**Methods**

**Participants.** We analyzed the data of two cohorts of children. Both cohorts contained subjects who were previously institutionalized (PI) in either orphanages or foster care, and comparisons (COMP) or controls who had not. One cohort, those with whom models were trained, were sampled from Los Angeles (N = 151) and included 61 PI and 90 COMP participants. The other cohort were sampled from New York City (N = ) and included 30 PI and 41 COMP. The generalizability of our models and findings were validated with New York City sample. The mean age in the train cohort is 10.6 (SD = 3.53) years, and 58% are female. In the validation cohort, the mean age is 12.27 (SD = 3.32) years, and 61% are female.

**Statistical Analysis**

To rigorously investigate the effect of early life institutionalization on subcortical grey matter development, we trained traditional multivariable and computational multivariate statistical models to identify subjects who had been previously institutionalized from those who had not. Specifically, we fit a multivariable logistic regression model, created an ensemble of such logistic regressions, and used the gradient boosting algorithm to fit an ensemble of gradient boosted decision tree models (i.e., gradient boosted machines). Each model was fit with the LA cohort to identify subjects who belonged to the groups defined above, and predictions were made for subjects in the NYC sample. We assessed the predictive validity of the models of group differences in subcortical grey matter volume by evaluating the predicted probabilities that each subject belonged to their respective groups, as well as through non-parametric permutation tests of significance for the overall Area Under the receiver-operating characteristic Curve (AUC) of each model. To further assess the validity of the relationship between specific brain regions and past exposure to institutionalization, we randomly permuted the volumetric measurements of each subcortical region one at a time, and recalculated the predictions both for subjects in the LA and NYC samples to assess the “variable importance”, or in this case the strength of the relationship between the grey matter volume and the grouping factor. We repeated these steps after *first* randomly permuting the grouping factor labels, in order to conduct non-parametric permutation tests of the variable importance for each subcortical region described above. Finally, we conducted these analyses with two separate versions of our subcortical grey-matter volume measurements: in one set of analysis we conducted the above procedure without adjusting the subcortical grey-matter volume, and in a second set of analyses we replaced each subcortical feature with the residuals from a linear regression in which we regressed the subcortical feature on the subjects’ age and gender to control for these variables in later analyses. In so doing, the that classified groups using the adjusted subcortical grey matter assess the relationship between early life institutionalization and brain volume after accounting for the linear effect age and gender. Since it is possible that such effects are non-linear, however, we report and compare the results from the models with both the adjusted and unadjusted predictors.

**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 generates 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). Since the mode at each iteration is fit to minimizing the existing residual error, the observations (e.g., the subjects) associated with the largest residuals will be most influential in the fitting of each subsequent model. As a result, the method is often better able to generate more accurate predictions, relative to parametric linear models, for subjects whose predicted values yield the largest residuals. This amounts to most efficiently minimizing the loss function by iteratively progressing along its gradient through the use of boosting, hence the term “Gradient Boosted” Machines (GBMs; Friedman, 2001).

**Discovery of Brain Features related to Previous Institutionalization.** To robustly identify regions of the brain where SGMV differed between PI and COMP subjects, we fit GBMs to model the probability that subjects belonged to each group based on their SGMV features. We did so using the ‘gbm’ software package (Greenwell, Boehmke, Cunningham, et al., 2020), written for the R statistical computing language and available at https://cran.r-project.org/web/packages/gbm/ (R Core Team, 2020). We trained our models with the scans collected from subjects in the SB study. To control for the effect that age, group sample size imbalance, and random resampling could have on the accuracy of our model performance estimate, we generated 100 training datasets from SB consisting of the 90 PI subjects and an age-matched subsample of 90 COMP subjects (Snoek, Miletić, & Scholte, 2019). From each of the 100 resultant training datasets, we randomly selected 30% of PI and 30% of COMP subjects into a tuning dataset. To maintain independence between all partitions of the data, sampling occurred on the basis of subject ID, so that any participant with multiple scans would appear in only the training or testing dataset, or neither (Poldrack, Huckin, & Varoquaux, 2020).

Using each of the 100 pairs of training and tuning datasets, we fit a GBM to model the probability of group membership for subjects in the training dataset given the SGMV features, and we assessed the predictive accuracy of the model at each boosting iteration with the subjects in the tuning dataset. The number of trees used in the final model is therefore the iteration at which prediction accuracy is maximized for subjects in the tuning dataset (Greenwell, Boehmke, Cunningham, et al., 2020). We held other hyperparameters constant at the levels suggested by the authors of the ‘gbm’ package, with one exception: since we were interested in exploring potential interactions between SGMV features, we allowed the tree at each iteration to build up to at most six split points, rather than one. In addition, we generated equal weights for each subject that were directly inversely proportionate to the number of scans collected on a subject, in order to prevent the algorithm from overfitting to subjects with a greater number of scans.

