Multimodal-Neural Predictive Models of Children's General Intelligence That Are Stable Across Two Years of Development

Narun Pornpattananangkul[a], Yue Wang[a], Argyris Stringaris[b]

[a]Department of Psychology, University of Otago, New Zealand
[b]Section on Clinical and Computational Psychiatry, National Institute of Mental Health, USA

Corresponding author:
Narun Pornpattananangkul, PhD
Department of Psychology, University of Otago
William James Building
275 Leith Walk
Dunedin 9016, New Zealand
Email: narun.pat@otago.ac.nz

Acknowledgments:
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 9-10 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 U01DA041022, U01DA041028, U01DA041048, U01DA041089, U01DA041106, U01DA041117, U01DA041120, U01DA041134, U01DA041148, U01DA041156, U01DA041174, U24DA041123, U24DA041147, U01DA041093, and U01DA041025. 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/scientists/workgroups/. 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. We thank the developers of several R libraries, including semTools (Sunthud Pornprasertmanit), eNetXplorer (Julián Candia), and ggseg (Athanasia M. Mowinckel), for their technical advice.

Declaration of Interest: The authors declare no competing interests.

Abstract

Children's general intelligence, or G-Factor, is associated with important life outcomes, including educational attainment, employment, health, and mortality. Thus, determining the neural predictor of children's G-Factor has become one of the main tasks in neuroscience. Here we aim to build neuroimaging-based predictive models of children's general intelligence that are longitudinally stable across two years. To achieve this goal, we used large-scale, longitudinal data with multiple neuroimaging modalities from the Adolescent Brain Cognitive Development (ABCD) study (n ~11k). We first computed modality-specific models from six MRI modalities (three task-based, resting-state, structural, and diffusion tensor imaging) of the baseline data using Elastic Net. We, then, combined the predicted values of all modality-specific models using Random Forest Opportunistic Stacking. The stacked model allowed us to predict the G-Factor of unseen, same-age (9-10-year-old) children at Pearson's r=.44, better than any other modality-specific models. Based on permutation tests, this prediction was predominantly driven by activity in the parietal and frontal areas during the N-Back working-memory task. Importantly, this model was generalizable to unseen children from the follow-up data who were two years older at r=.41. Moreover, this model allowed for missingness in the data, making it possible for us to maintain around 73% of the data that would otherwise be excluded due to different artifacts occurring to any modalities. Accordingly, we developed an MRI-multimodal predictive model for children's G-Factor that is 1) stable across years, 2) interpretable and 3) able to handle missing values.

Keywords: general intelligence, multimodal neuroimaging, machine learning, children, longitudinal, large-scale data

Introduction

Children's general intelligence, known as the G-Factor[1], captures individual differences in the ability to perform across cognitive domains, such as language, mental flexibility, and memory. Children's G-factor is associated with many meaningful life outcomes, including academic achievement[1], job attainment and performance[2], health[3], and mortality[4]. Given these associations, neuroscientists have long been striving to identify the neural predictive models of children's G-factor using brain Magnetic Resonance Imaging (MRI), but so far with modest success[5,6]. Indeed, examining the neural predictive models of children's G-Factor faces several challenges. First, given that children on-average improve their G-Factor as they grow older[7], the predictive models should not only be generalizable to out-of-sample children but also be longitudinally stable[6]. That is, the predictive models should consistently capture children's G-factor across years during development. Second, MRI poses specific challenges inherent to young age, especially the noise due to movements during scan[8]. Thus the predictive models should handle missing values due to the noise. Third, to ensure that biological understanding of the G-Factor can be gained through predictive modeling, the models should be interpretable[9]. Here using large-scale, longitudinal MRI data from the Adolescent Brain Cognitive Development (ABCD) study[10], we take a multimodal, data-driven approach to develop the predictive models that address these three challenges.

Over the past decades, research has associated the G-Factor (and other intelligence-related variables) with MRI data from different modalities in adults and children[6]. While earlier studies focused mainly on in-sample association between the G-Factor and MRI indices[10,11], more recent studies started to develop models that are generalizable to unseen, out-of-sample participants[13–15]. Nonetheless, to the best of our knowledge, all work on the G-Factor has developed predictive models based on cross-sectional data. Using longitudinal data would allow us to test the stability of the models across time, which is particularly important for generalizing prediction across years during child development. Moreover, most studies have used a unimodal approach, i.e., focusing on one-single MRI modality at a time[6]. Different brain modalities offer different insights into brain structure and physiology. Without a direct comparison in a single study, it is unclear which modality provides a better prediction than others. Integrating multiple modalities could potentially provide a more comprehensive view into the predictive model[6]. It is possible, for instance, that the G-Factor depends not only on the activity of certain areas during certain cognitive tasks, such as working memory[11,12], (task-based functional MRI; task-fMRI) but also on the intrinsic functional connectivity between different areas[13–15] (resting-state fMRI; rs-fMRI) as well as the anatomy of grey[16] (structural MRI; sMRI) and white[17,18] (Diffusion Tensor Imaging; DTI) matter. Recent studies in adults[19,20] have shown advantages of the multimodal approach in the prediction of the G-Factor that might outweigh the higher complexity of the models. Still, few, if any, multimodal studies have been done in children. Moreover, task-based fMRI data that offer activity specific to tasks/cognitive processes are rarely integrated into the multimodal model.

Developing longitudinally predictive models from multimodal data for children is challenging as children's MRI data can be noisy, partly due to their movements during scanning[8]. For instance, the ABCD study recommended a set of quality control variables for detecting noisy data from each modality[10,21], resulting in an exclusion of 17% to over 50% of data depending on a modality. If we were to exclude 9-10-year-olds who have noisy data from any single modality, we would strictly limit the generalizability of our model to children with highly clean

data. Fortunately, a recently developed, opportunistic stacking framework in machine learning can deal with missingness in multimodal modeling[22]. Here, researchers use two training layers. The first training layer separately fits data from each modality to predict a response variable via penalized regression, such as Elastic Net[23]. In the second training layer, researchers then compute the predicted values of a response variable based on each modality-specific model from the first layer and use these predicted values as features via ensemble modeling, such as Random Forest[24]. Importantly, in the second training layer, researchers create two duplicated features for each modality: one with missing values coded as an arbitrarily large number or the other with an arbitrarily small number. Thus, as long as there is at least one modality available, then the data can be kept, leaving more data in the model building process.

To gain neurobiological insights of the G-Factor, the predictive models have to be interpretable[9], allowing researchers to demonstrate which neural indices contribute to the prediction. For a typical unimodal analysis, we[25] recently proposed a framework, known as eNetXplorer[26], that applies permutation along with Elastic Net to enable a statistical inference for each brain feature. Briefly, researchers first fit two sets of many Elastic Net models: one for predicting the true response variable (target models) and the other for predicting a randomly permuted response variable (null models). Researchers can then compare the magnitude of the coefficient estimates of a certain brain feature in target models to null models to make a statistical inference of this particular feature[27,28]. This approach can be applied to the first layer training of the opportunistic stacking model. As for the second layer training, we would need a method to infer which of the modalities drive the prediction of the random forests. Conditional permutation importance[29] allows this inference. Briefly, researchers randomly permute one feature (i.e., MRI modality in this case), while keeping other features in the same order to predict the G-Factor for the out-of-bag observations. If permuting certain features leads to dramatic decreases in prediction accuracy, then these features are important for the model. To control for correlated features, researchers further constrain the feature permutation to be within partitions of other features. Together, modern machine learners have developed techniques to assist with the interpretation of the predictive models from multimodal data.

Our goal is to develop longitudinally predictive models for the G-Factor from children's MRI data of different modalities. These models should be 1) stable across years, 2) able to handle missing values and 3) easy to interpret. We used the ABCD Release 3.0[10] that included full baseline data (age 9-10 years old) and half follow-up data (age 11-12 years old). We applied a higher-order confirmatory factor analysis (CFA) to extract the G-Factor from six cognitive tasks, collected outside of MRI. We then used the baseline data from children whose follow-up data have not been released yet as the training set, which was further split into first- and second-layer training sets. For the first-layer training set, we used Elastic Net to compute modality-specific models from six MRI modalities, including three task-based fMRI (working-memory "N-Back", reward "Monetary Incentive Delay; MID" and inhibitory control "Stop Signal"), rs-fMRI, sMRI, and DTI. At the second-layer training set, we applied Random Forests to integrate the predicted values from all modality-specific models and handled missing values with Opportunistic Stacking. We then tested the predictive performance of modality-specific (using the first layer only) and stacked modeling (using both layers) on the unseen baseline and follow-up data (see Figure 1). We next applied permutation tests (eNetXplorer[26] and conditional permutation importance[29]) to interpret the

feature importance of the final models. Finally, to ensure the generalisability of the models across sites of data collection, we performed leave-one-site-out cross-validation (i.e., splitting the data based on sites and evaluating the model performance on one hold-out site). We found that the stacked model was longitudinally stable and generalizable well across sites, better than the best performing modality-specific model, the N-Back fMRI task.

