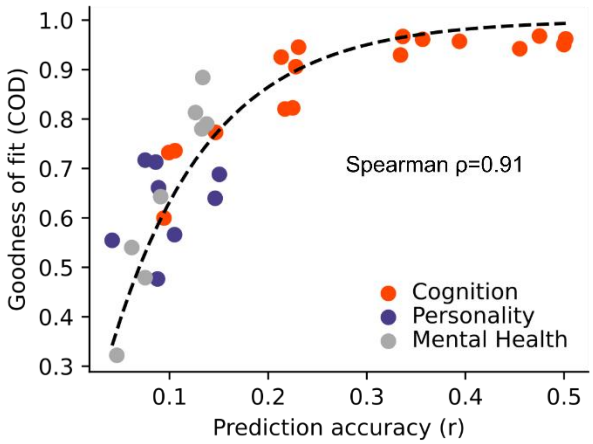


Goodness of fit of model against prediction accuracy for each phenotype

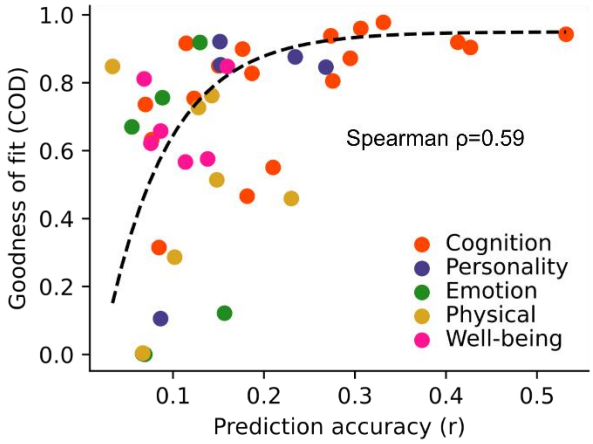
(C) ABCD (Logarithm)



(D) ABCD (Theoretical)



(E) HCP (Logarithm)



(F) HCP (Theoretical)

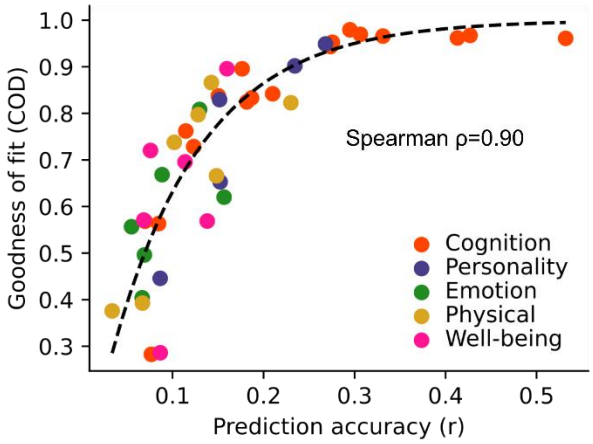


Figure 4. Logarithmic and theoretical models work better for well-predicted phenotypes. (A) Scatter plot of prediction accuracy against total scan duration for an exemplary phenotype with high prediction accuracy. (B) Scatter plot of prediction accuracy against total scan duration for an exemplary phenotype with low prediction accuracy. (C) Scatter plot of logarithmic model goodness-of-fit (coefficient of determination or COD) against prediction accuracies of different ABCD phenotypes. COD (also known as R^2) is a measure of explained variance. Here, we considered phenotypes whose prediction accuracies (Pearson's r) were positive in at least 90% of all combinations of sample size *N* and scan time *T*, yielding 42 HCP phenotypes and 33 ABCD phenotypes. Prediction accuracy (horizontal axis) was based on maximum scan time and sample size. For visualization, we plot a dashed black line by fitting to a monotonically increasing function. (D) Same as panel C but using theoretical (instead of logarithmic) model. (E) Same as panel C but using HCP (instead of ABCD) dataset. (F) Same as panel C, but using HCP (instead of ABCD) and using theoretical (instead of logarithmic) model. For all panels, logarithmic model fit was performed using up to 20 minutes of scan time per participant. For theoretical model fit, the maximum scan time per participant was used.

State effects during fMRI scans weaken model adherence

The theoretical model better matched the empirical data than the logarithmic model. However, there remain discrepancies, particularly in the HCP dataset, which sometimes showed decreases in prediction accuracy with increasing scan time (Figure S7). As noted above, some phenotypes likely fail to match the logarithmic or theoretical models because of intrinsically poor predictability. However, there were phenotypes that were reasonably well-predicted yet still exhibited a low fit to both logarithmic and theoretical models. For example, "Anger: Aggression" was reasonably well-predicted in the HCP dataset, but prediction accuracy was primarily improved by sample size and not scan time (Figure 5A). As scan time per participant increased, prediction accuracy appeared to increase, decrease and then increase again. This pattern was remarkably consistent across sample sizes (Figure 5A).

We hypothesize that this might be due to state changes between or during resting-state scans (Wang et al., 2016; Bijsterbosch et al., 2017; Orban et al., 2020). For example, participants may come to scan sessions under different conditions that can affect brain measurements (e.g., fasted/fed, caffeinated or not, quality of sleep; Laumann et al., 2015; Poldrack et al., 2015; Yeo et al., 2015). Further, it is well known that arousal generally decreases during resting-state scans (Tagliazucchi & Laufs, 2014), which might increase or decrease prediction accuracy depending on the phenotype. To test this hypothesis, we randomized the fMRI run order for each participant and repeated the analysis (see Methods). In the case of "Anger: Aggression", the prediction accuracies were now well-explained by the logarithmic and theoretical models (Figure 5B), although the diminishing returns of scan time still existed for certain phenotypes (Figure S13 and S14).

