Published in final edited form as:
Neuroimage. 2021 August 01; 236: 118044. doi:10.1016/j.neuroimage.2021.118044.

Resample aggregating improves the generalizability of connectome predictive modeling

David O'Connor[a,*], **Evelyn M.R. Lake**[b], **Dustin Scheinost**[a,b,d,e], **R. Todd Constable**[a,b,c]

[a]Department of Biomedical Engineering, Yale University, United States

[b]Department of Radiology and Biomedical Imaging, Yale School of Medicine, United States

[c]Department of Neurosurgery, Yale School of Medicine, United States

[d]Deparment of Statistics & Data Science, Yale University, United States

[e]Child Study Center, Yale School of Medicine, United States

Abstract

It is a longstanding goal of neuroimaging to produce reliable, generalizable models of brain behavior relationships. More recently, data driven predictive models have become popular. However, overfitting is a common problem with statistical models, which impedes model generalization. Cross validation (CV) is often used to estimate expected model performance within sample. Yet, the best way to generate brain behavior models, and apply them out-of-sample, on an unseen dataset, is unclear. As a solution, this study proposes an ensemble learning method, in this case resample aggregating, encompassing both model parameter estimation and feature selection. Here we investigate the use of resampled aggregated models when used to estimate fluid intelligence (fIQ) from fMRI based functional connectivity (FC) data. We take advantage of two large openly available datasets, the Human Connectome Project (HCP), and the Philadelphia Neurodevelopmental Cohort (PNC). We generate aggregated and non-aggregated models of fIQ in the HCP, using the Connectome Prediction Modelling (CPM) framework. Over various test-train splits, these models are evaluated in sample, on left-out HCP data, and out-of-sample, on PNC data. We find that a resample aggregated model performs best both within- and out-of-sample. We also find that feature selection can vary substantially within-sample. More robust feature selection methods, as detailed here, are needed to improve cross sample performance of CPM based brain behavior models.

[*]Corresponding author. dave.oconnor@yale.edu (D. O'Connor).

Credit authorship contribution statement
David O'Connor: Conceptualization, software, formal analysis, data curation, visualization, writing – original draft, writing- review and editing. **Evelyn Lake:** Conceptualization, writing – original draft, writing- review and editing. **Dustin Scheinost:** Conceptualization, writing – original draft, writing- review and editing. **Todd Constable:** Conceptualization, writing – original draft, writing- review and editing, funding acquisition, supervision.

Supplementary materials
Supplementary material associated with this article can be found, in the online version, at doi: 10.1016/j.neuroimage.2021.118044.

1. Introduction

A longstanding goal of neuroimaging research has been to establish generalizable links between brain structure/function and behavior or traits (Woo et al., 2017). A general approach is to identify discriminating imaging features which, when incorporated into a statistical model, can be used either for inference into potential causal links, or to reliably estimate an observable phenotype for novel participants (Bzdok and Ioannidis, 2019). The ultimate aim is to derive clinically actionable diagnoses or intervention strategies from imaging data (Insel et al., 2010). In order to achieve a clinically actionable model, both feature engineering and model building are important. A simple model may perform well with robust features that have a large effect size, yet even the most complex model may underperform when given poorly curated features. In neuroimaging generally, and in particular functional Magnetic Resonance Imaging (fMRI), an ideal feature set is seldom seen due to a combination of factors including site effects (Badhwar et al., 2020), physiological noise (Keilholz et al., 2017; Liu, 2016), hardware noise (Triantafyllou et al., 2005), and small sample sizes (Button et al., 2013), not to mention the complexity of the underlying neural activity itself. In order for the combination of a feature set and model to be clinically actionable these challenges, which are ubiquitous, need to be overcome. In other words, in virtually all cases, model performance and generalization are constrained by the feature quality. fMRI based functional connectivity (FC) is a commonly used feature set for developing brain imaging-based models of observable phenotypes. Though initial investigations using fMRI focused on functional specialization of brain regions, it has been shown that widespread neural processes can contribute to higher level brain function. This implies that functional integration, rather than discrete specialization, is likely the key to characterizing more complex phenotypes (Turk-Browne, 2013; Rissman et al., 2020; Horien et al., 2020). As a result, FC based models commonly use the functional connectome, a map of the functional connections between all pairs of brain regions (or nodes, defined by an atlas), as the starting point for feature selection (Castellanos et al., 2013; Dadi et al., 2019; Arbabshirani et al., 2017; Pervaiz et al., 2020). This framework for analyzing brain function mostly incorporates undirected, first order, pair-wise estimates of connectivity between brain regions (Sporns, 2013). Even still, dependent upon the atlas used to define nodes, these approaches typically yield upwards of 70,000 candidate features. Such massive, and often disparate, sets of features are both hard to interpret, and easily allow for overfitting, particularly when working with small samples (Bzdok et al., 2019).

Approaches to avoid overfitting depend upon the goals of the analysis, and the type of model being built, whether it be a predictive or explanatory model. These strategies are not mutually exclusive (Rosenberg et al., 2018), but can be contrasted as being biased towards using a feature set to predict future observations with high accuracy (predictive), or developing a mechanistic understanding of the relationship between the features and the observation (explanatory) (Shmueli, 2010). Here, we will focus on the former. Recently there has been much interest in successful predictive modeling approaches that estimate brain-behavior relationships (Pervaiz et al., 2020; Finn et al., 2015; Smith et al., 2015; Rosenberg et al., 2017; Liem et al., 2017; Gao et al., 2019). These strategies, are geared toward model performance-based assessments, and attempt to identify reproducible models using cross validation (CV) to yield reliable estimates of model performance within sample (Varoquaux et al., 2017), and out of sample (Abraham et al., 2017). However, as is the case with predictive frameworks in general, outcomes are agnostic to the consistency of model parameters or selected features, with respect to changes in the underlying data. Within sample predictive approaches in particular do not guarantee a model will translate to other datasets. Notably, within-sample performance estimates can vary, especially when datasets are small (< 100 participants) (Varoquaux, 2018). This is not to say that models built within one dataset using CV to estimate within-sample performance will always fail to generalize. Some studies have shown good performance out-of-sample (Rosenberg et al., 2020; Yip et al., 2019; Greene et al., 2018; Lake et al., 2019), when it was a requirement that a given feature be selected in a minimum number of CV folds before being included in the external application of the model. This out-of-sample application of the model begins to resemble an ensemble learning method.

Ensemble learning refers to a group of statistical methods that act to combine multiple models, in order to boost performance (Opitz and Maclin, 1999). Ensemble learning can be used to alleviate overfitting. One form of ensemble learning is bootstrap aggregating, or bagging (Varoquaux et al., 2017; Breiman, 1996). Bagging is based on training models in multiple subsets of a training data set, or bootstraps, and aggregating model performance across the models. It is also possible to aggregate model parameters and features across subsets (De Bin et al., 2016). The CV models cited above (Rosenberg et al., 2020; Yip et al., 2019; Greene et al., 2018; Lake et al., 2019), which have performed well out-of-sample, have in essence used an ensemble approach in the out-of-sample application, though through subsampling and then aggregating rather than bootstrap aggregating. Bagging has been applied to fMRI data before to determine brain state (Richiardi et al., 2011), and for brain parcellation (to define an atlas) (Bellec et al., 2010; Nikolaidis et al., 2020). With respect to predictive modeling, it been shown to boost within-sample performance of resting state FC based brain-behavior regression models (Wei et al., 2020), and has also been applied in a classification approach (Hoyos-Idrobo et al., 2018).

Our goal is to build a more generalizable FC based brain-behavior regression model using an ensemble learning approach, building within sample on data from one dataset. In this paper, we applied four CV and three resampling approaches, including a bagged modeling approach, using connectome predictive modeling (CPM) as the model framework. We develop brain-behavior models relating working memory FC and fluid intelligence (fIQ) using the typical approach which implements one feature selection and model training step. We compare these results to alternative resampling approaches that combine information from multiple feature selection and model building iterations. We show, using out-of-sample testing, that resample aggregating improves model generalizability. Our results may provide a framework for building more translatable brain behavior models.

2. Methods

2.1. Datasets

Data from the Human Connectome Project (HCP), specifically the S900 dataset, (Van Essen et al., 2013) and the Philadelphia Neurodevelopmental Cohort (PNC), $N = 1000$ (Calkins et al., 2015), were used. From these datasets, *N* = 827 participants from the HCP, ages 21–35, and *N* = 788 participants from the PNC, ages 8–21, were included based on the availability of preprocessed T_1-weighted images, working memory fMRI scans, and fIQ measures.