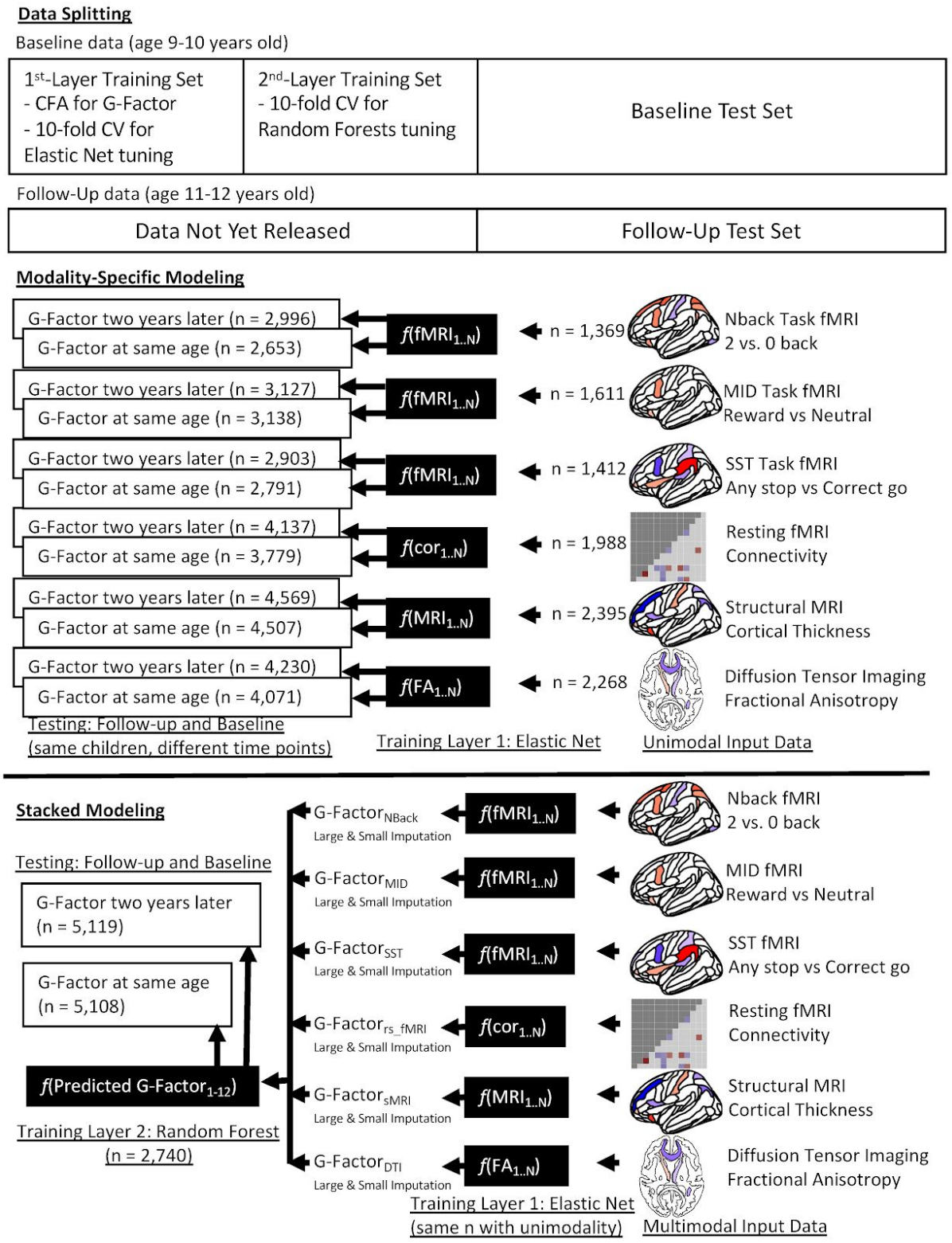

Figure 1. Longitudinally Predictive Modeling Approach. We split the data into four sets: first-layer training, second-layer training, baseline test, and follow-up test. We used the same participants in the baseline test and follow-up test sets. Modality-specific modeling only used the first-layer training set, while stacked modeling used both training sets to combine predicted values across modalities. The number of observations was different depending on the quality control of data from each modality. Stacked modeling imputed missing value to a large number (1000) and a small number (-1000) in two separate features. CFA = Confirmatory Factor Analysis; CV = Cross-Validation; cor = correlation; FA = fractional anisotropy.

Results

<u>The G-Factor</u>

Figure 2 shows the CFA of the G-Factor. The higher-order model of the G-Factor shows a good fit (*scaled, robust CFI*=.995*, TLI*=.988 and *RMSE*= .029 (*90%CI*=.015-.043), and the G-Factor has high internal consistency (*OmegaL2=.78*). The distribution of the G-Factor factor scores was similar across data splits.

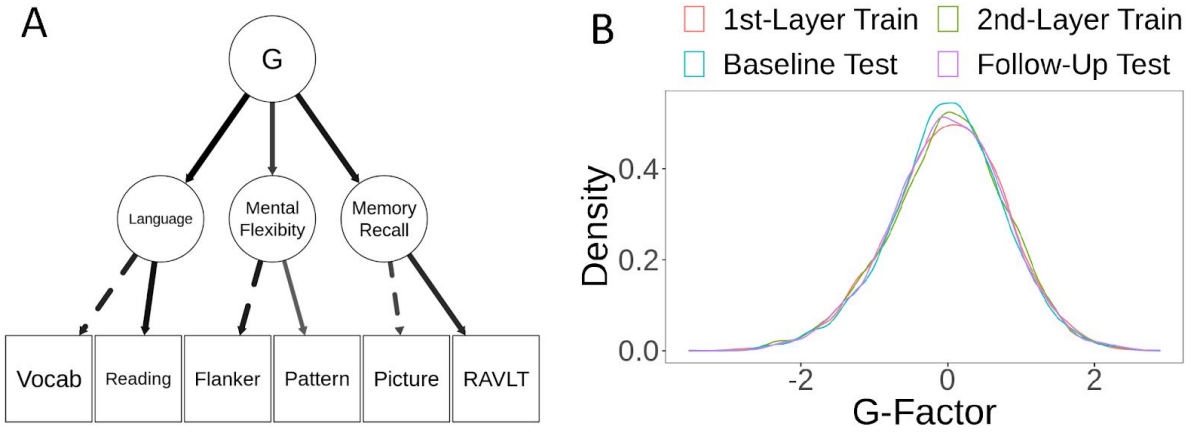

Figure 2. Confirmatory Factor Analysis (CFA) of the G-Factor. 2A shows the higher-order model for the G-Factor. Line thickness reflects the magnitude of standardized parameter estimates. The dotted lines indicate marker variables that were fixed to 1. Vocab = Picture Vocabulary; Reading = Oral Reading Recognition; Pattern = Pattern Comparison Processing; Picture = Picture Sequence Memory; RAVLT = Rey-Auditory Verbal Learning. 2B shows the distribution of the G-Factor factor score across the four data splits.

Longitudinally Predictive Models

Figure 1 shows the number of participants left in each modality per data split after quality control. Figure 3 shows the proportion of missing data in the two test sets. sMRI had the lowest missing observations, while the three task-based fMRI had the highest. Missing observations in the stacked model with at least one modality present, Stacked All, were around 3-6%, while those in the stacked model with all modalities present, Stacked Complete, were up to 78.79%.

From the first-layer training set, we found the best-tuned Elastic Net's mixture was 0 for N-Back task-based fMRI, rs-fMRI and DIT, .1 for sMRI, and 1 for SST and MID task-based fMRI. Thus, whether to have brain features shrunk together (the Ridge solution, mixture close to 0) or to have certain features from all features selected (the Lasso solution, mixture close to 1) depended on the modality. The averaged Elastic Net's penalty was .339 (SD = .45). From the second-layer training set, we found the best-tuned Random Forests' mtry (the number of features randomly sampled at each path split) at 4 and min_n (the minimum number of observations in a node per path split) at 170.

Table 1 and Figure 3 summarises the predictive ability of the longitudinally predictive models. Overall the predictive ability was similar across baseline and follow-up test sets. Comparing across modality-specific models, N-Back task-based fMRI provided the best out-of-sample predictive ability for both baseline and follow-up test sets, while SST task-based fMRI provided the worst for both baseline and follow-up. Other modalities had predictive ability around r=.2. Given that the N-Back task-based fMRI had the highest performance among modality-specific models, we set the missing values of the Stacked Best to be the same as those of the N-Back task-based fMRI. Performance of Stacked All, Stacked Complete and Stacked Best was among the top.