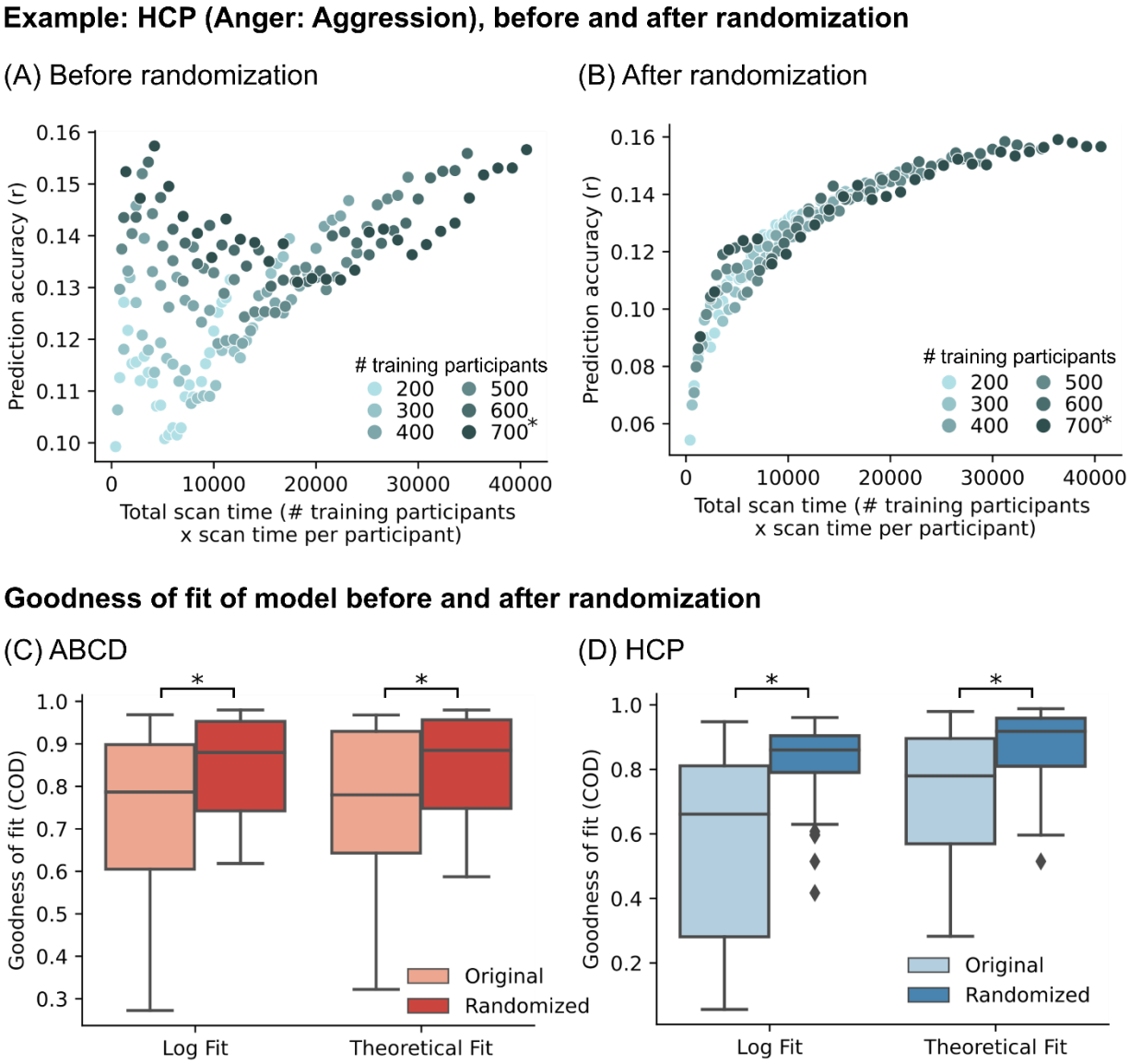


Figure 5. State effects during resting-state fMRI weaken adherence to logarithmic and theoretical models. (A) Scatter plot of prediction accuracy against total scan duration for the "Anger: Aggression" phenotype in the HCP dataset. Despite relatively high accuracy, the phenotype improved with larger sample size, but not scan time. As scan time per participant increases, prediction accuracy appeared to increase, decrease, then increase again. (B) Scatter plot of prediction accuracy against total scan duration for the "Anger: Aggression" phenotype in the HCP dataset after randomizing fMRI run order for each participant. Observe that the prediction accuracy now adheres strongly to the logarithmic and theoretical models. (C) Box plots showing goodness of fit to logarithmic and theoretical models before and after randomizing fMRI run order for a larger set of phenotypes in the ABCD dataset. Here, we considered all phenotypes whose prediction accuracies (Pearson's r) were positive in at least 90% of all combinations of N and T. * indicates that goodness-of-fits were significantly different (after FDR correction with $q < 0.05$). (D) Same as panel C, but in the HCP dataset. For all panels, model fit was performed using the maximum scan time per participant.

Many studies would benefit from longer scan time per participant

We have shown that investigators have the flexibility of attaining a specified prediction accuracy through different combinations of sample size and scan time per participant. To derive a reference for future studies, we fitted the theoretical model to the 17 HCP and 19 ABCD phenotypes, yielding 89% average explained variance (Table S1B). For each phenotype, the model was normalized by its maximum achievable accuracy (estimated by the theoretical model), yielding a fraction of maximum achievable prediction accuracy for every combination of sample size and scan time per participant. The fraction of maximum achievable prediction accuracy was then averaged across the 36 phenotypes (Figure 6A).

For the purpose of study design, we need to consider the fundamental asymmetry between sample size and scan duration per participant because of inherent fixed overhead cost associated with each participant (including recruitment effort and non-fMRI scanning time), which can be substantial. Figure 6B illustrates the prediction accuracy that can be achieved with different total fMRI budgets, costs per hour of scan time and overhead cost per participant. The solid circles indicate the optimal scan time per participant leading to the highest prediction accuracy.

There are three main observations (Figure 6B). First, larger total fMRI budgets, lower scan cost per hour and lower overhead cost per participant allowed for greater achievable prediction accuracy. Second, the optimal scan time increases with larger overhead cost per participant, lower total fMRI budget and lower scan cost per hour. Third, the optimal sample size decreases with larger overhead cost per participant, lower total fMRI budget and lower scan cost per hour.

As an example, when total fMRI budget was $100K, scan cost was $500 per hour and overhead cost was $500 for each participant, the optimal prediction accuracy was achieved by scanning 105 participants for 54 minutes per participant. As another example, when total fMRI budget was $10M, scan cost was $500 per hour and overhead cost was $500 for each participant, the optimal prediction accuracy was achieved by scanning 12,500 participants for 36 minutes per participant. Thus, many studies, including very large-scale studies, might have benefited from increasing scan time per participant than typically assumed.

Similar conclusions were reached if we only considered a subset of 13 phenotypes that exhibited strong agreement with the theoretical model without serious over-shoot or under-shoot (Figure S15), all 36 phenotypes after randomizing the run order (Figure S16) and a subset of 17 phenotypes that exhibited strong agreement with the theoretical model without serious over-shoot or under-shoot after randomizing run order (Figure S17).