For the HCP dataset, fIQ was measured using a 24-item version of the Penn Progressive Matrices assessment, scores ranged from 4 to 24, with a mean ± standard deviation (SD) of 16.78 ± 4.7 (Bilker et al., 2012). In the PNC dataset, the 24- and 18-item versions of the Penn Matrix Reasoning Test were used (Moore et al., 2015; Gur et al., 2010). Scores ranged from 0 to 23, with a mean ± SD of 11.85 ± 4.06. In both datasets, the score corresponds to the number of correct responses.

For the HCP, MRI data were acquired on a 3T Siemens Skyra. The fMRI scans were collected using a slice-accelerated, multiband, gradient-echo, echo planar imaging (EPI) sequence (TR = 720 ms, TE = 33.1 ms, flip angle = 52°, resolution = 2.0mm^3, multiband factor = 8, left-right phase encoding, scan duration = 5:01). The T_1-weighted structural scans were collected using a Magnetization Prepared Rapid Gradient Echo (MPRAGE) sequence (TR = 2400 ms, TE = 2.14 ms, TI = 1000 ms, resolution = 0.7mm^3) (Van Essen et al., 2012). For the PNC, the MRI data were acquired on a 3T Siemens TIM Trio. fMRI scans were collected using a multi-slice, gradient-echo EPI sequence (TR = 3000 ms, TE = 32 ms, flip angle = 90°, resolution = 3mm^3, scan duration = 11:39). T_1-weighted structural scans were collected using an MPRAGE sequence (TR = 1820 ms, TE = 3.5 ms, TI = 1100 ms, resolution = 0.9375 × 0.9375 × 1 mm) (Satterthwaite et al., 2014).

2.2. Preprocessing

For the HCP, the HCP minimal preprocessing pipeline was used on these data (Glasser et al., 2013), which includes artifact removal, motion correction, and registration to MNI space. All subsequent preprocessing was performed in BioImage Suite (Joshi et al., 2011) and included standard preprocessing procedures (Finn et al., 2015), including removal of motion-related components of the signal; regression of mean time courses in white matter, cerebrospinal fluid, and gray matter; removal of the linear trend; and low-pass filtering.

For the PNC, structural scans were skull stripped using an optimized version of the FMRIB's Software Library (FSL) pipeline (Smith et al., 2004). Slice time and motion correction were performed in SPM8 (Frackowiak et al., 2003). The remainder of image preprocessing was performed in BioImage Suite (Joshi et al., 2011) and included linear and nonlinear registration to the MNI template; regression of mean time courses in white matter, cerebrospinal fluid, and gray matter; and low-pass filtering.

FC matrices were generated from the working memory fMRI data using the Shen atlas (Shen et al., 2013) (which defines 268 cortical and subcortical nodes), and Pearson's correlation as the time course similarity metric. Following this the matrices were z-transformed. All further analyses were performed in python using custom scripts. These scripts have been made publicly available, see section on data and code availability. Figures were generated in python using matplotlib (Hunter, 2007) and seaborn (Waskom et al., 2017), with the flowchart generated using app.diagrams.net.

2.3. Modeling protocol

For our modeling framework we chose to use Connectome Predictive Modelling (CPM). CPM was performed as in (Shen et al., 2017), with the FC matrices as the explanatory variable, and fIQ as the target variable, with one exception; partial correlation was used at the feature selection step (Hsu et al., 2018). In full, the CPM process was as follows:

1. Across the training set, correlate each element in the FC matrix, often referred to as an edge, (predictive variables) with fIQ (target variable). In this work, partial correlation was used. First, mean frame-to-frame displacement in the fMRI scan is regressed out of both the edge values and fIQ. Then Pearson’s correlation is computed from the residuals.
2. Select positively correlated edges with, where $p < 0.01$.
3. For each participant or scan, sum the connectivity scores for all selected edges.
4. Fit a linear model, without regularization, between the sum of connectivity scores and fIQ summary score.
5. Apply the model to unseen participants and estimate performance.

Overall, we perform 20 data splits on the HCP dataset, into test and train samples, to assess the impact of participant inclusion in the training set on model performance. The following protocol explains what happens within each data split. The protocol is also shown in Fig. 1.

Model Training: 800 subjects are randomly selected from the $N = 827$ HCP subjects, and split equally into a train and test sample. This randomization was done with respect to the family structure within the dataset, ensuring that family members were never split across train and test samples. Supplementary Figure 1 shows the results of accounting for family structure in the test-train split, when compared to a purely random test-train split. It did not significantly impact model performances. The test sample is used exclusively to test within-sample performance, and none of the participants in the train sample are included in testing, see Fig. 1. The train sample is used to build three types of models: (1) resample aggregated models, (2) CV models, and (3) a single model trained on the whole train sample.

1. Three varieties of resample aggregated models are generated. All 400 participants in the train sample are used in each split. 100 iterations of feature selection and model building are performed on resampling of different numbers of randomly selected participants to create three different models: (i) bootstrapping, where 400 subjects are selected, with replacement, (ii) models with a subsample of 300 subjects selected without replacement, and (iii) models with subsamples of 200 subjects selected without replacement. These models are termed the bagged model (i), the subsample 300 model (ii) and the subsample 200 model (iii). In each case, model parameters, the slope and intercept, and features are aggregated to form one model and feature set per split. The slope and intercept are the mean across bootstraps/subsamples. A mean feature vector is formed, which is the frequency with which each feature (edge) is selected across all resampling.

2. CV is performed four different ways: (i) using split-half, (ii) fivefold, (iii) tenfold and (iv) leave-one-out (LOO) resampling. 100 iterations of (i-iii), are performed. The models generated by (iv) are unaffected by different iterations. Thus, the following number of models per data split of each CV type are generated: (i) split-half: 200, (ii) five-fold: 500, (iii) tenfold: 1000, and (iv) LOO: 400.

3. A single model generated using all the participants in the train sample at once. Herein, termed a train-only model.

Twenty test-train splits are performed. Thus, our full protocol yields 20 of each of the resample aggregated models (type 1), 4000 split-half, 10,000 fivefold, 20,000 tenfold, and 8000 LOO CV models (type 2), and 20 train-only models (type 3).

Model Testing: The HCP test sample is used to assess the within-sample performance of all models, and the PNC data is used to assess out-of-sample performance. In both within-sample, and out-of-sample testing, each model is tested on 100 random subsamples of 200 participants from the HCP test sample and 100 random subsamples of 200 participants from the external PNC data set. This framework yields 100 measures of performance for each model within-sample, and 100 measures of performance for each model out-of-sample. In addition to subsamples of 200 participants, models were tested on subsamples of size 300, and the full test set. The results of this are shown in the Supplementary Figure 2. The model performance shown in the main manuscript focuses on subsamples of 200.

In generating the resample aggregated models, a feature vector is created, in which each element corresponds to the frequency with which a given feature passes feature selection, across resampling. This allows the possibility of imposing a minimum frequency threshold, and filtering the features used in the model. The resample aggregated models are first tested including every feature which occurred at least once in any single resample. Following this, the feature vector is thresholded. First, to include features which occur in 10% or more resamples, then 20% or more, rising in 10% increments, to those that occur in 90% or more of resamples. For each threshold, the resample aggregated models' performance is assessed within- and out-of-sample for each data split, as described above.

Model performance is quantified as the variance explained by the predicted fIQ, with respect to the actual fIQ. Differences in model performance are assessed using the Wilcoxon signed rank test.

3. Results

The performance of the resample aggregated models (with all features included), CV models, and the train-only models, in predicting fIQ, within- and out-of-sample are shown in Fig. 2. Performance metrics are separated based on the 20 test-train HCP data set splits. Performance of all models vary depending on the test train split. In terms of mean performance, the subsample 300 models perform best within- and out-of-sample, where they are roughly equivalent to the subsample 200 models.