Table 1. Predictive Performance of the longitudinally predictive model and leave one-site out cross-validation. R squared = coefficient of determination; MAE = mean absolute error; RMSE = root mean squared error.

Predictive Performance from the baseline test set

Models	Pearson's r	R squared	MAE	RMSE
Stacked_All	0.439	0.191	0.699	0.895
Stacked_Complete	0.429	0.183	0.61	0.78
Stacked_Best	0.442	0.195	0.62	0.798
Stacked_NoBest	0.296	0.085	0.783	0.987
NBack	0.402	0.072	0.664	0.857
SST	0.129	-0.033	0.744	0.95
MID	0.202	0.013	0.738	0.944
rs_fMRI	0.233	0.042	0.749	0.955
sMRI	0.208	0.04	0.763	0.969
DTI	0.19	0.033	0.757	0.972

Predictive Performance from the follow-up test set

Models	Pearson's r	R squared	MAE	RMSE
Stacked_All	0.414	0.166	0.719	0.913
Stacked_Complete	0.427	0.168	0.651	0.829
Stacked_Best	0.438	0.175	0.666	0.846
Stacked_NoBest	0.317	0.1	0.794	1
NBack	0.383	0.118	0.687	0.875
SST	0.145	-0.004	0.76	0.961
MID	0.148	-0.003	0.757	0.955
rs_fMRI	0.251	0.055	0.754	0.954
sMRI	0.226	0.049	0.764	0.965
DTI	0.212	0.045	0.771	0.978

Mean (SD) of Out-Of-Site Predictive Performance

Models	Pearson's r	R squared	MAE	RMSE
Stacked	0.46 (0.057)	0.21 (0.052)	0.698 (0.023)	0.888 (0.029)
NBack	0.408 (0.069)	0.167 (0.055)	0.718 (0.028)	0.91 (0.031)
MID	0.227 (0.096)	0.05 (0.05)	0.772 (0.021)	0.973 (0.025)
SST	0.139 (0.071)	0.019 (0.024)	0.783 (0.014)	0.988 (0.012)
rs_fMRI	0.255 (0.061)	0.064 (0.03)	0.765 (0.015)	0.966 (0.016)
sMRI	0.248 (0.092)	0.061 (0.046)	0.763 (0.024)	0.967 (0.024)
DTI	0.223 (0.076)	0.049 (0.037)	0.766 (0.016)	0.974 (0.019)

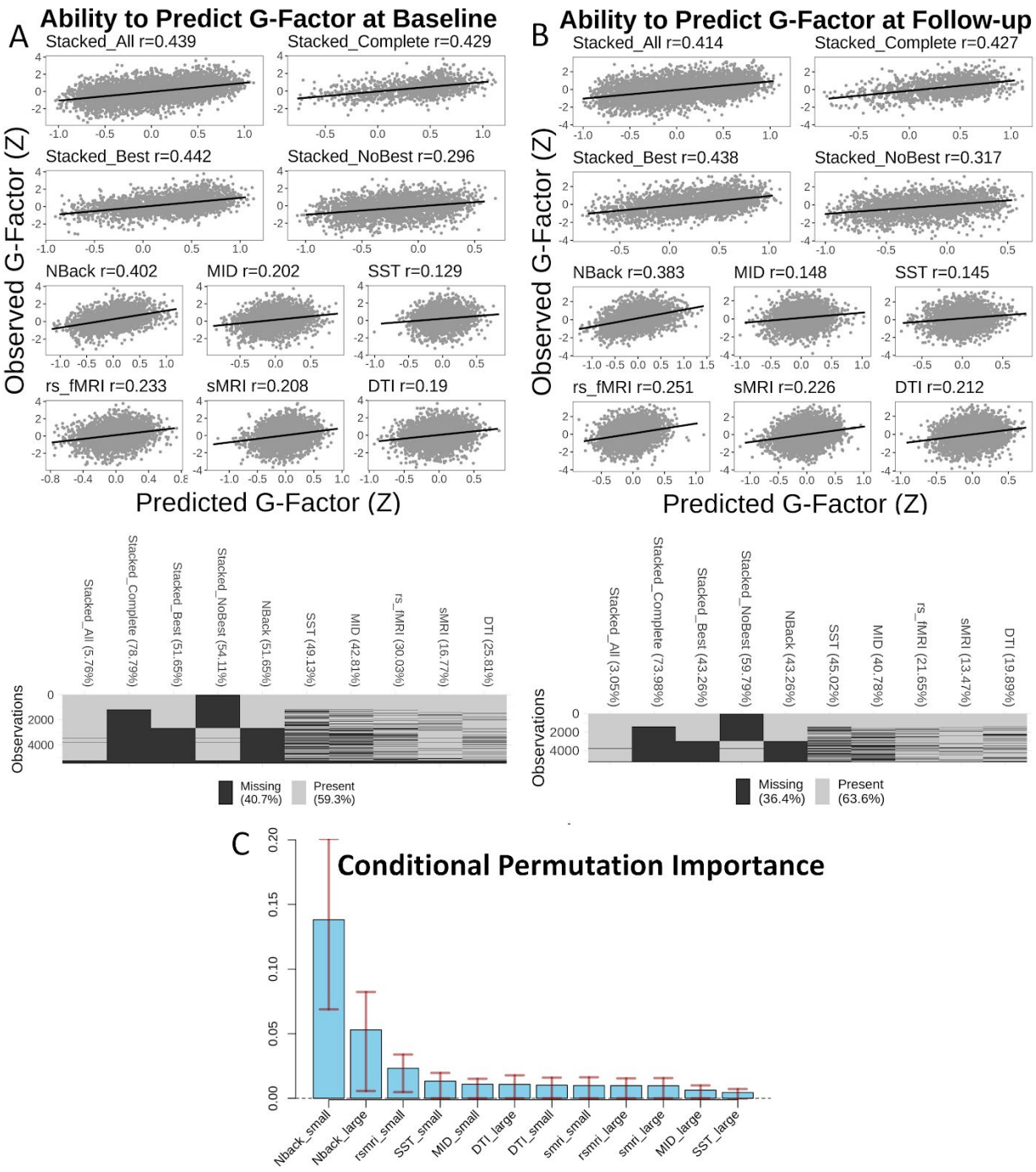

Figure 3. Predictive Ability of the Longitudinally Predictive Models and Missing Observations as a Function of Modalities in the Test Sets (3A for the baseline test set and 3B for the follow-up test set) and the Conditional Permutation Importance (CPI) of the Stacked Model (3C). Stacked All required data with at least one modality present. Stacked Complete required data with all modalities present. Stacked Best had the same missing values with the modality with the best prediction (N-Back task-based fMRI). Stacked No Best did not have any data from the modality with the best prediction and had at least one modality present. The CPI was computed based on the second-layer training set. Error bars in the bar plot show an interval between .25 and .75 quantiles of the CPI for each tree in the Random Forests. The "_large" and "_small" suffixes indicate whether the missing values were coded as a large (1000) or small (-1000) number, respectively.

Figure 4 compared the predictive ability between the Stacked Best and N-Back task-based fMRI using bootstrapped differences. The Stacked Best had significantly higher performance in both baseline and follow-up test sets, indicated by higher Pearson's *r* and R^2 and lower MAE and RMSE.

