The Challenges and Prospects of Brain-based Prediction of Behavior

Jianxiao Wu[1,2*], Jingwei Li[1,2], Simon B. Eickhoff[1,2], Dustin Scheinost[3,4,5,6,7], and Sarah Genon[1,2*]

[1]Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Center Jülich, Jülich, Germany

[2]Institute for Systems Neuroscience, Medical Faculty, Heinrich-Heine University Düsseldorf, Düsseldorf, Germany

[3]Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, USA

[4]Department of Statistics and Data Science, Yale University, New Haven, USA

[5]Child Study Center, Yale School of Medicine, New Haven, USA

[6]Interdepartmental Neuroscience Program, Yale School of Medicine, USA

[7]Department of Biomedical Engineering, Yale School of Engineering and Applied Sciences, New Haven, USA

* Corresponding Author. Email address: j.wu@fz-juelich.de; s.genon@fz-juelich.de

Abstract

Relating individual brain patterns to behavior is fundamental in system neuroscience. Recently, the predictive modeling approach has become increasingly popular, largely due to the recent availability of large open datasets and access to computational resources. This means that we can use machine learning models, and interindividual differences at the brain level represented by neuroimaging features to predict interindividual differences in behavioral measures. By doing so, we could identify biomarkers and neural correlates in a data-driven fashion. Nevertheless, this budding field of neuroimaging-based predictive modelling is facing issues that may limit its potential applications. Here, we review these existing challenges, as well as those that we anticipate as the field develops. We focus on the impact of these challenges on brain-based predictions. We suggest potential solutions to address the resolvable challenges, while keeping in mind that some general and conceptual limitations may also underlie the predictive modeling approach.

The study of the relationships between individual differences in brain phenotypes and individual behaviors is fundamental in neuroscience, both from a basic scientific perspective and an applied perspective. The term 'predictive modeling' refers to the use of machine learning techniques to build a statistical model for the estimation of behavioral variables from brain-based neuroimaging data, either structural or functional[1,2]. More precisely, a prediction model is trained to predict particular behavioral variables from brain-based data from a number of individuals (the training set), and its performance is then evaluated on unseen data (test set).

The potential practical applications promised by such prediction approaches in precision medicine, healthcare, human resources and education[1,3-5] are certainly exciting. Potential future applications may include prediction of individual treatment outcomes to guide treatment choices and dosage, classification of clinical subgroups with different brain pathology and thus different treatment requirements, as well as prediction of future cognitive abilities and mental health at developmental stage. As concrete examples that could be envisioned, brain-based predictions may provide objective biomarkers when evaluating the effect of cognitive training or cognitive-behavior therapies (e.g., for mild functional cognitive alterations and anxio-depressive phenotypes, respectively). While the effect of these interventions could be more readily investigated with standard cognitive tests and interview/questionnaires respectively, such approaches are prone to many biases (e.g., practice effects, subjectivity biases, expectations biases). As a recent working example, a prediction model of sustained attention provided a neuromarker of sustained attention[6]. This neuromarker can be used both for predicting attention deficit symptoms, and for localizing targets of potential brain-based treatments. Ultimately, brain-based prediction could be expected to provide objective biomarkers that can inform us about the brain mechanisms behind the effects under scientific investigations. Aided by the publicly available large neuroimaging datasets, accessible computational resources, as well as code sharing practices, predictive modeling has become a powerful tool towards these future outlooks.

Among the various types of neuroimaging data, functional data may be an intuitive choice for relating brain organization to behavioral functions. In particular, task-free resting-state functional Magnetic Resonance Imaging (rs-fMRI) scans can be readily collected for large groups of subjects[7], making them popular choices for neuroimaging-based predictions. In the last ten years, RSFC has been the most popular input features to brain-behavior prediction models[2], in predictions of various phenotypes including fluid intelligence[8-10], attention[6,11,12], and working memory[13-15]. Brain-based psychometric prediction using other features such as task-based functional connectivity, gray matter volume, cortical thickness, and structural connectivity has also been investigated in predictions of general cognitive abilities[16-18], attentional control[19], and working memory[20,21]. However, and although this may be expected to change in the future, as far, the majority of studies forming the scientific literature have used RSFC alone or in combination with other features, for psychometric prediction[2].

As a budding and growing field, brain-based psychometric predictions remain to be improved and validated. Many reviews have analyzed methodological options based on the current state of the field and given guidance for future studies[1,2,4,5,22-24]. Practical tutorials have also been published for guidance on specific implementation details[22,25]. Nevertheless, the field also faces general and conceptual issues that are likely to limit the future usefulness of predictive modeling.

In this review, we discuss the current and anticipated future challenges in psychometric prediction based on neuroimaging features. For each challenge, we identify both inherent limitations in brain-based psychometric predictions which may not be readily solved based on current resources and aspects that could be addressed with potential solutions. In the following sections, we discuss the general challenges of low prediction accuracies, followed by two core issues, generalizability and interpretability. Finally, we briefly discuss the potential vulnerability of brain-based prediction models to enhancement and adversarial attacks.

Low prediction accuracies

Low prediction accuracies limit any potential application of the model. The general procedure of prediction model development and validation is described in Fig 1. A prediction model is assessed by applying it in a validation sample separate from the training sample, and by measuring the similarity or dissimilarity between the values predicted for the subjects in this sample and the truly observed values of the psychometric variable for these subjects (Box 1). Fig. 2 shows three examples of the most commonly used measure of model accuracy (Pearson's correlation coefficient), and the predicted-observed relationships underlying the accuracies. This measure indicates the global linear trend between predicted and observed values, but cannot identify systematic biases and size of errors. Presently, prediction accuracies of various psychometric variables have been reported from as low as 0.06 to as high as 0.908[1,2]. This wide range of accuracies with both low and high values close to the value bound reflects the complexity of brain-based psychometric prediction study design, as model accuracy can be affected by methodological decision and data characteristics (e.g., the amount of relevant variance in behavioral and/or brain data). While many studies that showed high prediction accuracies also appear to have used very small samples, in studies using large samples, the prediction accuracies are usually reported in the range of 0.2 to 0.4[26-29], implying a generally lower accuracy when evaluating brain-based predictions in population-representative samples. A recent literature survey have evidenced a correlation of r=-0.265 between the sizes of the training sample and the reported prediction accuracies, demonstrating the generality of this trend.

While big data and deep learning has enabled substantial successes in many fields, neither has been particularly helpful in improving the performance of brain-based prediction models. To begin with, even the easy-to-collect rs-fMRI data are considerably more difficult to collect than pictures or texts typically used in the field of computer vision and natural language processing, respectively. The lack of truly big data in cognitive neuroscience may explain why deep learning has often been reported to not outperform simpler models[1,24,27]. The potential of deep learning as more powerful models would thus depend on the possibility of collecting truly big neuroimaging datasets.

