There is no fundamental trade-off between prediction accuracy and feature importance reliability

Jianzhong Chen[1,2,3,4*], Leon Qi Rong Ooi[1,2,3,4,5*], Jingwei Li[6,7], Christopher L. Asplund[1,2,4,8,9,10], Simon B Eickhoff[6,7], Danilo Bzdok[11,12], Avram J Holmes[13], B.T. Thomas Yeo[1,2,3,4,5,14]

1: Centre for Sleep and Cognition, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
2: Centre for Translational MR Research, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
3: Department of Electrical and Computer Engineering, National University of Singapore, Singapore
4: N.1 Institute for Health & Institute for Digital Medicine (WisDM), National University of Singapore, Singapore
5: Integrative Sciences and Engineering Programme (ISEP), National University of Singapore
6: Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Center Jülich, Jülich, Germany
7: Institute for Systems Neuroscience, Medical Faculty, Heinrich-Heine University Düsseldorf, Düsseldorf, Germany
8: Division of Social Sciences, Yale-NUS College, Singapore
9: Department of Psychology, National University of Singapore, Singapore
10: Duke-NUS Medical School, Singapore
11: Department of Biomedical Engineering, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada
12: Mila - Quebec AI Institute, Montreal, Canada
13: Yale University, Departments of Psychology and Psychiatry, New Haven, CT, USA
14: Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, USA

* These authors contribute equally to this work

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

Abstract

There is significant interest in using neuroimaging data to predict behavior. The predictive models are often interpreted by the computation of feature importance, which quantifies the predictive relevance of an imaging feature. Tian and Zalesky (2021) suggest that feature importance estimates exhibit low test-retest reliability, pointing to a potential trade-off between prediction accuracy and feature importance reliability. This trade-off is counter-intuitive because both prediction accuracy and test-retest reliability reflect the reliability of brain-behavior relationships across independent samples. Here, we revisit the relationship between prediction accuracy and feature importance reliability in a large well-powered dataset across a wide range of behavioral measures. We demonstrate that, with a sufficient sample size, feature importance (operationalized as Haufe-transformed weights) can achieve fair to excellent test-retest reliability. More specifically, with a sample size of about 2600 participants, Haufe-transformed weights achieve average intra-class correlation coefficients of 0.75, 0.57 and 0.53 for cognitive, personality and mental health measures respectively. Haufe-transformed weights are much more reliable than original regression weights and univariate FC-behavior correlations. Intriguingly, feature importance reliability is strongly positively correlated with prediction accuracy across phenotypes. Within a particular behavioral domain, there was no clear relationship between prediction performance and feature importance reliability across regression algorithms. Finally, we show mathematically that feature importance reliability is necessary, but not sufficient, for low feature importance error. In the case of linear models, lower feature importance error leads to lower prediction error (up to a scaling by the feature covariance matrix). Overall, we find no fundamental trade-off between feature importance reliability and prediction accuracy.

1. Introduction

Neuroimaging provides a non-invasive means to study human brain structure and function. *In vivo* imaging features have been linked to many clinically relevant phenotypes when contrasting populations of patients and healthy controls (Greicius *et al.* 2004, Kennedy *et al.* 2006). However, these group-level studies ignore inter-individual differences within and across patient populations (Zhang *et al.* 2016, Xia *et al.* 2018, Zabihi *et al.* 2019, Tang *et al.* 2020, Wolfers *et al.* 2020). As a result, there is an increasing interest in the field to shift from group differences to accurate individual-level predictions (Dosenbach *et al.* 2010, Finn *et al.* 2015, Hsu *et al.* 2018, Nostro *et al.* 2018, Kong *et al.* 2019).

One goal of neuroimaging-based behavioral prediction is clinical usage to forecast practically useful clinical endpoints (Gabrieli *et al.* 2015). This ambition requires users to have trust in the predictive models, which often rests on a given models' interpretability (Bussone *et al.* 2015, Price 2018, Anderson and Anderson 2019, Diprose *et al.* 2020, Hedderich and Eickhoff 2020). Indeed, the recently enacted European Union Global Data Protection Regulation (GDPR) states that patients have a right to "meaningful information about the logic involved" when automated decision-making systems are used (Vasey *et al.* 2022a, 2022b). Furthermore, in many studies, the derived predictive models are often interpreted to gain insights into the predictive principles and inter-individual differences that underpin observed brain-behavior relationships (Finn *et al.* 2015, Greene *et al.* 2018, Chen *et al.* 2022). Therefore, while many studies in the neuroimaging literature have focused on prediction accuracy (Dadi *et al.* 2019, He *et al.* 2020, Pervaiz *et al.* 2020, Schulz *et al.* 2020, Abrol *et al.* 2021), enhancing model interpretability remains an important issue.

One approach to interpret predictive models is the computation of feature-level importance, which quantifies the relevance of an imaging feature in the predictive model. In the case of linear models, most previous studies have interpreted the regression weights (Jiang *et al.* 2020, Sripada *et al.* 2020, Cropley *et al.* 2021, Xiao *et al.* 2021) of predictive models, which can be highly misleading (Haufe *et al.* 2014). Instead, Haufe and colleagues suggested that it is necessary to perform a simple transformation of the linear models to yield better interpretation (Haufe *et al.* 2014), which we will refer to as the Haufe transform.

A recent study suggested that in the context of behavioral predictions from functional connectivity (FC), the reliability of feature-level importance (original regression weights and Haufe-transformed weights) across independent samples was poor (Tian and Zalesky 2021). Because the study utilized a maximum sample size of 400 and predicted only a small selection of cognitive measures and sex, it remains unclear whether the results generalize to other sample sizes and behavioral domains. Tian and Zalesky also found that a higher-resolution parcellation led to better prediction accuracy but lower feature importance reliability, thus suggesting a potential trade-off between feature importance reliability and prediction accuracy. However, this trade-off is counter-intuitive given that both feature importance reliability and prediction accuracy should reflect the reliability of brain-behavior relationship across independent datasets. More specifically, if the brain-behavior relationships in two independent data samples are highly similar, then we would expect that a model trained on one dataset to generalize well to the other dataset (i.e., high prediction accuracy). We would also expect the models trained on both datasets to be highly similar, leading to high test-retest feature importance reliability. Therefore, we hypothesize that there might not be a fundamental trade-off between prediction accuracy and feature importance reliability.

In the present study, we used the Adolescent Brain Cognitive Development (ABCD) study to investigate the relationship between prediction accuracy and feature importance reliability. Resting-state functional connectivity was used to predict a wide range of 36 behavioral measures across cognition, personality (related to impulsivity), and mental health. We considered three commonly used prediction models: kernel ridge regression (KRR), linear ridge regression (LRR), and least absolute shrinkage and selection operator (LASSO) models. Consistent with Tian and Zalesky (2021), we found that Haufe-transformed weights were more reliable than regression weights and univariate FC-behavior correlations. However, for sufficiently large sample sizes, we found fair to excellent test-retest reliability for the Haufe-transformed weights. Intriguingly, feature importance reliability was strongly correlated with prediction accuracy across behavioral measures. Within a particular behavioral domain, there was no clear relationship between prediction performance and feature importance reliability across regression algorithms. We show mathematically that test-retest feature importance reliability is necessary, but not sufficient, for low feature importance error. In the case of linear models, prediction error closely reflects feature importance error. Overall, we demonstrate that there is not a fundamental trade-off between feature importance reliability and prediction accuracy.

2. Methods

2.1 Dataset

The Adolescent Brain Cognitive Development (ABCD) dataset (2.0.1 release) was used for its large sample size, as well as its rich imaging and behavioral measures. The Institutional Review Board (IRB) at the University of California, San Diego approved all aspects of the ABCD study (Auchter *et al.* 2018). Parents or guardians provided written consent while the child provided written assent (Clark *et al.* 2018).

After quality control and excluding siblings, the final sample consisted of 5260 unrelated participants. Consistent with our previous studies (Chen *et al.* 2022, Ooi *et al.* 2022), each participant had a 419 × 419 FC matrix as the imaging features, which were used to predict 36 behavioral measures across the behavioral domains of cognition, personality, and mental health.

2.2 Image preprocessing

Images were acquired across 21 sites in the United States with harmonized imaging protocols for GE, Philips, and Siemens scanners (Casey *et al.* 2018). We used structural T1 and resting-fMRI. For each participant, there were four resting-fMRI runs. Each resting-fMRI run was 300 seconds long. Preprocessing followed our previously published study (Chen *et al.* 2022). For completeness, the key preprocessing steps are summarized here.

Minimally preprocessed T1 data were used (Hagler *et al.* 2019). The structural data were further processed using FreeSurfer 5.3.0 (Dale *et al.* 1999, Fischl, Sereno, and Dale 1999, Fischl, Sereno, Tootell, *et al.* 1999, Fischl *et al.* 2001, Ségonne *et al.* 2004, 2007), which generated accurate cortical surface meshes for each individual. Individuals' cortical surface meshes were registered to a common spherical coordinate system (Fischl, Sereno, and Dale 1999, Fischl, Sereno, Tootell, *et al.* 1999). Individuals who did not pass recon-all quality control (Hagler *et al.* 2019) were removed.

