bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Title: Machine Learning Classification of Attention-Deficit/Hyperactivity Disorder Using Structural MRI Data

Authors: Yanli Zhang-James[1], Emily C Helminen[2], Jinru Liu[3], the ENIGMA-ADHD working group*, Barbara Franke[4,5], Martine Hoogman[5,6], Stephen V Faraone[1,7]

Affiliations:

1. Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, New York
2. Department of Psychology, Syracuse University, Syracuse, New York
3. University of Illinois at Urbana-Champaign,
4. Department of Psychiatry, Radboud university medical center, Nijmegen, The Netherlands
5. Department of Human Genetics, Radboud university medical center, Nijmegen, The Netherlands
6. Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands
7. Department of Neuroscience and Physiology, SUNY Upstate Medical University, Syracuse, New York

Corresponding author:

Stephen V. Faraone, Ph. D.

SUNY Upstate Medical University

750 E Adam St,

Syracuse, NY, 13210

315-464-3113, 315-849-1839 (fax)

sfaraone@childpsychresearch.org

Short running title: Image Classification for ADHD

Key words: Attention-Deficit/Hyperactivity Disorder, biomarkers, classification, machine learning, MRI

Word counts:

Abstract: 246

Main Text: 3981

Numbers of Figures: 5

Numbers of Tables: 4

Supplementary Material: 1 Supplementary Figure and 2 Supplementary Tables

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgment:

Dr. Faraone is supported by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 602805, the European Union's Horizon 2020 research and innovation programme under grant agreements No 667302 & 728018 and NIMH grants 5R01MH101519 and U01 MH109536-01. Dr. Franke is supported by a personal Vici grant (016-130-669) and Dr. Hoogman from a personal Veni grant (91619115), both from the Netherlands Organization for Scientific Research (NWO). We thank Margaret Mariano for administrative assistance and proofreading the manuscript.

The ENIGMA Working Group gratefully acknowledges support from the NIH Big Data to Knowledge (BD2K) award (U54 EB020403 to Paul Thompson). **The ENIGMA-ADHD working group** contributing members: Ambrosino, S., Anikin, A., Asherson, P., Banaschewski, T., Baranov, A., Baumeister, S., Baur-Streubel, R., Bellgrove, M., Biederman, J., Bralten, J., Bramati, I., Brandeis, D., Brem, S., Buitelaar, J., Castellanos, F., Chantiluke, K., Christakou, A., Coghill, D., Conzelmann, A., Cubillo, A., Dale, A., de Zeeuw, P., Doyle, A., Durston, S., Earl, E., Epstein, J., Ethofer, T., Fair, D., Fallgatter, A., Frodl, T., Gogberashvili, T., Haavik, J., Hartman, C., Heslenfeld, D., Hoekstra, P., Hohmann, S., Høvik, M., Jahanshad, N., Jernigan, T., Kapilushniy, D., Kardatzki, B., Karkashadze, G., Kelly, C., Khadka, S., Kohls, G., Konrad, K., Kuntsi, J., Lesch, K., Lundervold, A. J, Malpas, C., Mattos, P., McCarthy, H., Mehta, M., Mostert, J., Namazova-Baranova, L., Nigg, J., Novotny, S., Oberwelland Weiss, E., O'Gorman Tuura, R., Oosterlaan, J., Oranje, B., Paloyelis, Y., Pauli, P., Plessen, K., Ramos-Quiroga, J., Reif, A., Reneman, L., Rubia, K., Schrantee, A., Schwarz, L., Schweren, L., Seitz, J., Shaw, P., Silk, T., Skokauskas, N., Soliva Vila, J., Solovieva, A., Stevens, M., Sudre, G., Tamm, L., Thompson, P., Tovar-Moll, F., van Erp, T., Vance, A., Vilarroya, O., Vives-Gilabert, Y., von Polier, G., Walitza, S., Wolfers, T., Yoncheva, Y., Zentis, M., Ziegler, G., Zierhut, K.