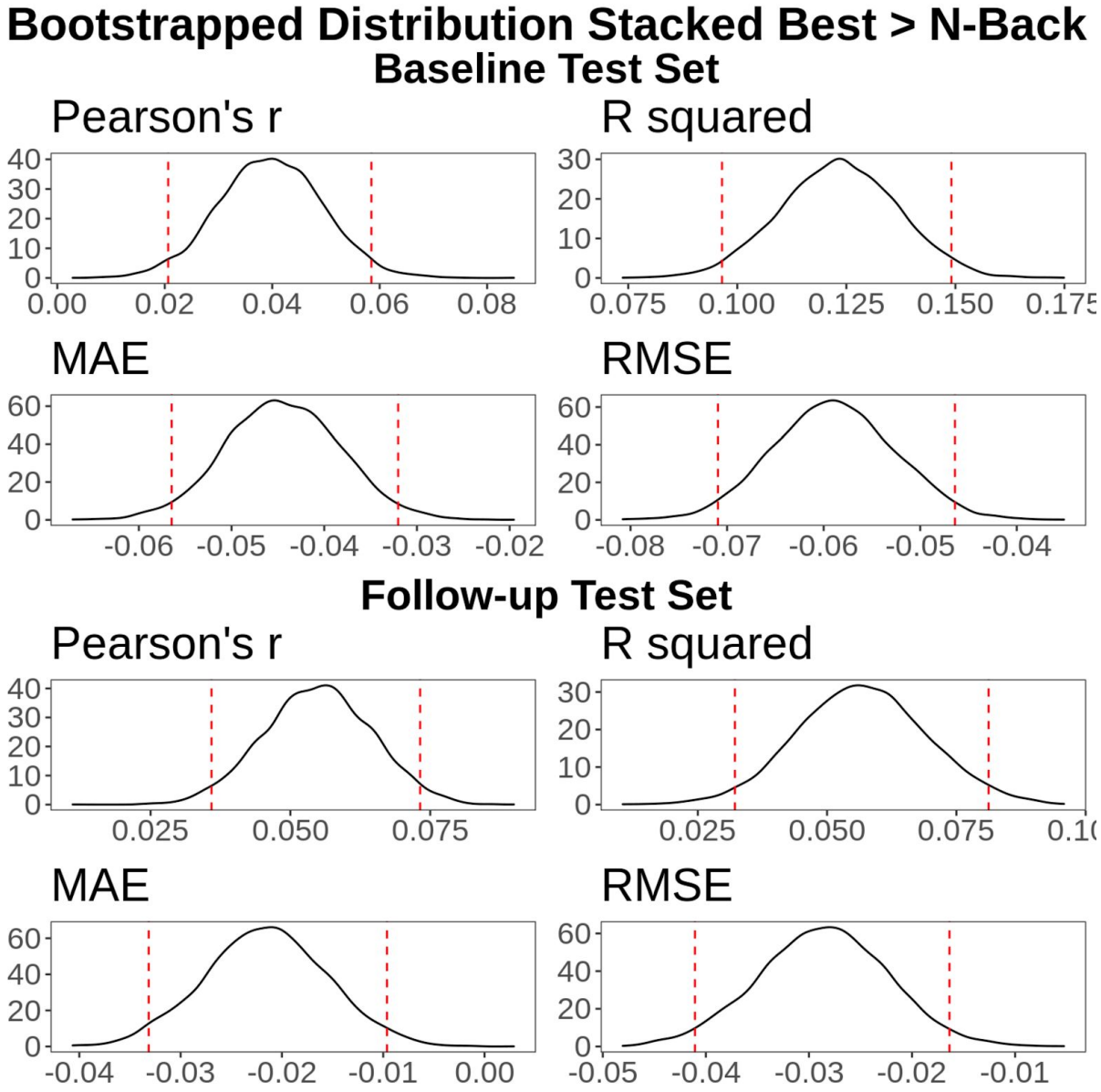

Figure 4. The Bootstrapped Distribution of the Differences in Predictive Ability between the Stacked Best and N-Back Task-Based fMRI. Dotted lines indicate 95% confidence intervals. We subtracted the predictive ability of N-Back Task-Based fMRI from that of the Stacked Best. R squared = coefficient of determination; MAE = mean absolute error; RMSE = root mean squared error.

Feature Importance

Figure 3C shows the Conditional Permutation Importance of the stacked model. N-Back task-based fMRI had the highest importance score. Figure 5 shows brain features that significantly (empirical $p < .05$) contributed to the prediction of the modality-specific model, based on eNetXplorer[26] permutation. For N-Back task-based fMRI, the G-Factor prediction was driven by activity in areas, such as the precuneus, sulcus intermedius primus, superior frontal sulci, and dorsal cingulate. For MID task-based fMRI, the prediction was driven by activity in several areas in the parietal, frontal and temporal regions. For SST, the prediction was contributed by activity in areas such as the supramarginal gyrus and inferior precentral sulcus. For rs-fMRI, the prediction was driven by connectivity within cinguloparietal and sensory-motor-hand as well as between networks that were connected with frontoparietal, default-mode, and sensory-motor-hand networks. For sMRI, the prediction was driven by the volume/thickness at several areas, such as the insula, middle frontal gyrus, and lingual sulcus. For DTI, the prediction was driven by FA at several white matter tracts, such as the superior longitudinal fasciculus, forceps minor, uncinate and parahippocampal cingulum.

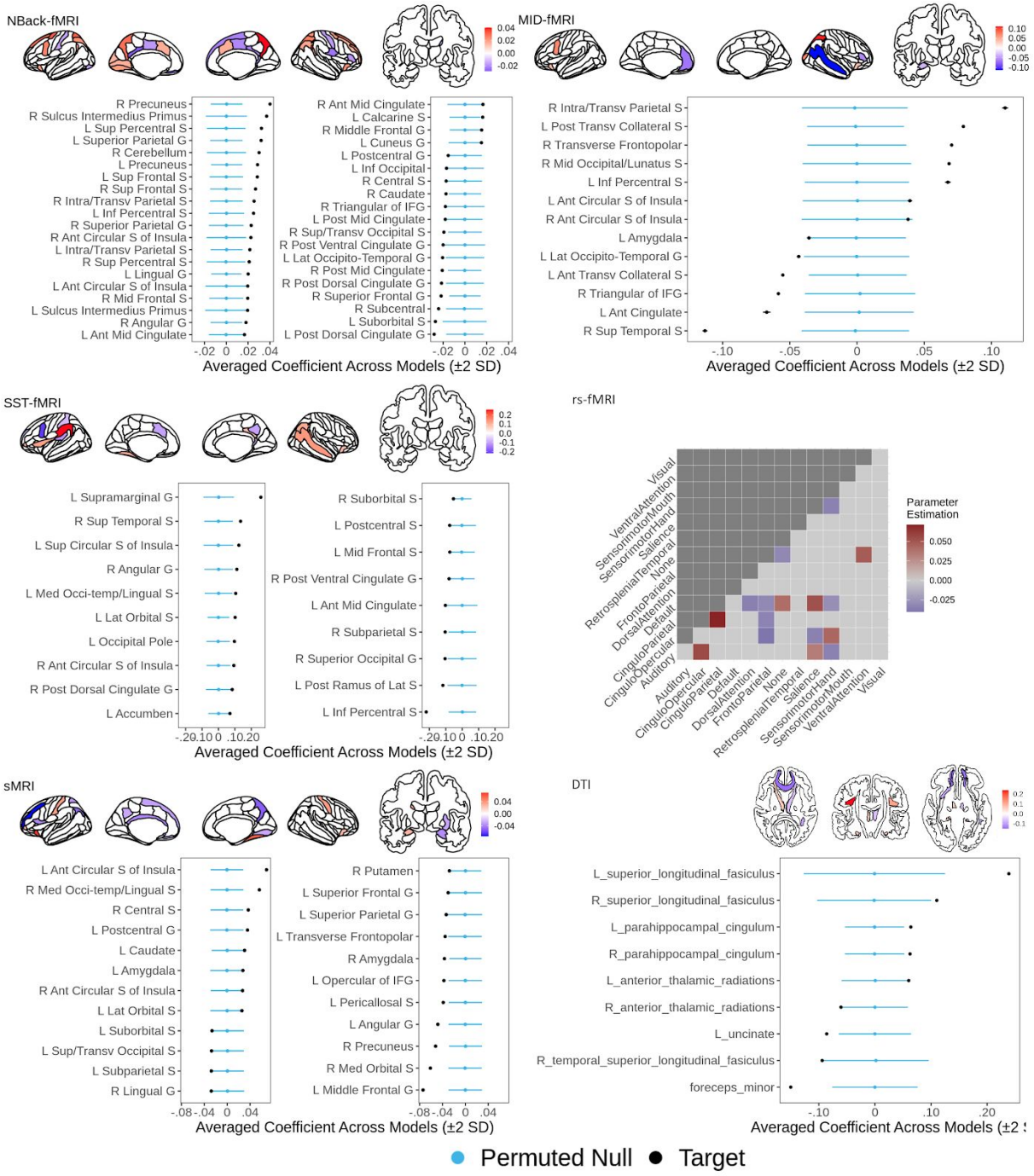

Figure 5. Feature importance for modality-specific models. We only plotted brain features with empirical p < .05 based on eNetXplorer[26] permutation. Sup=Superior; Ant=Anterior; Lat=lateral; Med=Medial; S=Sulcus; G=Gyrus; IFG=Inferior Frontal Gyrus; L=left; R=right.

<u>Leave-one-site-out cross-validation</u>

Table 1 and Figure 6 show the results of the leave-one-site-out cross-validation. Across 21 sites, the out-of-site predictive ability of the stacked model was highest, followed by N-Back task-based fMRI.

