**Methods**

**Statistical Analysis**

*Gradient Boosting Machines*

Gradient boosting refers to the process of iteratively modeling a response variable by sequentially adding simple models (e.g., shallow decision trees) to an ensemble that cumulatively generate a final prediction. Each new model predicts the residuals of the existing in order to further minimize a loss function (e.g., misclassification rate or mean-squared error). At every iteration, a new model is added that maximally decreases the prediction error from the previous iteration. As a result, the observations (e.g., the subjects) associated with the largest residuals will be most influential in the fitting of each subsequent model. This amounts to most efficiently minimizing the loss function by iteratively progressing along its gradient. Gradient Boosted Machines (GBMs; Friedman, 2001) can be therefore generalized to predict a variety of categorical or continuous response types, minimize any differentiable loss function, and often perform well in prediction and classification problems despite noisy and heterogeneous data.

*Classification of Previously Institutionalized Group Status*

To identify robust differences between the previously institutionalized (PI) and comparison (COMP) groups within the domain of SGMV, we trained, tuned, and validated 100 GBM classifiers on 100 novel resamples of the data. Since our goal is to build a model that is sensitive and specific in its identification of PI group status, each resample began by randomly assigning 70% of the PI group to the training dataset (train). To control for the effect that age and group sample size imbalance might have on model performance, we added an equal number of randomly selected, age-matched COMPs to train (Snoek, Miletić, & Scholte, 2019). The remaining 30% of PI subjects and approximately 45% of COMP subjects were assigned to the testing dataset (test). Finally, we generated a tuning dataset (tune) by randomly extracting 40% of the subjects in train. To maintain independence between train, tune, and test, all resamples were generated by subject/participant ID (Poldrack, Huckin, & Varoquaux, 2020), so subjects with repeat scans would either have both scans in the same dataset, or one scan in a single dataset and the other scan dropped. Finally, to control for the confounding effect of total Intra-Cranial Volume (ICV), we built a regression model to predict all SGMV features on ICV using train (Snoek, Miletić, & Scholte, 2019). The regression model parameters were used to adjust SGMV features in tune and test to avoid data leakage (Poldrack, Huckins, & Varoquaux, 2020).

For each of the 100 resamples, we trained a GBM to classify group membership using the ‘gbm’ function in the R package ‘gbm’ (Greenwell et al., 2019). The models contained an ensemble of gradient boosted decision trees and predicted the probability of membership in the PI group using a logit link function. Optimal hyperparameter values were assessed and selected using a grid search. We investigated the hyperparameters that determine the minimum number of subjects allowed in each tree’s terminal nodes, the weight applied to the residuals from each new tree, and the number of splits permitted by each tree, denoted *n.minobsinnode*, *shrinkage,* and *interaction.depth* by the ‘gbm’ authors respectively. For each combination of hyperparameter values, we fit a candidate model with train data and calculate the area under the receiver operating characteristic curve (AUC) for its classifications on tune. We retained the final model, and hyperparameters, that maximized this metric. Following model selection, we obtained the predictions of each final model for its respective test set. We computed the final vector of predicted probabilities of group status by averaging the predicted probabilities made for each subject when each subject occupied a given test set. Therefore, the cross-validation estimate of predictive performance is the AUC computed from these averaged predictions with respect to the true group labels.

To assess the statistical significance of the out-of-sample AUC, we conducted permutation testing. To do so, we independently randomized the response vectors of the train, tune, and test sets 1000 times. We then fit models on train to classify these random group labels from the corresponding set of SGMV features. Thus, for each of the 100 resamples, we fit 1000 null models, which generated 1000 null prediction vectors for the given test set. As in the case of the true models, the predictions from these null models were averaged across the resamples, resulting in 1000 vectors of averaged null predictions. From these vectors we calculated a distribution of null AUCs with which to assess our observed value. We compute and report the permutation p-value for the observed cross-validated AUC as the proportion of null AUCs greater than that observed.

*Computation of Cross-Validated Permutation Variable Importance*

We obtained each SGMV feature’s cross-validated permutation variable importance (CVPVI) and corresponding p-value using the R package ‘vip’ (**CITATION**). We compute the variable importance by measuring its contribution to the AUC in prediction of the test data (Janitza et al., 2015); specifically, the variable importance is the percentage that the model improves prediction, in terms of test set AUC, when we add the given variable to the model. We obtain the statistical significance of each variable’s CVPVI by comparing it to the distribution of CVPVIs for that variable that we obtained when we fit the 100,000 null models (Altmann, 2013).