In Fig. 3, performance metric distributions are shown without differentiating by test-train split. Within-sample, the subsample 300 models perform best, better than the next best performing CV models (LOO). Using a Wilcoxon signed rank test to compare, the result is $W = 772{,}076{,}502$, $p = 0.003$. This corresponded to the subsample 300 models performing better than LOO in 54% of within-sample test cases. Out-of-sample, the subsample 300 models perform best, though with similar mean performance to the subsample 200 models. Both the subsample 300 and subsample 200 models perform better than the best performing CV models, LOO, with test statistics: $W = 661{,}298{,}771$, $p < 0.001$ and $W = 676{,}212{,}100$, $p < 0.001$ respectively. This corresponded to the subsample 300 models performing better than LOO models in 79% of out-of-sample test cases, and the subsample 200 models performing better than LOO models in 72% of out-of-sample test cases. Density plots showing the distribution of differences in model performances for these models are shown in Supplementary Figure 3. In addition to the Wilcoxon signed rank tests, a Fishers exact test was performed for each set of model comparisons. In each case 2 × 2 contigency tables were built, comparing model performance differences. The first row contained the number of times the first model performed better in sample, followed by the number of times the second model performed better in sample. The second row contained the number of times the first model performed better out of sample, followed by the number of times the second performed better out of sample. Comparing the subsample 300 and LOO models yielded $p < 0.001$, and comparing the subsample 200 and LOO models also yielded $p < 0.001$.

The distribution of feature occurrence in the bagged models is shown in Fig. 4, left panel. On average 48.7% ± 0.8%, or 17,423 ± 302 edges out of a possible 35,778, are selected at least once across bootstraps from our undirected FC matrix (268 node atlas). The corresponding values for the subsample 300 and 200 models are: 11.7% ± 1%, or 4169 ± 356 edges, and 20% ± 1.2%, or 7184 ± 420 edges, respectively. On average, in the bagged models, only 46 ± 16 edges, reoccur in > = 90% of bootstraps. The corresponding numbers for subsample 300 and 200 models are: 58 ± 20 edges, and 1.95 ± 1.5 edges, respectively.

With CPM, edges (features) are selected based on how strongly the edge strength between any two nodes varies with the target variable across individuals within the training sample. The selection step of whether or not to include an edge depends on a statistical significance threshold on the correlation between edge strength and behavior. When creating the resample aggregated models, the R values at this feature selection step are indicative of whether a feature occurs frequently across bootstraps/subsamples, as shown in Fig. 5.

The number of features passing the selection criteria may be related to the number of participants included for a given resampling method. The bagged models have the greatest number of features selected at least once overall. This may be as a result of having the highest number of participants included for feature selection (400). This also may impact the performance of the bagged model as it leads to the more frequent selection of low strength edges. In this scenario relatively small effect sizes are sufficient to pass a p-based threshold, when compared to a sample size of 200 or 300. While the bagged model has the greatest number of features selected overall, the subsample 300 models have the greatest proportion of features passing higher frequency thresholds (from > = 20% and up), see Fig. 5. Though the bagged models nominally include more participants, resampling is performed with replacement. Thus, only 63.2%, or 253 out of 400, are unique participants in a given bootstrap on average (Pathak and Rao, 2013). On the other hand, the subsample 200 and 300 models are created without replacement, and therefore only contain unique participants. Thus, the subsample 300 models have the highest number of unique participants included in any resampling. This may explain why the subsample 300 models perform well.

The effect of thresholding the feature vector on the performance of the resample aggregated models assessed within-sample, and out-of-sample, is shown in Fig. 6. This reflects the performance as assessed on subsamples of 200 participants in each test set. The performance was also evaluated on subsamples of 300 participants, and on the full test sets. These results are shown in supplemental Figures 4 and 5 respectively.

Within sample.

As described above, with all features included, the subsample 300 models perform best within-sample, followed by subsample 200 models which perform marginally better than the best non-aggregated models (LOO). The no-threshold bagged models perform worse than these non-aggregated models. However, the bagged models' mean performance initially increases with thresholding. The performance of the thresholded bagged models' remains higher than the no-threshold performance (including all features, 17,423 ± 302) up until a 60% threshold in-sample (413 ± 94 features). Beyond 60%, the mean performance drops below the no-threshold level. Both the subsample 300 and 200 models experience a drop in performance with any thresholding. The drop off is more severe for the subsample 200 models, presumably because many fewer edges passed the more stringent thresholds. With higher thresholds, the bagged models' within-sample performance surpasses the no-threshold performance of both the subsample 300 and 200 models (whose performance decrease with thresholding), and the best performing non-aggregated models. At a threshold of > = 40%, the bagged model performs best within-sample amongst all models generated (and across all feature vector frequency thresholds).

While the mean performance of the subsample 200 models decreases precipitously as the feature selection threshold increases, perhaps as a result of fewer features being included at the feature selection step, the subsample 300 and bagged models exhibit much less decline. Illustratively, the mean within-sample performance of bagged models is 0.11 at a feature vector threshold > = 90%. This is 74% of the bagged models' top performance (which occurs at a feature vector threshold of > = 40%). Notably, at a feature vector threshold > = 90%, the bagged models contain an average of 46 edges compared to an average of 1147 edges at *a* > = 40% threshold. Similarly, for the subsample 300 models, mean performance is 0.12 at *a* > = 90% feature selection threshold (average model size of 58 edges). This performance is 81% of the models' top performance level which occurs with no-threshold (including an average of 4170 edges).

Out-of-sample.

With all features included, the subsample 300 and 200 models perform better than the best non-aggregated models out-of-sample: the train-only models. The no-threshold bagged models perform as well as the train-only models. Though, once again, the bagged models' mean performance initially increases with thresholding out-of-sample, beyond the no-threshold subsample 300 and 200 models, which is the best performance for those models across all feature frequency thresholds. The performance of the thresholded bagged models' remains higher than the no-threshold (including all features, 17,423 ± 302) up until a 50% threshold (713 ± 131 features). The > = 10% threshold bagged model performs best out-of-sample amongst all models generated, across all feature frequency thresholds.

The trend of model performance dropping at higher feature frequency thresholds was also observed out-of-sample, though with a slightly larger decrease in performance. At a 90% threshold, the out-of-sample bagged models mean performance is 0.047, also with 46 edges on average. This is 64% of the bagged models top out-of-sample performance at a threshold of > = 10%, with 5572 edges on average. Out-of-sample subsample 300 models mean performance at a 90% threshold is 0.049, also with 58 edges on average. This is 66% of the subsample models top out-of-sample performance at a threshold of > 0%, with 4170 edges on average.

4. Discussion

In this study we show that an ensemble method, in this case a resample aggregated (subsample or bootstrap aggregated) model, can improve the generalizability (out-of-sample performance) of CPM based brain-behavior models for prediction of fIQ, when compared with CPM models generated as part of standard CV. Three types of model are evaluated: resample aggregated models, CV models, and models trained on the complete train sample (train-only). Within-sample evaluation, on a held test sample from the same data set, shows that a bagged model with a 50% feature frequency threshold performs best, with the no threshold subsample 200 and 300 models also performing well. Out-of-sample testing is used as a proxy for generalizability. Out-of-sample, the 10% threshold bagged model perform best with the no threshold subsample 300 and 200 models again performing well. The resample aggregated are good alternatives to models generated as part of CV when generalizability is of key concern. Additionally, the subsample 300 and bagged models show relatively good performance at high thresholds given the parsimonious number of features.

Generation of the resample aggregate models allows an estimate of how frequently a given feature is selected across resamples. High variance is exhibited, with almost half of all possible features (48.7%) being selected at least once when a bagged modeling approach is implemented. This fraction is lower for the subsample 200 (20%) and 300 (11.7%) models. There is a strong relationship between effect size (R) and how many bootstraps/subsamples a given feature is selected within. This varies based on both resample size, and number of unique participants included, and may affect the performance of the resample aggregated models with increasing feature frequency thresholds.

Our study is performed in response to a general increase in research which focuses on predictive modeling (Varoquaux and Thirion, 2014; Gabrieli et al., 2015; Bzdok and Yeo, 2017), and in particular approaches that utilize FC measures to uncover meaningful brain-behavior relationships (Pervaiz et al., 2020). Ultimately, with these studies, a clinical application is desirable (Castellanos et al., 2013). However, present FC based predictive models are yet to perform well enough in terms of accuracy, sensitivity or specificity to have an impact on clinical practice. Generalizability is of particular importance in a clinical setting, and presently a major problem for predictive models in neuroimaging. In some cases, sophisticated models which have shown much success in other fields, such as convolutional neural networks, have been applied to model brain-phenotype relationships (Pervaiz et al., 2020), (Jiang et al., 2020). However, they are yet to provide a generalizable model with sufficiently high-performance for clinical application. Given the failure of highly flexible and complex models to fit these data in a generalizable manner, and the myriad of potential confounds that exist in raw neuroimaging data, models themselves may not be the barrier generalization. This work suggests that the feature selection step may be a critical component in developing generalizable models.