Alternatively, techniques such as few-shot learning could inspire new solutions to utilize deep learning without acquiring big data. From the data perspective, the few-shot learning strategy called data augmentation can be employed to artificially increase the sample size. Furthermore, simulated rs-fMRI and RSFC data have been used to generate additional datasets recently[30-33].

Their applications for predictive modeling of behavior remain to be further investigated. From the parameter perspective, the meta-learning paradigm of few-short learning can be useful by training a generalized model on a large dataset, which can be used for prediction of different targets in smaller datasets[34]. Nevertheless, both strategies impose some requirements and may not appear beneficial for all types of brain-based predictions. Augmented or simulated data are limited by the characteristics of the existing data used for augmentation or simulation. Accordingly, a nonrepresentative dataset (e.g., including only a certain age group or ethnicity) cannot become population representative through augmentation. As for the meta-learning strategy, its performance depends on the similarity of the prediction target in the large dataset and the prediction target in the smaller dataset[34]. This means that the meta-learning model would only be beneficial for smaller datasets which use the same or very similar instrument for behavioral measurement as those existing in the larger datasets in which the original model is developed.

It may instead be more feasible to capitalize on existing data, including neuroimaging features from multiple modalities, to boost prediction accuracies. Structural, functional, and diffusion MRI probe different neurobiological aspects, offering complementary information for psychometric prediction. In prediction studies based on functional MRI, resting-state and task functional MRI features are often combined[13,35-37]. However, the benefit of combining these features in terms of prediction performance has not been comprehensively investigated. Prediction studies using multimodal data have found different type of features to contribute to the prediction, including local connectome[18], cortical area[18], cortical thickness[17], gray matter volume[21], RSFC[17,38-40], and task functional connectivity[39,41]. Some studies reported that integrating multimodal MRI data did not actually improve the prediction performance than using a single modality[21,39]. Furthermore, combining multimodal features inevitably increases the feature dimension and in turn the risk of overfitting, requiring feature selection or reduction techniques, such as stacking[18,38,41]. Generally, a systematic evaluation of multimodal psychometric prediction across multiple distinct cohorts, with an extensive set of neuroimaging features, psychometric measures, and model design, would be an important next step for validating this research direction.

Moreover, psychometric prediction accuracies are dependent on the target psychometric variable to predict. For behavioral traits in cognition and socioaffective domains, the definition of the abstract constructs measured by many behavioral variables and relatedly the construct validity of these variables are still debated[42-44]. The reliability and validity of these behavioral traits require improvement through both theoretical and experimental validations. Interestingly, many studies have reported higher prediction accuracies for cognitive measures compared to mental health traits[5,34,38,45,46]. It may be assumed that prediction of mental health would be particularly difficult in healthy population because the participants would show very limited variations in mental health measures. As the largest and highest neuroimaging quality datasets open to the research community include mainly healthy population, studies attempting to develop predictive models of mental health may be limited either by data availability and quality for clinical populations, or lower prediction accuracies when using easily accessible data. Relatedly, low test-retest reliability of functional MRI measures may be another source of poor prediction accuracies[47,48]. As the reliability of connectivity features computed may depend on data collection protocols[49-51], the selection of reliable data would further restrict the available sample size.

One pessimistic view is that current modeling approaches may not be able to handle the heterogeneity of the population-representative samples, or that the brain-behavior relationships captured in neuroimaging datasets may simply be too weak[52-55]. Neuroimaging patterns may be a reduced summary of endogenous factors and the exposome that has a limited power to explain interindividual variability in behavior. Crucially, brain-based prediction models need to be justified based on the additional predictive power not already provided by non-neuroimaging features that can be easily collected by questionnaires and interviews, especially considering the high cost of MRI. From a practical standpoint, it may be useful to investigate the prediction performance of hybrid models making use of all types of data available in a realistic situation. For instance, RSFC patterns may vary with age due to developmental effects in younger population and due to aging in older population. Similarly, cognitive measures may be affected by age differently in different age subgroups. Allowing the prediction model to learn these interactions across a large age range would thus help the model to predict the target variable more accurately in general. Finally, the variability of brain-behavior association patterns across different subgroups brings forth another crucial challenge: model generalizability.

Generalizability of prediction models

The utility of a prediction model depends on its generalizability. That is, its ability to make accurate predictions on unseen data, firstly the test set data and ultimately data from the broader population. In the context of brain-based psychometric predictions, we discuss generalizability both in terms of generalizing to completely new cohorts and of generalizing to different subgroups of the population.

Cross-cohort generalizability

Cross-cohort generalizability can be defined as the prediction performance of a model in a different dataset from the training dataset (Fig. 1). Generalizable models are important both for discovering neurobiological insights general to the population and for deploying prediction models to broader settings. In most present brain-based prediction studies, the training and test sets are drawn from the same cohort under a cross-validation scheme[2]. While cross-validation helps to evaluate model performance without requiring additional datasets, to rigorously test the cross-cohort generalizability of a model, it is necessary to evaluate the model on completely unrelated datasets. Among studies which employed both internal and external validation, many studies found similar prediction accuracies in internal and external test sets[9,13,56-59]. Nevertheless, most of these studies had small external test samples ($N < 200$), calling into question the representativeness of these test cohorts. In two studies with large test cohorts ($N \sim 1000$), drops in prediction accuracies were observed when generalizing to new cohorts[26,60]. It has been suggested that reproducible brain-behavior association may only be found using samples with thousands of participants[55,61]. However, it has also been shown that generalizable associations and predictions can be achieved with much smaller samples in some specific cases[62,63]. Additionally, it should be noted that generalizability of the statistical model is not a direct indication of the generalizability of brain-behavior association derived from the model, the latter showing a low to moderate extent of generalizability across cohorts[60].

At present, the main challenge from the perspective of cross-cohort generalizability is the lack of awareness from scientific investigators and hence the lack of assessment. The need for large external test cohorts for evaluating prediction models is often overlooked during the planning phase of a study, and later dismissed on the grounds that such large cohorts are not available for the specific psychometric measure investigated. More generally, cross-cohort generalizability of prediction models may be affected and limited by the similarity of data collection and processing protocols in the different cohorts[60]. The need for large datasets has led to researchers' reliance on whatever data is provided by the several publicly (or semi-publicly) shared datasets. Many studies have trained and evaluated prediction models using the Human Connectome Project Young Adult data, which were processed with a specific pipeline not always adopted or viable in other datasets[64,65]. Ideally, standardizing data collection protocols and processing pipelines would improve model generalizability in both research and practical situations. However, imaging conditions in samples involving children or older adults would often result in lower scan duration, making it difficult to achieve the same standards that can be set in healthy young samples[66]. The need for large cohorts and varied data specification may not be fully reconcilable. Partial solutions would be more robust preprocessing strategies and prediction models to harmonize data differences or to extract generalizable information despite the data differences.