To derive the vector containing the CVPVI for the *p* variables in each model, we perturb the model’s respective test set by permuting one SGMV feature at a time, and then recompute its predictions and resultant AUC. The true AUC (i.e., the AUC computed on the unperturbed test set) is then divided by the recomputed AUC and the quotient is denoted the permuted variable’s CVPVI. A CVPVI larger (smaller) than 1 indicates that the variable increases (decreases) the model’s AUC in prediction of the test set. To avoid introducing noise into our estimates of each CVPVI due to the randomness of permutation, we repeat the permutation step 100 times and take the average of the resultant CVPVIs. This is repeated for every variable in each model, such that we obtain 100 observed CVPVI vectors and 100,000 nulls. To estimate the true CVPVI and its null distribution, we average the observed and null vectors across the 100 resamples resulting in one vector of length *p* containing an average CVPVI for each variable, and one *p* x 1000 matrix containing the null distribution of the averaged CVPVI for each variable. Therefore, the CVPVI for each variable *i* in the set of SGMV predictors is defined,

where *R* represents the number of resamples, *p* refers to the number of permutations of each variable, *AUCR* is the prediction AUC for the model trained and validated with the *R*-th (cross-validation or null) resample, and represents the prediction AUC following the *p*-th permutation of the *i*-th variable. The p-value associated with each CVPVI is the proportion of null CVPVI (for that variable) exceeding the observed value.

**2 Results**

**2.1 Prediction of Groups and Model Assessment**

To assess the generalizability of our ability to predict group status from SGMV features, we examined the cross-validation AUC and its corresponding permutation test p-value. Figure 1 shows the null distribution for the cross-validation AUC relative to the observed value and Figure 2 provides the AUC curve of the final, averaged predictions in a 95% bootstrap confidence interval. The observed value of .529 [.429, .630] was significantly greater than the permuted null, p < .001. As shown in Figure 3, this result is largely driven by accurate classifications of COMP participants, as the model true negative rate is high (75.7%), and false positives are low (24.3%).

Figure 1

*Null Distribution and observed cross-validation AUCs*

A screenshot of a social media post

Description automatically generated

Note. The gaussian kernel density curve (solid line) shows along the y-axis the density of AUCs values, where said values are explicated on the x-axis. The observed value from the final predictions, (AUC = .529) is shown in the dotted line, and exceeds all null values.

Figure 2

*Confusion matrix for final predictions*

*A screenshot of a cell phone

Description automatically generated*

Note. Across the rows are the predictions for all subjects in each group (such that the means of the rows equal the group sizes). COMP appear in the top row, and PI across the bottom. Within each row, the number in the first column corresponds to the number of predictions made that a participant is in the COMP group, the second column is analogous for PI group.

Figure 3

*Receiver-Operating Characteristic Curve within a 95% Bootstrap Confidence Interval*