Minimally preprocessed fMRI data (Hagler *et al.* 2019) were further processed with the following steps: (1) removal of initial frames, with the number of frames removed depending on the type of scanner (Hagler *et al.* 2019); and (2) alignment with the T1 images using boundary-based registration (Greve and Fischl 2009) with FsFast (http://surfer.nmr.mgh.harvard.edu/fswiki/FsFast). Functional runs with boundary-based registration (BBR) costs greater than 0.6 were excluded. Framewise displacement (FD) (Jenkinson *et al.* 2002) and voxel-wise differentiated signal variance (DVARS) (Power *et al.* 2012) were computed using fsl_motion_outliers. Respiratory pseudomotion was filtered out using a bandstop filter (0.31-0.43 Hz) before computing FD (Power *et al.* 2019, Fair *et al.* 2020, Gratton *et al.* 2020). Volumes with FD > 0.3 mm or DVARS > 50, along with one volume before and two volumes after, were marked as outliers and subsequently censored. Uncensored segments of data containing fewer than five contiguous volumes were also censored (Gordon *et al.* 2016, Kong *et al.* 2019). Functional runs with over half of their volumes censored and/or max FD > 5mm were removed. Individuals who did not have at least 4 minutes of data were also excluded from further analysis.

The following nuisance covariates were regressed out of the fMRI time series: global signal, six motion correction parameters, averaged ventricular signal, averaged white matter signal, and their temporal derivatives (18 regressors in total). Regression coefficients were estimated from the non-censored volumes. We chose to regress the global signal because we were interested in behavioral prediction, and global signal regression has been shown to improve behavioral prediction performance (Greene *et al.* 2018, Li *et al.* 2019). The brain scans were interpolated across censored frames using least squares spectral estimation (Power *et al.* 2014), band-pass filtered (0.009 Hz ≤ f ≤ 0.08 Hz), projected onto FreeSurfer fsaverage6 surface space and smoothed using a 6 mm full-width half maximum kernel.

2.3 Functional connectivity

We used a whole-brain parcellation comprising 400 cortical regions of interest (ROIs) (Schaefer *et al.* 2018) and 19 subcortical ROIs (Fischl *et al.* 2002). For each participant and each fMRI run, functional connectivity (FC) was computed as the Pearson's correlations between the average time series of each pair of ROIs. FC matrices were then averaged across runs, yielding a 419□×□419 FC matrix for each participant. Correlation values were converted to z-scores using Fisher's r-to-z transformation prior to averaging and converted back to correlation values after averaging. Censored frames were ignored when computing FC.

2.4 Behavioral data

Following our previous study (Chen *et al.* 2022), we considered 16 cognitive, 11 mental health, and 9 impulsivity-related personality measures. The cognitive measures were vocabulary, attention, working memory, executive function, processing speed, episodic memory, reading, fluid cognition, crystallized cognition, overall cognition, short delay recall, long delay recall, fluid intelligence, visuospatial accuracy, visuospatial reaction time, and visuospatial efficiency. The mental health measures were anxious depressed, withdrawn depressed, somatic complaints, social problems, thought problems, attention problems, rule-breaking behavior, aggressive behavior, total psychosis symptoms, psychosis severity, and mania. The impulsivity-related personality measures were negative urgency, lack of planning, sensation seeking, positive urgency, lack of perseverance, behavioral inhibition, reward responsiveness, drive, and fun seeking.

Participants who did not have all behavioral measures were excluded from further analysis. As recommended by the ABCD consortium, individuals from Philips scanners were also excluded due to incorrect preprocessing. Finally, by excluding siblings, the main analysis utilized data from 5260 unrelated children.

2.5 Split-half cross-validation

ABCD is a multi-site dataset. To reduce sample size variability across sites, smaller sites were combined to create 10 "site-clusters", each containing at least 300 individuals (Table S1). Thus, participants within a site were in the same site-cluster.

A split-half cross-validation procedure was utilized to evaluate the prediction performance and the test-retest reliability of feature importance. For each split, 5 site-clusters were selected as the training set and the remaining 5 were selected as the test set. Prediction models were trained on the training set and evaluated on the test set. Here, we considered kernel ridge regression (KRR), linear ridge regression (LRR), and least absolute shrinkage and selection operator (LASSO) models for prediction. Hyperparameters were tuned using cross-validation within the training set (Chen et al., 2022).

Prediction accuracy was defined as the Pearson's correlation between the predicted and observed behavior of test participants. Feature importance of the regression models was computed in the training set (see Section 2.7). After the prediction model was trained and evaluated, the training and test sets were swapped. The model training and evaluation procedure were then repeated. Thus, for a given regression approach and interpretation method, each data split yielded two prediction accuracies and two sets of feature importance.

For each data split, the two accuracy numbers were averaged yielding an overall prediction accuracy for the split. The two sets of feature importance were used to compute test-retest reliability, defined as the intra-class correlation coefficient (ICC) (Noble *et al.* 2019, Tian and Zalesky 2021). To ensure stability, the data split was repeated 126 (the number of unique ways to split ten site-clusters into two halves, which is 10 choose 5 divided by 2) times.

2.6 Reliability across different sample sizes

The procedure in the previous section utilized the full sample size. To evaluate feature importance reliability across different sample sizes, the previous procedure (Section 2.5) was repeated, but the participants were subsampled for each split-half cross-validation to achieve a desired sample size *N*. More specifically, we considered sample sizes of 200, 400, 1000, and 1500. For each sample size *N*, we first split the 10 site-clusters into two halves, each containing 5 site-clusters (Section 2.5). *N*/10 samples were then randomly sampled from each site-cluster. The procedure was repeated 126 (the number of unique ways to split ten site-clusters into two halves, which is 10 choose 5 divided by 2) times.

2.7 Original and Haufe-transformed weights

We used KRR, LRR and LASSO to predict 36 behavioral measures from FC features. In particular, the lower triangular entries of the FC matrix were used as input for the regression models. LRR and LASSO are commonly used in the literature. We have previously demonstrated that KRR is a powerful approach for resting-FC behavioral prediction (He *et al.* 2020).

Since KRR is less commonly used in the literature, we will provide a high-level explanation here. Briefly, let y_i and FC_i be the behavioral measure and FC of training individual i. Let y_t and FC_t be the behavioral measure and FC of a test individual. Then, kernel regression would predict the test individual's behavior as the weighted average of the training individuals' behavior, i.e. $y_t \approx \sum_{i \in training\ set} Similarity(FC_i, FC_t) y_i$, where $Similarity(FC_i, FC_t)$ was defined as the Pearson's correlation between FC_i and FC_t. Thus, kernel regression assumed that individuals with more similar FC exhibit more similar behavior. To reduce overfitting, an l_2-regularization term was included, which was tuned in the training set (Kong *et al.* 2019, Li *et al.* 2019, He *et al.* 2020).

To interpret the trained models, we considered both the regression weights and Haufe-transformed weights. Since LRR and LASSO are linear models, the regression weights were straightforward to obtain. In the case of KRR, the kernel regression model was converted to an equivalent linear regression model, yielding one regression weight for each feature (Liu *et al.* 2007, Chen *et al.* 2022). We note that this conversion was possible because we used the correlation kernel, which is linear when the input features are pre-normalized.

Each prediction model was also inverted using the Haufe's transform (Haufe *et al.* 2014). Briefly, Haufe defined feature importance as the covariance between the predicted behavior and imaging feature in the training set (Chen *et al.* 2022).

2.8 Mass univariate associations

Besides predictive models, we also examined the test-retest reliability of mass univariate associations between FC and behavioral measures, which is sometimes referred to as brain-wide association analysis (Marek *et al.* 2022). We note that mass univariate associations are often used for feature selection in neuroimaging predictive models (Finn *et al.* 2015). The selected features are then used to interpret the model (Finn *et al.* 2015, Shen *et al.* 2017). Therefore, mass univariate associations are a good proxy for such approaches. Here, univariate association is defined as the correlation between each FC feature and each behavioral measure. To study the test-reliability of univariate associations, we performed the same split-half procedure (Sections 2.5 and 2.6). However, instead of training a predictive model in the training set, we correlated the FC features and the behavioral measures of the training participants to obtain one t-statistic for each feature and each behavioral measure. This procedure was repeated for the test participants. Test-retest reliability was defined as the ICC of the t-statistic values between the two halves of the dataset (i.e., training and test sets).

2.9 Data and code availability

The ABCD data are publicly available via the NIMH Data Archive (NDA). Processed data from this study have been uploaded to the NDA. Researchers with access to the ABCD data will be able to download the data: LINK_TO_BE_UPDATED. Analysis code specific to this study was can be found on GitHub: LINK_TO_BE_UPDATED. Co-author LQRO reviewed the code before merging it into the GitHub repository to reduce the chance of coding errors.

3. Results

3.1 Haufe-transformed weights exhibit fair to excellent test-retest reliability with large sample sizes

We computed resting-state functional connectivity (RSFC) among 400 cortical (Schaefer *et al.* 2018) and 19 subcortical (Fischl *et al.* 2002) regions for 5260 participants from the ABCD dataset (Casey *et al.* 2018). The lower triangular entries of the 419 × 419 RSFC matrix were then vectorized to predict 36 behavioral scores that span across 3 domains: cognition, personality, and mental health.

Feature importance of KRR predictive models was interpreted using two approaches: regression weights and Haufe-transformed weights. For comparison, t-statistics from mass univariate associations were also computed. We used a split-half procedure to compute the test-retest reliability of feature importance. For each split, we fit the KRR model on each half and obtain the feature importance. The test-retest reliability is defined as the intraclass correlation coefficient (ICC) of the feature importance values between the two halves.

Figure 1 shows the split-half test-retest reliability of the two interpretation methods and mass univariate associations across 126 splits for different sample sizes and behavioral domains. Consistent with previous studies, test-retest reliability of feature importance increases with larger sample sizes across all behavioral domains and interpretation methods (Tian and Zalesky 2021, Marek *et al.* 2022). The Haufe-transformed weights were consistently more reliable than univariate associations (t-statistics), which were in turn more reliable than the regression weights. Haufe-transformed weights at a sample size of 200 were more reliable than the original regression weights at a sample size of 2630.