Generalizability across subgroups (within a single dataset)

Typically, the test set for evaluating a prediction model is randomly selected from the cohort or taken from an external validation cohort. The composition of the test set may be completely random or stratified for balanced distributions of age, gender, and other variables of interest. While the model performances reflect the average performance in the test cohort population, they are not informative of potential prediction biases between test subjects. In both medical and non-medical applications, model bias has been reported for potential mistreatments of subgroups based on gender, ethnicity and socioeconomic status[67-69]. In connectivity-based prediction, ethnicity-based bias has been reported where prediction accuracies were lower in African American subjects in comparison to White American subject, even if models were trained on only African American subjects[70]. Moreover, models tend to predict lower cognitive scores and higher negative social behavior scores for African American subjects[70], demonstrating the potential biases in applications of the prediction models. Such robust biases call for more balanced samples in scientific approaches, including not only more data collection in underrepresented population, but also the development of brain templates, atlases, and preprocessing tools based on balanced samples.

Common concepts used to define population subgroups like gender and ethnicity are complex notions themselves often entangled with socioeconomic factors. Relatedly, brain-based prediction models do not see the population divided into distinct gender-based or ethnic groups but have been shown to learn complex profiles relating brain measures, covariates, and psychometric variables[70]. It was recently demonstrated that individuals that do not follow the majority trend of brain-phenotype relationships in the training sample can cause consistent prediction failure[71]. For instance, if most older subjects in the training sample scored lower for a cognitive test, a few older subjects in the validation sample who scored high for the cognitive test would become outliers and lead to prediction failures. In other words, model bias may be caused by any form of stereotypical brain-behavior relationships in the training sample, not specific to an ethnic or gender group. This could lead to further difficulty in collecting balanced samples since these stereotypical relationships can hardly be anticipated during data collection phase.

In the case where differences in brain-behavior relationships can be assumed across different subgroups in the sample, group-specific models have been used to improve prediction accuracies within certain subgroups or provide insights into the differences in brain-behavior association across subgroups[17,37,72,73]. Nevertheless, the validity and potential bias in subgroup definition, for instance ambiguity in ethnicity reporting, could limit the validity of any insights generated. Furthermore, brain-behavior relationships inferred from group-specific models should not be simplified in terms of causal relations with the subgroups, lest we fall into the trap of model bias and mistreatment again[70]. Alternatively, an ensemble learning technique called boosting may be useful for capturing different brain-behavior relationships without defining subgroups. In boosting, a sequence of models is trained where each model assigns more importance to subjects that were wrongly predicted by previous models, thereby automatically identifying the outlying subjects.

From a basic neuroscience perspective, the insights gained from a biased prediction model may lead to false conclusions regarding behavior and social identities, while from a practical perspective, a biased model deployed for social applications would easily lead to inequitable treatment of target populations. In order to develop a fair prediction model, both dedicated study design and model transparency are vital. This hence calls for more population-representative samples, clearly documented study and model parameters, as well as interpretable models.

Model interpretability

While accuracy and generalizability are requirements of any predictive model, interpretability is another crucial goal, if less easy to quantify. From a basic neuroscience perspective, prediction models need to be interpretable to contribute to our knowledge about brain-behavior relationships, while from a practical perspective, interpretability is required to evaluate the neurobiological validity of the model and, relatedly, its trustworthiness. A model with lower accuracy but higher interpretability may be preferred to a black-box model with higher accuracy, as the transparency of the former model allows assessments of the model trustworthiness. For instance, model bias against an ethnic minority could be identified earlier if the model can be interpreted easily. Nonetheless, achieving good model interpretability is not trivial and sometimes requires compromise in prediction performance[22].

Many early studies provided an illusion of interpretability by treating regression weights from machine learning models as feature importance for neuroscientific interpretation. Later studies have demonstrated that these weights are neither stable across cross-validation folds[46,74], nor conceptually valid as brain feature importance[75]. It may still be possible to interpret the regression weights after transforming them into corresponding forward model weights using the Haufe transform[75]. While stable predictive networks may be identified for cognition[45], the stability in cross-validation and generalizability to new cohorts of the transformed weights were still reported to be low[60,74]. The reliability of transformed weights may improve with larger sample size[76], making this technique potentially suitable in large cohorts. Nevertheless, when using functional connectivity as features, it may be difficult to align the connectivity edges to brain mapping literature, or to summarize the feature importance values into practically useful information. Feature importance of connectivity edges may be more easily visualized and interpreted by grouping the connectivity edges in networks or finding the top connections. For instance, Fig. 3a shows groups of important connections for predicting cognition within the visual network, within the default mode network, as well as between the default mode network and other networks, while Fig. 3b shows that the most important connections for predicting fluid intelligence tend to be cross-hemispheric between medial regions or between temporal regions.

Many other solutions have been proposed for interpreting prediction models. Using a feature-dropping concept used in random forests[77], feature importance for each feature can be quantified as the decrease in prediction performance when that feature is removed from the feature set[62,78-80]. This has been sometimes referred to as a 'virtual lesion' approach in the computational neuroimaging field. Such simple implementations may not, however, scale well to large feature sets as each feature is delt with independently. Alternatively, using sparse regression models, only a small subset of features is selected by the regression algorithm for prediction. This leads to an inbuilt binary interpretation where only the small set of selected features is considered important. For instance, Fig. 3c shows the feature importance assignment for predicting novelty seeking by a sparse algorithm, helping the model interpretations to focus on frontal-subcortical, parietal-frontal, and within-frontal connections. This approach identifies predictive features in a data-driven manner, albeit limited to research questions where sparsity can be safely assumed. When using highly correlated features like functional connectivity, some algorithms may fail to include all important features that are correlated to each other[81]. Considering a large set of features without feature selection, it may still be possible to assess feature importance using Shapley Additive exPlanation (SHAP)[82,83]. SHAP determines each feature's contribution similar to the 'virtual lesion' approach, but in all possible subsets of features, providing a distribution of feature importance for each feature. Finally, using a recently proposed region-wise framework, each brain region's features set can be assessed instead of individual features. Concretely, a region-wise model is trained and tested to provide a model accuracy specific to the brain region[60]. Interpretations based on region-wise models are easy to illustrate (Fig. 3d) and to some extent align with the brain mapping literature. Nevertheless, the distributed aspect of brain organization is not modeled by the region-wise models, limiting strong interpretations to mostly region-specific properties.

Ultimately, useful model interpretations are reliant on the prediction accuracy and generalizability of the model. With very low accuracies, the interpretations generated from the models may be arbitrary at best, while with low generalizability the interpretations may be valid only for the training sample. The challenge hence lies in designing models where interpretability can be achieved with minimal or no compromise in accuracy. Potential directions may include more powerful generative models, more informative priors, and interpretable deep neural networks. Generative models and deep neural network models may be combined into deep generative models to bring forth the benefits of both interpretability and accuracy, with specific approaches including variational autoencoders[84], generative adversarial networks[85], and autoregressive models[86,87]. With traditional machine learning models, feature importance based on existing models can help to reduce feature dimensionality in new models in new cohorts, which offers new interpretations to validate against the existing model's interpretation. In this way, a positive reinforcement loop may exist between boosting prediction accuracy and interpretability, reducing the need to sacrifice one for the other.