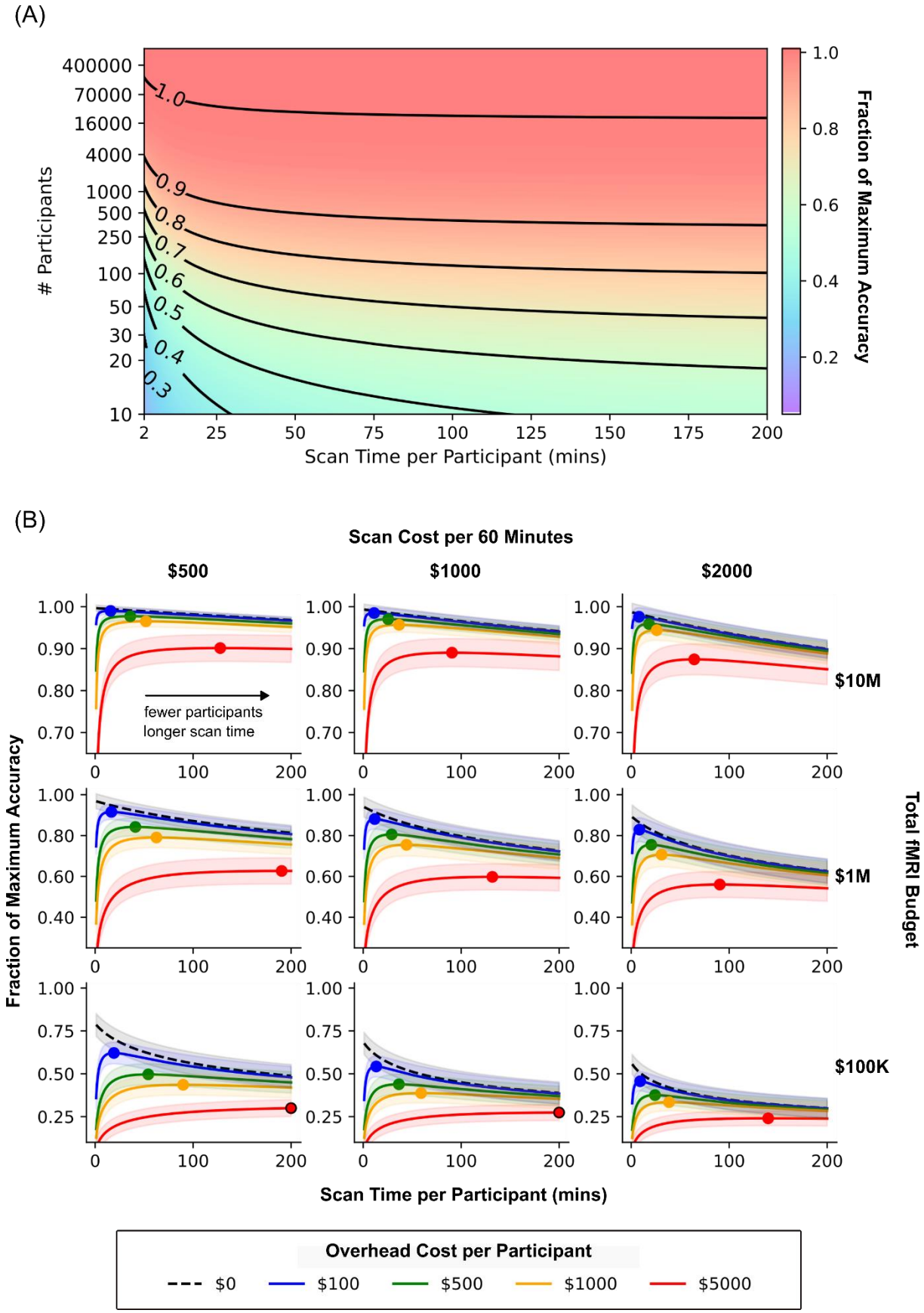


Figure 6. Empirical reference for balancing sample size and scan duration while accounting for fixed costs per participant to optimally design BWAS. (A) Fraction of maximum achievable prediction accuracy as a function of sample size and scan time per participant. The theoretical model was fitted to 36 HCP and ABCD phenotypes, yielding 89% average explained variance (Table S1B). For each phenotype, the model was normalized by its maximum achievable accuracy (based on the theoretical model), yielding a fraction of maximum achievable prediction accuracy for every combination of sample size and scan time per participant. The fraction of maximum achievable prediction accuracy was then averaged across the 36 phenotypes yielding the plot. (B) Fraction of maximum achieve prediction accuracy as a function of total fMRI budget, scan cost per hour and overhead cost per participant. The solid circles indicate location of the maximum prediction accuracy. Black contour around a circle indicates that the optimal combination of sample size and scan time was beyond the edge of the graph (i.e., more than 200 minutes of scan time). As an example, when total fMRI budget was $10M, scan cost was $500 per hour and overhead cost was $500 for each participant, the optimal prediction accuracy was achieved by scanning 12,500 participants for 36 minutes per participant. Overall, this suggests that many existing studies, including very large-scale studies, might have benefitted from increasing scan time per participant.

We note that these results (Figure 6) do not account for second-order effects. For example, certain populations (e.g., children) might not be able to handle more than 1 hour of MRI scanning at a time, so longer scans would need to be broken up into multiple sessions, yielding an overhead cost associated with each session, and so on. As another example, beyond a certain sample size, multi-site data collection becomes necessary, which increases overhead cost per participant. Our web application (WEB_APPLICATION_LINK) allows for more flexible usage.

Conclusions are similar between prediction accuracy & BWAS reliability

We next turn our attention to the effects of sample size and scan duration per participant on the reliability of BWAS (Marek et al., 2022) using a previously established split-half procedure (Figure S18A; Tian & Zalesky, 2021; Chen et al., 2023). Similar conclusions were obtained for both univariate and multivariate BWAS reliability, except that diminishing returns of scan time occurred beyond 10 minutes per participant, instead of 20-30 minutes of scan time for prediction accuracy (Figures S18 to S32).

However, we strongly recommend that prediction accuracy, instead of reliability should be prioritized during study design. The reason is that reliability does not imply validity (Schmidt et al., 2000; Noble et al., 2019). For example, hardware artifacts may appear reliably in measurements without having any biological relevance. In the case of resting-state fMRI, reliable BWAS features might not be actually predictive of individual-level phenotypes.

Discussion

Neuroimaging studies are always confronted with the difficult decision of how to allocate fixed resources for an optimal study design. Here, we systematically investigate the trade-off between maximising scan duration and sample size in the context of predicting phenotypes from resting state fMRI data. We found that sample size and scan time per participant are broadly interchangeable. Prediction accuracy was explained remarkably well by a simple logarithmic model and a more complex theoretical model. The model fits were consistent across many phenotypes across multiple phenotypic domains and two different datasets, suggesting strong generalizability of these findings. When accounting for overhead cost per participant, we found that future study designs might benefit from longer scan durations per participant than those employed in existing studies.

Overall, our results suggest an advantage to flexibly modifying study designs based on population- and site-specific characteristics. For example, a researcher seeking to study unmedicated 9-10 year-old children with autism spectrum disorder who are able to stay still during MRI scans (i.e., higher overhead cost per participant) might find it more economical to increase the scan time for each participant in order to achieve the maximum possible prediction accuracy. Another researcher facing particularly high per-hour scan charges might choose to decrease scan times and increase sample sizes.

More broadly, our results strongly argue against the common practice of employing traditional power analyses, whose only inputs are sample size, to inform BWAS design. Because such power analyses inevitably point towards maximizing sample size, scan times then become minimized under budget constraints. The resulting prediction accuracies are likely lower than would be produced with alternate designs, thus impeding scientific discovery.

To more accurately enable flexible decision making under varying constraints and inform study planning, we have provided a web application (WEB_APPLICATION_LINK) that estimates the fraction of maximum prediction accuracy that can be achieved with different sample size, scan duration per participant and overhead cost per participant, together with additional factors. For example, certain demographic and patient populations might not be able to tolerate longer scans, so an additional factor will be the maximum scan duration in each MRI session. As another example, beyond a certain sample size, multi-site data collection becomes a necessity, resulting in higher overhead costs.

An important limitation is that the empirically informed reference is less useful for poorly predicted phenotypes, which predominantly included non-cognitive phenotypes (Figure 4) There are two non-exclusive reasons for poorly predicted phenotypes. One reason is that the measurement of the phenotype might not be reliable or valid (Uher, 2015; Nikolaidis et al., 2022; Gell et al., 2023), suggesting the need to improve the measurement of the phenotype. A second reason is that there may only be a weak relationship between the phenotype and resting-state fMRI, in which case, other imaging modalities might be worth exploring.

Another caveat is that the empirically informed reference is less useful for phenotypes whose prediction accuracies are highly influenced by inter-individual differences in brain states during resting-state fMRI (Figure 5), potentially due to level of arousal (Bijsterbosch et al., 2017). This

appeared to be a larger problem for the HCP dataset, which involved significantly longer scan time than the ABCD dataset and was obtained in two different scan sessions. By exploring the interactions between brain states and these phenotypes, future work could potentially develop better brain-based predictions of these phenotypes.

Furthermore, it is important to note that in addition to economic factors, the representativeness and diversity of the data sample and their generalizability to subpopulations are also important to consider (Benkarim et al., 2022; Greene et al., 2022; Li et al., 2022; Kopal et al., 2023). Finally, not all studies are interested in cross-sectional relationships between brain and non-brain-imaging phenotypes. For example, the use of individual-level networks for brain stimulation treatment of mental disorders (Cash et al., 2021; Lynch et al., 2022) or neurosurgical planning (Boutet et al., 2021), might require higher per-participant quantities of resting-state fMRI data for accurate individual-level network estimation (Laumann et al., 2015; Braga & Buckner, 2017; Gordon et al., 2017).

Conclusion

We find that sample size and scan time per participant are broadly interchangeable for brain-wide association studies (BWAS), although there are eventually diminishing returns of scan time per participant with respect to sample sizes. When accounting for fixed overhead costs per participant, we find that most studies (including large-scale studies) might benefit from greater scan time per participant than previously assumed. Our findings establish an empirically informed reference for calibrating scan times and sample sizes to optimize the study of how inter-individual variation in brain network architecture is related with individual differences in behavior.