Applying more focus on feature sets, the edges in connectome-based modeling, and requiring that they are consistently related to a target variable, in this case a behavior, can be seen as more of an explanatory approach. Explanatory approaches are usually contrasted with predictive approaches, whereby explanatory approaches are more focused on elucidating the brain-behavior link. Predictive approaches, on the other hand, are agnostic to the underlying relationship between a feature set and a target variable. However, these two approaches are not mutually exclusive (Rosenberg et al., 2018), and emphasis can be placed on identifying reproducible features, while still defining success based on prediction performance. Ensemble modeling methods that incorporate the feature selection step, for example (De Bin et al., 2016), allow for the derivation of a stable estimate of both model parameters and features within sample, and within a predictive framework. Our work shows that one can achieve better generalization, in terms of out-of-sample model performance, using resample aggregated models. And one can also derive a small and much more tractable group of features that are consistently linked to the target variable in question, potentially shedding light on the underlying brain-behavior relationship. These methods may provide a pathway to better generalization.

A prerequisite for generalization is within-sample performance estimation. A common method for reliable within-sample performance estimation is CV (Scheinost et al., 2019; Poldrack et al., 2020). Though there are exceptions (Rosenberg et al., 2020; Yip et al., 2019; Greene et al., 2018; Lake et al., 2019), models using CV are more commonly built and evaluated in one relatively homogenous sample and sometimes subsequently tested out-of-sample. While CV is necessary to get a balanced measure of performance within-sample it does not guarantee replication of that performance out-of-sample. Some studies which have used CV within-sample and demonstrated successful replication have essentially implemented a form of ensemble learning through resampling-based feature aggregation (Rosenberg et al., 2020; Yip et al., 2019; Greene et al., 2018; Lake et al., 2019). These studies tested the models built as part of CV on an out-of-sample data-set using only the most commonly occurring features across CV folds. This is akin to the methods described in this paper, however, in these implementations, the maximum number of feature selection iterations was 10. Given that with the 100 bootstraps, almost 50% of the features occur at least once in the bagged protocol shown here, and indeed 20% of features with subsample 200 and 11% of features with subsample 300, resample aggregation may provide an improvement on identifying noisy features compared to 10-fold cross validation, as well as provide more insight into feature variation. A previous study by Wei et al. using bagged models for prediction has also recommended ~100 bootstraps (Wei et al., 2020). An additional factor to note is that the greater the fraction of subjects from the training set that are included in the feature selection step, the less likely the features are to vary across folds. In the case of 10-fold CV this would include 90% of subjects. This is generally an advantage but it can also be important to obtain information on how smaller subsamples influence the variance in feature selection, provided the overall sample size is sufficient.

A potential downside of a resampling approach is that feature occurrence is a fairly simple way of uncovering within-sample variance. Hypothetically, it is possible that some features are highly related and do not add unique information. This would indicate using dimensionality reduction (Mwangi et al., 2014; Barron et al., 2019) and/or a regularized modeling approach (Gao et al., 2019) as a first step. However, it is possible to integrate these methods into an ensemble approach, and if this is done, it would be prudent to assess the impact of resampling on the parameters generated.

Despite increased performance out-of-sample, there was still a discrepancy with respect to within-sample performance for all models. This may suggest that even the resampled aggregate models are somewhat overfit to the training data, especially in light of the number of features included in the bagged model at the lower levels of feature thresholding. Overfitting is defined as “The production of an analysis which corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably” (Overfitting | Meaning of Overfitting by Lexico 2020). This is commonly thought to mean that the model has fit too much noise, or signal of non-interest related to that particular sample. In the case of FC based models, overfitting can occur as a result of modeling signal of non-interest, or as a result of fitting a signal of interest which does not generalize. That is to say, a FC model of fIQ built in the HCP dataset might capture meaningful signal related to fIQ, but it is specific to the experimental conditions and participant demographics of the HCP study. To simply ascribe the difference in performance to overfitting signal of non-interest, would be to ignore studies suggesting age and sex-based differences in FC (Liem et al., 2017), (Satterthwaite et al., 2015; Zhang et al., 2018; Betzel et al., 2014; Dosenbach et al., 2010), along with a host of other factors. One of the strengths of ensemble learning is that it can incorporate variation within dataset through resampling, but it can only do that to the extent to which the training data exhibits said variance. This may explain why the subsample models’ performance decreased out-of-sample with feature thresholding, and why the bagged models’ performance peaked at an earlier threshold out-of-sample, compared to within-sample. At a low threshold for feature inclusion, it is likely that we capture some features relevant to estimating fIQ in the out-of-sample test set, particularly given that nearly half of all possible features were included. Increasing the threshold for feature inclusion could equate to including features that are more relevant to the training sample and not the test sample. In this study, models are trained on HCP data, which is comprised of adults (ages 21–35), and tested on PNC data, which includes adolescents (ages 8–21). In addition to capturing noise specific to the training sample, our HCP trained model is likely not capturing features that are specific to adolescent cognition that are undoubtedly present in the out-of-sample test set from the PNC (Betzel et al., 2014). Another confounding factor is that the working memory tasks in the HCP and PNC data sets have non-negligible differences. The experimental designs of each sample use different visual stimuli, and working memory loads (Ragland et al., 2002; Barch et al., 2013). Additionally, the target variables are derived from different, though related, assessments of fIQ (Bilker et al., 2012; Moore et al., 2015; Gur et al., 2010). These factors likely contribute to the difference in performance observed between within- and out-of-sample testing across models. It is possible to mitigate this performance difference, using data harmonization methods (Yu et al., 2018). While these types of methods can reduce inter site variability in FC, they are unlikely to account fully for the demographic and experimental heterogeneity described above. Our results suggest there is a tradeoff between fitting the variance we are interested in (adult cognition) and providing a generalizable model of fIQ. Generalizability encompasses both model parameters and features that are neither too “noisy”, nor too specific. In this respect, the bagged model seems to capture features that improve generalizability, but not sufficiently to match the level of within-sample performance. It is likely that overfitting and an imperfect match between samples hampered generalization.

Though we are in an era where predictive modeling is the focus of much research, we do not yet have a complete understanding of how the features we identify vary across cohorts; in this case FC data obtained during a working memory task in two different populations. How we should adapt our approaches to account for noise and sample-specific features is evolving. While the models used in this study are fairly simple linear models, and how our approach would generalize to more complicated predictive modeling schema is unclear, we assert that an ensemble approach to predictive modeling, such as the one detailed here, can help discern features that are relevant to model generalizability.

Conclusions

Resample aggregated models allow for greater model performance and generalizability, within the context of CPM. The within-sample boost in performance, in light of including all features in performance assessment, are suggestive of overfitting. However, out-of-sample performance was also significantly better than non-resample aggregate models, suggesting better model generalizability. The resampling procedure also provides an estimate of the stability of feature selection across training resampling. Bagged models increase and maintain performance when decreasing the number of features up to a point, within- and out-of-sample. Subsample aggregated model performance decreases as feature selection becomes more stringent.

Data and code availability statement

The HCP data that support the findings of this study are publicly available on the ConnectomeDB database (https://db.humanconnectome.org). The PNC data that support the findings of this study are publicly available on the database of Genotypes and Phenotypes (dbGaP, accession code phs000607.v1.p1); a data access request must be approved to protect the confidentiality of participants. Code for conducting the analyses described here can be found at https://github.com/YaleMRRC/CPMBaggingAnalysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by R01 MH121095. Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. The remainder of the data used in this study were provided by the Philadelphia Neurodevelopmental Cohort (Principal Investigators: Hakon Hakonarson and Raquel Gur; phs000607.v1.p1). Support for the collection of the data sets was provided by grant RC2MH089983 awarded to Raquel Gur and RC2MH089924 awarded to Hakon Hakonarson. All subjects were recruited through the Center for Applied Genomics at The Children’s Hospital in Philadelphia.

References

Woo C-W, Chang LJ, Lindquist MA, Wager TD, 2017. Building better biomarkers: brain models in translational neuroimaging. Nat. Neurosci 20 (3), 365–377 3. [PubMed: 28230847]

Bzdok D, Ioannidis JPA, 2019. Exploration, Inference, and Prediction in Neuroscience and Biomedicine. Trends Neurosci. 42 (4), 251–262 Elsevier Ltd 01-4-. [PubMed: 30808574]

Insel T, et al., 2010. Research domain criteria (RDoC): toward a. Am. J. Psychiatry Online 748–751 no. 7.