Enhancement and adversarial attacks

Enhancement and adversarial attacks can threaten the trustworthiness of neuroimaging-based predictive models. Enhancement attacks are those where purposeful data alterations can lead to falsely enhanced model performance, while adversarial attacks are those where specifically designed noise are added to the data to cause a model to fail[88]. An artificially enhanced model may be the result of scientific malpractice or fraud which, if not discovered, could lead to large amount of time and resources wasted in the wrong research direction. Successful adversarial attacks on deployed models mean that prediction outcomes would become unreliable. In biomedical development for example, the effect of a treatment or of a drug could be faked or exaggerated with data manipulation of the machine learning model to mislead financial investors. . Similar to the issue of generalizability, the main challenge of these attacks in the field of neuroimaging-based psychometric prediction is the lack of awareness. As practical applications of neuroimaging-based prediction models are still far-fetched at present, there is a lack of motivation for researchers to anticipate models that are robust to these attacks. Furthermore, replication studies that might detect enhancement attacks are still rather lacking in the field. While there is no evidence of existing enhancement or adversarial attacks in the field and no practical solution proposed against them currently, these are crucial issues to address in the perspective of the deployment of brain-based prediction models for applications in the society.

Simple data enhancement can be done by biased subject selection. Subjects may be retroactively selected based on their individual prediction outcome, or only chosen if they follow certain brain-phenotype stereotypes. Such manipulations can be detected if data characteristics and exclusion criteria are reported faithfully, especially when outliers are excluded based on a threshold. A more advanced approach involves adding patterns correlated to the behavioral variable of interest to the imaging features, boosting the prediction accuracies to almost perfect accuracies without causing the features to become significantly different from the original features[88]. Furthermore, it is possible to design data enhancements to cause machine learning models to learn brain-behavior relationships not existing in the original data. This means that enhancement attacks may also be detrimental from a basic neuroscience perspective as conclusions drawn would not be valid. This type of attack may be detected when a replication study fails to generalize the model to new cohorts but can only really be confirmed if the raw data and data processing code can be openly examined.

The effects of adversarial attacks in machine learning models for clinical applications have been investigated[89,90]. For brain-based prediction models in healthy population, it has also been shown that very minor data manipulations can cause the classification accuracy to drop to 0%[88]. To design this type of attack, the model parameters must be known, hence bringing forth additional challenges in achieving both open science and practical utility. Data validation to identify manipulated data, if possible, may become paramount in the future of adversarial attacks.

Potentially, machine learning models employed in practical applications can make use of online learning where a trained model continues to receive new batches of data for additional training, while only sharing the model at baseline for scientific purposes.

In face of potential enhancement and adversarial attacks, model and study reproducibility enabled by open science is necessary to detect and address these data manipulations. With transparent study design and provenance tracking, the field can benefit from multiple aspects including easier replication, enhancement attack monitoring, comparison across studies, and results pooling[22,88].

Conclusions

Many challenges lie in the way of brain-based predictive modeling of behavior before it can be substantially useful for understanding complex brain-behavior relationships or for practical applications. While some limitations are inherent, such as smaller sample sizes in studies interested in phenotypic measure uncommon in large open datasets, others are solvable, such as assessment and improvement of generalizability. By acknowledging this and addressing the solvable issues, brain-based psychometric predictions can steadily progress towards scientific and practical utility. We encourage more comprehensive study design, comprising multiple cohorts to cover more population-representative samples, and ensuring model validity with careful confound handling. Furthermore, we advocate for model evaluation based on both accuracies and generalizability. Predictive modeling in neuroscience is a necessarily interdisciplinary field, which requires combinations of neuroscientific knowledge, statistical concepts, and machine learning techniques to achieve its potential. Beyond this interdisciplinarity, transparent models, diverse data, and rigorous study designs are the keys to move forward.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, GE 2835/2–1, EI 816/4–1), the Helmholtz Portfolio Theme ‘Supercomputing and Modelling for the Human Brain’ and the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement No. 720270 (HBP SGA1) and Grant Agreement No. 785907 (HBP SGA2).