References

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage*, *145*(Pt B), 137-165. https://doi.org/10.1016/j.neuroimage.2016.02.079

Benkarim, O., Paquola, C., Park, B.-y., Kebets, V., Hong, S.-J., Vos de Wael, R., Zhang, S., Yeo, B. T. T., Eickenberg, M., Ge, T., Poline, J.-B., Bernhardt, B. C., & Bzdok, D. (2022). Population heterogeneity in clinical cohorts affects the predictive accuracy of brain imaging. *PLOS Biology*, *20*(4), e3001627. https://doi.org/10.1371/journal.pbio.3001627

Bijsterbosch, J., Harrison, S., Duff, E., Alfaro-Almagro, F., Woolrich, M., & Smith, S. (2017). Investigations into within- and between-subject resting-state amplitude variations. *Neuroimage*, *159*, 57-69. https://doi.org/10.1016/j.neuroimage.2017.07.014

Boutet, A., Madhavan, R., Elias, G. J. B., Joel, S. E., Gramer, R., Ranjan, M., Paramanandam, V., Xu, D., Germann, J., Loh, A., Kalia, S. K., Hodaie, M., Li, B., Prasad, S., Coblentz, A., Munhoz, R. P., Ashe, J., Kucharczyk, W., Fasano, A., & Lozano, A. M. (2021). Predicting optimal deep brain stimulation parameters for Parkinson's disease using functional MRI and machine learning. *Nature Communications*, *12*(1), 3043. https://doi.org/10.1038/s41467-021-23311-9

Braga, R. M., & Buckner, R. L. (2017). Parallel Interdigitated Distributed Networks within the Individual Estimated by Intrinsic Functional Connectivity. *Neuron*, *95*(2), 457-471.e455. https://doi.org/10.1016/j.neuron.2017.06.038

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376. https://doi.org/10.1038/nrn3475

Bzdok, D., & Ioannidis, J. P. A. (2019). Exploration, Inference, and Prediction in Neuroscience and Biomedicine. *Trends Neurosci*, *42*(4), 251-262. https://doi.org/10.1016/j.tins.2019.02.001

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 223-230. https://doi.org/10.1016/j.bpsc.2017.11.007

Cash, R. F. H., Weigand, A., Zalesky, A., Siddiqi, S. H., Downar, J., Fitzgerald, P. B., & Fox, M. D. (2021). Using Brain Imaging to Improve Spatial Targeting of Transcranial Magnetic Stimulation for Depression. *Biol Psychiatry*, *90*(10), 689-700. https://doi.org/10.1016/j.biopsych.2020.05.033

Chen, J., Ooi, L. Q. R., Tan, T. W. K., Zhang, S., Li, J., Asplund, C. L., Eickhoff, S. B., Bzdok, D., Holmes, A. J., & Yeo, B. T. T. (2023). Relationship between prediction accuracy and feature importance reliability: An empirical and theoretical study. *Neuroimage*, *274*, 120115. https://doi.org/https://doi.org/10.1016/j.neuroimage.2023.120115

Eickhoff, S. B., & Langner, R. (2019). Neuroimaging-based prediction of mental traits: Road to utopia or Orwell? *PLoS Biol*, *17*(11), e3000497. https://doi.org/10.1371/journal.pbio.3000497

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychol Sci*, *31*(7), 792-806. https://doi.org/10.1177/0956797620916786

Feng, P., Jiang, R., Wei, L., Calhoun, V. D., Jing, B., Li, H., & Sui, J. (2023). Determining four confounding factors in individual cognitive traits prediction with functional connectivity: an exploratory study. *Cerebral Cortex*, *33*(5), 2011-2020. https://doi.org/10.1093/cercor/bhac189

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., & Constable, R. T. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, *18*(11), 1664-1671. https://doi.org/10.1038/nn.4135

Gabrieli, J. D. E., Ghosh, S. S., & Whitfield-Gabrieli, S. (2015). Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron*, *85*(1), 11-26. https://doi.org/10.1016/j.neuron.2014.10.047

Gell, M., Eickhoff, S. B., Omidvarnia, A., Küppers, V., Patil, K. R., Satterthwaite, T. D., Müller, V. I., & Langner, R. (2023). The Burden of Reliability: How Measurement Noise Limits Brain-Behaviour Predictions. *bioRxiv*, 2023.2002.2009.527898. https://doi.org/10.1101/2023.02.09.527898

Gordon, E. M., Chauvin, R. J., Van, A. N., Rajesh, A., Nielsen, A., Newbold, D. J., Lynch, C. J., Seider, N. A., Krimmel, S. R., Scheidter, K. M., Monk, J., Miller, R. L., Metoki, A., Montez, D. F., Zheng, A., Elbau, I., Madison, T., Nishino, T., Myers, M. J., Kaplan, S., Badke D'Andrea, C., Demeter, D. V., Feigelis, M., Ramirez, J. S. B., Xu, T., Barch, D. M., Smyser, C. D., Rogers, C. E., Zimmermann, J., Botteron, K. N., Pruett, J. R., Willie, J. T., Brunner, P., Shimony, J. S., Kay, B. P., Marek, S., Norris, S. A., Gratton, C., Sylvester, C. M., Power, J. D., Liston, C., Greene, D. J., Roland, J. L., Petersen, S. E., Raichle, M. E., Laumann, T. O., Fair, D. A., & Dosenbach, N. U. F. (2023). A somato-cognitive action network alternates with effector regions in motor cortex. *Nature*, *617*(7960), 351-359. https://doi.org/10.1038/s41586-023-05964-2

Gordon, E. M., Laumann, T. O., Gilmore, A. W., Newbold, D. J., Greene, D. J., Berg, J. J., Ortega, M., Hoyt-Drazen, C., Gratton, C., Sun, H., Hampton, J. M., Coalson, R. S., Nguyen, A. L., McDermott, K. B., Shimony, J. S., Snyder, A. Z., Schlaggar, B. L., Petersen, S. E., Nelson, S. M., & Dosenbach, N. U. F. (2017). Precision Functional Mapping of Individual Human Brains. *Neuron*, *95*(4), 791-807.e797. https://doi.org/10.1016/j.neuron.2017.07.011

Greene, A. S., Shen, X., Noble, S., Horien, C., Hahn, C. A., Arora, J., Tokoglu, F., Spann, M. N., Carrión, C. I., Barron, D. S., Sanacora, G., Srihari, V. H., Woods, S. W., Scheinost, D., & Constable, R. T. (2022). Brain–phenotype models fail for individuals who defy sample stereotypes. *Nature*, *609*(7925), 109-118. https://doi.org/10.1038/s41586-022-05118-w

He, T., Kong, R., Holmes, A. J., Nguyen, M., Sabuncu, M. R., Eickhoff, S. B., Bzdok, D., Feng, J., & Yeo, B. T. T. (2020). Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *Neuroimage*, *206*, 116276. https://doi.org/https://doi.org/10.1016/j.neuroimage.2019.116276

Kharabian Masouleh, S., Eickhoff, S. B., Hoffstaedter, F., Genon, S., & Alzheimer's Disease Neuroimaging, I. (2019). Empirical examination of the replicability of associations between brain structure and psychological variables. *eLife*, *8*, e43464. https://doi.org/10.7554/eLife.43464

Kopal, J., Uddin, L. Q., & Bzdok, D. (2023). The end game: respecting major sources of population diversity. *Nature Methods*. https://doi.org/10.1038/s41592-023-01812-3