**A picture containing text, map

Description automatically generated**

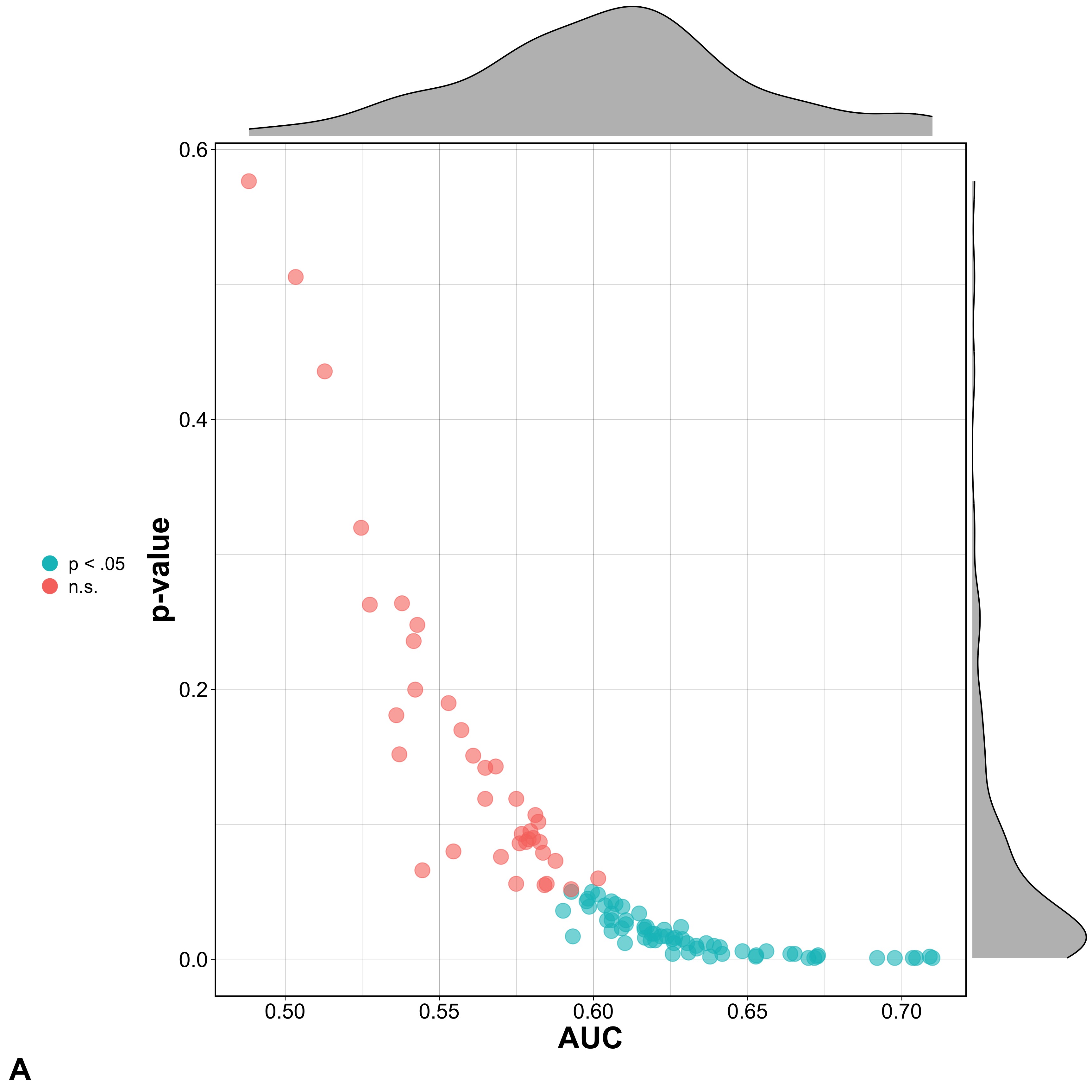
**2.2 Assessment of Variable Importance**