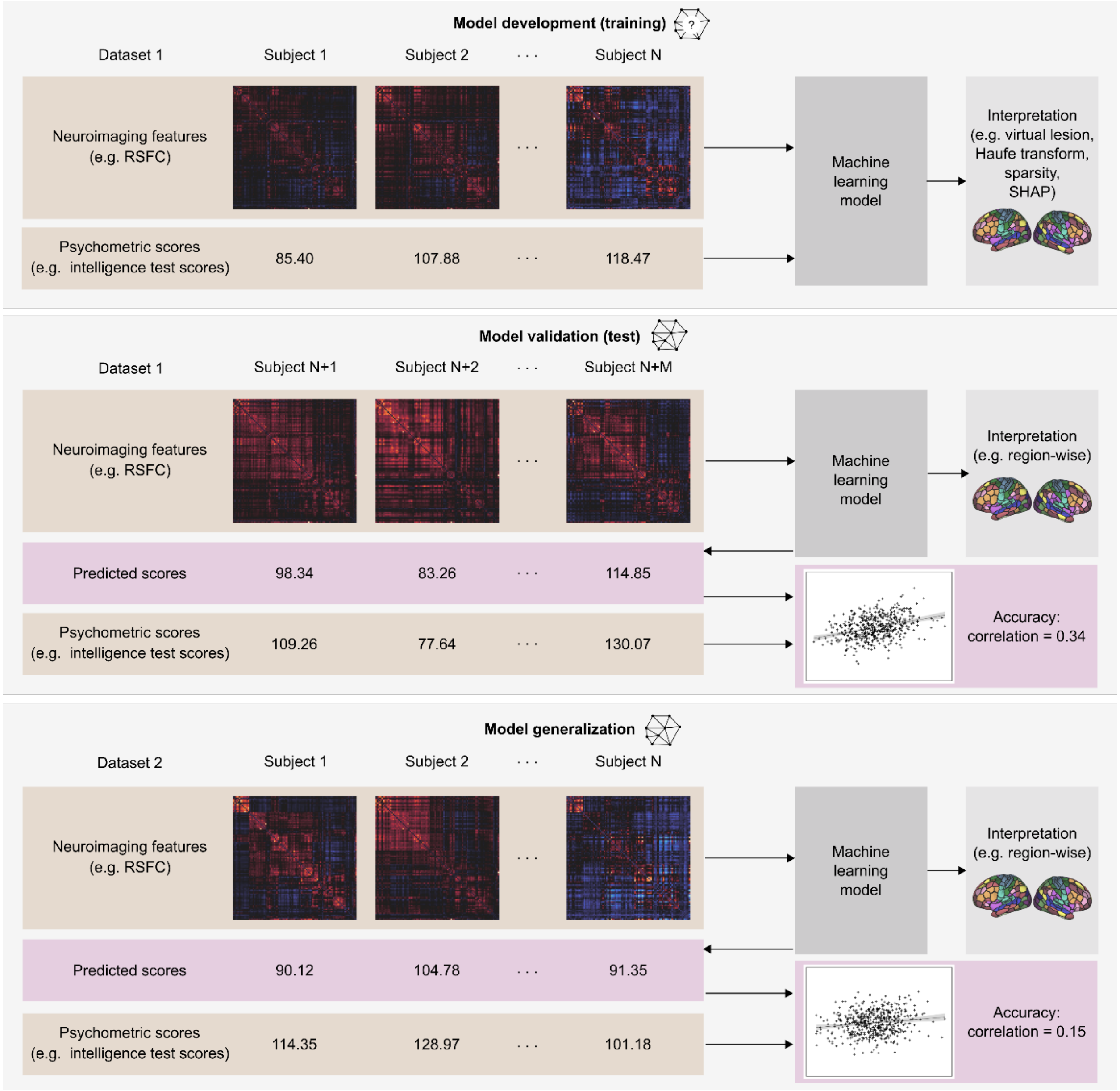

Fig. 1 | Model development and validation for neuroimaging-based psychometric predictions. A machine learning model is first trained using neuroimaging features and psychometric scores from subjects 1 to N (from the training set). The model learns a relationship between the neuroimaging features and the psychometric scores. For validation, the model takes in neuroimaging features from subjects N+1 to N+M (from the test set), and outputs predicted values for the psychometric scores. The predicted scores can then be compared to the actual scores using various accuracy measures (see Box 1) to evaluate the performance of the model. To assess the generalizability of the model, the model needs to be applied to a new dataset in a similar way to its application in the test set.

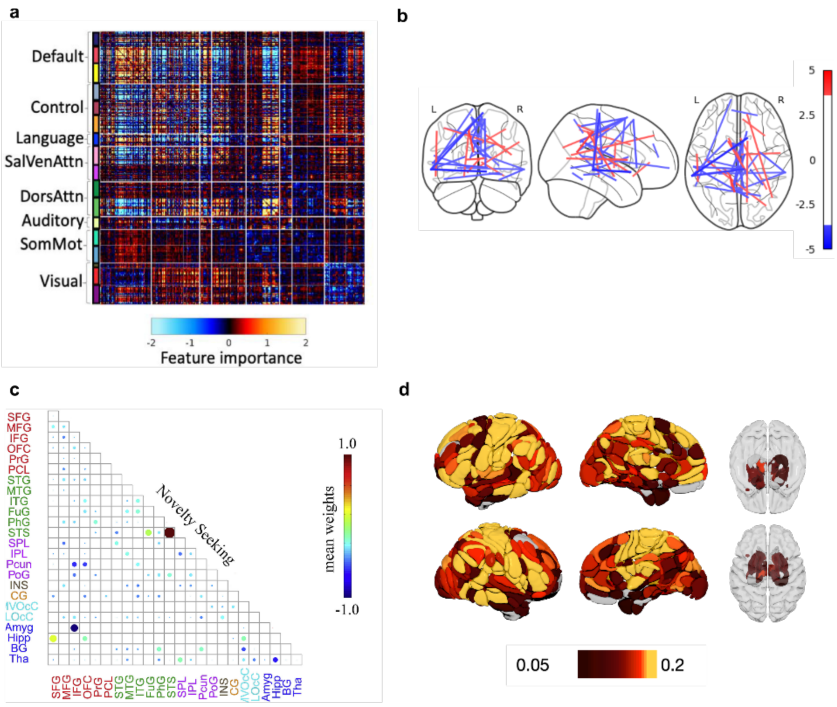

Fig. 2 | Prediction accuracies measured by Pearson's correlation. **a**, Scatter plot of observed and predicted openness trait, with a Pearson's correlation accuracy of 0.24[91]. Blue line shows the fitted line between observed and predicted values, while black dashed line marks the line with unit slope and zero intercept. It can be noted that, while (standardized) observed values have a wide range of variations (roughly between -20 and 15, predicted values remain tightly scattered around zero. **b**, Scatter plot of observed and predicted scores of meaning and purpose, with a Pearson's correlation accuracy of 0.17 in African American subjects and 0.049 in White American subjects[70]. Blue and green lines show the fitted line between observed and predicted values in African American and White American subjects respectively. The correlation appears slightly higher in African American than White American, while the prediction errors may actually be greater in the former group. **c**, Scatter plot of observed and predicted visual working memory performance, with a Pearson's correlation accuracy of 0.402[21]. Blue line shows the fitted line between observed and predicted values. Overall, from all three plots, it can be observed that the Pearson's correlation coefficient is higher when the fitted line has a slope closer to one. It is also noteworthy that predicted values in all cases tend to have smaller variances compared to the observed values. This reflects the tendency of machine learning algorithms to generate predictions closer to the sample mean. Finally, outliers or prediction failures can be observed in all plots even when correlation accuracies are moderate. As the correlation accuracies measure the relative goodness-of-fit, they are less affected by (or reflective of) outliers compared to accuracy measures based on absolute errors.