Badhwar AP, et al., 2020. Multivariate consistency of resting-state fMRI connectivity maps acquired on a single individual over 2.5 years, 13 sites and 3 vendors. Neuroimage 205, 116210 1. [PubMed: 31593793]

Keilholz SD, Pan W-J, Billings J, Nezafati M, Shakil S, 2017. Noise and non-neuronal contributions to the BOLD signal: applications to and insights from animal studies. Neuroimage 154, 267–281 7. [PubMed: 28017922]

Liu TT, 2016. Noise contributions to the fMRI signal: an overview. Neuroimage 143, 141–151 12. [PubMed: 27612646]

Triantafyllou C, et al., 2005. Comparison of physiological noise at 1.5 T, 3 T and 7 T and optimization of fMRI acquisition parameters. Neuroimage 26 (1), 243–250 5. [PubMed: 15862224]

Button KS, et al., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci 14 (5), 365–376 5. [PubMed: 23571845]

Turk-Browne NB, 2013. Functional interactions as big data in the human brain. Science 342 (6158), 580–584 American Association for the Advancement of Science. [PubMed: 24179218]

Rissman J, Gazzaley A, D’esposito M, and Wheeler HH, 2020 “Measuring functional connectivity during distinct stages of a cognitive task.”

Horien C, Greene AS, Constable RT, Scheinost D, 2020. Regions and connections: complementary approaches to characterize brain organization and function. Neuro-scientist 26 (2), 117–133 SAGE Publications Inc. 01-4.

Castellanos FX, Di Martino A, Craddock RC, Mehta AD, Milham MP, 2013. Clinical applications of the functional connectome. Neuroimage 80, 527–540 10. [PubMed: 23631991]

Dadi K, et al., 2019. Benchmarking functional connectome-based predictive models for resting-state fMRI. Neuroimage 192, 115–134 5. [PubMed: 30836146]

Arbabshirani MR, Plis S, Sui J, Calhoun VD, 2017. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. Neuroimage 145, 137–165 1. [PubMed: 27012503]

Pervaiz U, Vidaurre D, Woolrich MW, Smith SM, 2020. Optimising network modelling methods for fMRI. Neuroimage 211, 116604 5. [PubMed: 32062083]

Sporns O, 2013. Network attributes for segregation and integration in the human brain. Curr. Opin. Neurobiol 23 (2), 162–171 4. [PubMed: 23294553]

Bzdok D, Nichols TE, Smith SM, 2019. Towards algorithmic analytics for large-scale datasets. Nat. Mach. Intell 1 (7), 296–306 Nature Research 01-7. [PubMed: 31701088]

Rosenberg MD, Casey BJ, Holmes AJ, 2018. Prediction complements explanation in understanding the developing brain. Nat. Commun 9 (1), 589 12. [PubMed: 29467408]

Shmueli G, 2010. To explain or to predict? Stat. Sci 25 (3), 289–310 8.

Finn ES, et al., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat. Neurosci 18 (11), 1664–1671 11. [PubMed: 26457551]

Smith SM, et al., 2015. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. Nat. Neurosci 18 (11), 1565–1567 11. [PubMed: 26414616]

Rosenberg MD, Finn ES, Scheinost D, Constable RT, Chun MM, 2017. Characterizing attention with predictive network models. Trends Cogn. Sci 21 (4), 290–302 Elsevier Ltd 01-4. [PubMed: 28238605]

Liem F, et al., 2017. Predicting brain-age from multimodal imaging data captures cognitive impairment. Neuroimage 148, 179–188 3. [PubMed: 27890805]

Gao S, Greene AS, Constable RT, Scheinost D, 2019. Combining multiple connectomes improves predictive modeling of phenotypic measures. Neuroimage 201, 116038 11. [PubMed: 31336188]

Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B, 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. Neuroimage 145, 166–179 1. [PubMed: 27989847]

Abraham A, et al., 2017. Deriving reproducible biomarkers from multi-site resting-state data: an Autism-based example. Neuroimage 147, 736–745 2. [PubMed: 27865923]

Varoquaux G, 2018. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage 180, 68–77 Academic Press Inc. 15-10. [PubMed: 28655633]

Rosenberg MD, et al., 2020. Functional connectivity predicts changes in attention observed across minutes, days, and months. Proc. Natl. Acad. Sci. U.S.A 117 (7), 3797–3807 2. [PubMed: 32019892]

Yip SW, Scheinost D, Potenza MN, Carroll KM, 2019. Connectome-based prediction of cocaine abstinence. Am. J. Psychiatry 176 (2), 156–164 2. [PubMed: 30606049]

Greene AS, Gao S, Scheinost D, Constable RT, 2018. Task-induced brain state manipulation improves prediction of individual traits. Nat. Commun 9 (1) 12.

Lake EMR, et al., 2019. The functional brain organization of an individual allows prediction of measures of social abilities transdiagnostically in autism and attention-d-eficit/hyperactivity disorder. Biol. Psychiatry 86 (4), 315–326 8. [PubMed: 31010580]

Opitz D, Maclin R, 1999. Popular Ensemble Methods: an Empirical Study. J. Artif. Intell. Res 11, 169–198 8.

Breiman L, "Bagging predictors," 1996.

De Bin R, Janitza S, Sauerbrei W, Boulesteix A-L, 2016. Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. Biometrics 72 (1), 272–280 3. [PubMed: 26288150]

Richiardi J, Eryilmaz H, Schwartz S, Vuilleumier P, Van De Ville D, 2011. Decoding brain states from fMRI connectivity graphs. Neuroimage 56 (2), 616–626 5. [PubMed: 20541019]

Bellec P, Rosa-Neto P, Lyttelton OC, Benali H, Evans AC, 2010. Multi-level bootstrap analysis of stable clusters in resting-state fMRI. Neuroimage 51 (3), 1126–1139 7. [PubMed: 20226257]

Nikolaidis A, Heinsfeld AS, Xu T, Bellec P, Vogelstein J, Milham M, 2020. Bagging improves reproducibility of functional parcellation of the human brain. Neuroimage, 116678 2.

Wei L, Jing B, Li H, 2020. Bootstrapping promotes the RSFC-behavior associations: an application of individual cognitive traits prediction. Hum. Brain Mapp p. hbm.24947, 3.

Hoyos-Idrobo A, Varoquaux G, Schwartz Y, Thirion B, 2018. FReM – Scalable and stable decoding with fast regularized ensemble of models. Neuroimage 180, 160–172 10. [PubMed: 29030104]

Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K, 2013. The WU-Minn human connectome project: an overview. Neuroimage 80, 62–79 10. [PubMed: 23684880]

Calkins ME, et al., 2015. The Philadelphia neurodevelopmental cohort: constructing a deep phenotyping collaborative. J. Child Psychol. Psychiatry Allied Discip 56 (12), 1356–1369 12.

Bilker WB, Hansen JA, Brensinger CM, Richard J, Gur RE, Gur RC, 2012. Development of abbreviated nine-item forms of the Raven’s standard progressive matrices test. Assessment 19 (3), 354–369. [PubMed: 22605785]

Moore TM, Reise SP, Gur RE, Hakonarson H, Gur RC, 2015. Psychometric properties of the penn computerized neurocognitive battery. Neuropsychology 29 (2), 235–246 3. [PubMed: 25180981]

Gur RC, et al., 2010. A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: standardization and initial construct validation. J. Neurosci. Methods 187 (2), 254–262 3. [PubMed: 19945485]

Van Essen DC, et al., 2012. The human connectome project: a data acquisition perspective. Neuroimage 62 (4), 2222–2231 10. [PubMed: 22366334]

Satterthwaite TD, et al., 2014. Neuroimaging of the Philadelphia neurodevelopmental cohort. Neuroimage 86, 544–553 Academic Press Inc. 01-2. [PubMed: 23921101]

Glasser MF, et al., 2013. The minimal preprocessing pipelines for the human connectome project. Neuroimage 80, 105–124 10. [PubMed: 23668970]

Joshi A, et al., 2011. Unified framework for development, deployment and robust testing of neuroimaging algorithms. Neuroinformatics 9 (1), 69–84 3. [PubMed: 21249532]

Smith SM, et al., 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23, S208–S219 no. SUPPL. 1. [PubMed: 15501092]

Frackowiak R, et al., 2003. Human Brain Function 2nd Edition. Chapter 6. Morphometry.