We validated each of the 100 models by generating and evaluating their predictions for the subjects in the ELFK dataset. To do so, we generated 100 vectors of predicted values, within which each value pertained to the predicted probability that an ELFK subject belonged to the PI group (e.g., a predicted value of .66 indicates that the model predicts with 66% certainty that the subject belongs to the PI group). We averaged each of these vectors subject-wise to obtain a final prediction for each ELFK participant. To calculate final model accuracy, we rounded predictions to the nearest integer to determine if the average model predicts a subject to be in the COMP group (i.e., if the average prediction rounds to 0) or PI group (i.e., if the average prediction rounds to 1).With the unrounded final predicted probabilities we also calculated the ROC-AUC.

To assess the statistical significance of the out-of-sample accuracy and ROC-AUC, we conducted permutation testing. We generated a null distribution by repeating our entire pipeline 10,000 times after randomly shuffling the response vectors of each dataset. Specifically, after resampling our 100 train and tune sets from the SB population, we permuted the group labels of train and tune, independently, 100 times. We then built 100 models to predict these random response vectors from subjects’ SGMV, again using tune to identify the optimal boosting iteration. For each such model, we also permuted the response variable (i.e., the group labels) in ELFK, and validated the null models with that permutation of the test data. We concatenated the accuracy and ROC-AUC resulting from the 100 models built at each of the 100 cross-validation iterations, to form a null distribution of the model accuracy and ROC-AUC, each containing 10,000 null values. These values were compared to the accuracy and ROC-AUC obtained from the averaged predicted probability or classification vector of the 100 true models. We compute and report the permutation p-value for the observed cross-validation accuracy and ROC-AUC of the final averaged predictions as the proportion of null values greater than their respective observed value.

In addition to the overall performance of the model, we were interested in whether the SGMV features provided any additional predictive power after controlling for the covariates of age, total intra-cranial (cortical) brain volume (ICV), and sex. To do so, we used the one true and 10,000 null prediction vectors to fit 10,001 logistic regression models of ELFK subjects’ group labels. We predict the actual ELFK group labels from age, then ICV, then sex, and lastly we add the term for either the true or null predicted values from the GBM pipeline described above. From each logistic regression, we conduct ANOVA to obtain the reduction in residual deviance attributable to the predicted values from the (true or null) GBM model(s). The p-value associated with the decrease in residual deviance attributable to the GBM model is defined as the proportion of the 10,000 null differences in residual deviance that exceed the observed value.

**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 ‘vip’ software package (Greenwell, Boehmke, & Gray, 2020), written for the R statistical computing language and available at https://cran.r-project.org/web/packages/vip/ (R Core Team, 2020). We compute the variable importance by measuring its contribution to the ROC-AUC in prediction of the ELFK data averaged across the iterations of cross-validation (Janitza et al., 2015). Specifically, the variable importance is the average percentage, across the 100 optimal models, by which the ROC-AUC improves when a given SGMV feature is added to a model. To compute a single permutation variable importance within one cross-validation iteration for one variable, we permute (here, 100 times) the data of the SGMV feature from within the set of predictors in the test data, and each time recalculate the prediction accuracy metrics for the model (i.e., ROC-AUC). These 100 values are averaged to estimate the performance of the model without the permuted variable. We then obtain the ratio of the true cross-validation performance of the model, *AUC*, to this new value, *AUC*\*, and subtract it from 100% to convert it to the percent *improvement* of having had the *un-*permuted variable in the model relative to not. This is repeated for all variables, and we average the resultant CVPVI obtained from each model to obtain the final CVPVI of each SGMV predictor. The CVPVI for the *i*-th SGMV feature is therefore:

where *R* represents the number of resamples, *p* refers to the number of permutations of each variable, *AUCR* is the prediction ROC-AUC for the *R*-th optimal model, and represents the prediction ROC-AUC following the *p*-th permutation of the *i*-th variable.

We obtain the statistical significance of each variable’s CVPVI by comparing it to a distribution of null CVPVIs that we obtained after first permuting the group labels of the test data (Altmann, 2013). To do so, we generate calculated the CVPVI for each of th 10,000 null models described above. For each one, after permuting the group labels of the test data, we calculate the CVPVI of each SGMV predictor in the same way that we do for the true values. The p-value for each variable’s CVPVI is the percentage of null CVPVI, for the same variable, that are larger than that variable’s true CVPVI.

For the sake of comparison, we repeat this entire procedure, including model training and computation of CVPVI, using logistic regression models instead to generate all predictions. For each logistic regression model, we used the full 180 participant training dataset, since tuning was unnecessary. We fit the models to estimate the probability that a subject belonged to the PI group using the ‘glm’ function in R (R Core Team, 2020). In order to avoid biasing our cross-validation estimates of model performance due to class imbalance in the testing dataset, we did not estimate the intercept parameter. As in the previous models, we employed weighted least squares to restrain the model to consider all subjects equally, rather than all scans equally. Following model training, predictions are generated for the subjects in ELFK, accuracies and CVPVI are computed identically to the above.

Finally, following the classification model training and validation, we conducted a linear regression to determine the implications for the results with respect to three metrics of psychopathology: CBCL Total *t*-scores, CBCL Internalizing *t*-scores, and CBCL Externalizing *t*-scores. To do so, we regressed the CBCL scores on the possible confounders of age, ICV, sex, and true group labels before adding terms for only those SGMV features found to significantly distinguish groups in the above models.