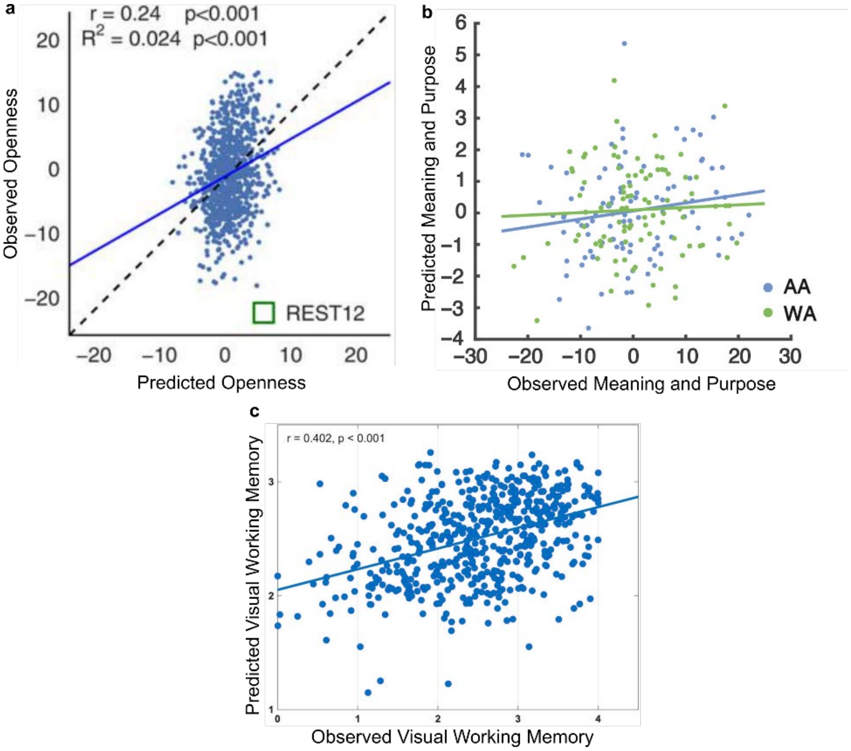

Fig. 3 | Visualizations of model interpretations. a, Feature importance of all RSFC edges for predicting general cognition in a young adult cohort with parcels grouped under networks[38]. Colors correspond to the Haufe transformed weight values. Important connections can be found within the visual network, within the default mode network, between the default mode network and the control network, as well as between the default mode network and the attention networks. **b**, Feature importance of top RSFC edges for predicting fluid intelligence in a young adult cohort shown in their corresponding positions in the brain[60]. Colors correspond to the Haufe transformed weight values. Most top connections can be found between medial regions or temporal regions across the hemispheres. **c**, Feature importance of all RSFC edges for predicting novelty seeking in a young adult cohort when a sparse algorithm was used. Colors correspond to the mean weight values across cross-validation splits[92]. The sparse set of selected features mostly include frontal-subcortical, parietal-frontal, and within-frontal connections. **d**, Brain region importance for predicting fluid cognition in an aging cohort based on the RSFC features using the region-wise approach[60]. Colors correspond to the prediction accuracies achieved using brain regional connectivity profiles. The relatively more predictive regions can be identified in the cingulate cortex, the peripheral visual area, the right supramarginal gyrus, the right anterior insula, the central sulcus, and the right lateral frontal cortex.

Box 1 | Measures of model accuracy

In order to evaluate a model in a validation sample, its predictions need to be compared against the actual values of the psychometric variable. The closer the predicted values are to the actual values, the more accurate the model is. This degree of closeness can be represented either by correlation metrics examining the linear trend between all predicted and observed values, or by error metrics examining the absolute differences between each pair of predicted and observed values.

The most common metric in the literature is the Pearson's correlation coefficient (r) between predicted and observed values[2]**Fehler! Textmarke nicht definiert.**, measuring the normalized covariance between the two variables ($r = \frac{cov(pred,obs)}{\sigma_{pred}\sigma_{obs}}$). This correlation coefficient is an indication of the extent to which a given increase or decrease in one variable is associated with a similar increase or decrease in the other variable . Similarly, Spearman's correlation can be used to measure the ranked correlation between predicted and observed values, providing an indication of how well the two groups of values are monotonically related.

Common error metrics include mean absolute error, mean squared error (MSE), and root mean squared error, measuring the average difference between predicted and observed values in the validation sample in slightly different manners. In general, the error values should be normalized by the standard deviation (or the range of predicted values for absolute errors) of the validation sample, so that they are comparable to standardized measures from other samples[25]**Fehler! Textmarke nicht definiert.**.

While a high correlation suggests that predicted values are generally higher when observed values are higher, it does not mean that predicted values are numerically close to the observed values. As a result, the correlation metrics cannot detect systematic biases where the predicted values are consistently higher (or lower) than the observed value. It may be recommended that high correlation accuracies should be validated with error-based accuracies to check for systematic bias. On the other hand, correlation metrics might be more useful when generalizing a model to new data where the psychometric variables are similar to but not the same as those in the training sample and numerical closeness between predicted and observed values may not be required.

Finally, a useful metric for model evaluation is the coefficient of determination (or R^2), providing a measure of goodness-of-fit of the model. A simple form of R^2 is r^2 may also be computed as the square of the correlation coefficient from the correlation metric. It should be noted that this r^2 measures the goodness-of-fit between the predicted-observed relation and its fitted line, and hence is not a direct measure of model fit itself. Using error metrics such as MSE, the more general R^2 can be computed as $R^2 = 1 - \frac{MSE}{\sigma^2}$, measuring the goodness-of-fit of the regression equation estimated by the prediction model to the validation data. The R^2 values can also be interpreted as the ratio of explained variance by the model to the total variance in the sample, offering an intuitive way to explain the accuracies measured.

References

1. Sui, J., Jiang, R., Bustillo, J. & Calhoun, V. Neuroimaging-based Individualized Prediction of Cognition and Behavior for Mental Disorders and Health: Methods and Promises. *Biol. Psychiatry* **88**, 818-828 (2020).
2. Yeung, A.W.K., More, S., Wu, J. & Eickhoff, S.B. Reporting details of neuroimaging studies on individual traits prediction: a literature survey. *NeuroImage* **256**, 119275 (2022).
3. Cirillo, D. & Valencia, A. Big data analytics for personalized medicine. Curr. Opin. *Biotechnol.* **58**, 161-167 (2019).
4. Dadi, K. et al. Benchmarking functional connectome-based predictive models for resting-state fMRI. *NeuroImage* **192**, 115-134 (2019).
5. Dhamala, E., Yeo, B.T.T. & Holmes, A.J. One Size Does Not Fit All: Methodological Considerations for Brain-Based Predictive Modelling in Psychiatry. *Biol. Psychiatry* **93**, 717-728 (2023).
6. Rosenberg, M.D. et al. A neuromarker of sustained attention from whole-brain functional connectivity. *Nat. Neurosci*. **19**, 165-171 (2016).
7. Lee, M.H., Smyser, C.D. & Shimony, J.S. Resting-state fMRI: a review of methods and clinical applications. *Am. J. Neuroradiol*. **34**, 1866-1872 (2013).
8. Finn, E.S. et al. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat. Neurosci*. **18**, 1664-1671 (2015).
9. Ferguson, M.A., Anderson, J.S., Spreng, R.N. Fluid and flexible minds: Intelligence reflects synchrony in the brain's intrinsic network architecture. *Netw. Neurosci*. **1**, 192-207 (2017).
10. Li, J. et al. A neuromarker of individual general fluid intelligence from the white-matter functional connectome. *Transl. Psychiatry* **10**, 147 (2020).
11. Kumar, S. et al. An information network flow approach for measuring functional connectivity and predicting behavior. *Brain Behav*. **9**, e01346 (2019).
12. Rosenberg, M.D. et al. Functional connectivity predicts changes in attention observed across minutes, days, and months. *Proc. Natl. Acad. Sci. USA* **117**, 3797-3807 (2020).
13. Avery, E.W. et al. Distributed Patterns of Functional Connectivity Predict Working Memory Performance in Novel Healthy and Memory-impaired Individuals. *J. Cogn. Neurosci*. **32**, 241-255 (2020).
14. Pläschke, R.N. et al. Age differences in predicting working memory performance from network-based functional connectivity. *Cortex* **132**, 441-459 (2020).
15. Zhang, H. et al. Do intrinsic brain functional networks predict working memory from childhood to adulthood? *Hum. Brain Mapp*. **41**, 4574-4586 (2020).
16. Girault, J.B. et al. White matter connectomes at birth accurately predict cognitive abilities at age 2. *NeuroImage* **192**, 145-155 (2019).
17. Jiang, R. et al. Multimodal data revealed different neurobiological correlates of intelligence between males and females. *Brain Imaging Behav*. **14**, 1979-1993 (2020).
18. Rasero, J., Sentis, A.I., Yeh, F.C. & Verstynen, T. Integrating across neuroimaging modalities boosts prediction accuracy of cognitive ability. *PLoS Comput. Bio*l. **17**, e1008347 (2021).
19. Wei, L., et al. Grey matter volume in the executive attention system predict individual differences in effortful control in young adults. *Brain Topogr*. **32**, 111-117 (2019).
20. Kaufmann, T. et al. Task modulations and clinical manifestations in the brain functional connectome in 1615 fMRI datasets. *NeuroImage* **147**, 243-252 (2017).
21. Xiao, Y. et al. Predicting visual working memory with multimodal magnetic resonance imaging. *Hum. Brain Mapp*. **42**, 1446-1462 (2021).