Shen X, Tokoglu F, Papademetris X, Constable RT, 2013. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. Neuroimage 82, 403–415 11. [PubMed: 23747961]

Hunter JD, 2007. Matplotlib: a 2D graphics environment. Comput. Sci. Eng 9 (3), 90–95.

Waskom M et al., “Mwaskom/seaborn: v0.8.1 (September 2017).” 9-2017.

Shen X, et al., 2017. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. Nat. Protoc 12 (3), 506–518 2. [PubMed: 28182017]

Hsu W−T, Rosenberg MD, Scheinost D, Constable RT, Chun MM, 2018. Resting-s-tate functional connectivity predicts neuroticism and extraversion in novel individuals. Soc. Cogn. Affect. Neurosci 13 (2), 224–232 1. [PubMed: 29373729]

Pathak PK, Rao CR, 2013. The sequential bootstrap. In: Handbook of Statistics, 31. Elsevier B.V., pp. 2–18.

Varoquaux G, Thirion B, 2014. How machine learning is shaping cognitive neuroimaging. Gigascience 3 (1) BioMed Central Ltd., 17-11.

Gabrieli JDE, Ghosh SS, Whitfield-Gabrieli S, 2015. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. Neuron 85 (1), 11–26 Cell Press 07-1. [PubMed: 25569345]

Bzdok D, Yeo BTT, 2017. Inference in the age of big data: future perspectives on neuroscience. Neuroimage 155, 549–564 7. [PubMed: 28456584]

Jiang H, et al., 2020. Predicting brain age of healthy adults based on structural MRI parcellation using convolutional neural networks. Front. Neurol 10, 1346 1. [PubMed: 31969858]

Scheinost D, et al., 2019. Ten simple rules for predictive modeling of individual differences in neuroimaging. Neuroimage 193, 35–45 Academic Press Inc. 01-6. [PubMed: 30831310]

Poldrack RA, Huckins G, Varoquaux G, 2020. Establishment of best practices for evidence for prediction: a review. JAMA Psychiatry 77 (5), 534–540 American Medical Association 01-5. [PubMed: 31774490]

Mwangi B, Tian TS, Soares JC, 2014. A review of feature reduction techniques in Neuroimaging. Neuroinformatics 12 (2), 229–244 Humana Press Inc. [PubMed: 24013948]

Barron DS et al., “Task-based functional connectomes predict cognitive phenotypes across psychiatric disease,” bioRxiv, p. 638825, 5 2019.

“Overfitting | Meaning of Overfitting by Lexico.” [Online]. Available: https://www.lexico.com/definition/overfitting. [Accessed: 28-May-2020] 2020.

Satterthwaite TD, et al., 2015. Linked sex differences in cognition and functional connectivity in youth. Cereb. Cortex 25 (9), 2383–2394 9. [PubMed: 24646613]

Zhang C, Dougherty CC, Baum SA, White T, Michael AM, 2018. Functional connectivity predicts gender: evidence for gender differences in resting brain connectivity. Hum. Brain Mapp 39 (4), 1765–1776 4. [PubMed: 29322586]

Betzel RF, Byrge L, He Y, Goñi J, Zuo XN, Sporns O, 2014. Changes in structural and functional connectivity among resting-state networks across the human lifespan. Neuroimage 102 (P2), 345–357 11. [PubMed: 25109530]

Dosenbach NUF, et al., 2010. Prediction of individual brain maturity using fMRI. Science (80-.) 329 (5997), 1358–1361 9.

Ragland JD, et al., 2002. Working memory for complex figures: an fMRI comparison of letter and fractal n-back tasks. Neuropsychology 16 (3), 370–379. [PubMed: 12146684]

Barch DM, et al., 2013. Function in the human connectome: task-fMRI and individual differences in behavior. Neuroimage 80, 169–189 10. [PubMed: 23684877]

Yu M, et al., 2018. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. Hum. Brain Mapp 39 (11), 4213–4227 11. [PubMed: 29962049]

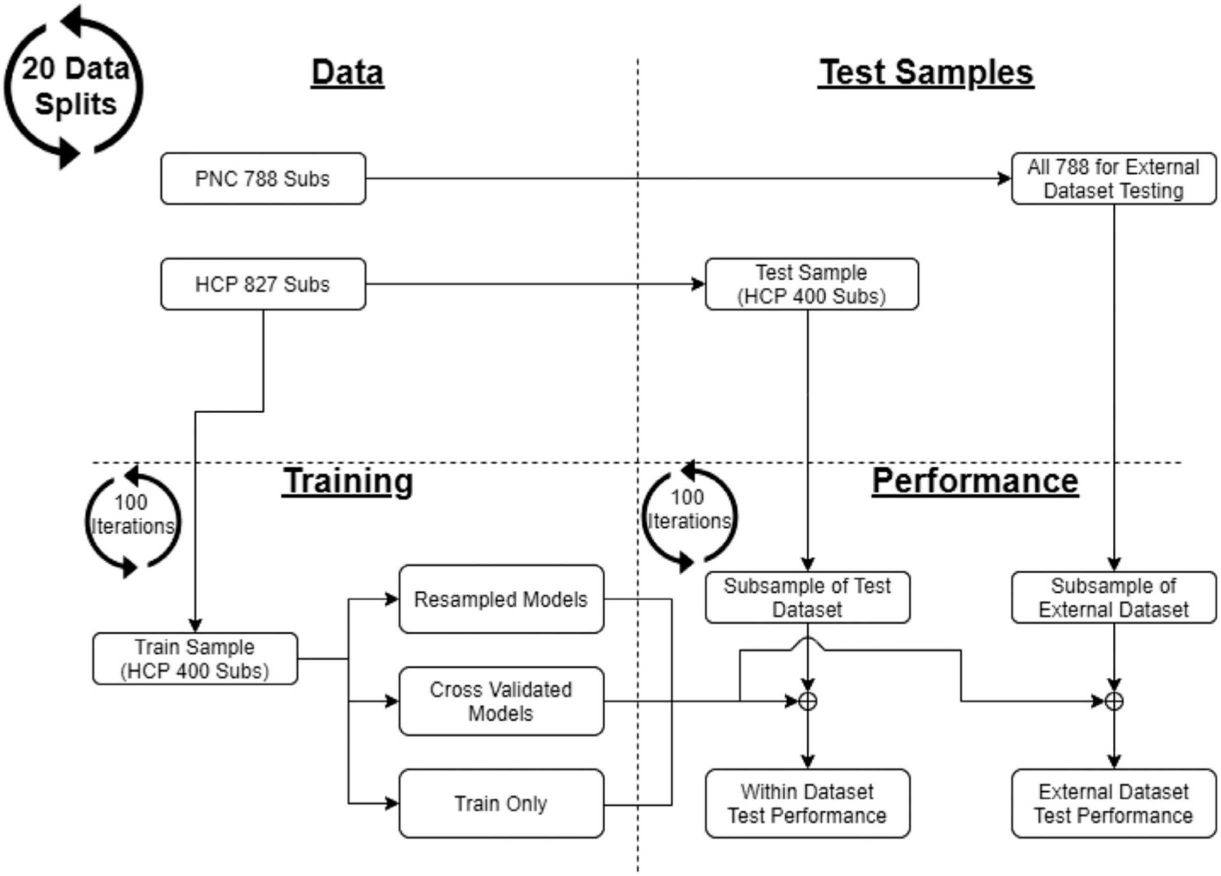
20 Data Splits
Data
Test Samples
PNC 788 Subs
All 788 for External Dataset Testing
HCP 827 Subs
Test Sample (HCP 400 Subs)
100 Iterations
Training
100 Iterations
Performance
Resampled Models
Train Sample (HCP 400 Subs)
Cross Validated Models
Train Only
Subsample of Test Dataset
Subsample of External Dataset
Within Dataset Test Performance
External Dataset Test Performance

Fig. 1.

Analytic workflow. The PNC data is only used for out-of-sample testing. The HCP data is split into train and test samples. The train sample (400 subjects) is used to train 3 types of models: (1) resample aggregated models, (2) CV models, and (3) train-only models. All models are then tested within-sample on the test HCP sample, and out-of-sample on the PNC dataset.