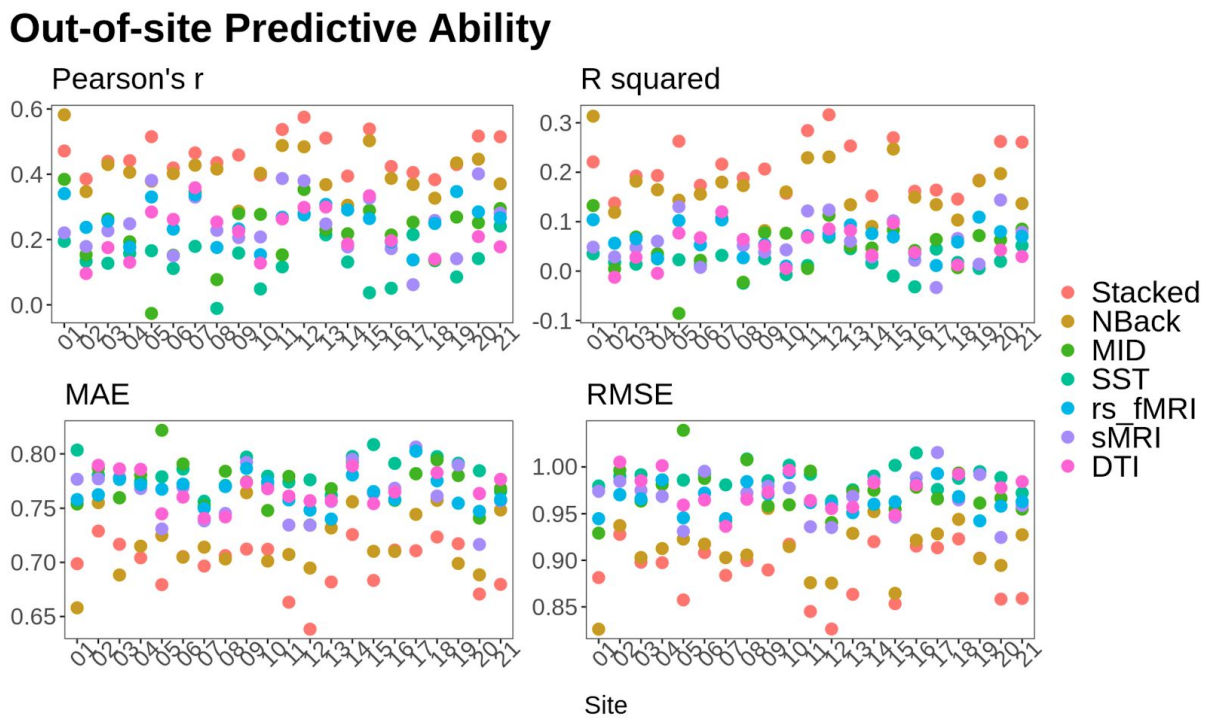

Figure 6. Leave-one-site-out cross-validation. We evaluated out-of-site predictive ability using Pearson's r between predicted vs. observed G-Factor in the held-out site. Note that DTI data were not available from 3 sites (sites 1, 17, and 19). R squared = coefficient of determination; MAE = mean absolute error; RMSE = root mean squared error.

Discussion

Here we developed longitudinally predictive models for children's G-Factor from MRI data of different modalities. We built models from the baseline data and tested them on unseen children at the same age and two years older. We found similar predictive ability across these two test sets for all modality-specific and stacked models. That is, the models that had high out-of-sample prediction on same-age children also had high out-of-sample prediction on older children, suggesting the longitudinal stability of MRI for many modalities. The best model across all performance indices was the stacked model that incorporated all six modalities, which was followed closely by the N-back task-related fMRI model. Apart from the SST task-related fMRI model, other models (including the MID tasked-related, rs-fMRI, sMRI, and DTI) performed moderately well. We also found a similar magnitude of predictive ability based on leave-one-site-out cross-validation, suggesting the generalisability of MRI not only across ages but also across data collection sites. Overall, the stacked model predicted around 20% of the variance in the children's G-Factor. This made the stacked model the most generalizable to out-of-sample children and the most longitudinally stable.

The stacked model improved predictive ability over and above the best modality, which was the N-Back task-based fMRI. This is based on bootstrapping of the difference in performance indices between the N-Back task-based fMRI and stacked model with the same participants (i.e., the Stacked Best). Accordingly, combining across modalities, as long as they are available, in one model seems to capture additional variance in the test data that

cannot be achieved even by the best modality in the model. This is consistent with previous studies showing the enhanced predictive power of the stacked mode[19,22].

Beyond generalisability across ages and sites, the stacked model also allowed us to handle missingness in the data[22]. This is especially important for children's MRI data given high levels of noise in certain modalities[8]. If we were to use data only from children with all modalities present (i.e., the Stacked Complete), the model would not apply to around 80% of the children. The Opportunistic Stacking allowed us to use the data as long as one modality was present (i.e., the Stacked All), leaving the exclusion to just around 5%. Importantly, the predictive performance of Stacked Complete and Stacked All were both relatively high, ensuring the ability of Opportunistic Stacking to deal with the missing data. Moreover, in the case when the best modality was not available, using the stacked model (i.e., the Stacked No Best) could be helpful. While the predictive ability of the Stacked No Best was not as strong as the Stacked Complete, Stacked All, and Stacked Best, its performance measures of variance (Pearson's r and R squared) appeared stronger in magnitude than any other non-optimal modalities by itself. Accordingly, in settings when not all of the modalities are available, researchers/practitioners can still take advantage of the boosted predictive ability of the stacked models over unimodal models.

Our use of permutation allowed easy-to-interpret models, highlighting the neurobiological bases of children's G-Factor. Conditional Permutation Importance[29] allowed us to infer that prediction from the stacked model was driven primarily by N-back task-related fMRI. This indicates the important role of working memory. eNetXplorer permutation[26] further showed us that contribution from fMRI activity in the parietal and frontal areas during the N-back task drove the prediction. These areas were similar to the areas previously found in an early association study[9]. Similarly, we also found brain indices from other modalities, from activity during other tasks to the cortical thickness and white matter density, that contributed to the prediction of the G-Factor, albeit with lower predictive performance.

Unlike previous unimodal studies[11–18] and recent multimodal studies[19,20], we were able to compare the ability of task-based fMRI with other modalities in predicting the G-Factor. We found that one of the three task-based fMRI models, the N-Back, performed exceptionally well. Based on the Conditional Permutation Importance[29], the N-Back task-related fMRI appeared to drive the prediction of the stacked model. This finding is contradictory to recent work that showed low-rank stability of region-specific activity of task-based fMRI across times, compared to sMRI[30]. We argue that, first, capturing individual differences of cognitive performance outside of the MRI scanner with task-based fMRI is achievable via a predictive modeling framework. That is, while region-specific task-based fMRI may have low-rank stability, building a model from activity across the whole brain enabled us to predict the G-Factor collected outside of the MRI scanner during the visit. Contrary to sMRI, task-based fMRI may vary from time to time, but so does cognitive performance. Showing that task-based fMRI could capture cognitive performance across a two-year gap provided a promising outlook for the use of task-based fMRI as a predictive tool.

It is important to note that not all fMRI tasks were suitable for predicting certain targets. The N-Back task and SST, for instance, were designed to capture working memory[31,32] and inhibitory control[31,33], respectively. Accordingly, both should be related to the G-Factor, especially on memory recall and mental flexibility portions of the G-Factor. Yet, only the

N-Back task showed good predictive ability. This may be due to different cognitive processes in each task (i.e., working memory vs. inhibitory control) or to different task configurations. It is entirely possible, for instance, that the block design used in the N-Back, as opposed to the event-related design used in the SST, allowed the N-Back to have higher predictive power. Accordingly, while task-based fMRI can have high predictive power, systematic comparisons are required in future research to better understand the characteristics of some tasks that make them more suitable for predicting the G-Factor and other individual differences.

Our study is not without limitations. We relied on the ABCD study's curated, preprocessed data[10,21,31]. This provided certain advantages. For instance, given its standardization, other studies that wish to apply our model of G-Factor to the ABCD data can readily do so without concerns about differences in preprocessing steps. Preprocessed data also enabled us to apply the manual quality control done by the study, a process that required time and well-trained labour[10,21,31]. Preprocessing large-scale multi-modal data ourselves would not only demand significant computer power and time but is prone to error. However, using the preprocessed data only allowed us to follow the choices of processing done by the study. For example, ABCD Release 3 only provided Freesurfer's parcellation[34,35] for task-based fMRI. While this popular method allowed us to interpret task-based activity on subject-specific anatomical landmarks, the regions are relatively large compared to other parcellations. Future studies will need to examine if smaller and/or different parcellations would improve predictive performance.