At the largest sample size of 2630, an average ICC of 0.75 was achieved for Haufe-transformed weights of models predicting cognitive measures, which is considered “excellent” test-retest reliability (Cicchetti 1994). On the other hand, an average ICC of 0.57 and 0.53 were achieved for personality and mental health at the full sample size, which are considered “fair” test-retest reliabilities (Cicchetti 1994). Under the same sample size and interpretation method, the test-retest reliability of feature importance for mental health and personality was consistently lower than that of cognition.

Similar conclusions were obtained with linear ridge regression (Figure 2) and LASSO (Figure S1). Note that univariate associations (tstats) were computed independent of regression models and are therefore the same across Figures 1, 2 and S1. Overall, we found that Haufe-transformed weights achieved fair to excellent test-retest reliability with sufficiently large samples.

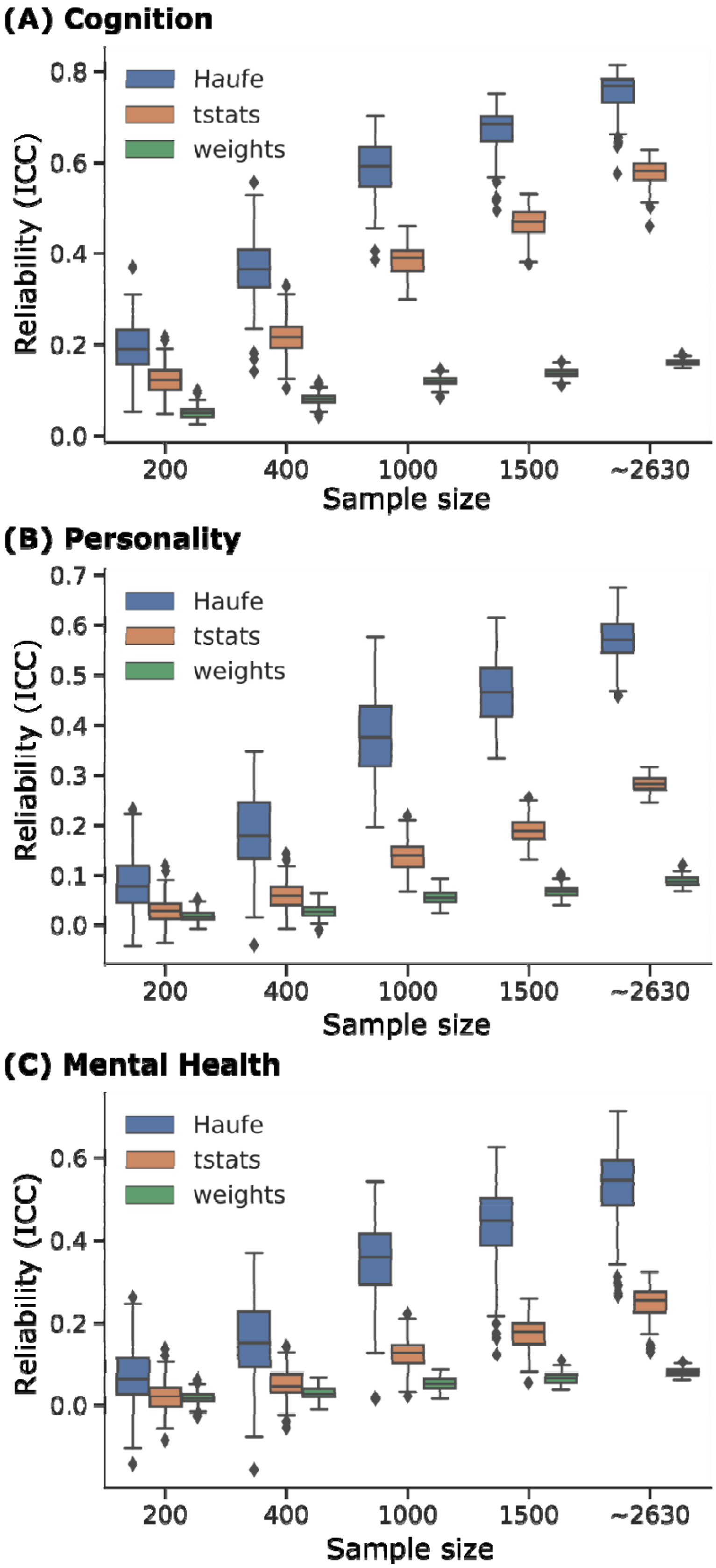
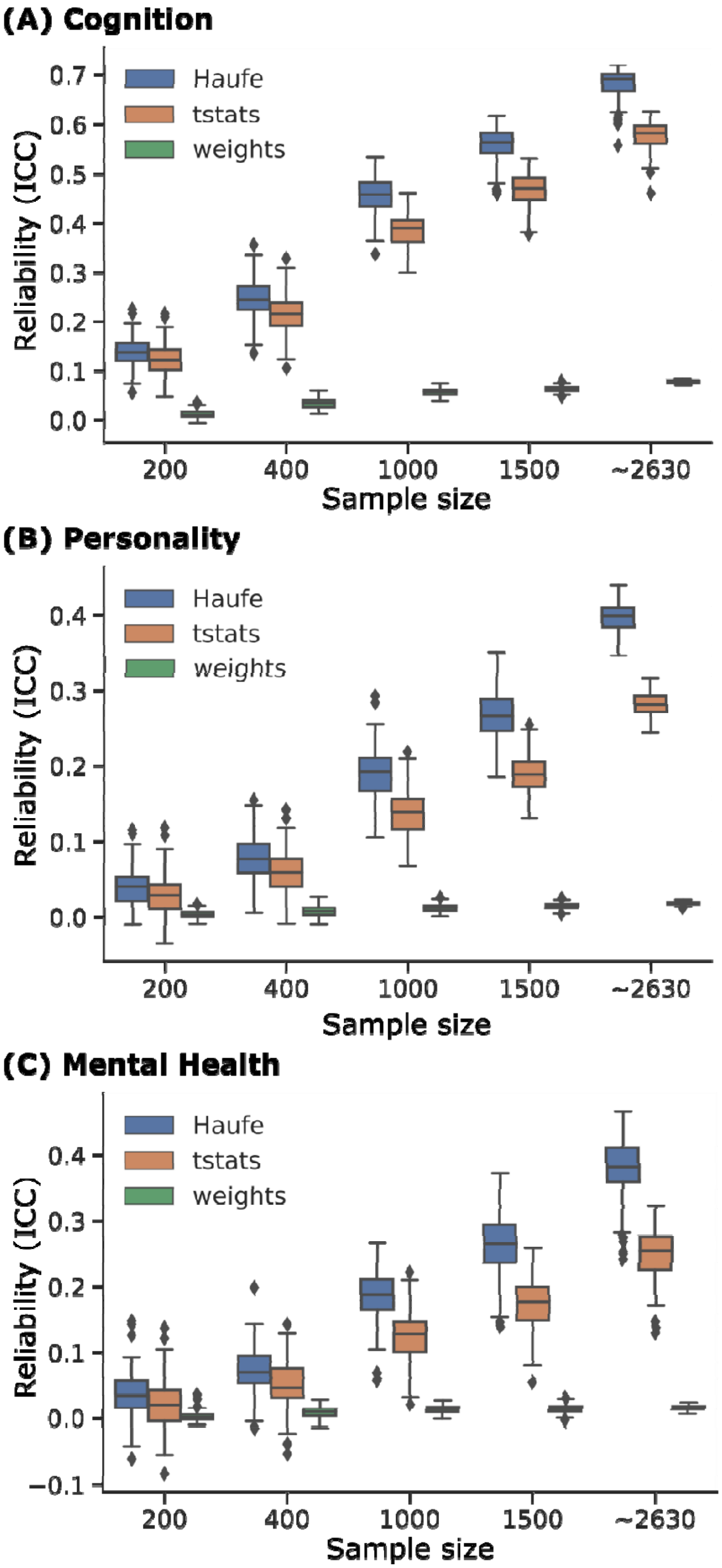

(A) Cognition
Reliability (ICC)
0.8
0.6
0.4
0.2
0.0
Haufe
tstats
weights
200
400
1000
1500
~2630
Sample size
(B) Personality
Reliability (ICC)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Haufe
tstats
weights
200
400
1000
1500
~2630
Sample size
(C) Mental Health
Reliability (ICC)
0.6
0.4
0.2
0.0
Haufe
tstats
weights
200
400
1000
1500
~2630
Sample size

Figure 1. Test-retest reliability of feature importance of kernel ridge regression (KRR) models across different sample sizes, interpretation methods, and behavioral domains: (A) cognition, (B) personality, and (C) mental health. Test-retest reliability was computed as interclass correlation coefficients (ICC) of feature importance obtained from two non-overlapping split-halves of the ABCD participants. After splitting, participants were randomly subsampled to show the effect of sample size on feature importance reliability. Full data without subsampling was reported as a sample size of ~2630. "~" was used because the two halves have similar (but not exactly the same) sample sizes that summed to 5260 (total number of participants). ICC values were reported for Haufe-transformed model weights (Haufe), mass univariate associations (tstats), and original regression weights (weights). Boxplots show the distribution of average ICC within each behavioral domain across 126 split-half pairs. For each boxplot, the box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extend from the box to show the data range (excluding outliers). Outliers are defined as data points beyond 1.5 times the interquartile range and shown as flier points past the whiskers. Overall, across different sample sizes and behavioral domains, Haufe-transformed weights were more reliable than mass univariate associations (tstats), which were in turn more reliable than regression weights. Similar conclusions were obtained with linear ridge regression (Figure 2) and LASSO (Figure S1).