**Results**

The overall results of the

**Model Performance Classifying Groups**

The overall accuracy of the The cross-validation ROC-AUC and accuracy derived from the average predicted values and probabilities from the 100 GBM models are plotted in Figure 1. The models showed a high degree of sensitivity and specificity, achieving a ROC-AUC of .752, *p* < .0001. The model was 64.79% accurate, which was significantly higher than the permuted nulls at *p* < 01.

Figure 2 shows the CVPVI for each variable and their respective null distributions. Left hippocampus demonstrated the highest CVPVI, improving the ROC-AUC of the model on average by 22.94%, which was significant at *p* < .01. The next strongest predictor is the right ventral diencephalon, which improved the model an average of 6.35% in terms of ROC-AUC, though for which the permutation *p* value is .055.

Figure 3 shows the both of which reached statistical significance. Specifically, left hippocampus contributed an average of . and generated a ROC-AUC of , also significant at was significantand logistic regression models are presented in Figure 1 and 2, respectively, along with the corresponding 10,000 null values for each, and the final performance value derived from the averaged predictions. For the GBM models, the final averaged predictions resulted in significant cross-validation accuracy and ROC-AUC (64.71%, boot-CI [53.14%, 76.27%], *p =* .0171; 0.77, boot-CI [0.65, 0.88], *p =* .001). The logistic regression models were not significantly better than their permuted nulls (51.47%, boot-CI [39.36%, 38.30%], *p =* .6949; 0.52, boot-CI [0.38, 0.67], *p =* .3546). The CVPVI for the significant GBM models is provided in Figure 3. The variables whose CVPVI reached significance, at alpha equal to .05, relative to their permuted nulls included the left hippocampus (CVPVI = 22.39%, *p =* .01) and the right ventral diencephalon (CVPVI = 6.71%, *p =* .05).

As presented in Table 1, The confound regression for the predictions of the GBM models showed that SGMV features significantly distinguished the PI and COMP groups in ELFK, after controlling for age, sex, and total cranial volume (Chi-Sq. = 16.41, df = 1, *p =* .0001). The same was not true for the logistic regression model (Chi-Sq. = 1.13, df = 1, *n.s.*). Despite the ability for the GBM classifier to identify groups significantly, Table 3 shows that the regions responsible for these accurate predictions did not significantly predict mental health outcomes, as measured by the CBCL Internalizing (*F* = 0.862 (6, 45), *n.s.*), Externalizing (*F* = 1.197 (6, 45), *n.s.*), or Total *t*-scores (*F* = 1.615 (6, 45), *n.s.*) when controlling for group differences, age, sex, and total cortical volume.

Tables and Figures

Table 1. ANOVA and regression estimates for the model predicting group labels from GBM model predictions and confounds

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Beta | | Chi-Square | |  |
| Term | Est. | *p* | Est. | *p* | Residual Deviance |
| Cortical Volume | < 0.001 | .161 | 2.16 | .142 | 88.93 |
| Age | .232 | < .05 \* | 5.01 | < .05 | 87.17 |
| Sex | -.373 | .586 | 0.30 | .583 | 86.85 |
| GBM Predictions | 15.5 | < .01 \*\* | 16.41 | < .0001\*\*\*\* | 69.58 |
| Null | - | - | - | - | 91.365 |

Table 2. Regression estimates for the model predicting CBCL *t*-scores from data-driven ROIs and confounds.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | CBCL Internalizing | | CBCL Externalizing | | CBCL Total | |
| Term | Beta | *p* | Beta | *p* | Beta | *p* |
| Group | 5.74 | .278 | 5.31 | .112 | 5.75 | .094 |
| Age | .374 | .442 | 0.21 | .686 | .607 | .248 |
| Sex | -0.03 | .990 | -1.33 | .657 | -2.97 | .337 |
| Cortical Volume | < 0.001 | .442 | < 0001 | .95 | < 0.001 | .658 |
| Left Hippocampus | 0.005 | .378 | -0.004 | .503 | -0.004 | .521 |
| Right Ventral Diencephalon | 0.001 | .806 | -0.0007 | .882 | -0.0009 | .852 |

Chart

Description automatically generatedFigure 1. The distribution of misclassification rate and ROC-AUC scores for the GBM models at each cross-validation iteration (in blue) and their corresponding null distributions (in red). The dotted black line marks the score of the averaged predictions.

Chart, surface chart

Description automatically generated

Figure 2. The distribution of misclassification rate and ROC-AUC scores for the logistic regression models at each cross-validation iteration (in blue) and their corresponding null distributions (in red). The dotted black line marks the score of the averaged predictions.

Table, Excel

Description automatically generated

Figure 3. The black dots indicate the CVPVI of each variable, representing the % the variable improves from adding the given feature to the model. The colored dots reflect the null distribution of those values for each variable’s CVPVI. The left hippocampus and right ventral diencephalon are the only variables to have exceeded their null distribution values over 95% of the time.

Chart, surface chart

Description automatically generated