In conclusion, we developed a multimodal, predictive model for children's G-Factor that was 1) generalizable and consistent across two years, 2) robust against missingness in different modalities and 3) easy to interpret. Our model enhances biological insights about children's G-Factor through different neuroimaging lenses, from task activation, resting connectivity to grey and white-matter anatomy. While we only accounted for 20% of the variance in children's G-Factor, our approach should pave the way for future researchers to employ multimodal MRI as predictive tools for children's G-Factor.

Online Methods

We employed the ABCD Study Curated Annual Release 3.0[10], which included 3T MRI data and cognitive tests from 11,758 children (female=5,631) at the baseline (9-10 years old) and 5693 children (female = 2,617) at the two-year follow-up (11-12 years old). The study recruited the children from 21 sites across the United States[36] according to the ethical oversight as detailed elsewhere[37]. We further excluded 54 children based on Snellen Vision Screener[38,39]. These children either could not read any line, could only read the 1st (biggest) line, or could read up to the 4th line but indicated difficulty in reading stimuli on the iPad used for administering cognitive tasks (see below).

Longitudinally Predictive Model

Data Splitting

We split the data into 4 parts (Figure 1): 1) first-layer training set (n=3,041), 2) second-layer training set (n=3,042), 3) baseline test set (n=5,622) and 4) follow-up test set (n=5,656). Especially noteworthy is that children who were in the baseline test set were also in the follow-up test set. In other words, none of the children in the first-layer and second-layer training sets were in either of the test sets.

Target: the G-Factor

We modeled the G-Factor using children's performance from six cognitive tasks. These six tasks, collected on an iPad during a 70-min in-session visit outside of MRI[39,40], were available in both baseline and follow-up datasets. First, the Picture Vocabulary measured vocabulary comprehension and language[41]. Second, the Oral Reading Recognition measured reading and language decoding[42]. Third, the Franker measured conflict monitoring and inhibitory control[43]. Fourth, the Pattern Comparison Processing measured the speed of processing[44]. Fifth, the Picture Sequence Memory measured episodic memory[45]. Sixth, the Rey-Auditory Verbal Learning measured auditory recognition, recall, and learning[46].

Similar to the previous work[40,47,48], we applied a higher-order G-Factor model using confirmatory factor analysis (CFA) to encapsulate the G-factor as the latent variable underlying performance across cognitive tasks. More specifically, in our higher-order G-Factor model (Figure 2), we had the G-Factor as the 2nd-order latent variable. We also had three 1st-order latent variables in the model: language (capturing the Picture Vocabulary and Oral Reading Recognition), mental flexibility (capturing the Franker and Pattern Comparison Processing), and memory recall (capturing the Picture Sequence Memory and Rey-Auditory Verbal Learning).

To prevent data leakage, we fit the CFA model to the observations in the first-layer training set and then computed factor scores of the G-Factor on all training and test sets. We fixed latent factor variances to one and applied robust maximum likelihood estimation (MLR) with robust (Huber-White) standard errors and scaled test statistics. To demonstrate model fit, we used scaled and robust comparative fit index (CFI), Tucker-Lewis Index (TLI), root mean squared error of approximation (RMSEA) with 90% CI and internal consistency, OmegaL2[49], of the G-Factor. For CFA, we used lavaan[50] (version=.6-6) and semTools[49] along with semPlot[51] for visualization.

Features: multimodal MRI

We used MRI data from six modalities: three task-based fMRI, rs-fMRI, sMRI, and DTI. The ABCD Study provided detailed procedures on data acquisition and MRI image processing elsewhere[10,21,31]. We strictly followed their recommended exclusion criteria based on automated and manual QC review for each modality, listed under the *abcd_imgincl01* table[10]. These criteria involved not only the image quality but also MR neurological screening, behavioral performance, number of TRs left after scrubbing, the integrity of task presentation among others. We removed participants with an exclusion flag at any MRI indices, separately for each modality. We also applied the three IQR rule with listwise deletion to remove observations with outliers in any indices within each modality. Additionally, to adjust for between-site variability, we used an Empirical Bayes method, ComBat[52,53]. We applied ComBat to all modalities except for task-based fMRI, given that between-site variability was found to be negligible for task-based contrasts[53]. To prevent data leakage, we applied the IQR rule and Combat separately for first-layer training, second-layer training, baseline test, and follow-up test sets.

1-3) *Three Task-Based fMRI*

We used task-based fMRI from three tasks. First, in the working-memory "N-Back" task[31,32], children saw pictures of houses and emotional faces. Depending on the blocks, children reported if a picture matched either: (a) a picture that is shown 2 trials earlier (2-back), or (b) a picture that is shown at the beginning of the block (0-back). To focus on working-memory-related activity, we use the [2-back vs 0-back] linear contrast (i.e., high vs. low working memory load). Second, in the Monetary Incentive Delay (MID) task[31,54], children needed to respond before the target disappeared. And doing so would provide them with a reward, if and only if the target followed the "reward cue" (but not in "neural cue").To focus on reward anticipation-related activity, we used the [Reward Cue vs Neutral Cue] linear contrast. Third, in the Stop-Signal Task (SST)[31,33], children needed to withhold or interrupt their motor response to a "Go" stimulus when it is followed unpredictably by a Stop signal. To focus on inhibitory control-related activity, we used the [Any Stop vs Correct Go] linear contrast. Note that, for the SST, we used two additional exclusion criteria, *tfmri_sst_beh_glitchflag,* and *tfmri_sst_beh_violatorflag,* to address glitches in the task as recommended by the study[55,56]. For all tasks, we used the average contrast values across two runs. These values were embedded in the brain parcels based on FreeSurfer[57]'s Destrieux[35] and ASEG[34] atlases (148 cortical surface and 19 subcortical volumetric regions, resulting in 167 features for each task-based fMRI task).

4) *Resting-State fMRI (rs-fMRI)*

During rs-fMRI collection, the children viewed a crosshair for 20 minutes. The ABCD's preprocessing strategy has been published elsewhere[21]. Briefly, the study parcellated regions into 333 cortical-surface regions[58] and correlated their time-series[21]. They then grouped these correlations based on 13 predefined large-scale networks[58]: auditory, cingulo-opercular, cingulo-parietal, default-mode, dorsal-attention, frontoparietal, none, retrosplenial-temporal, salience, sensorimotor-hand, sensorimotor-mouth, ventral-attention, and visual networks. Note that "none" refers to regions that do not belong to any networks. After applying Fisher r to z transformation, the study computed mean correlations between pairs of regions within each large-scale network (n=13) and between large-scale networks (n=78) and provided these mean correlations in their Releases[10]. This resulted in 91 features for the rs-fMRI. Given that the correlations between (not within) large-scale

networks were highly collinear with each other (e.g., the correlation between auditory and cingulo-opercular was collinear with that between auditory and default-mode), we further decorrelated them using partial correlation. We first applied inverse Fisher r to z transformation, then partial correlation transformation, and then reapplied Fisher r to z transformation.

5) *Structural MRI (sMRI)*

The ABCD study processed sMRI, including cortical reconstruction and subcortical volumetric segmentation, using FreeSurfer[57]. Here we considered FreeSurfer-derived Destrieux[35] regional cortical thickness measures (n=148 cortical surface) and ASEG[34] regional subcortical volume measures (n=19), resulting in 167 features for sMRI. We also adjusted regional cortical thickness and volumetric measures using mean cortical thickness and total intracranial volume, respectively.

6) *Diffusion Tensor Imaging (DTI)*

Here we focused on fractional anisotropy (FA)[59]. FA characterizes the directionality of the distribution of diffusion within white matter tracts, which can indicate the density of fiber packing[59]. The ABCD study segmented major white matter tracts using AtlasTrack[21,60]. Here we considered FA of 23 major tracks, 10 of which were separately labeled for each hemisphere. These tracks included corpus callosum, forceps major, forceps minor, cingulate and parahippocampal portions of cingulum, fornix, inferior frontal occipital fasciculus, inferior longitudinal fasciculus, pyramidal/corticospinal tract, superior longitudinal fasciculus, temporal lobe portion of superior longitudinal fasciculus, anterior thalamic radiations and uncinate. This left 23 features for DTI.

Predictive Model Fitting

We started with the first-layer training set. Here we used standardized features from each modality to separately predict the G-Factor via the Elastic Net algorithm[23] from the glmnet package[61] (see Figure 1). As a general form of penalized regression, the Elastic Net requires two hyperparameters. First, the 'penalty' determines how strong the feature's slopes are regularised. Second, the 'mixture' determines the degree to which the regularisation is applied to the sum of squared coefficients (known as Ridge) vs. to the sum of absolute values of the coefficients (known as LASSO). We tuned these two hyperparameters using a 10-fold cross-validation grid search and selected the model with the lowest Mean Absolute Error (MAE). In the grid, we used 200 levels of the penalty from 10^-10 to 10, equally spaced on the logarithmic-10 scale and 11 levels of the mixture from 0 to 1 on the linear scale.