22. Scheinost, D. et al. Ten simple rules for predictive modeling of individual differences in neuroimaging. *NeuroImage* **193**, 35-45 (2019).
23. Gabrieli, J.D.E., Ghosh, S.S., Whitfield-Gabrieli, S. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron* **85**, 11-26 (2015).
24. Pervaiz, U., Vidaurre, D., Woolrich, M.W., Smith, S.M. Optimising network modelling methods for fMRI. *NeuroImage* **221**, 116604 (2020).
25. Poldrak, R.A., Huckins, G. & Varoquax, G. Establishment of Best Practices for Evidence for Prediction: A Review. *J. Am. Med. Associ. Psychiatry* **77**, 534-540 (2020).
26. Sripada, C. et al. Prediction of neurocognition in youth from resting state fMRI. *Mol. Psychiatry* **25**, 3413-3421 (2019).
27. He, T. et al. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage* **206**, 116276 (2020).
28. He, L. et al. Functional Connectome Prediction of Anxiety Related to the COVID-19 Pandemic. *Am. J. Psychiatry* **178**, 530-540 (2021).
29. Gao, S., Greene, A.S., Constable, R.T. & Scheinost, D. Combining multiple connectomes improves predictive modeling of phenotypic measures. *NeuroImage* **201**, 116038 (2019).
30. Zalesky, A., Fornito, A., Cocchi, L., Gollo, L.L. & Breakspear, M. Time-resolved resting-state brain networks. *Proc. Natl. Acad. Sci. USA* **111**, 10341-10346 (2014).
31. Bahg, G., Evans, D.G., Galdo, M. & Turner, B.M. Gaussian process linking functions for mind, brain, and behavior. *Proc. Natl. Acad. Sci. USA* **117**, 29398-29406 (2020).
32. Mihalik, A. et al. Canonical Correlation Analysis and Partial Least Squares for identifying brain-behaviour associations: a tutorial and a comparative study. *Biol. Psychiatry* **7**, 1055-1067 (2022).
33. Gal, S., Tik, N., Bernstein-Eliav, M. & Tavor, I. Predicting individual traits from unperformed tasks. *NeuroImage* **249**, 118920 (2022).
34. He, T. et al. Meta-matching as a simple framework to translate phenotypic predictive models from big to small data. *Nat. Neurosci* **25**, 795-804 (2022).
35. Takagi, Y., Hirayama, J.I., Tanaka, S.C. State-unspecific patterns of whole-brain functional connectivity from resting and multiple task states predict stable individual traits. *NeuroImage* **201**, 116036 (2019).
36. Burr, D.A. et al. Functional connectivity predicts the dispositional use of expressive suppression but not cognitive reappraisal. *Brain Behav.* **10**, e01493 (2020).
37. Jiang, R. et al. Task-induced brain connectivity promotes the detection of individual differences in brain-behavior relationships. *NeuroImage* **207**, 116370 (2020).
38. Ooi, L.Q.R. et al. Comparison of individualized behavioral predictions across anatomical, diffusion and functional connectivity MRI. *NeuroImage* **263**, 119636 (2022).
39. Dhamala, E., Jamison, K.W., Jaywant, A., Dennis, S. & Kuceyeski, A. Distinct functional and structural connections predict crystallised and fluid cognition in healthy adults. *Hum. Brain Mapp.* **42**, 3102-3118 (2021).
40. Mansour, L.S, Tian, Y., Yeo, B.T.T., Cropley, V. & Zalesky, A. High-resolution connectomic fingerprints: Mapping neural identity and behavior. *NeuroImage* **229**, 117695 (2021).
41. Pat, N. et al. Longitudinally stable, brain-based predictive models mediate the relationships between childhood cognition and socio-demographic, psychological and genetic factors. *Hum. Brain Mapp.* **43**, 5520-5542 (2022).
42. Hurtz, G.M. & Donovan, J.J. Personality and job performance: The Big Five revisited. *J. Appl. Psychol.* **85**, 869-879 (2000).