(A) Cognition

(B) Personality

(C) Mental Health

Figure 2. Test-retest reliability of feature importance of linear ridge regression (LRR) models across different sample sizes, interpretation methods, and behavioral domains: (A) cognition, (B) personality, and (C) mental health. Same as Figure 1, except using LRR as the prediction model. Test-retest reliability was computed as interclass correlation coefficients (ICC) of feature importance obtained from two non-overlapping split-halves of the dataset. After splitting, data were randomly subsampled to show the effect of sample size on feature importance reliability. Full data without subsampling was reported as a sample size of ~2630. "~" was used because the two halves have similar (but not exactly the same) sample sizes that summed to 5260 (total number of participants). Note that mass univariate associations (tstats) were computed independent of regression models and are therefore the same across Figures 1 and 2. Overall, across different sample sizes and behavioral domains, Haufe-transformed weights were more reliable than mass univariate associations (tstats), which were in turn more reliable than original regression weights.

3.2 Haufe-transformed weights are highly consistent across prediction models

The previous section investigated the reliability of feature importance across different data samples. Here, we seek to examine the reliability of feature importance across different prediction models in the full sample of 5260 participants. For each split-half of the 5260 participants, we computed the similarity (Pearson's correlation) of feature importance (original regression weights or Haufe-transformed weights) across the 3 prediction models: KRR, LRR, and LASSO.

Figure 3 shows the similarity of feature importance across prediction models. Consistent with Tian and Zalesky (2021), we found that Haufe-transformed weights showed better consistency than the original regression weights. Unlike Tian and Zalesky (2021), because of our significantly larger sample size, excellent consistency was observed for the Haufe-transformed weights (max = 0.97, min = 0.76).

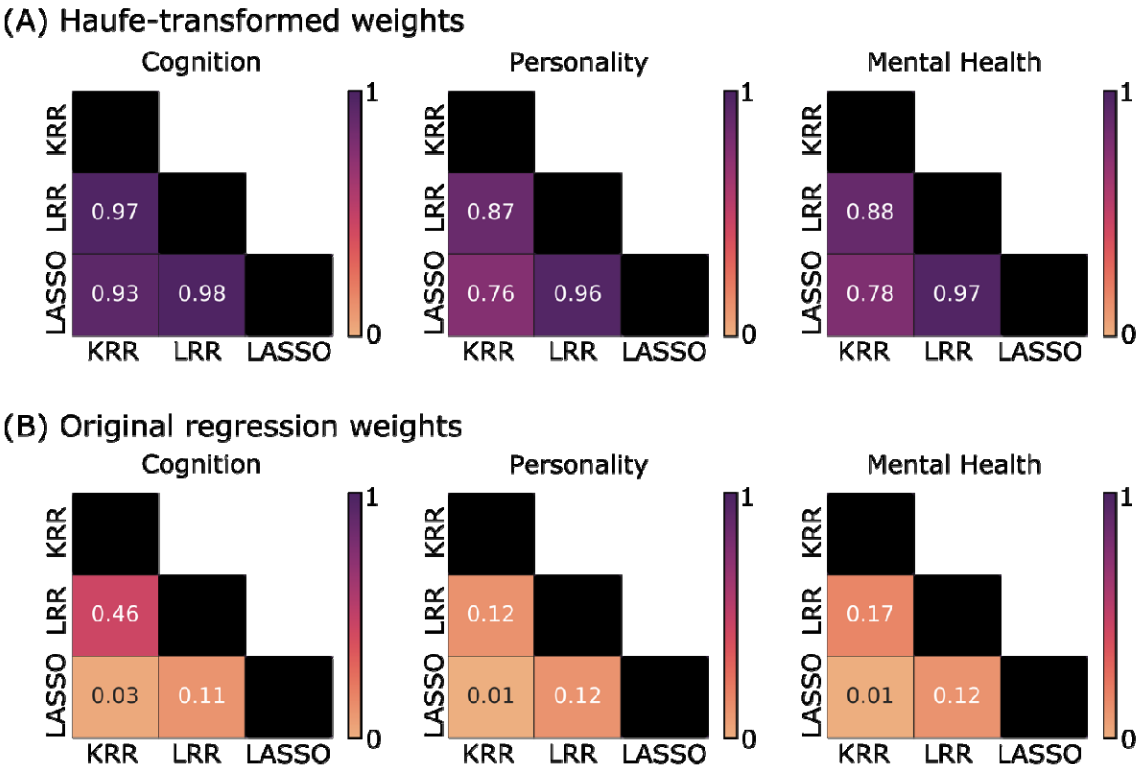


Figure 3. Similarity of feature importance across three predictive models in the full sample of 5260 participants. (A) Consistency of feature importance for Haufe-transformed weights. (B) Consistency of feature importance for original regression weights. Similarity was computed as the Pearson's correlation between the original regression weights (or Haufe-transformed weights) across different predictive models (KRR, LRR and LASSO). Similarity was computed for each split-half and then averaged across the 126 data splits. Excellent consistency was observed for the Haufe-transformed weights.

3.3 Feature importance reliability is strongly positively correlated with prediction accuracy across behavioral measures

So far, our results have been largely consistent with Tian and Zalesky (2021), except our larger sample sizes led to better test-retest reliability of the Haufe-transformed weights. Next, we investigated the relationship between prediction accuracy and test-retest reliability of feature importance using the full sample of 5260 participants.

Test-retest reliability and prediction accuracy of each behavioral score were computed for each split-half of the dataset, followed by averaging across the 126 data splits. Figure 4A shows the correlation between feature importance reliability and prediction accuracy across the 36 behavioral measures for KRR. Prediction accuracy was highly correlated with test-retest reliability of Haufe-transformed model weights (r = 0.78), t-statistics (r = 0.94) and original regression weights (r = 0.97). This suggests that a behavioral measure that was predicted with higher accuracy also enjoyed better feature importance reliability.

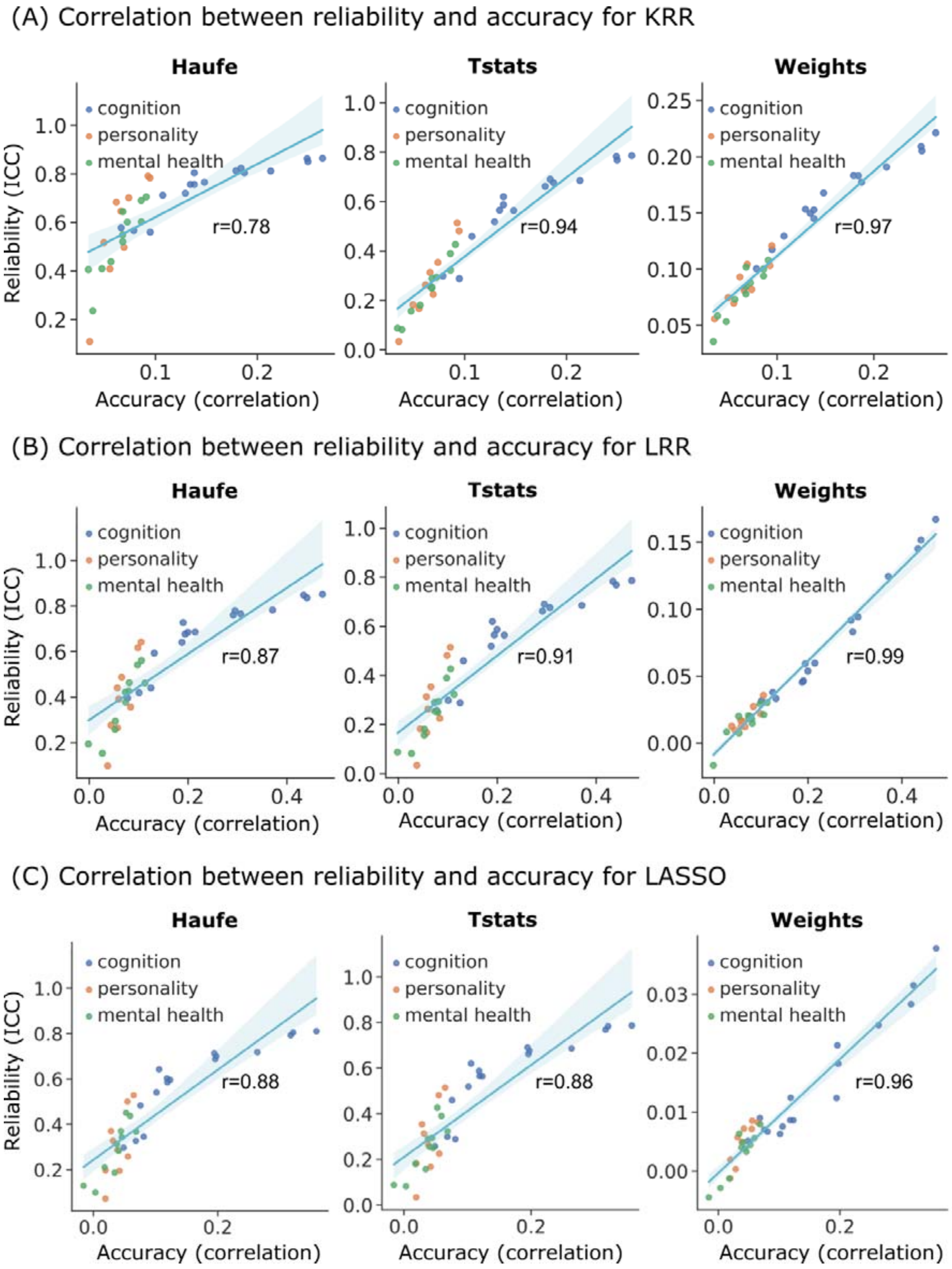
Similar conclusions were obtained with linear ridge regression (Figure 4B) and LASSO (Figure 4C). Overall, we found a strong positive relationship between feature importance reliability and prediction accuracy.