Laumann, Timothy O., Gordon, Evan M., Adeyemo, B., Snyder, Abraham Z., Joo, Sung J., Chen, M.-Y., Gilmore, Adrian W., McDermott, Kathleen B., Nelson, Steven M., Dosenbach, Nico U. F., Schlaggar, Bradley L., Mumford, Jeanette A., Poldrack, Russell A., & Petersen, Steven E. (2015). Functional System and Areal Organization of a Highly Sampled Individual Human Brain. *Neuron*, *87*(3), 657-670. https://doi.org/https://doi.org/10.1016/j.neuron.2015.06.037

Li, J., Bzdok, D., Chen, J., Tam, A., Ooi, L. Q. R., Holmes, A. J., Ge, T., Patil, K. R., Jabbi, M., Eickhoff, S. B., Yeo, B. T. T., & Genon, S. (2022). Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci Adv*, *8*(11), eabj1812. https://doi.org/10.1126/sciadv.abj1812

Lynch, C. J., Elbau, I. G., Ng, T. H., Wolk, D., Zhu, S., Ayaz, A., Power, J. D., Zebley, B., Gunning, F. M., & Liston, C. (2022). Automated optimization of TMS coil placement for personalized functional network engagement. *Neuron*, *110*(20), 3263-3277.e3264. https://doi.org/10.1016/j.neuron.2022.08.012

Lynch, C. J., Power, J. D., Scult, M. A., Dubin, M., Gunning, F. M., & Liston, C. (2020). Rapid Precision Functional Mapping of Individuals Using Multi-Echo fMRI. *Cell Rep*, *33*(12), 108540. https://doi.org/10.1016/j.celrep.2020.108540

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., Moore, L. A., Conan, G. M., Uriarte, J., Snider, K., Lynch, B. J., Wilgenbusch, J. C., Pengo, T., Tam, A., Chen, J., Newbold, D. J., Zheng, A., Seider, N. A., Van, A. N., Metoki, A., Chauvin, R. J., Laumann, T. O., Greene, D. J., Petersen, S. E., Garavan, H., Thompson, W. K., Nichols, T. E., Yeo, B. T. T., Barch, D. M., Luna, B., Fair, D. A., & Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), 654-660. https://doi.org/10.1038/s41586-022-04492-9

Nee, D. E. (2019). fMRI replicability depends upon sufficient individual-level data. *Communications Biology*, *2*(1), 130. https://doi.org/10.1038/s42003-019-0378-6

Newbold, D. J., Laumann, T. O., Hoyt, C. R., Hampton, J. M., Montez, D. F., Raut, R. V., Ortega, M., Mitra, A., Nielsen, A. N., Miller, D. B., Adeyemo, B., Nguyen, A. L., Scheidter, K. M., Tanenbaum, A. B., Van, A. N., Marek, S., Schlaggar, B. L., Carter, A. R., Greene, D. J., Gordon, E. M., Raichle, M. E., Petersen, S. E., Snyder, A. Z., & Dosenbach, N. U. F. (2020). Plasticity and Spontaneous Activity Pulses in Disused Human Brain Circuits. *Neuron*, *107*(3), 580-589.e586. https://doi.org/10.1016/j.neuron.2020.05.007

Nikolaidis, A., Chen, A. A., He, X., Shinohara, R., Vogelstein, J., Milham, M., & Shou, H. (2022). Suboptimal phenotypic reliability impedes reproducible human neuroscience. *bioRxiv*, 2022.2007.2022.501193. https://doi.org/10.1101/2022.07.22.501193

Noble, S., Scheinost, D., & Constable, R. T. (2019). A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *Neuroimage*, *203*, 116157. https://doi.org/10.1016/j.neuroimage.2019.116157

Ooi, L. Q. R., Chen, J., Zhang, S., Kong, R., Tam, A., Li, J., Dhamala, E., Zhou, J. H., Holmes, A. J., & Yeo, B. T. T. (2022). Comparison of individualized behavioral predictions across anatomical, diffusion and functional connectivity MRI. *Neuroimage*, *263*, 119636. https://doi.org/https://doi.org/10.1016/j.neuroimage.2022.119636

Orban, C., Kong, R., Li, J., Chee, M. W. L., & Yeo, B. T. T. (2020). Time of day is associated with paradoxical reductions in global signal fluctuation and functional connectivity. *PLoS Biol*, *18*(2), e3000602. https://doi.org/10.1371/journal.pbio.3000602

Poldrack, R. A., Huckins, G., & Varoquaux, G. (2020). Establishment of Best Practices for Evidence for Prediction: A Review. *JAMA Psychiatry*, *77*(5), 534-540. https://doi.org/10.1001/jamapsychiatry.2019.3671

Poldrack, R. A., Laumann, T. O., Koyejo, O., Gregory, B., Hover, A., Chen, M.-Y., Gorgolewski, K. J., Luci, J., Joo, S. J., Boyd, R. L., Hunicke-Smith, S., Simpson, Z. B., Caven, T., Sochat, V., Shine, J. M., Gordon, E., Snyder, A. Z., Adeyemo, B., Petersen, S. E., Glahn, D. C., Reese Mckay, D., Curran, J. E., Göring, H. H. H., Carless, M. A., Blangero, J., Dougherty, R., Leemans, A., Handwerker, D. A., Frick, L., Marcotte, E. M., & Mumford, J. A. (2015). Long-term neural and physiological phenotyping of a single human. *Nature Communications*, *6*(1), 8885. https://doi.org/10.1038/ncomms9885

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex*, *28*(9), 3095-3114. https://doi.org/10.1093/cercor/bhx179

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology*, *53*(4), 901-912. https://doi.org/10.1111/j.1744-6570.2000.tb02422.x

Schulz, M.-A., Yeo, B. T. T., Vogelstein, J. T., Mourao-Miranada, J., Kather, J. N., Kording, K., Richards, B., & Bzdok, D. (2020). Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nature Communications*, *11*(1), 4238. https://doi.org/10.1038/s41467-020-18037-z

Schulz, M. A., Bzdok, D., Haufe, S., Haynes, J. D., & Ritter, K. (2023). Performance reserves in brain-imaging-based phenotype prediction. *Cell Rep*, *43*(1), 113597. https://doi.org/10.1016/j.celrep.2023.113597

Tagliazucchi, E., & Laufs, H. (2014). Decoding Wakefulness Levels from Typical fMRI Resting-State Data Reveals Reliable Drifts between Wakefulness and Sleep. *Neuron*, *82*(3), 695-708. https://doi.org/https://doi.org/10.1016/j.neuron.2014.03.020

Tian, Y., & Zalesky, A. (2021). Machine learning prediction of cognition from functional connectivity: Are feature weights reliable? *Neuroimage*, *245*, 118648. https://doi.org/https://doi.org/10.1016/j.neuroimage.2021.118648

Uher, J. (2015). Developing "Personality" Taxonomies: Metatheoretical and Methodological Rationales Underlying Selection Approaches, Methods of Data Generation and Reduction Principles. *Integrative Psychological and Behavioral Science*, *49*(4), 531-589. https://doi.org/10.1007/s12124-014-9280-4

Varoquaux, G., & Poldrack, R. A. (2019). Predictive models avoid excessive reductionism in cognitive neuroimaging. *Current Opinion in Neurobiology*, *55*, 1-6. https://doi.org/https://doi.org/10.1016/j.conb.2018.11.002

Wang, C., Ong, J. L., Patanaik, A., Zhou, J., & Chee, M. W. L. (2016). Spontaneous eyelid closures link vigilance fluctuation with fMRI dynamic connectivity states. *Proceedings of the National Academy of Sciences*, *113*(34), 9653-9658. https://doi.org/10.1073/pnas.1523980113

Woo, C.-W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: brain models in translational neuroimaging. *Nature Neuroscience*, *20*(3), 365-377. https://doi.org/10.1038/nn.4478

Yeo, B. T., Tandi, J., & Chee, M. W. (2015). Functional connectivity during rested wakefulness predicts vulnerability to sleep deprivation. *Neuroimage*, *111*, 147-158. https://doi.org/10.1016/j.neuroimage.2015.02.018

Methods

Datasets, phenotypes and participants

Following previous studies, we considered 58 HCP phenotypes (Kong et al., 2019; Li et al., 2019) and 36 ABCD phenotypes (Chen et al., 2022; Chen et al., 2023). We additionally consider a cognition factor score derived from all phenotypes from each dataset (Ooi et al., 2022), yielding a total of 59 HCP and 37 ABCD phenotypes (see Tables S3 and S4).