Potential conflicts of Interest:

Yanli Zhang-James, Emily C Helminen, Jinru Liu and Martine Hoogman declare no conflict of interest. Barbara Franke has received educational speaking fees from Shire and Medice. Dr. Stephen Faraone received income, travel expenses and/or research support from and/or has been on an Advisory Board for Pfizer, Ironshore, Shire, Akili Interactive Labs, CogCubed, Alcobra, VAYA Pharma, Neurovance, Impax, NeuroLifeSciences and research support from the National Institutes of Health (NIH). With his institution, he has US patent US20130217707 A1 for the use of sodium-hydrogen exchange inhibitors in the treatment of ADHD. In previous years, he received consulting fees or was on Advisory Boards or participated in continuing medical education programs sponsored by: Shire, Alcobra, Otsuka, McNeil, Janssen, Novartis, Pfizer and Eli Lilly. Dr. Faraone receives royalties from books published by Guilford Press: Straight Talk about Your Child's Mental Health, Oxford University Press: Schizophrenia: The Facts and Elsevier, ADHD: Non-Pharmacologic Treatments.

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract:

Background: Clinical symptoms-based ADHD diagnosis is considered "subjective". Machine learning (ML) classifiers have been explored to develop objective diagnosis of ADHD using magnetic resonance imaging (MRI) biomarkers.

Methods: We reviewed previous literature and developed ensemble classifiers using the ENIGMA-ADHD dataset, with the implementation of data balancing to control for age, sex, diagnostic groups, and sample sites and a held-out test set for independent evaluation.

Results: Our review showed that classification accuracies reported previously using cross-validation (CV) samples were inflated and did not generalize well to independent test samples. Our results showed a significant discrimination between ADHD and control samples for both adult and children, but the accuracies were modest (the area under the receiver operating characteristic curve (AUC): 66% and 67% respectively). We found that child samples were informative for predicting adult ADHD, and *vice versa.* The most important brain MRI structures for prediction were intracranial volume (ICV), followed by surface area and some subcortical volumes. The cortical thickness measurements were the least useful.

Conclusions: Although previous ML classification studies reported overly optimistic accuracies and suffered methodological limitations, our results suggest that clinically useful classification of ADHD may be possible with larger samples. In contrast to prior reports of ENIGMA-ADHD studies, our work finds ADHD-related sMRI differences in adults and shows that the brain differences between cases and controls seen in youth can be useful in discriminating adults with and without ADHD. This provides additional evidence for the continuity of ADHD's pathophysiology from childhood to adulthood.

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Clinicians diagnose attention-deficit/hyperactivity disorder (ADHD) by evaluating symptoms and impairments. Despite of the concurrent and predictive validity of clinical diagnosis (1, 2), many raised concerns about the possibilities of over-diagnosing ADHD in the community (3, 4) because it relies on clinicians' "subjective" evaluation of responses from patients, parents, and/or informants. Concerns also exist about the under-diagnosis of ADHD (5, 6), especially in girls and women. The misdiagnosis of ADHD is also a serious concern, with an estimated misdiagnosis rate as high as 20% in the US (7). Those who are inappropriately diagnosed with the disorder may be unnecessarily exposed to chronic use of medications. Those who have ADHD and are not diagnosed will continue to have impaired functioning leading to increased risks for other health and social problems (8). When people who have ADHD are incorrectly diagnosed with another disorder, they may be exposed to unnecessary treatments and still face many impairments associated with ADHD.

In response to such concerns, researchers have sought to develop objective measures. Measures examined in the past years included peripheral biochemical markers(9, 10) and measures of oxidative stress (11), neuropsychological (12), electroencephalographic (EEG) (13), actigraphy (14), eye vergence (15), interactive gaming (16) and continuous performance tests (CPTs) (e.g. 17, 18, 19). Although many significantly differentiated subjects according to their ADHD diagnosis, none met the criteria of a "useful" biomarker defined by the World Federation of ADHD, which accordingly, must exceed 80% sensitivity and specificity, be reliable, reproducible, inexpensive, non-invasive, easy to use, and confirmed by at least two independent studies(20).

Magnetic resonance (MRI) data has also been examined for their potential to provide objective biomarkers for ADHD (21). The enthusiasm was further kindled by the ADHD-200 Global Competition (22), which provided an opportunity for researchers to compete for the best diagnostic classifier using a dataset much larger than any existing neuroimaging biomarker studies at that time, consisting of 776 children (63% healthy controls, 37% ADHD) contributed from eight sites(23-25). Although no predictive biomarkers were observed from those studies (26), the ADHD-200 dataset continues to be used by researchers to look for better brain-based biomarkers (27). It was later incorporated into a larger consortium by the Enhancing Neuro Imaging Genetics Through Meta Analysis (ENIGMA) ADHD Working Group. By Aug 2017, The ENIGMA-ADHD dataset contained 3,377 subjects, including >1,000 adults, from 23 participating sites. The initial report from ENIGMA-ADHD found small but significant and widespread regional volumetric differences between ADHD patients and healthy controls for children but not adults. These differences included volumetric reductions in intracranial volume (ICV), amygdala, caudate nucleus, nucleus accumbens, hippocampus, and cortical surface areas from many brain regions (28). The largest effect was found for total surface area: Cohen's d= -0.21, pFDR=<0.001 and the effect was larger in the youngest tertile (4-9 years, d =-0.35, pFDR=<0.001)(29).