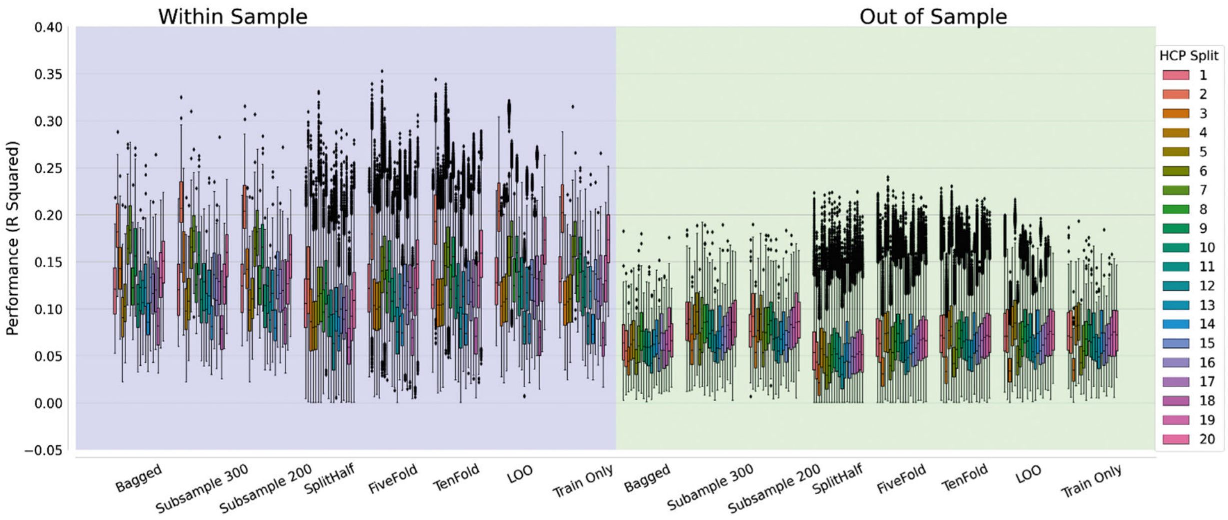

Fig. 2.

Within and out-of-sample model performance, stratified by data split. In the left panel (purple), the first three columns show performance of resample aggregated models within-sample, columns 4–7 show the CV models, and the eight column shows performance of the train-only models. Each column has 20 boxplots, color-coded (in rainbow) by train/test split. All models are tested within-sample on 100 random subsamples of 200 subjects from the HCP test sample. The second panel (green) shows the performance of the same models (same order as the left panel) out-of-sample using random subsamples of 200 subjects from the PNC data set.

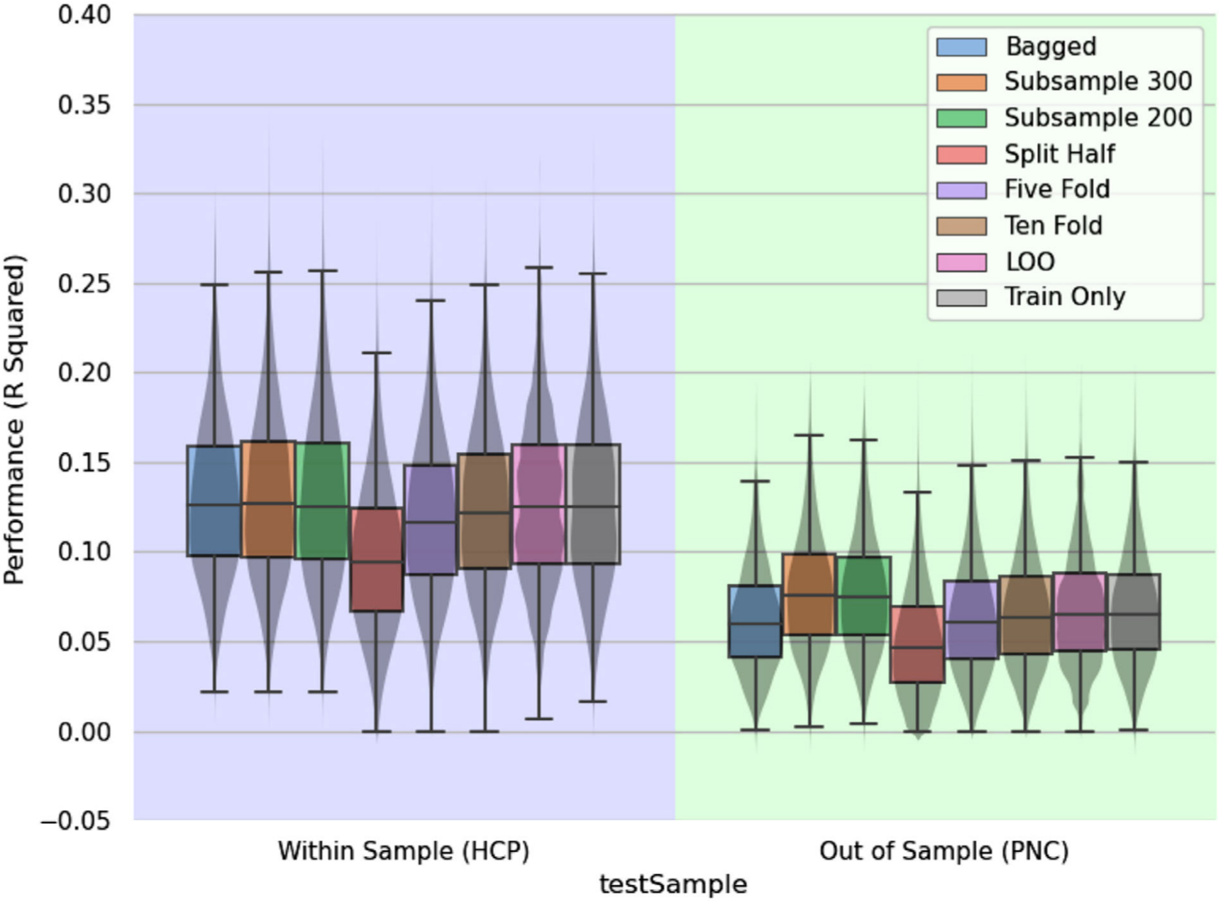

Fig. 3.

Within and out-of-sample model performance. Column one (shaded in purple) shows performance of all models, across all data splits, within-sample. All models are tested on 100 random subsamples of 200 subjects from the test sample of the HCP data set. The second column (shaded in green) shows the performance of the same models tested on random subsamples of 200 subjects from the PNC data set. The box and whisker plots show the median, interquartile range, and 5%–95% markers of the performance distribution. The underlying shaded violin plots show the shape of the model performance distribution.

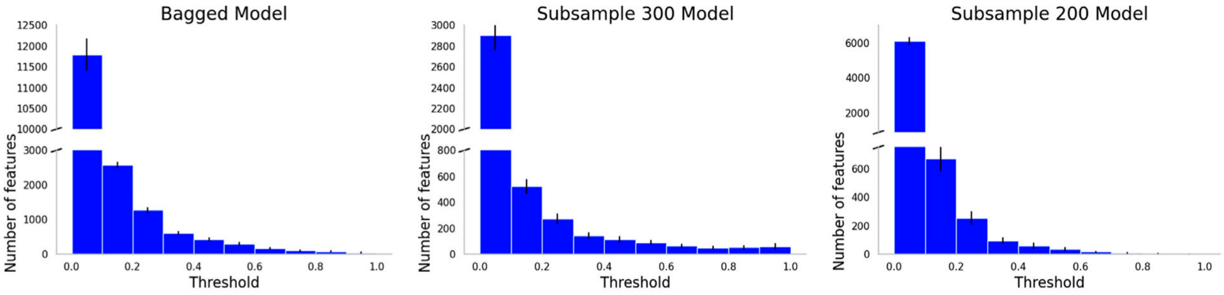

Fig. 4.

Distribution of feature (edge) occurrence across subsamples for the ensemble models. For the bagged model (left), nearly 11,851 features occur in between 0% and 10% of bootstraps, compared to 6056 for the subsample 200 model (right) and 2839 for the subsample 300 model (center).