In this study, we used participants from the HCP WU-Minn S1200 release. We filtered participants from Li's set of 953 participants (Li et al., 2019), excluding participants who did not have at least 40 minutes of uncensored data (censoring criteria are discussed under "Image Processing") and did not have the full set of the 59 non-brain-imaging phenotypes (henceforth referred to as phenotypes) that we investigated. This resulted in a final set of 792 participants with demographics found in Table S5.

We additionally considered participants from the ABCD 2.0.1 release. We filtered participants from Chen's set of 5260 participants (Chen et al., 2023). We excluded participants who did not have at least 15 minutes of uncensored resting-fMRI data (censoring criteria are discussed under "Image Processing") and did not have the full set of the 37 phenotypes that we investigated. This resulted in a final set of 2565 participants with demographics found in Table S5.

Image processing

For the HCP dataset, the MSMAll ICA-FIX resting state scans were used (Glasser et al., 2013). Global signal regression has been shown to improve behavioral prediction (Li et al., 2019), so we further applied global signal regression (GSR) and censoring, consistent with our previous studies (Li et al., 2019; He et al., 2020; Kong et al., 2021). The censoring process entailed flagging frames with either FD > 0.2mm or DVARS > 75. The frame immediately before and after flagged frames were marked as censored. Additionally, uncensored segments of data consisting of less than 5 frames were also censored during downstream processing.

For the ABCD dataset, the minimally processed resting state scans were utilized (Hagler et al., 2019). Processing of functional data was performed in line with our previous study (Chen et al., 2022). Specifically, we additionally processed the minimally processed data with the following steps. (1) The functional images were aligned to the T1 images using boundary-based registration (Greve & Fischl, 2009). (2) Respiratory pseudomotion motion filtering was performed by applying a bandstop filter of 0.31-0.43Hz (Fair et al., 2020). (3) Frames with FD > 0.3mm or DVARS > 50 were flagged. The flagged frame, as well as the frame immediately before and two frames immediately after the marked frame were censored. Additionally, uncensored segments of data consisting of less than 5 frames were also censored. (4) Global, white matter and ventricular signals, 6 motion parameters, and their temporal derivatives were regressed from the functional data. Regression coefficients were estimated from uncensored data. (5) Censored frames were interpolated with the Lomb-Scargle periodogram (Power et al., 2014). (6) The data underwent bandpass filtering (0.009Hz – 0.08Hz). (7) Lastly, the data was then projected onto FreeSurfer fsaverage6 surface space and smoothed using a 6 mm full-width half maximum kernel.

We derived a 419 × 419 RSFC matrix for each HCP and ABCD participant using the first *T* minutes of scan time. The 419 regions consisted of 400 parcels from the Schaefer parcellation (Schaefer et al., 2018), and 19 subcortical regions of interest (Fischl et al., 2002). *T* was varied from 2 to the maximum scan time in intervals of 2 minutes. This resulted in 29 RSFC matrices per participant in the HCP dataset (as there were 29 RSFC matrices generated from using the minimum amount of 2 minutes to the maximum amount of 58 minutes in intervals of 2 minutes), and 10 RSFC matrices per participant in the ABCD dataset (as there are 10 RSFCs generated from using the minimum amount of 2 minutes to the maximum amount of 20 minutes in intervals of 2 minutes).

Prediction workflow

The RSFC generated from the first *T* minutes were used to predict each phenotypic measures (previous section) using kernel ridge regression (KRR; He et al., 2020) within an inner-loop (nested) cross-validation procedure.

Let us illustrate the procedure using the HCP dataset (Figure 1A). We began with the full set of participants. A 10-fold nested cross-validation procedure was used. Participants were divided in 10 folds (first row of Figure 1A). We note that care was taken so siblings were not split across folds, so the 10 folds were not exactly the same sizes. For each of 10 iterations, one fold was reserved for testing (i.e., test set), while the remainder was used for training (i.e., training set). Since there were 792 HCP participants, the training set size was roughly $792 \times 0.9 \approx 700$ participants. The KRR hyperparameter was selected via a cross-validation of the training set. The best hyperparameter was then used to train a final KRR model in the training set and applied to the test set. Prediction accuracy was measured using Pearson's correlation and coefficient of determination (Chen et al., 2022).

The above analysis was repeated with different training set sizes achieved by subsampling each training fold (second and third rows of Figure 1A), while the test set remained identical across different training set sizes, so the results are comparable across different training set sizes. The training set size was subsampled from 200 to 600 (in intervals of 100). Together with the full training set size of approximately 700 participants, there were 6 different training set sizes, corresponding to 200, 300, 400, 500, 600 and 700.

The whole procedure was repeated with different values of *T*. Since there were 29 values of *T*, there were in total 29 × 6 sets of prediction accuracies for each phenotypic measure. To ensure robustness, the above procedure was repeated 50 times with different splits of the participants into 10 folds to ensure stability (Figure 1A). The prediction accuracies were averaged across all test folds and all 50 repetitions.

The procedure for the ABCD dataset followed the same principle as the HCP dataset. However, the ABCD dataset comprised participants from multiple sites. Therefore, following our previous studies (Chen et al., 2022; Ooi et al., 2022), we combined participants across the 22 imaging sites, yielding 10 site-clusters (Table S6). Each site-cluster had a minimum of 227 participants. Instead of the 10-fold inner-loop (nested) cross-validation procedure in the HCP dataset, we performed a leave-3-site-clusters-out inner-loop (nested) cross-validation (i.e., 7 site-clusters are used for training and 3 site-clusters are used for testing) in the ABCD dataset. This procedure

was performed for every possible split of the site clusters, resulting in 120 replications. The prediction accuracies were averaged across all 120 replications.

Similar to the HCP, the analyses were repeated with different numbers of training participants, ranging from 200 to 1600 ABCD participants (in intervals of 200). Together with the full training set size of approximately 1800 participants, there were 9 different training set sizes. The whole procedure was repeated with different values of *T*. Since there were 10 values of *T* in the ABCD dataset, there were in total 10 × 9 values of prediction accuracies for each phenotype.

A full table of prediction accuracies for every combination of sample size and scan time per participant can be found in the supplementary spreadsheet.

Logarithmic fit of prediction accuracy with respect to total scan duration

By plotting total scan time (number of training participants × scan duration per participant) against prediction accuracy for each phenotypic measure, we observed that for most measures, scanning beyond 20-30 minutes per participant did not improve prediction accuracy.

Furthermore, visual inspection suggests that a logarithmic curve might fit well to each phenotypic measure when scan time per participant is 30 minutes or less. To explore the universality of a logarithmic relationship between total scan time and prediction accuracy, for each phenotypic measure p, we fitted the function $y_p = z_p log(t_p) + k_p$, where y_p was the prediction accuracy for phenotypic measure p, and t_p is the total scan time. z_p and k_p were estimated from data by minimizing the square error, yielding $\hat{z}_p$ and $\hat{k}_p$.