Furthermore, in the case of Haufe transform and univariate associations (t-stats), there appears to be a nonlinear relationship between prediction accuracies and ICC (Figure 4). More specifically, higher accuracies led to greater ICC, but with diminishing returns for behavioral measures with higher accuracies.

3.4 No clear relationship between prediction accuracy and feature importance reliability across predictive models.

Table 1 summarizes average prediction accuracies for cognitive, personality and mental health measures, as well as ICC of Haufe-transformed weights, original weights and univariate association (t-statistics). In general, KRR exhibited the highest ICC, but not necessarily the best prediction performance. LASSO generally had the worse prediction performance and the worst ICC. Finally, LRR exhibited the best prediction performance, but an intermediate level of ICC. Overall, there was no clear relationship between prediction performance and feature importance reliability.

Note that in our other studies (Chen *et al.* 2022, Ooi *et al.* 2022), the prediction performance of KRR was similar to (or slightly better) than LRR, suggesting that depending on the dataset (or even across different samples within the same dataset), prediction accuracies can vary across prediction approaches.


(A) Correlation between reliability and accuracy for KRR

(B) Correlation between reliability and accuracy for LRR

(C) Correlation between reliability and accuracy for LASSO

Figure 4. Test-retest reliability of feature importance is positively correlated with prediction accuracy across 36 behavioral measures for (A) kernel ridge regression (KRR), (B) linear ridge regression (LRR) and (C) LASSO. Test-retest reliability and prediction accuracy of each behavioral score were computed for each split-half of the dataset, followed by averaging across the 126 data splits.

Table 1. Summary of average prediction performance for cognitive, personality and mental health measures, as well as ICC of Haufe-transformed weights, original weights and univariate associations (t-statistics). In general, within a behavioral domain (e.g., cognition), lower (or higher) prediction performance for a given predictive model was not necessarily associated with lower (or higher) ICC.

Cognition	Corr	ICC (Haufe)	ICC (Weights)	ICC (Univariate association)
KRR	0.16	0.75	0.16	0.58
LRR	0.25	0.68	0.08	
LASSO	0.17	0.60	0.02	
Personality	Corr	ICC (Haufe)	ICC (Weights)	ICC (Univariate association)
KRR	0.07	0.57	0.09	0.28
LRR	0.07	0.40	0.02	
LASSO	0.04	0.30	0.01	
Mental Health	Corr	ICC (Haufe)	ICC (Weights)	ICC (Univariate association)
KRR	0.07	0.53	0.08	0.25
LRR	0.07	0.38	0.02	
LASSO	0.04	0.29	0.01	

3.5 Test-retest reliability is necessary, but not sufficient, for correct feature importance

We have shown a strong positive correlation between feature importance reliability and prediction accuracy (Figure 4). There is also a lack of relationship between prediction accuracy across prediction models and feature importance reliability. Overall, this appeared to contradict Tian and Zalesky (2021), who suggested a potential trade-off between feature importance reliability and prediction accuracy. In the remaining sections of this study, we will delve more deeply into the mathematical relationships among feature importance reliability, feature importance error and prediction error.

We begin by showing that test-retest feature importance reliability is necessary but not sufficient for obtaining the “correct” feature importance. Let f_G be the hypothetical ground-truth feature importance that might be derived assuming the correct generative process relating brain features and behavioral measures is known. However, in the following analysis, we do not assume the ground truth generative process is known and we make no assumption about how f_G can be computed even if the ground truth generative process is known.

Let f_S be the feature importance estimated from data sample S. Both f_G and f_S are $D \times 1$, where D is the number of features. The expected feature importance error can be defined as the expectation of the squared error across different data samples S: $E_S[(f_G - f_S)^T(f_G - f_S)]$. Let $\bar{f_S} = E_S[f_S]$ be the feature importance averaged across all possible data samples S. The feature importance error can then be decomposed into two terms:

$$E_S[(f_G - f_S)^T(f_G - f_S)] = \left(f_G - \bar{f}_S\right)^T\left(f_G - \bar{f}_S\right) + E_S\left[\left(f_S - \bar{f}_S\right)^T\left(f_S - \bar{f}_S\right)\right] \qquad (1)$$

The proof is provided in the Appendix. The decomposition of feature importance error as in Eq. (1) is similar in spirit (and derivation) to the classical bias-variance decomposition of prediction error.

The first term $\left(f_G - \bar{f}_S\right)^T\left(f_G - \bar{f}_S\right)$ in Eq. (1) measures the bias of the feature importance estimation procedure. The second term $E_S\left[\left(f_S - \bar{f}_S\right)^T\left(f_S - \bar{f}_S\right)\right]$ measures the variance of the estimated feature importance across different samples, which is the opposite of test-retest reliability. In other words, higher variance in feature importance estimation is the same as lower test-retest reliability. Therefore, from Eq. (1), we note that low feature importance variance (i.e., high feature importance reliability) is necessary but not sufficient for low feature importance error. Low feature importance variance must be coupled with low feature importance bias to achieve a small feature importance estimation error.

3.6 Prediction error reflects feature importance error for linear models

The previous section shows that the test-retest reliability of feature importance is not sufficient for low feature importance error. In this section, we show that when the ground truth data generation model is linear and feature importance is defined as regression weights (or Haufe-transformed weights), then the prediction error is directly related to the feature importance error.

A linear regression model assumes that the data is generated through a linear combination of features. For example, assume that a given data point (x_i, y_i) is generated by a linear model $y_i = x_i^T w_G + \epsilon$. Here, y_i is a scalar, x_i is a $D \times 1$ vector, w_G is the groundtruth $D \times 1$ regression weights, and D is the number of features. ϵ is an independent noise term with zero mean. Without loss of generality, we assume that the expectation of y across data samples is 0 and the expectation of x across data samples is 0 for every feature. In the case of FC prediction of behavioral traits, each data sample is a participant.

Suppose data sample $S = \{(x_1, y_1), \dots (x_N, y_N)\}$ is drawn as the training set. We can then train a linear regression model (e.g., LRR or LASSO) on S and obtain the regression weights w_S. The resulting prediction model will be $\hat{y} = x^T w_S$. Let the difference between the ground truth and estimated weights be $\Delta_w(S) = w_G - w_S$. Thus, the regression weights error (on average across different training sets S) can be defined as $E_S[(w_G - w_S)^T(w_G - w_S)] = E_S[\Delta_w(S)^T \Delta_w(S)]$.

On the other hand, the expected prediction error of the prediction algorithm can be defined as $E_S E_{x,y}[(y - x^T w_S)^2]$. Here, $E_{x,y}$ is the expectation of the squared prediction error over out-of-sample test data points sampled from the distribution of (x, y). We note that the test data points are sampled independently from the sampling of the training dataset S. Then, the expected test error can be decomposed into:

$$E_S E_{x,y}[(y - x^T w_S)^2] = Var(\epsilon) + E_S[\Delta_w(S)^T * COV(X) * \Delta_w(S)] \qquad (2)$$

The proof is found in the Appendix. In Eq. (2) the first term is the irreducible error $Var(\epsilon)$, which is the variance of the noise. The second term $E_S[\Delta_w(S)^T * COV(X) * \Delta_w(S)]$ is determined by both the regression weights error $\Delta_w(S)$ and the covariance of features $COV(X)$.

We can consider three different scenarios for the covariance matrix $COV(X)$. First, suppose $COV(X)$ is an identity matrix, which implies the features are independent and of unit variance. Then, the prediction error (Eq. (2)) can be written as $Var(\epsilon) + E_S[\Delta_w(S)^T \Delta_w(S)]$. Therefore, the prediction error is simply the sum of the regression weights error and the irreducible error.

Second, suppose $COV(X)$ is a diagonal matrix, i.e., $COV(X) = diag(\sigma_1, \sigma_2, \dots, \sigma_d)$, which implies the features are independent. In this case, the prediction error (Eq. (2)) can be written as $Var(\epsilon) + E_S[\sum_{d=1}^{D} \sigma_d \Delta_{w(d)}(S)^2]$. Here, $\Delta_{w(d)}(S)$ is the regression weight error of the d-th feature based on the training dataset S. In this scenario, a bigger regression weights error still leads to a bigger prediction error, but the weights error of features with larger variance results in a larger prediction error than features with small variance.

Third, suppose we do not make any independence assumptions about the features. Since $COV(X)$ is a symmetric matrix, we can decompose $COV(X)$ as $COV(X) = R^T DR$. Here, R is a rotation matrix where $R^T R$ is equal to an identity matrix and D is a diagonal matrix. Then, we can rewrite the prediction error (Eq. (2)) as:

$$Var(\epsilon) + E_S[\Delta_w(S)^T * R^T * D * R * \Delta_w(S)] \tag{3}$$

To summarize the three scenarios for $COV(X)$, regression weights errors of all features contribute to the prediction error, but features contributing more to the variance in the data (up to a rotation) have a bigger impact on the prediction error.

We can also establish a similar relationship between the Haufe-transformed weights error and the prediction error. Note that the Haufe-transformed weights can be computed as $COV(X_S) * w_S$. Here the w_S is the original regression weights and $COV(X_S)$ is the feature covariance of training sample S. Assuming that the sample covariance is close to the true covariance, i.e., $COV(X_S) \approx COV(X)$, then the Haufe-transformed weights error can be written as:

$$E_S[\Delta_w(S)^T * COV(X) * COV(X) * \Delta_w(S)] = E_S[\Delta_w(S)^T * R^T D^2 R * \Delta_w(S)] \tag{4}$$

Comparing the Haufe-transformed weights error (Eq. (4)) with the prediction error (Eq. (3)), we see that the Haufe-transformed weights error is closely related to the prediction error, given that equations 3 and 4 only differ by the square of the diagonal matrix D.