43. Kane, M.J., Conway, A.R.A., Miura, T.K. & Colflesh, G.J.H. Working memory, attention control, and the n-back task: A question of construct validity. *J. Exp. Psychol.* **33**, 615-622 (2007).
44. Sanchez-Cubillo, I. et al. Construct validity of the Trail Making Test: Role of task-switching, working memory, inhibition/interference control, and visuomotor abilities. *J. Int. Neuropsychol. Soc.* **15**, 438-450 (2009).
45. Chen, J. et al. Shared and unique brain network features predict cognitive, personality, and mental health scores in the ABCD study. *Nat. Commun*. **13**, 2217 (2022).
46. Wu, J. et al. A Connectivity-Based Psychometric Prediction Framework for Brain-Behavior Relationship Studies. *Cereb. Cortex* **31**, 3732-3751 (2021).
47. Noble, S., Scheinost, D. & Constable, R.T. A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *NeuroImage* **203**, 116157 (2019).
48. 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).
49. Patriat, R. et al. The effect of resting condition on resting-state fMRI reliability and consistency: A comparison between resting with eyes open, closed, and fixated. *NeuroImage* **78**, 463-473 (2013).
50. Birn, R.M. et al. The effect of scan length on the reliability of resting-state fMRI connectivity estimates. *NeuroImage* **83**, 550-558 (2013).
51. Bennett, C.M. & Miller, M.B. fMRI reliability: Influences of task and experimental design. *Cogn. Affect. Behav. Neurosci.* **13**, 690-702 (2013).
52. Cremers, H.R., Wager, T.D., Yarkoni, T. The relation between statistical power and inference in fMRI. *PLoS One* **12**, e0184923 (2017).
53. Kharabian Masouleh, S., Eickhoff, S.B., Hoffstaedter, F., Genon, S. & Alzheimer's Disease Neuroimaging Initiative. Empirical examination of the replicability of associations between brain structure and psychological variables. *eLife* **8**, e43464 (2019).
54. Genon, S., Eickhoff, S.B., Kahrabian, S. Linking interindividual variability in brain structure to behaviour. *Nat. Rev. Neurosci*. **23**, 307-318 (2022).
55. Marek, S. et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654-660 (2022).
56. Beaty, R.E. et al. Robust prediction of individual creative ability from brain functional connectivity. *Proc. Natl. Acad. Sci. USA* **115**, 1087-1092 (2018).
57. Liu, P. et al. The functional connectome predicts feeling of stress on regular days and during the COVID-19 pandemic. *Neurobiol. Stress* **14**, 100285 (2021).
58. Ren, Z. et al. Connectome-based Predictive Modeling of Creativity Anxiety. *NeuroImage* **225**, 117469 (2021).
59. Fong, A.H.C. et al. Dynamic functional connectivity during task performance and rest predicts individual differences in attention across studies. *NeuroImage* **188**, 14-25 (2019).
60. Wu, J. et al. Cross-cohort replicability and generalizability of connectivity-based psychometric prediction patterns. *NeuroImage* **262**, 119569 (2022).
61. Tervo-Clemmens, B. et al. Reply to: Multivariate BWAS can be replicable with moderate sample sizes. *Nature* **615**, E8-E12 (2023).

62. Rosenberg, M.D. & Finn, E.S. How to establish robust brain-behavior relationships without thousands of individuals. *Nat. Neurosci.* **25**, 835-837 (2022).

63. Spisak, T., Bingel, U. & Wager, T. Replicable multivariate BWAS with moderate sample sizes. Preprint at https://www.biorxiv.org/content/10.1101/2022.06.22.497072v1 (2022).

64. Van Essen, D.C. et al. The WU-Minh Human Connectome Project: an overview. *NeuroImage* **80**, 62-79 (2013).

65. Glasser, M.F. et al. The Minimal Preprocessing Pipelines for the Human Connectome Project. *NeuroImage* **80**, 105-124 (2013).

66. Harms, M.P. et al. Extending the Human Connectome Project across ages: imaging protocols for the lifespan development and aging projects. *NeuroImage* **183**, 972-984 (2018).

67. Chouldechova, A., Benavides-Prado, D., Fialko, O. & Vaithianathan, R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proc. Mach. Learn. Res.* **81**, 134-148 (2018).

68. Martin, A.R. et al. Clinical use of current polygenic risk scores may exacerbate healthy disparities. *Nat. Genet.* **51**, 584-591 (2019).

69. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447-453 (2019).

70. Li, J. et al. Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci. Adv.* **8**, eabj1812 (2022).

71. Greene, A.S. et al. Brain-phenotype models fail for individuals who defy sample stereotypes. *Nature* **609**, 109-118 (2022).

72. Greene, A.S., Gao, S., Scheinost, D. & Constable, R.T. Task-induced brain state manipulation improves prediction of individual traits. *Nat. Commun.* **9**, 2807 (2018).

73. Nostro, A.D. et al. Predicting personality from network-based resting-state functional connectivity. *Brain Struct. Funct.* **223**, 2699-2719 (2018).

74. Tian, Y. & Zalesky, A. Machine learning prediction of cognition from functional connectivity: Are feature weights reliable? *NeuroImage* **245**, 118648 (2021).

75. Haufe, S. et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage* **87**, 96-110 (2014).

76. Chen J. et al. Relationship between prediction accuracy and feature importance reliability: An empirical and theoretical study. *NeuroImage* **274**, 120115 (2023).

77. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5-32 (2001).

78. Yip, S.W., Kiluk, B., & Scheinost, D. Towards Addiction Prediction: An Overview of Cross-Validated Predictive Modeling Findings and Considerations for Future Neuroimaging Research. *Biol. Psychiatry Cogn Neuroscie. Neuroimaging* **5**, 748-758 (2020).

79. Jiang, R., Woo, C.W., Qi, S., Wu, J. & Sui, J. Interpreting Brain Biomarkers: Challenges and solutions in interpreting machine learning-based predictive neuroimaging. *IEEE Signal Process. Mag.* **39**, 107-118 (2022).

80. Chormai, P., Pu, Y., Hu, H., Fisher, S.E., Francks, C. & Kong, X.Z. Machine learning of large-scale multimodal brain imaging data reveals neural correlates of hand preference. *NeuroImage*, **262**, 119534 (2022).

81. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. Royal Statistical Soc. B* **67**, 301-320 (2005).

82. Lundberg, S.M. & Lee, S.-I. A unified approach to interpreting model prediction. *Adv. Neural Inf. Process. Syst.* **30** (2017).
83. Pat, N., Wang, Y., Bartonicek, A., Candia, J. & Stringaris, A. Explainable machine learning approach to predict and explain the relationship between task-based fMRI and individual differences in cognition. *Cereb. Cortex* **33**, 2682-2703 (2023).
84. Kingma, D.P. & Welling, M. Auto-Encoding Variational Bayes. Preprint at https://arxiv.org/abs/1312.6114 (2013).
85. Goodfellow, I.J. et al. Generative Adversarial Networks. Preprint at https://arxiv.org/abs/1406.2661 (2014).
86. van den Oord, A., Kalchbrenner, N. & Kavukcuoglu, K. Pixel Recurrent Neural Networks. Preprint at https://arxiv.org/abs/1601.06759 (2016).
87. Fried, D. et al. Speaker-Follower Models for Vision-and-Language Navigation. In *NeuroIPS* (2018).
88. Rosenblatt, M. et al. Connectome-based machine learning models are vulnerable to subtle data manipulations. *Patterns* (In press).
89. Finlayson, S.G. et al. Adversarial attacks on medical machine learning. *Science* **363**, 1287-1289 (2019).
90. Finlayson, S.G., Chung, H.W., Kohane, I.S. & Beam, A.L. Adversarial Attacks Against Medical Deep Learning Systems. Preprint at https://doi.org/10.48550/arXiv.1804.05296 (2019).
91. Dubois, J., Galdi, P., Han, Y., Paul, L.K. & Adolphs, R. Resting-state functional brain connectivity best predicts the personality dimension of openness to experience. *Personal. Neurosci*. **1**, E6 (2018).
92. Jiang, R. et al. Connectome-based individualized prediction of temperament trait scores. *NeuroImage* **183**, 366-374 (2018).

Competing interests

The authors declare no competing interests.