In addition to fitting the logarithmic curve to different phenotypic measures, the fitting can also be performed with different prediction accuracy measures (Pearson's correlation or coefficient of determination) and different predictive models (kernel ridge regression and linear ridge regression). Assuming the datapoints are well explained by the logarithmic curve, the normalized accuracies $(y_b - \hat{k}_b)/\hat{z}_b$ should follow a standard (t) curve across phenotypic measures, prediction accuracies, predictive models, and datasets. As an example, Figure 2B shows the normalized prediction performance of the cognitive factors for different prediction accuracy measures (Pearson's correlation or coefficient of determination) and different predictive models (kernel ridge regression and linear ridge regression) across HCP and ABCD datasets.

Fit of theoretically-motivated model of prediction accuracy, sample size and scan time

We observed that sample size and scan time per participant did not contribute equally to prediction accuracy, with sample size playing a slightly more important role than scan time. To explain this observation, we derived a mathematical relationship relating the expected Pearson's correlation between noisy brain measurements and non-brain-imaging phenotype with scan time and sample size.

Based on a linear regression model with no regularization and assumptions including (1) stationarity of fMRI (i.e., autocorrelation in fMRI is the same at all timepoints), and (2) prediction errors are uncorrelated with errors in brain measurements, we found that

$$E(\hat{\rho}) \approx K_0 \sqrt{\frac{1}{1 + \frac{K_1}{N} + \frac{K_2}{NT}}},$$

where $E(\hat{\rho})$ is the expected correlation between regression weights estimated from noisy brain measurements and the observed phenotype. K_0 is related to the ideal association between brain measurements and phenotypes, attenuated by phenotype reliability. K_1 is related to the true association between brain measurements and phenotype. K_2 is related to brain-phenotype prediction errors due to brain measurement inaccuracies. Full derivations can be found in Supplementary Methods Sections 1.1 and 1.2.

Based on the above equation, we fitted the following function $y_p = K_{0,p}\sqrt{\frac{1}{1+K_{1,p}/N+K_{2,p}/(NT)}}$, where y_p was the prediction accuracy for phenotypic measure p, N was the sample size and T was the scan time per participant. $K_{0,p}, K_{1,p}$ and $K_{2,p}$ were estimated by minimizing the mean squared error between the above functional form and actual observation of y_p using gradient descent.

Analysis of state effects

In the original analysis, FC matrices were generated with increasing time *T* based on the original run order. To account for the possibility of state effects, we randomized the order in which the runs were considered for each participant. Since both HCP and ABCD datasets contained 4 runs of resting-fMRI, we generated FC matrices from all 24 possible permutations of run order. For each cross-validation split, the FC matrix for a given participant was randomly sampled from one of the 24 possible permutations. We note that the randomization was independently performed for each participant.

Brain-wide association reliability workflow

To explore the reliability of univariate brain-wide association analyses (BWAS; Marek et al., 2022), we followed a previously established split-half procedure (Tian & Zalesky, 2021; Chen et al., 2023).

Let us illustrate the procedure using the HCP dataset (Figure S18A). We began with the full set of participants, which were then divided into 10 folds (first row of Figure S18A). We note that care was taken so siblings were not split across folds, so the 10 folds were not exactly the same sizes. The 10 folds were divided into two non-overlapping sets of 5 folds. For each set of 5 folds and each phenotype, we computed Pearson's correlation between each RSFC edge and phenotype across participants, yielding a 419 × 419 correlation matrix, which was then converted into a 419 × 419 t-statistic matrix. Split-half reliability between the (lower triangular portions of the symmetric) t-statistic matrices from the two sets of 5 folds was then computed using the intra-class correlation formula (Tian & Zalesky, 2021; Chen et al., 2023).

The above analysis was repeated with different sample sizes achieved by subsampling each fold (second and third rows of Figure S18A). The split-half sample sizes were subsampled from 150 to 350 (in intervals of 50). Together with the full sample size of approximately 800 participants

(corresponding to a split-half sample size of around 400), there were 6 split-half sample sizes corresponding to 150, 200, 250, 300, 350 and 400 participants.

The whole procedure was also repeated with different values of *T*. Since there were 29 values of *T*, there were in total 29 × 6 univariate BWAS split-half reliability values for each phenotype. To ensure robustness, the above procedure was repeated 50 times with different split of the participants into 10 folds to ensure stability (Figure 18A). The reliability values were averaged across all 50 repetitions.

The same procedure was followed in the case of the ABCD dataset, except as previously explained, the ABCD participants were divided into 10 site-clusters. Therefore, the split-half reliability was performed between two sets of 5 non-overlapping site-clusters. In total, this procedure was repeated 126 times since there were 126 ways to divide 10 site-clusters into two sets of 5 non-overlapping site-clusters.

Similar to the HCP, the analyses were repeated with different numbers of split-half participants, ranging from 200 to 1000 ABCD participants (in intervals of 200). Together with the full training set size of approximately 2400 participants (corresponding to a split-half sample size of approximately 1200 participants, there were 6 split-half sample sizes, corresponding to 200, 400, 600, 800, 1000, 1200.

The whole procedure was also repeated with different values of *T*. Since there were 10 values of *T* in the ABCD dataset, there were in total 10 × 6 values univariate BWAS split-half reliability values for each phenotype.

Previous studies have suggested the Haufe-transformed coefficients from multivariate prediction are significantly more reliable than univariate BWAS (Tian & Zalesky, 2021; Chen et al., 2023). Therefore, we repeated the above analyses by replacing BWAS with the multivariate Haufe-transform.

A full table of split-half BWAS reliability for each given combination of sample size and scan time per participant can be found in the supplementary spreadsheet.

Data and Code Availability

Both HCP (https://www.humanconnectome.org/) and ABCD (https://abcdstudy.org/) datasets are publicly available. Data used in the ABCD portion of the analysis is available on the NIMH Data Archive (NDA) website (LINK_TO_BE_UPDATED).

Code for this study is publicly available in the GitHub repository maintained by the Computational Brain Imaging Group (https://github.com/ThomasYeoLab/CBIG). Processing pipelines of the fMRI data can be found here (https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/preprocessing/CBIG_fMRI_Preproc2016).

Code specific to the analyses in this study can be found here (LINK_TO_BE_UPDATED). Code related to this study was reviewed by co-author TWKT to reduce the chance of coding errors.

References (Methods)

Chen, J., Ooi, L. Q. R., Tan, T. W. K., Zhang, S., Li, J., Asplund, C. L., Eickhoff, S. B., Bzdok, D., Holmes, A. J., & Yeo, B. T. T. (2023). Relationship between prediction accuracy and feature importance reliability: An empirical and theoretical study. *Neuroimage*, *274*, 120115. https://doi.org/https://doi.org/10.1016/j.neuroimage.2023.120115

Chen, J., Tam, A., Kebets, V., Orban, C., Ooi, L. Q. R., Asplund, C. L., Marek, S., Dosenbach, N. U. F., Eickhoff, S. B., Bzdok, D., Holmes, A. J., & Yeo, B. T. T. (2022). Shared and unique brain network features predict cognitive, personality, and mental health scores in the ABCD study. *Nature Communications*, *13*(1), 2217. https://doi.org/10.1038/s41467-022-29766-8

Fair, D. A., Miranda-Dominguez, O., Snyder, A. Z., Perrone, A., Earl, E. A., Van, A. N., Koller, J. M., Feczko, E., Tisdall, M. D., van der Kouwe, A., Klein, R. L., Mirro, A. E., Hampton, J. M., Adeyemo, B., Laumann, T. O., Gratton, C., Greene, D. J., Schlaggar, B. L., Hagler, D. J., Watts, R., Garavan, H., Barch, D. M., Nigg, J. T., Petersen, S. E., Dale, A. M., Feldstein-Ewing, S. W., Nagel, B. J., & Dosenbach, N. U. F. (2020). Correction of respiratory artifacts in MRI head motion estimates. *Neuroimage*, *208*, 116400. https://doi.org/https://doi.org/10.1016/j.neuroimage.2019.116400