Overall, we conclude that higher original regression weights errors (Eq. (3)) and higher Haufe-transformed errors (Eq. (4)) leads to greater prediction error up to a scaling by the feature covariance matrix.

Discussion

In this study, we provide empirical and theoretical evidence that there is no fundamental trade-off between prediction accuracy and feature importance reliability.

Haufe-transformed model weights are more reliable than original regression weights and univariate FC-behavior correlations

Consistent with Tian and Zalesky (2021), we found that Haufe-transformed weights were more reliable than original regression weights. In our experiments, we note that even with a sample size of ~2630 participants, the original kernel regression weights achieved an ICC of less than 0.2 when predicting cognitive measures, which is less than the ICC of Haufe-transformed weights with a sample size of 200 (Figure 1A). This is perhaps not surprising since it has been empirically shown that regression weights contain more noise than the Haufe-transformed weights (Haufe *et al.* 2014). Furthermore, for predictive models with sparse regularization (e.g., LASSO), it is well-known that noise in the features can lead to very different features being selected, which will lead to low test-rest reliability in the regression weights.

Also consistent with Tian and Zalesky (2021), we found that Haufe-transformed weights were more reliable than univariate brain-behavior correlations. In our experiments, we note that with a sample size of ~2630 participants, the univariate FC-behavior correlations achieved an ICC of less than 0.6 for cognitive measures, which is less than the ICC of Haufe-transformed weights with a sample size of 1000 (Figure 1A). The higher ICC of Haufe-transformed weights over univariate associations is somewhat surprising. A previous study has suggested that the predicted outcomes of predictive models is substantially more reliable than the functional connectivity features themselves (Taxali *et al.* 2021). Here, we speculate that the predicted behavioral measures might even be more reliable than the raw behavioral measures themselves. The reason is that the regularization of many predictive models serves to “shrink” the predicted outcomes towards the population mean, which should increase reliability. If predicted behavioral measures are more reliable than raw behavioral measures, then the covariance of the predicted behavioral measures with FC (i.e., haufe-transformed weights) should be more reliable than the correlation between raw behavioral measures and FC (i.e., univariate associations).

It is also worth mentioning that Tian and Zalesky (2021) found that the ICC of Haufe-transformed weights remained lower than 0.4 across split-half of 800 participants (i.e., two groups of 400 participants), which is consistent with our results (see sample size of 400 in our Figures 1 and 2). Not surprisingly, we obtained higher reliability with larger sample sizes. More specifically, with a sample size of about 2600 participants, Haufe-transformed weights achieve average intra-class correlation coefficients of 0.75, 0.57 and 0.53 for cognitive, personality and mental health measures respectively (Figure 1). Overall, the use of Haufe-transformed weights might help to alleviate reliability issues in neuroimaging studies (Kharabian Masouleh *et al.* 2019, Marek *et al.* 2022).

There is not an empirical trade-off between feature importance reliability and prediction accuracy

We found that behavioral measures that are predicted better also enjoy better feature importance reliability (Figure 4). This appears to contradict Tian and Zalesky (2021), who found that FC-based prediction using lower resolution atlases (compared with higher resolution atlases) had higher feature importance reliability but lower prediction accuracy, thus suggesting a potential trade-off between prediction accuracy and feature importance reliability. While we do not dispute their results, we believe that their conclusion on a trade-off is premature. For example, as can be seen in Figure 2 of Tian and Zalesky (2021), kernel ridge regression enjoyed better prediction accuracy *and* feature importance reliability than connectome-based predictive modelling. In our current study, within a behavioral domain, there was no clear relationship between prediction performance and feature importance reliability across regression algorithms (Table 1). Overall, these empirical results show that it is possible to achieve high prediction accuracy *and* high feature importance reliability, suggesting that there is not necessarily a trade-off between prediction accuracy and feature importance reliability.

There is not a theoretical trade-off between feature importance reliability and prediction accuracy

Eq. (1) shows that feature importance reliability is necessary but not sufficient for obtaining the “correct” feature importance (or low feature importance error). More specifically, feature importance error can be decomposed into a bias term and a variance term, where the variance term is the opposite of feature importance reliability. Consequently, low feature importance variance (i.e., high feature importance reliability) is necessary but not sufficient for low feature importance error.

This result echoes previous studies in neuroimaging (Noble *et al.* 2017), as well as other areas of quantitative research (Kirk and Miller 1986), demonstrating that reliability is not the same as validity. To give an extreme example, if we utilized an extremely strong regularization in our regression models, the regression weights would be driven to zero. In this scenario, the feature importance (regression weights) would be highly reliable across data samples, but the feature importance would not be valid or close to the ground truth values (derived from the ground truth generative process).

In the case of linear models, we further showed in Eq. (2) that higher feature importance error (operationalized by original regression weights) leads to worse prediction accuracy, up to a rotation and scaling by the feature covariance matrix. In Eq. (4), we showed that higher feature importance error (operationalized by Haufe-transformed weights) leads to worse prediction accuracy, up to a scaling of the eigenvalues of the feature covariance matrix.

Overall, these theoretical results suggest that at least in the case of linear models, there is no fundamental trade-off between feature importance reliability and prediction accuracy. In fact, improving prediction performance might even reduce feature importance error and potentially improve feature importance reliability. However, given finite sample sizes, this might not always empirically be true (Table 1).

Acknowledgements

Our research is currently 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), and the USA NIH (R01MH120080). 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 or the Singapore NMRC.

Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development^{SM} (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.

Appendix

Proof of Eq. (1): Bias variance decomposition of feature importance error

In this appendix, we will provide proof of Eq. (1), which decomposes the feature importance error $E_S[(f_G - f_S)^T(f_G - f_S)]$ into a bias term $\left(f_G - \bar{f}_S\right)^T\left(f_G - \bar{f}_S\right)$ and a variance term $E_S\left[\left(f_S - \bar{f}_S\right)^T\left(f_S - \bar{f}_S\right)\right]$.

$$
\begin{aligned}
&E_S[(f_G - f_S)^T(f_G - f_S)] \\
&= E_S\left[\left((f_G - \bar{f}_S) - (f_S - \bar{f}_S)\right)^T\left((f_G - \bar{f}_S) - (f_S - \bar{f}_S)\right)\right] \\
&= E_S\left[\left(f_G - \bar{f}_S\right)^T\left(f_G - \bar{f}_S\right)\right] + E_S\left[\left(f_S - \bar{f}_S\right)^T\left(f_S - \bar{f}_S\right)\right] \\
&\quad -2E_S\left[\left(f_G - \bar{f}_S\right)^T\left(f_S - \bar{f}_S\right)\right] \\
&= \left(f_G - \bar{f}_S\right)^T\left(f_G - \bar{f}_S\right) + E_S\left[\left(f_S - \bar{f}_S\right)^T\left(f_S - \bar{f}_S\right)\right].
\end{aligned}
$$

where the last equality is true because $E_S\left[\left(f_G - \bar{f}_S\right)^T\left(f_S - \bar{f}_S\right)\right] = \left(f_G - \bar{f}_S\right)^T\left(\bar{f}_S - \bar{f}_S\right) = 0$.

Proof of Eq. (2): Relationship between prediction error and regression weights error

In this appendix, we will provide proof of Eq. (2), which establishes the relationship between the prediction error $E_S E_{x,y}[(y - x^T w_S)^2]$ and regression weights error $\Delta_w(S)$, assuming an underlying linear model.

$$
\begin{aligned}
&E_S E_{x,y}[(y - x^T w_S)^2] \\
&= E_S E_{x,y}[(x^T w_G + \epsilon - x^T w_S)^2] \\
&= E_S E_{x,y}[(x^T \Delta_w(S) + \epsilon)^2], \text{where } \Delta_w(S) = w_G - w_S \\
&= E_S E_{x,y}[\epsilon^2] + E_S E_{x,y}\left[\left(x^T \Delta_w(S)\right)^2\right] + 2 * E_S E_{x,y}[\epsilon * x^T \Delta_w(S)]
\end{aligned}
$$

$$= Var(\epsilon) + E_S E_{x,y}[\Delta_w(S)^T x x^T \Delta_w(S)], \text{ because } E_{x,y}(\epsilon * x^T) = E_{x,y}(\epsilon) E_{x,y}(x^T) = 0$$

$$= Var(\epsilon) + E_S[\Delta_w(S)^T * COV(X) * \Delta_w(S)],$$

References

Abrol, A., Fu, Z., Salman, M., Silva, R., Du, Y., Plis, S., and Calhoun, V., 2021. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nature communications*, 12 (1), 353.

Anderson, M. and Anderson, S.L., 2019. How should AI be developed, validated, and implemented in patient care? *AMA journal of ethics*, 21 (2), E125-130.

Auchter, A.M., Hernandez Mejia, M., Heyser, C.J., Shilling, P.D., Jernigan, T.L., Brown, S.A., Tapert, S.F., and Dowling, G.J., 2018. A description of the ABCD organizational structure and communication framework. *Developmental cognitive neuroscience*, 32, 8–15.

Bussone, A., Stumpf, S., and O'Sullivan, D., 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. *In*: *2015 International Conference on Healthcare Informatics*. ieeexplore.ieee.org, 160–169.