Once we obtained the final modality-specific models from the first-layer training set, we fit these models to data in the second-layer training set. This gave us six predicted values of the G-Factor from six modalities, and these are the features to predict the G-Factor in the second-layer training set. To handle missing observations when combining these modality-specific features, we applied an Opportunistic Stacking approach[22] by creating duplicates of each modality-specific feature. After standardization, we coded missing observations in one as an arbitrarily large value of 1000 and in the other as an arbitrarily small value of -1000, resulting in 12 features. That is, as long as a child had at least one modality available, we would be able to include this child in Stacked modeling.

We then used the Random Forests algorithm[24] from the ranger package[62] to predict the G-Factor from these 12 features[22,63]. Random Forests use a multitude of decision trees on various sub-samples of the data and implement averaging to enhance prediction and to control over-fitting. We used 1000 trees and turned two hyperparameters. First 'mtry' is the number of features randomly sampled at each split. Second 'min_n' is the minimum number of observations in a node needed for the node to be split further. We implemented a 10-fold cross-validation grid search and selected the model with the lowest Root Mean Squared Error (RMAE). In the grid, we used 12 levels of the mtry from 1 to 12, and 101 levels of the min_n from 1 to 1000, both on the linear scale. This resulted in the "stacked" model that incorporated data across modalities.

We examined the predictive ability of the models between predicted vs. observed G-Factor, using Pearson's correlation (r), coefficient of determination (R^2, calculated using the sum of square definition), mean absolute error (MAE), and root mean squared error (RMSE). To investigate the prediction of the modality-specific models, we used the models tuned from the first-layer training set. To investigate the prediction of the stacked model, we used the model tuned from the second-layer training set. We used the baseline test set for out-of-sample, same-age predictive ability, while we used the follow-up test sets for out-of-sample, longitudinally predictive ability.

To test the stacked model's performance, we further split the test sets based on the presence of each modality. First, Stacked All required data with at least one modality present. This allowed us to examine the stacked model's performance when the missing values were all arbitrarily coded. Second, Stacked Complete required data with all modalities present. This represents the situation when the data were as clean as possible. Third, Stacked Best had the same missing values with the modality with the best prediction. This allowed us to make a fair comparison in performance between the stacked model and the model with the best modality, given their same noise level from missing value. We applied bootstrapping with 5,000 iterations to examine the difference in performance indices. Fourth, Stacked No Best did not have any data from the modality with the best prediction and had at least one modality present. This represents the highest level of noise possible. For the machine learning workflow, we used 'tidymodels' (www.tidymodels.org).

Feature Importance

To understand which features contribute to the prediction of the modality-specific, Elastic Net model, we applied permutation from the eNetXplorer[64] package to the first-layer training set. We first chose the best mixture from the previously run grid and fit two sets of several Elastic Net models. The first "target" models used the true G-Factor as the target, while the second "null" models used the randomly permuted G-Factor as the target. eNetXplorer split the data into 10 folds 100 times/runs. For each run, eNetXplorer performed cross-validation by repeatedly training the target models on 9 folds and tested on the leftover fold. Also in each cross-validation run, eNetXplorer trained the null models 25 times. eNetXplorer then used the mean of non-zero model coefficients across all folds in a given run as a coefficient for each run, k^r. Across runs, eNetXplorer weighted the mean of a model coefficient by the frequency of obtaining a non-zero model coefficient per run. Formally, we defined an empirical p-value as:

$$p = \frac{1}{1 + nr*np} \left\{ 1 + \sum_{r=1}^{nr} \sum_{p=1}^{np} \Theta \left(|k|_{null}^{r,p} - |k|_{target}^{r} \right) \right\} ,(1)$$

where p is an empirical p-value, r is a run index, nr is the number of runs, p is a permutation index, np is the number of permutation, Θ is the right-continuous Heaviside step function, and $|k|$ is the magnitude of feature' coefficient. That is, to establish statistical significance for each feature, we used the proportion of runs in which the null models performed better than the target models. We plotted the target models' coefficients with $p < .05$ on the brain images using the ggseg[65] package.

To understand which modalities contributed strongly to the prediction of the stacked, Random Forests model, we applied conditional permutation importance (CPI) to the second-layer training set using the 'permimp' package[66]. The implementation was documented in detail elsewhere[66]. Briefly, the original permutation importance[24] shuffled the observations of one feature at a time while holding the target and other features in the same order. Researchers then examined decreases in prediction accuracy in the out-of-bag observations due to the permutation of some features. Stronger decreases are then assumed to reflect the importance of such features. However, this method has shown to be biased when there are correlated features[67]. CPI corrected for this bias by constraining the feature permutation to be within partitions of other features, which was controlled by the threshold 's' value. We used the default s value at .95, which assumed dependencies among features[66].

Leave-one-site-out cross-validation

In addition to our main analyses on longitudinally predictive models, we also applied leave-one-site-out cross-validation to the baseline data. This allowed us to examine the generalisability of the modality-specific and stacked models on different data collection sites. Different sites involved different fMRI machines, experimenters as well as demographics across the US[36]. Here we held out data from one site as a test set and divided the rest to be first- and second-layer training sets. We then cross-validated predictive ability across these held-out sites. We applied the same modeling approach with the longitudinally predictive models, except for two configurations to reduce the amount of ram used and computational time. Specifically, in our grid search, we used 100 levels of penalty (as opposed to 200) for Elastic Net and limited the maximal min_n to 500 (as opposed to 1000) for Random Forests. For the stacked model, we tested its predictive ability on children with at least one modality (i.e., stacked all). We examined the out-of-site prediction between predicted vs. observed G-Factor at each held-out site using r, R^2, MAE, and RMSE.

Code Availability

We made the code for this manuscript along with its detailed outputs available at https://narunpat.github.io/GFactorModelingABCD3/GFactorModelingABCD3.html

References

1. Jensen, A. R. *The g factor: the science of mental ability*. (Praeger, 1998).
2. Schmidt, F. L. & Hunter, J. General Mental Ability in the World of Work: Occupational Attainment and Job Performance. *Journal of Personality and Social Psychology* **86**, 162–173 (2004).
3. Der, G., Batty, G. D. & Deary, I. J. The association between IQ in adolescence and a range of health outcomes at 40 in the 1979 US National Longitudinal Study of Youth. *Intelligence* **37**, 573–580 (2009).
4. Batty, G. D., Deary, I. J. & Gottfredson, L. S. Premorbid (early life) IQ and Later Mortality Risk: Systematic Review. *Annals of Epidemiology* **17**, 278–288 (2007).
5. Rosenberg, M. D., Casey, B. J. & Holmes, A. J. Prediction complements explanation in understanding the developing brain. *Nature Communications* **9**, 589 (2018).
6. Sui, J., Jiang, R., Bustillo, J. & Calhoun, V. Neuroimaging-based Individualized Prediction of Cognition and Behavior for Mental Disorders and Health: Methods and Promises. *Biological Psychiatry* **88**, 818–828 (2020).
7. Tucker-Drob, E. M. Differentiation of cognitive abilities across the life span. *Developmental Psychology* **45**, 1097–1118 (2009).
8. Fassbender, C., Mukherjee, P. & Schweitzer, J. B. Minimizing noise in pediatric task-based functional MRI; Adolescents with developmental disabilities and typical development. *NeuroImage* **149**, 338–347 (2017).
9. Scheinost, D. *et al.* Ten simple rules for predictive modeling of individual differences in neuroimaging. *NeuroImage* **193**, 35–45 (2019).
10. Yang, R. & Jernigan, Terry; Adolescent Brain Cognitive Development Study (ABCD) - Annual Release 3.0. doi:10.15154/1519007.
11. Gray, J. R., Chabris, C. F. & Braver, T. S. Neural mechanisms of general fluid intelligence. *Nature Neuroscience* **6**, 316–322 (2003).
12. Waiter, G. D. *et al.* Exploring possible neural mechanisms of intelligence differences

using processing speed and working memory tasks: An fMRI study. *Intelligence* **37**, 199–206 (2009).