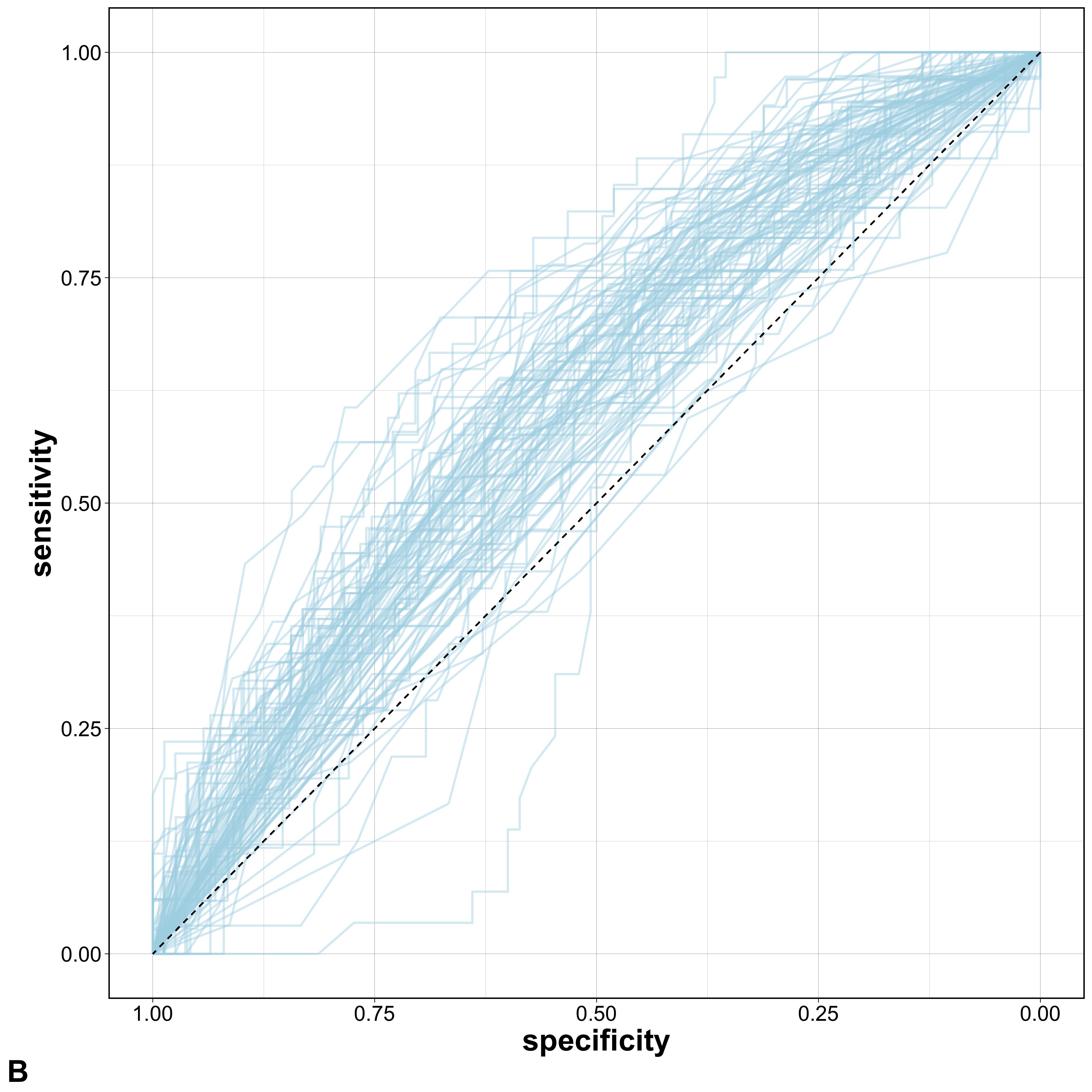
To examine which SGMV features most significantly contributed to accurate predictions of group status in new data, we show the CVPVI and its corresponding permutation p-value for each variable in Figure 4. Though not significant at alpha = .05, the strongest contributors to the prediction of test set group status include the right pallidum (CVPVI = 1.027, p = .08), right ventral diencephalon (CVPVI = 1.012, p = .10), followed by the left accumbens (CVPVI = 1.010, p = .13), right caudate (CVPVI = 1.009, p = .12), and right amygdala (CVPVI = 1.008, p = .17).