Casey, B.J., Cannonier, T., Conley, M.I., Cohen, A.O., Barch, D.M., Heitzeg, M.M., Soules, M.E., Teslovich, T., Dellarco, D.V., Garavan, H., Orr, C.A., Wager, T.D., Banich, M.T., Speer, N.K., Sutherland, M.T., Riedel, M.C., Dick, A.S., Bjork, J.M., Thomas, K.M., Chaarani, B., Mejia, M.H., Hagler, D.J., Jr, Daniela Cornejo, M., Sicat, C.S., Harms, M.P., Dosenbach, N.U.F., Rosenberg, M., Earl, E., Bartsch, H., Watts, R., Polimeni, J.R., Kuperman, J.M., Fair, D.A., Dale, A.M., and ABCD Imaging Acquisition Workgroup, 2018. The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Developmental cognitive neuroscience*, 32, 43–54.

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., and 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.

Cicchetti, D.V., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, 6 (4), 284–290.

Clark, D.B., Fisher, C.B., Bookheimer, S., Brown, S.A., Evans, J.H., Hopfer, C., Hudziak, J., Montoya, I., Murray, M., Pfefferbaum, A., and Yurgelun-Todd, D., 2018. Biomedical ethics and clinical oversight in multisite observational neuroimaging studies with children and adolescents: The ABCD experience. *Developmental cognitive neuroscience*, 32, 143–154.

Cropley, V.L., Tian, Y., Fernando, K., Mansour L, S., Pantelis, C., Cocchi, L., and Zalesky, A., 2021. Brain-Predicted Age Associates With Psychopathology Dimensions in Youths. *Biological psychiatry. Cognitive neuroscience and neuroimaging*, 6 (4), 410–419.

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., Varoquaux, G., and Alzheimer's Disease Neuroimaging Initiative, 2019. Benchmarking functional connectome-based predictive models for resting-state fMRI. *NeuroImage*, 192, 115–134.

Dale, A.M., Fischl, B., and Sereno, M.I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, 9 (2), 179–194.

Diprose, W.K., Buist, N., Hua, N., Thurier, Q., Shand, G., and Robinson, R., 2020. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association: JAMIA*, 27 (4), 592–600.

Dosenbach, N.U.F., Nardos, B., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., Nelson, S.M., Wig, G.S., Vogel, A.C., Lessov-Schlaggar, C.N., Barnes, K.A., Dubis, J.W., Feczko, E., Coalson, R.S., Pruett, J.R., Jr, Barch, D.M., Petersen, S.E., and Schlaggar, B.L., 2010. Prediction of individual brain maturity using fMRI. *Science (New York, N.Y.)*, 329 (5997), 1358–1361.

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., Jr, Watts, R., Garavan, H., Barch, D.M., Nigg, J.T., Petersen, S.E., Dale, A.M., Feldstein-Ewing, S.W., Nagel, B.J., and Dosenbach, N.U.F., 2020. Correction of respiratory artifacts in MRI head motion estimates. *NeuroImage*, 208, 116400.

Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., and Constable, R.T., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*, 18 (11), 1664–1671.

Fischl, B., Liu, A., and Dale, A.M., 2001. Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE transactions on medical imaging*, 20 (1), 70–80.

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., and Dale, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33 (3), 341–355.

Fischl, B., Sereno, M.I., and Dale, A.M., 1999. II: Inflation, Flattening, and a Surface-Based Coordinate System. *NeuroImage*, 9, 195–207.

Fischl, B., Sereno, M.I., Tootell, R.B., and Dale, A.M., 1999. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human brain mapping*, 8 (4), 272–284.

Gabrieli, J.D.E., Ghosh, S.S., and Whitfield-Gabrieli, S., 2015. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron*, 85 (1), 11–26.

Gordon, E.M., Laumann, T.O., Adeyemo, B., Huckins, J.F., Kelley, W.M., and Petersen, S.E., 2016. Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations. *Cerebral cortex* , 26 (1), 288–303.

Gratton, C., Dworetsky, A., Coalson, R.S., Adeyemo, B., Laumann, T.O., Wig, G.S., Kong, T.S., Gratton, G., Fabiani, M., Barch, D.M., Tranel, D., Miranda-Dominguez, O., Fair, D.A., Dosenbach, N.U.F., Snyder, A.Z., Perlmutter, J.S., Petersen, S.E., and Campbell, M.C., 2020. Removal of high frequency contamination from motion estimates in single-band fMRI saves data without biasing functional connectivity. *NeuroImage*, 217, 116866.

Greene, A.S., Gao, S., Scheinost, D., and Constable, R.T., 2018. Task-induced brain state manipulation improves prediction of individual traits. *Nature communications*, 9 (1), 2807.

Greicius, M.D., Srivastava, G., Reiss, A.L., and Menon, V., 2004. Default-mode network activity distinguishes Alzheimer’s disease from healthy aging: evidence from functional MRI. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (13), 4637–4642.

Greve, D.N. and Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48 (1), 63–72.

Hagler, D.J., Jr, 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., and Dale, A.M., 2019. Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. *NeuroImage*, 202, 116091.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., and Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87, 96–110.

He, T., Kong, R., Holmes, A.J., Nguyen, M., Sabuncu, M.R., Eickhoff, S.B., Bzdok, D., Feng, J., and 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.

Hedderich, D.M. and Eickhoff, S.B., 2020. Machine learning for psychiatry: getting doctors at the black box? *Molecular psychiatry*, 26 (1), 23–25.

Hsu, W.-T., Rosenberg, M.D., Scheinost, D., Constable, R.T., and Chun, M.M., 2018. Resting-state functional connectivity predicts neuroticism and extraversion in novel individuals. *Social cognitive and affective neuroscience*, 13 (2), 224–232.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17 (2), 825–841.

Jiang, R., Calhoun, V.D., Fan, L., Zuo, N., Jung, R., Qi, S., Lin, D., Li, J., Zhuo, C., Song, M., Fu, Z., Jiang, T., and Sui, J., 2020. Gender differences in connectome-based predictions of individualized intelligence quotient and sub-domain scores. *Cerebral cortex* , 30 (3), 888–900.

Kennedy, D.P., Redcay, E., and Courchesne, E., 2006. Failing to deactivate: resting functional abnormalities in autism. *Proceedings of the National Academy of Sciences of the United States of America*, 103 (21), 8275–8280.

Kharabian Masouleh, S., Eickhoff, S.B., Hoffstaedter, F., Genon, S., and Alzheimer's Disease Neuroimaging Initiative, 2019. Empirical examination of the replicability of associations between brain structure and psychological variables. *eLife*, 8.

Kirk, J. and Miller, M.J., 1986. Reliability and validity in qualitative research. *SAGE Publications, Inc*.

Kong, R., Li, J., Orban, C., Sabuncu, M.R., Liu, H., Schaefer, A., Sun, N., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., and Yeo, B.T.T., 2019. Spatial Topography of Individual-Specific Cortical Networks Predicts Human Cognition, Personality, and Emotion. *Cerebral cortex* , 29 (6), 2533–2551.

Li, J., Kong, R., Liégeois, R., Orban, C., Tan, Y., Sun, N., Holmes, A.J., Sabuncu, M.R., Ge, T., and Yeo, B.T.T., 2019. Global signal regression strengthens association between resting-state functional connectivity and behavior. *NeuroImage*, 196, 126–141.

Liu, D., Lin, X., and Ghosh, D., 2007. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*, 63 (4), 1079–1088.

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., and Dosenbach, N.U.F., 2022. Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603 (7902), 654–660.

Noble, S., Scheinost, D., and Constable, R.T., 2019. A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *NeuroImage*, 203, 116157.

Noble, S., Spann, M.N., Tokoglu, F., Shen, X., Constable, R.T., and Scheinost, D., 2017. Influences on the Test–Retest Reliability of Functional Connectivity MRI and its Relationship with Behavioral Utility. *Cerebral cortex* , 27 (11), 5415–5429.

Nostro, A.D., Müller, V.I., Varikuti, D.P., Pläschke, R.N., Hoffstaedter, F., Langner, R., Patil, K.R., and Eickhoff, S.B., 2018. Predicting personality from network-based resting-state functional connectivity. *Brain structure & function*, 223 (6), 2699–2719.

Ooi, L.Q.R., Chen, J., Shaoshi, Z., Kong, R., Tam, A., Li, J., Dhamala, E., Zhou, J.H., Holmes, A.J., and Thomas Yeo, B.T., 2022. Comparison of individualized behavioral predictions across anatomical, diffusion and functional connectivity MRI. *bioRxiv*.

Pervaiz, U., Vidaurre, D., Woolrich, M.W., and Smith, S.M., 2020. Optimising network modelling methods for fMRI. *NeuroImage*, 211, 116604.

Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., and Petersen, S.E., 2012. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, 59 (3), 2142–2154.

Power, J.D., Lynch, C.J., Silver, B.M., Dubin, M.J., Martin, A., and Jones, R.M., 2019. Distinctions among real and apparent respiratory motions in human fMRI data. *NeuroImage*, 201, 116041.

Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., and Petersen, S.E., 2014. Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, 84, 320–341.

Price, W.N., 2018. Medical malpractice and black-box medicine. *In*: *Big Data, Health Law, and Bioethics*. Cambridge University Press, 295–306.

Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., and Yeo, B.T.T., 2018. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral cortex* , 28 (9), 3095–3114.

Schulz, M.-A., Yeo, B.T.T., Vogelstein, J.T., Mourao-Miranada, J., Kather, J.N., Kording, K., Richards, B., and Bzdok, D., 2020. Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nature communications*, 11 (1), 4238.

Ségonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., and Fischl, B., 2004. A hybrid approach to the skull stripping problem in MRI. *NeuroImage*, 22 (3), 1060–1075.