13. Sripada, C. *et al.* Prediction of neurocognition in youth from resting state fMRI. *Molecular Psychiatry* **25**, 3413–3421 (2020).
14. Pamplona, G. S. P., Santos Neto, G. S., Rosset, S. R. E., Rogers, B. P. & Salmon, C. E. G. Analyzing the association between functional connectivity of the brain and intellectual performance. *Front. Hum. Neurosci.* **9**, (2015).
15. Dubois, J., Galdi, P., Paul, L. K. & Adolphs, R. A distributed brain network predicts general intelligence from resting-state human neuroimaging data. *Philosophical Transactions of the Royal Society B: Biological Sciences* **373**, 20170284 (2018).
16. Narr, K. L. *et al.* Relationships between IQ and Regional Cortical Gray Matter Thickness in Healthy Adults. *Cerebral Cortex* **17**, 2163–2171 (2007).
17. Góngora, D., Vega-Hernández, M., Jahanshahi, M., Valdés-Sosa, P. A. & Bringas-Vega, M. L. Crystallized and fluid intelligence are predicted by microstructure of specific white-matter tracts. *Human Brain Mapping* **41**, 906–916 (2020).
18. Genç, E. *et al.* Diffusion markers of dendritic density and arborization in gray matter predict differences in intelligence. *Nature Communications* **9**, 1905 (2018).
19. Rasero, J., Sentis, A. I., Yeh, F.-C. & Verstynen, T. Integrating across neuroimaging modalities boosts prediction accuracy of cognitive ability. *bioRxiv* 2020.09.01.278747 (2020) doi:10.1101/2020.09.01.278747.
20. Jiang, R. *et al.* Multimodal data revealed different neurobiological correlates of intelligence between males and females. *Brain Imaging and Behavior* **14**, 1979–1993 (2020).
21. Hagler, D. J. *et al.* Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. *NeuroImage* **202**, 116091 (2019).
22. Engemann, D. A. *et al.* Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers. *eLife* **9**, e54055 (2020).
23. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of*

the Royal Statistical Society. Series B: Statistical Methodology **67**, 301–320 (2005).

24. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
25. Pornpattananangkul, N., Bartonicek, A., Wang, Y. & Stringaris, A. *An Omics-Inspired Elastic Net Approach Drastically Improves Out-of-Sample Prediction and Regional Inference of Task-Based fMRI*. http://biorxiv.org/lookup/doi/10.1101/2020.10.21.348367 (2020) doi:10.1101/2020.10.21.348367.
26. Candia, J. & Tsang, J. S. eNetXplorer: an R package for the quantitative exploration of elastic net families for generalized linear models. *BMC Bioinformatics* **20**, 189 (2019).
27. Helwig, N. E. Statistical nonparametric mapping: Multivariate permutation tests for location, correlation, and regression problems in neuroimaging. *Wiley Interdisciplinary Reviews: Computational Statistics* **11**, 1–24 (2019).
28. Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. & Nichols, T. E. Permutation inference for the general linear model. *NeuroImage* **92**, 381–397 (2014).
29. Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. Conditional variable importance for random forests. *BMC Bioinformatics* **9**, 307 (2008).
30. Elliott, M. L. *et al.* What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychol Sci* **31**, 792–806 (2020).
31. Casey, B. J. *et al.* The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience* **32**, 43–54 (2018).
32. Barch, D. M. *et al.* Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage* **80**, 169–189 (2013).
33. Whelan, R. *et al.* Adolescent impulsivity phenotypes characterized by distinct brain networks. *Nature Neuroscience* **15**, 920–925 (2012).
34. Fischl, B. *et al.* Whole Brain Segmentation. *Neuron* **33**, 341–355 (2002).
35. Destrieux, C., Fischl, B., Dale, A. & Halgren, E. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage* **53**, 1–15 (2010).
36. Garavan, H. *et al.* Recruiting the ABCD sample: Design considerations and procedures.

Developmental Cognitive Neuroscience **32**, 16–22 (2018).

37. Clark, D. B. *et al.* Biomedical ethics and clinical oversight in multisite observational neuroimaging studies with children and adolescents: The ABCD experience. *Developmental Cognitive Neuroscience* **32**, 143–154 (2018).
38. Snellen, H. Letterproeven Tot Bepaling Der Gezictsscherpte, Utrecht, Weyers (Dutch Edition). *Also published in many languages as Optotypi ad Visum Determinandum* (1862).
39. Luciana, M. *et al.* Adolescent neurocognitive development and impacts of substance use: Overview of the adolescent brain cognitive development (ABCD) baseline neurocognition battery. *Developmental Cognitive Neuroscience* **32**, 67–79 (2018).
40. Thompson, W. K. *et al.* The structure of cognition in 9 and 10 year-old children and associations with problem behaviors: Findings from the ABCD study's baseline neurocognitive battery. *Developmental Cognitive Neuroscience* **36**, 100606 (2019).
41. Gershon, R. C. *et al.* Language measures of the NIH toolbox cognition battery. *Journal of the International Neuropsychological Society* **20**, 642–651 (2014).
42. Bleck, T. P., Nowinski, C. J., Gershon, R. & Koroshetz, W. J. What is the NIH Toolbox, and what will it mean to neurology? *Neurology* **80**, 874–875 (2013).
43. Eriksen, B. A. & Eriksen, C. W. Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics* **16**, 143–149 (1974).
44. Carlozzi, N. E., Tulsky, D. S., Kail, R. V. & Beaumont, J. L. Vi. Nih Toolbox Cognition Battery (cb): Measuring Processing Speed. *Monographs of the Society for Research in Child Development* **78**, 88–102 (2013).
45. Bauer, P. J. *et al.* Iii. Nih Toolbox Cognition Battery (cb): Measuring Episodic Memory. *Monographs of the Society for Research in Child Development* **78**, 34–48 (2013).
46. Daniel, M. H. & Wahlstrom, D. Equivalence of Q-interactive™ and Paper Administrations of Cognitive Tasks: WISC®–V. 13 (2014).
47. Ang, Y.-S., Frontero, N., Belleau, E. & Pizzagalli, D. A. Disentangling vulnerability, state and trait features of neurocognitive impairments in depression. *Brain* awaa314 (2020)

doi:10.1093/brain/awaa314.

48. Pornpattananangkul, N. *et al.* Motivation and Cognitive Abilities as Mediators between Polygenic Scores and Psychopathology in Children. *medRxiv* 2020.06.08.20123877 (2020) doi:10.1101/2020.06.08.20123877.
49. Jorgensen, T. D. *et al.* semTools: Useful tools for structural equation modeling. *R package version 0.5-1* (2018).
50. Rosseel, Y. lavaan: An R Package for Structural Equation Modeling. *J. Stat. Soft.* **48**, (2012).
51. Epskamp, S. semPlot: Unified visualizations of structural equation models. *Structural Equation Modeling* **22**, 474–483 (2015).
52. Fortin, J.-P. *et al.* Harmonization of multi-site diffusion tensor imaging data. *NeuroImage* **161**, 149–170 (2017).
53. Nielson, D. M. *et al.* Detecting and harmonizing scanner differences in the ABCD study - annual release 1.0. *bioRxiv* 309260 (2018) doi:10.1101/309260.
54. Knutson, B., Westdorp, A., Kaiser, E. & Hommer, D. FMRI Visualization of Brain Activity during a Monetary Incentive Delay Task. *NeuroImage* **12**, 20–27 (2000).
55. Bissett, P. G., Hagen, M. P. & Poldrack, R. A. A cautionary note on stop-signal data from the Adolescent Brain Cognitive Development [ABCD] study. *bioRxiv* 2020.05.08.084707 (2020) doi:10.1101/2020.05.08.084707.
56. Garavan, H. *et al.* The ABCD Stop Signal Data: Response to Bissett et al. *bioRxiv* 2020.07.27.223057 (2020) doi:10.1101/2020.07.27.223057.
57. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage* **9**, 179–194 (1999).
58. Gordon, E. M. *et al.* Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations. *Cereb. Cortex* **26**, 288–303 (2016).
59. Alexander, A. L., Lee, J. E., Lazar, M. & Field, A. S. Diffusion tensor imaging of the brain. *Neurotherapeutics* **4**, 316–329 (2007).
60. Hagler, D. J. *et al.* Automated white-matter tractography using a probabilistic diffusion

tensor atlas: Application to temporal lobe epilepsy. *Human Brain Mapping* **30**, 1535–1547 (2009).

61. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**, 1–22 (2010).
62. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* **77**, 1–17 (2017).
63. Josse, J., Prost, N., Scornet, E. & Varoquaux, G. On the consistency of supervised learning with missing values. *arXiv:1902.06931 [cs, math, stat]* (2020).
64. Candia, J. & Tsang, J. S. ENetXplorer: An R package for the quantitative exploration of elastic net families for generalized linear models. *BMC Bioinformatics* **20**, 1–11 (2019).
65. Mowinckel, A. M. & Vidal-Piñeiro, D. Visualization of Brain Statistics With R Packages *ggseg* and *ggseg3d*. *Advances in Methods and Practices in Psychological Science* **3**, 466–483 (2020).
66. Debeer, D. & Strobl, C. Conditional permutation importance revisited. *BMC Bioinformatics* **21**, 307 (2020).
67. Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **8**, 25 (2007).