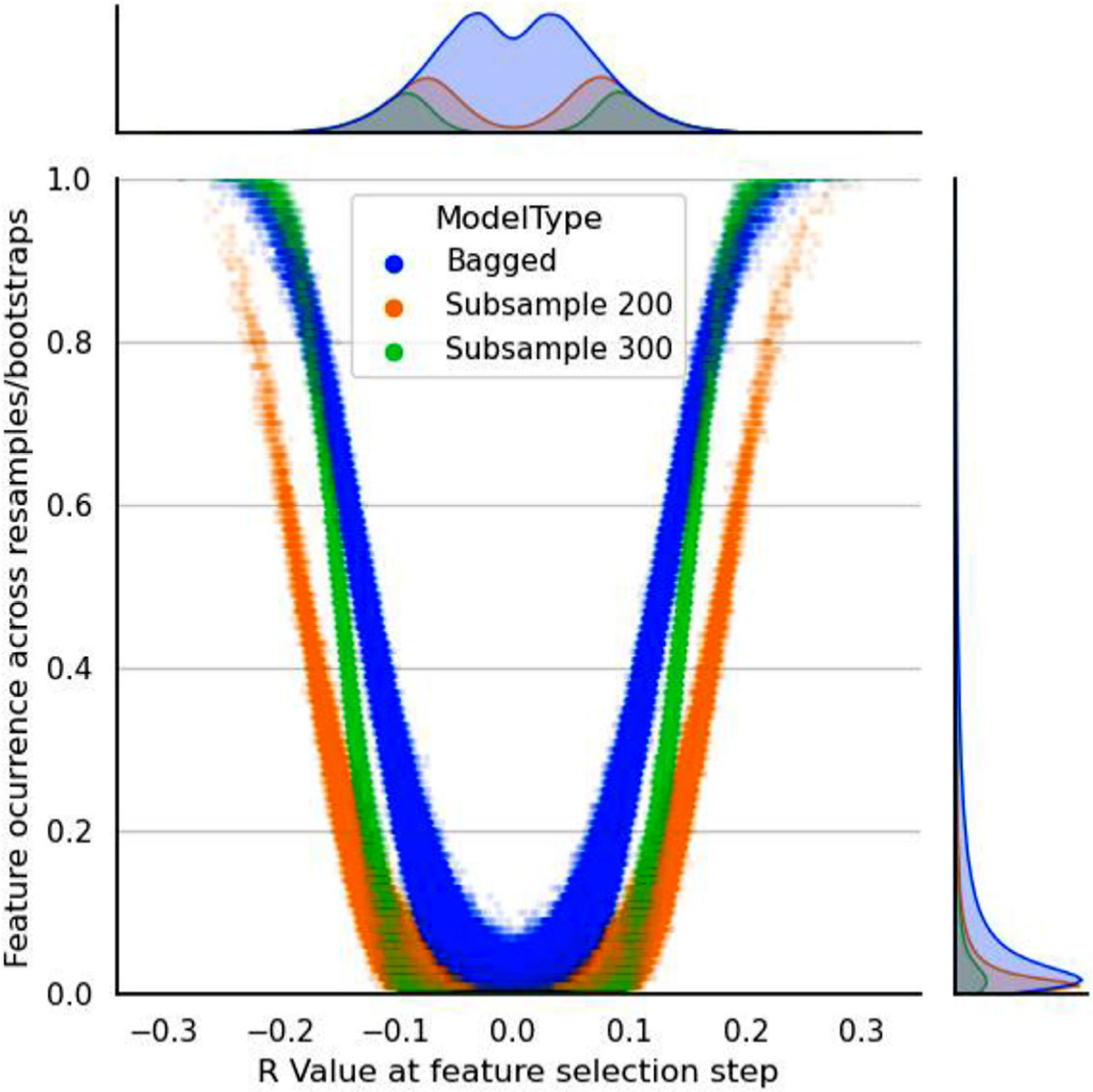

Fig. 5.

Relationship between effect size and feature occurrence for each aggregated model, across all resamples. The bagged models are shown in blue, the subsample 200 in orange, and the subsample 300 in green. The subplots on the right and top show probability density plots of the feature occurrence and effect size respectively.

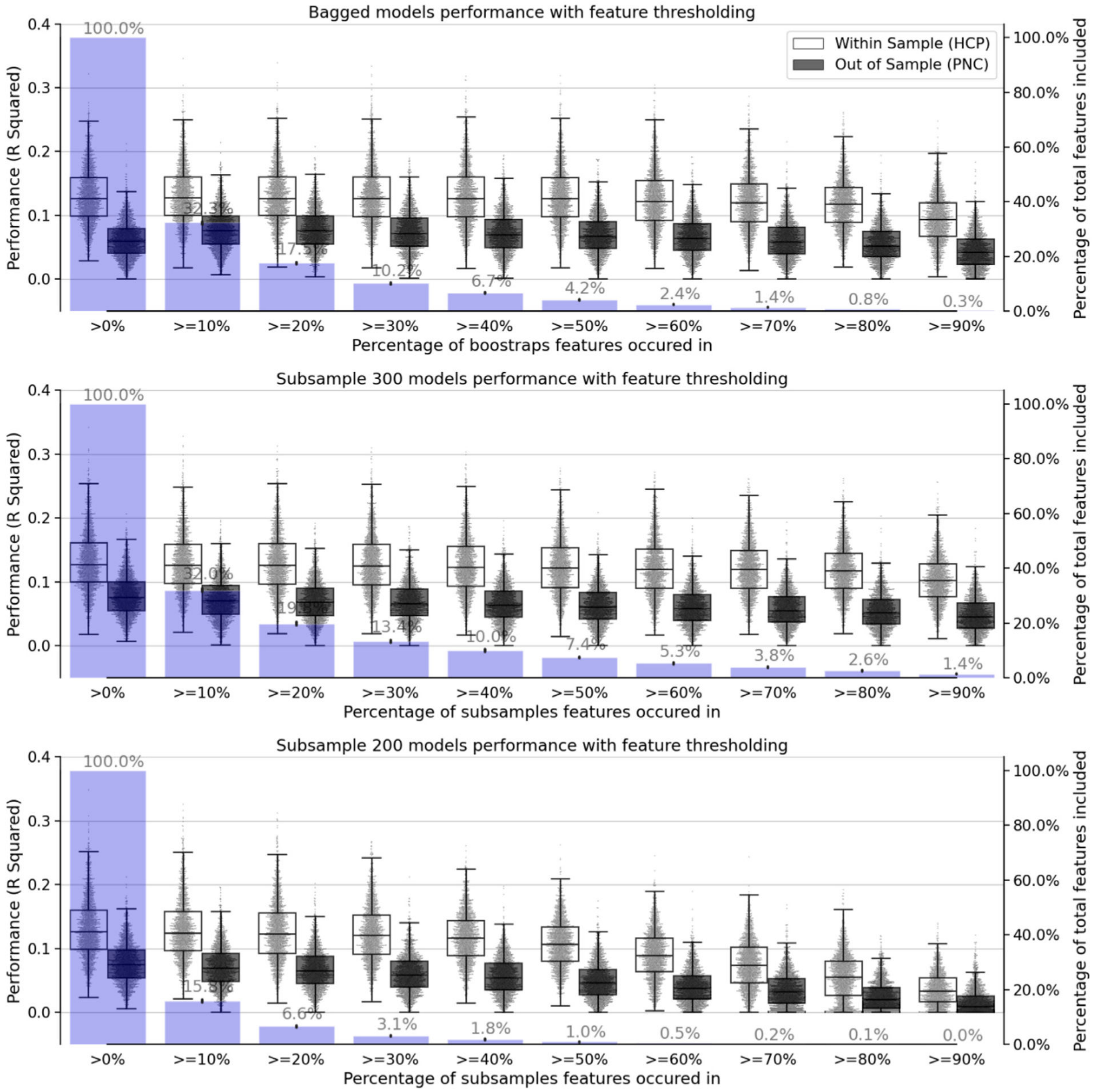

Fig. 6.

Resample aggregated model performance within sample (white boxplots), and out of sample (gray boxplots) across feature frequency thresholds. This reflects the performance as tested on subsamples of 200 participants. The box and whisker plots show the median, interquartile range, and 5% – 95% markers of the performance distribution. The underlying shaded distribution shows the individual data points. The top panel shows the performance of the bagged models, as the feature threshold is increased (reducing the number of features included). Middle shows the performance of the subsample 300 models, as the feature threshold is increased. Bottom shows the performance of the subsample 200 models, as the feature threshold is increased. In the background of all plots, a density-based histogram of the percent of features (as a function of all features selected for a given model) included is shown in blue.