Ségonne, F., Pacheco, J., and Fischl, B., 2007. Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. *IEEE transactions on medical imaging*, 26 (4), 518–529.

Shen, X., Finn, E.S., Scheinost, D., Rosenberg, M.D., Chun, M.M., Papademetris, X., and Constable, R.T., 2017. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nature protocols*, 12 (3), 506–518.

Sripada, C., Rutherford, S., Angstadt, M., Thompson, W.K., Luciana, M., Weigard, A., Hyde, L.H., and Heitzeg, M., 2020. Prediction of neurocognition in youth from resting state fMRI. *Molecular psychiatry*, 25 (12), 3413–3421.

Tang, S., Sun, N., Floris, D.L., Zhang, X., Di Martino, A., and Yeo, B.T.T., 2020. Reconciling Dimensional and Categorical Models of Autism Heterogeneity: A Brain Connectomics and Behavioral Study. *Biological psychiatry*, 87 (12), 1071–1082.

Taxali, A., Angstadt, M., Rutherford, S., and Sripada, C., 2021. Boost in Test-Retest Reliability in Resting State fMRI with Predictive Modeling. *Cerebral cortex* , 31 (6), 2822–2833.

Tian, Y. and Zalesky, A., 2021. Machine learning prediction of cognition from functional connectivity: Are feature weights reliable? *bioRxiv*.

Vasey, B., Nagendran, M., Campbell, B., Clifton, D.A., Collins, G.S., Denaxas, S., Denniston, A.K., Faes, L., Geerts, B., Ibrahim, M., Liu, X., Mateen, B.A., Mathur, P., McCradden, M.D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D.S.W., Watkinson, P., Weber, W., Wheatstone, P., and McCulloch, P., 2022a. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* , 377.

Vasey, B., Nagendran, M., Campbell, B., Clifton, D.A., Collins, G.S., Denaxas, S., Denniston, A.K., Faes, L., Geerts, B., Ibrahim, M., Liu, X., Mateen, B.A., Mathur, P., McCradden, M.D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D.S.W., Watkinson, P., Weber, W., Wheatstone, P., and McCulloch, P., 2022b. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature medicine*, 28 (5), 924–933.

Wolfers, T., Beckmann, C.F., Hoogman, M., Buitelaar, J.K., Franke, B., and Marquand, A.F., 2020. Individual differences v. the average patient: mapping the heterogeneity in ADHD using normative models. *Psychological medicine*, 50 (2), 314–323.

Xia, C.H., Ma, Z., Ciric, R., Gu, S., Betzel, R.F., Kaczkurkin, A.N., Calkins, M.E., Cook, P.A., García de la Garza, A., Vandekar, S.N., Cui, Z., Moore, T.M., Roalf, D.R., Ruparel, K., Wolf, D.H., Davatzikos, C., Gur, R.C., Gur, R.E., Shinohara, R.T., Bassett, D.S., and Satterthwaite, T.D., 2018. Linked dimensions of psychopathology and connectivity in functional brain networks. *Nature communications*, 9 (1), 3003.

Xiao, Y., Lin, Y., Ma, J., Qian, J., Ke, Z., Li, L., Yi, Y., Zhang, J., Cam-CAN, and Dai, Z., 2021. Predicting visual working memory with multimodal magnetic resonance imaging. *Human brain mapping*, 42 (5), 1446–1462.

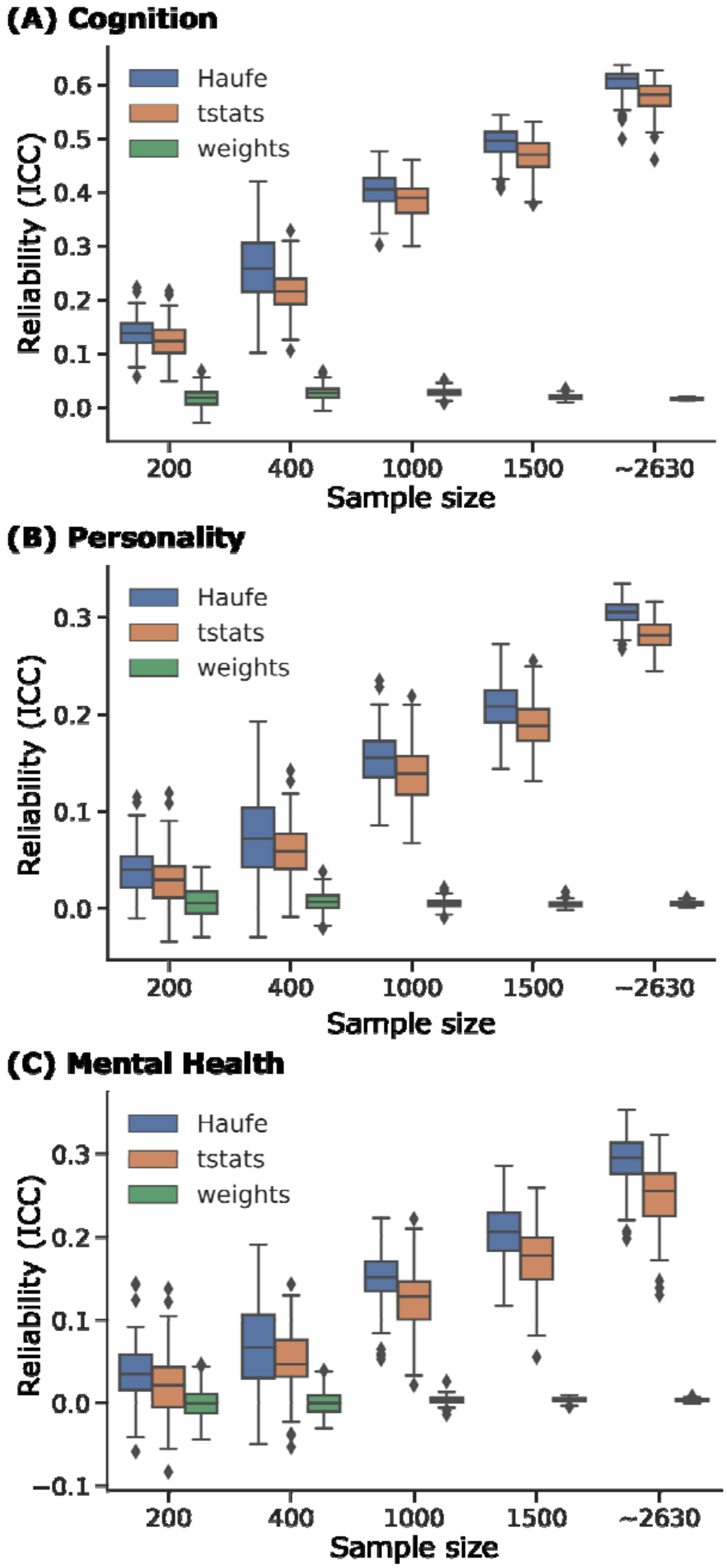
Zabihi, M., Oldehinkel, M., Wolfers, T., Frouin, V., Goyard, D., Loth, E., Charman, T., Tillmann, J., Banaschewski, T., Dumas, G., Holt, R., Baron-Cohen, S., Durston, S., Bölte, S., Murphy, D., Ecker, C., Buitelaar, J.K., Beckmann, C.F., and Marquand, A.F., 2019. Dissecting the heterogeneous cortical anatomy of autism spectrum disorder using normative models. *Biological psychiatry: cognitive neuroscience and neuroimaging*, 4 (6), 567–578.

Zhang, X., Mormino, E.C., Sun, N., Sperling, R.A., Sabuncu, M.R., Yeo, B.T.T., and Alzheimer’s Disease Neuroimaging Initiative, 2016. Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer’s disease. *Proceedings of the National Academy of Sciences of the United States of America*, 113 (42), E6535–E6544.

Supplementary Materials

Table S1. Distribution of the included samples (n=5260) by site and scanner

ABCD Site	**Make**	**Model**	**N**	**Site-cluster**
16	Siemens	Prisma	631	A
13	GE	Discovery MR750	389	B
4	GE	Discovery MR750	452	C
22	GE	Discovery MR750	26	C
14	Siemens	Prisma/Prisma fit	287	D
15	Siemens	Prisma fit	190	D
6	Siemens	Prisma fit	311	E
9	Siemens	Prisma fit	201	E
10	GE	Discovery MR750	403	F
11	Siemens	Prisma	228	F
3	Siemens	Prisma	356	G
5	Siemens	Prisma fit	183	G
2	Siemens	Prisma fit	220	H
7	Siemens	Prisma fit	172	H
8	GE	Discovery MR750	184	I
20	Siemens	Prisma/Prisma fit	286	I
12	Siemens	Prisma fit	283	J
18	GE	Discovery MR750	215	J
21	Siemens	Prisma fit/Prisma	251	J


(A) Cognition

(B) Personality

(C) Mental Health

Figure S1. Test-retest reliability of feature importance of LASSO models across different sample sizes, interpretation methods, and behavioral domains: (A) cognition, (B) personality, and (C) mental health. Same as Figure 1, except using lasso as the prediction model. Test-retest reliability was computed as interclass correlation coefficients (ICC) of feature importance obtained from two non-overlapping split-halves of the dataset. After splitting, data were randomly subsampled to show the effect of sample size on feature importance reliability. Full data without subsampling was reported as a sample size of ~2630. "~" was used because the two halves have similar (but not exactly the same) sample sizes that summed to 5260 (total number of subjects). Note that BWA t-statistics (tstats) were computed independent of regression models and are therefore the same across Figures 1, 2 and S1. Overall, across different sample sizes and behavioral domains, Haufe-transformed weights were more reliable than BWA t-statistics, which were in turn more reliable than regression weights.