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole Brain Segmentation. *Neuron*, *33*(3), 341-355. https://doi.org/10.1016/s0896-6273(02)00569-x

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., & Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage*, *80*, 105-124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, *48*(1), 63-72. https://doi.org/https://doi.org/10.1016/j.neuroimage.2009.06.060

Hagler, D. J., Hatton, S., Cornejo, M. D., Makowski, C., Fair, D. A., Dick, A. S., Sutherland, M. T., Casey, B. J., Barch, D. M., Harms, M. P., Watts, R., Bjork, J. M., Garavan, H. P., Hilmer, L., Pung, C. J., Sicat, C. S., Kuperman, J., Bartsch, H., Xue, F., Heitzeg, M. M., Laird, A. R., Trinh, T. T., Gonzalez, R., Tapert, S. F., Riedel, M. C., Squeglia, L. M., Hyde, L. W., Rosenberg, M. D., Earl, E. A., Howlett, K. D., Baker, F. C., Soules, M., Diaz, J., De Leon, O. R., Thompson, W. K., Neale, M. C., Herting, M., Sowell, E. R., Alvarez, R. P., Hawes, S. W., Sanchez, M., Bodurka, J., Breslin, F. J., Morris, A. S., Paulus, M. P., Simmons, W. K., Polimeni, J. R., Van Der Kouwe, A., Nencka, A. S., Gray, K. M., Pierpaoli, C., Matochik, J. A., Noronha, A., Aklin, W. M., Conway, K., Glantz, M., Hoffman, E., Little, R., Lopez, M., Pariyadath, V., Weiss, S. R., Wolff-Hughes, D. L., Delcarmen-Wiggins, R., Feldstein Ewing, S. W., Miranda-Dominguez, O., Nagel, B. J., Perrone, A. J., Sturgeon, D. T., Goldstone, A., Pfefferbaum, A., Pohl, K. M., Prouty, D., Uban, K., Bookheimer, S. Y., Dapretto, M., Galvan, A., Bagot, K., Giedd, J., Infante, M. A., Jacobus, J., Patrick, K., Shilling, P. D., Desikan, R., Li, Y., Sugrue, L., Banich, M. T., Friedman, N., Hewitt, J. K., Hopfer, C., Sakai, J., Tanabe, J., Cottler, L. B., Nixon, S. J., Chang, L., Cloak, C., Ernst, T., Reeves, G., Kennedy, D. N., Heeringa, S., Peltier, S., Schulenberg, J., Sripada, C., Zucker, R. A., Iacono, W. G., Luciana, M., Calabro, F. J., Clark, D. B., Lewis, D. A., Luna, B., Schirda, C., Brima, T., Foxe, J. J.,

Freedman, E. G., Mruzek, D. W., Mason, M. J., Huber, R., McGlade, E., Prescot, A., Renshaw, P. F., Yurgelun-Todd, D. A., Allgaier, N. A., Dumas, J. A., Ivanova, M., Potter, A., Florsheim, P., Larson, C., Lisdahl, K., Charness, M. E., Fuemmeler, B., Hettema, J. M., Maes, H. H., Steinberg, J., Anokhin, A. P., Glaser, P., Heath, A. C., Madden, P. A., Baskin-Sommers, A., Constable, R. T., Grant, S. J., Dowling, G. J., Brown, S. A., Jernigan, T. L., & Dale, A. M. (2019). Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. *Neuroimage*, *202*, 116091. https://doi.org/10.1016/j.neuroimage.2019.116091

He, T., Kong, R., Holmes, A. J., Nguyen, M., Sabuncu, M. R., Eickhoff, S. B., Bzdok, D., Feng, J., & Yeo, B. T. T. (2020). Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *Neuroimage*, *206*, 116276. https://doi.org/https://doi.org/10.1016/j.neuroimage.2019.116276

Kong, R., Li, J., Orban, C., Sabuncu, M. R., Liu, H., Schaefer, A., Sun, N., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2019). Spatial Topography of Individual-Specific Cortical Networks Predicts Human Cognition, Personality, and Emotion. *Cerebral Cortex*, *29*(6), 2533-2551. https://doi.org/10.1093/cercor/bhy123

Kong, R., Yang, Q., Gordon, E., Xue, A., Yan, X., Orban, C., Zuo, X.-N., Spreng, N., Ge, T., Holmes, A., Eickhoff, S., & Yeo, B. T. T. (2021). Individual-Specific Areal-Level Parcellations Improve Functional Connectivity Prediction of Behavior. *Cerebral Cortex*, *31*(10), 4477-4500. https://doi.org/10.1093/cercor/bhab101

Li, J., Kong, R., Liégeois, R., Orban, C., Tan, Y., Sun, N., Holmes, A. J., Sabuncu, M. R., Ge, T., & Yeo, B. T. T. (2019). Global signal regression strengthens association between resting-state functional connectivity and behavior. *Neuroimage*, *196*, 126-141. https://doi.org/https://doi.org/10.1016/j.neuroimage.2019.04.016

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., Moore, L. A., Conan, G. M., Uriarte, J., Snider, K., Lynch, B. J., Wilgenbusch, J. C., Pengo, T., Tam, A., Chen, J., Newbold, D. J., Zheng, A., Seider, N. A., Van, A. N., Metoki, A., Chauvin, R. J., Laumann, T. O., Greene, D. J., Petersen, S. E., Garavan, H., Thompson, W. K., Nichols, T. E., Yeo, B. T. T., Barch, D. M., Luna, B., Fair, D. A., & Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), 654-660. https://doi.org/10.1038/s41586-022-04492-9

Ooi, L. Q. R., Chen, J., Zhang, S., Kong, R., Tam, A., Li, J., Dhamala, E., Zhou, J. H., Holmes, A. J., & Yeo, B. T. T. (2022). Comparison of individualized behavioral predictions across anatomical, diffusion and functional connectivity MRI. *Neuroimage*, *263*, 119636. https://doi.org/https://doi.org/10.1016/j.neuroimage.2022.119636

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage*, *84*, 320-341. https://doi.org/10.1016/j.neuroimage.2013.08.048

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex*, *28*(9), 3095-3114. https://doi.org/10.1093/cercor/bhx179

Tian, Y., & Zalesky, A. (2021). Machine learning prediction of cognition from functional connectivity: Are feature weights reliable? *Neuroimage*, *245*, 118648. https://doi.org/https://doi.org/10.1016/j.neuroimage.2021.118648

Acknowledgements

Our research is supported by the Singapore National Research Foundation (NRF) Fellowship (Class of 2017), the NUS Yong Loo Lin School of Medicine (NUHSRO/2020/124/TMR/LOA), the Singapore National Medical Research Council (NMRC) LCG (OFLCG19May-0035), NMRC STaR (STaR20nov-0003), Singapore Ministry of Health (MOH) Centre Grant (CG21APR1009) and the USA NIH (R01MH120080, R01MH123245). Our computational work was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Singapore NRF, NMRC or MOH.

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

Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive DevelopmentSM (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 9-10 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, U24DA041123, 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 ABCD data repository grows and changes over time. The ABCD data used in this report came from http://dx.doi.org/10.15154/1504041.

Author Contributions

LQRO, CO, RK and BTTY conceptualized the study and designed the methodology. LQRO carried out the analysis. TEN derived the theoretical models in the study. TWKT and RK reviewed the code utilized in the study. LQRO, CO and BTTY wrote the original draft. All authors reviewed and edited the final manuscript.

Conflict of interest

DB is shareholder and advisory board member of MindState Design Labs, USA.