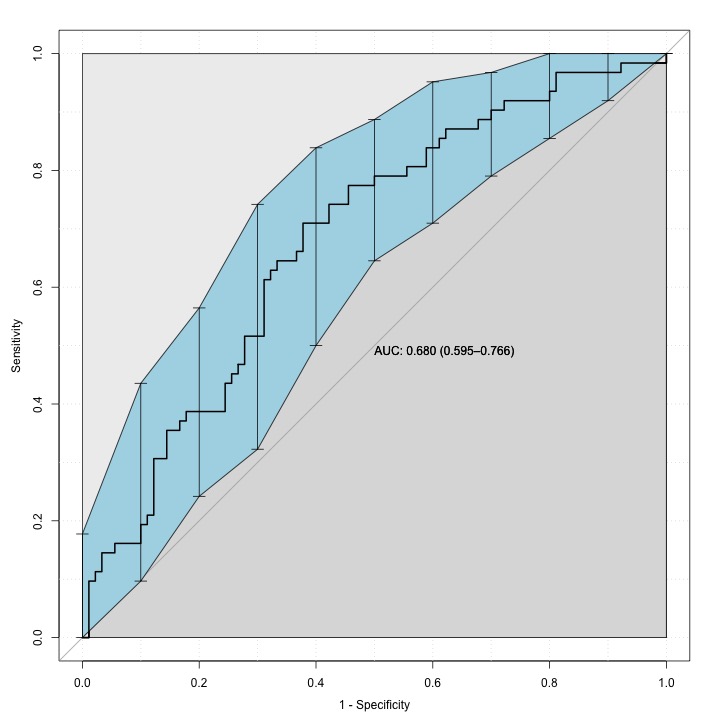
Figure 4

A screenshot of a cell phone

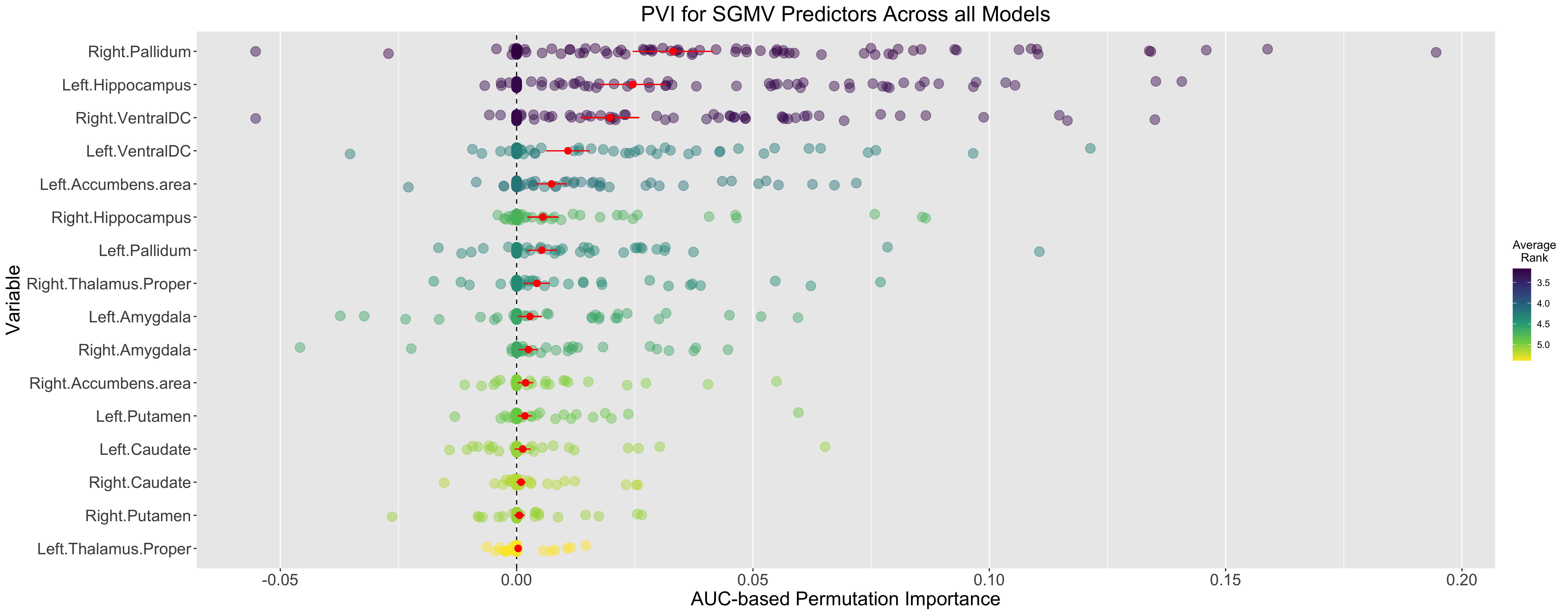
Description automatically generated

Figures

****

**Figure 1.** Plots of the performance of the model built within each cross-validation iteration. Graph **A** shows the AUC (x-axis) and p-values (y-axis) for each optimal model’s test-set predictions. *p*-values are computed by comparing the observed AUC to 1000 null values after permuting the test set’s group labels 1000 times. Graph B shows the 100 ROC curves for the test set predictions of each model; one blue line is plotted for each cross-validation iteration.

**Figure 2.** The ROC curve generated from the vector of each subject’s predicted probability of PI group membership, averaged across the model that excluded the subjects from the training or tuning subsets. The blue region shows the bootstrap 95% confidence interval from 2000 resamples, while the dark grey region quantifies the AUC. The 45% degree line bisecting the graph represents an AUC of .5, corresponding to random predictions.



**Figure 3.** Each variable’s 100 variable importance scores, computed at each cross-validation iteration on a given model’s respective test set. The red dot indicates the average PVI across the 100 cross-validation repetitions, accompanied by the 95% confidence interval of the standard error for the mean PVI estimate. Colors indicate the average rank for the variables across the 100 models. Within a given model, ranks can range from 1 to *i*, where *i* is the number of SGMV predictors and a rank of 1 corresponds to the largest PVI*i*.