The present study had three main objectives. We first sought to perform a systematic review of studies seeking to develop clinically useful classifiers for ADHD based on MRI data. Second, we applied ML to the ENIGMA-ADHD data with the goal of developing an improved neuroimaging classifier. Third, we used ML to test the hypothesis of continuity between childhood and adult ADHD (30-32). This idea has been challenged by recent studies (33). Given that symptoms and impairments persist into adulthood for a majority of children with ADHD (36), we hypothesized that ADHD-related brain structure differences in adults would be consistent with those in children and that ML methods may help uncover those differences.

Materials and Methods

Literature search and review

We searched PubMed using the key-words 'ADHD AND (classif*[ti] OR biomarker [ti]) AND machine learning' (up to May 1st, 2018) to identify studies that used neuroimaging to discriminate ADHD and non-ADHD groups. Additional studies were extracted by examining their cited references. We examined the relationship of logit-transformed accuracies reported (percent of correct classifications) with their cross-validation or testing methods, and sample sizes using a linear regression model and Pearson's correlation in STATA15. The analyses were weighted by the training sample sizes of the contributing studies.

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MRI Samples

The ENIGMA-ADHD project provided T1-weighted structural MRI (sMRI) data from 3,377 subjects from 23 participating sites (by Aug 2017) to the current study. Images were processed using the consortium's standard segmentation algorithms in FreeSurfer (V5.1 and V5.3) (28). Variables used included 72 cortical surface area and thickness measurements from each hemisphere, 14 subcortical regions, and intracranial volume (ICV). Subjects missing more than 50% of variables were removed. Remaining missing values and outliers (outside of 1.5 times the interquartile range (iqr 1.5)) were replaced with imputed values using multiple imputation with chained equations in STATA15. Four sites that provided only cases with no controls were excluded from ML model training. However, they were used in additional test for assessment of model generalizability. The final ML dataset consisted of 48.7% non-ADHD controls (n=1320, male to female ratio (m/f) = 1.44) and 51.4% ADHD participants (n=1393, m/f=2.52). Ages ranged from four to 63 years old; 62.8% were children (age<18 years) and 37.2% were adults (age≥18 years, Table 1).

Subjects were randomly assigned to training (~70%), validation (~15%), and test (~15%) subsets. The random splitting was carried out within each diagnosis, sex, age subgroup and site. Next, we balanced the case and control groups within each site and age group by random oversampling of under-represented diagnostic groups, a procedure commonly used to deal with class imbalance.

Feature preprocessing

Principal factors factor analysis (PFFA) with varimax rotation on sMRI features on the training set identified 16 factors that explained >80% of the variance. Factor scores were computed for all subjects based on the training set PFFA. Outliers of the factor score (irq 1.5) were replaced with their closest values. We included age and sex as predictors because: 1) they are readily available, and 2) given the known effects of age and sex on brain structures, they may interact with sMRI features and improve predictive accuracy. All input features were scaled based on the training set's minimum and maximum values.

ML Algorithms

We implemented ensemble classifier in Scikit-learn (37), by combining support vector machine (SVM), random forest (RF), K-Nearest Neighbors (KNN), and gradient boosting (GB) classifiers. The mathematical basis of the individual classifiers were described previously (38-41). We estimated each individual classifier using Scikit-learn's grid search function to find the best hyperparameters, including C and gamma for SVM, the number of features, maximum depth and number of estimators for RF, the number of neighbors (parameter *k*), p and the leaf size for KNN, the learning rate, number of estimators and maximum depth for GB. A second grid search was performed on the ensemble model with the best hyperparameters from individual classifiers to fine-tune their combinations and the weights for the individual classifiers. We used area under the curve (AUC) statistic from the training and the validation subsets for model optimization during this two-tiered grid search. The final models were selected based on the highest validation AUCs (to improve accuracy) and the smallest difference between the training and validation AUCs (to avoid overfitting), with preference for hyperparameters favoring less overfitting. Receiver operating characteristic (ROC) curves and AUC statistics from the test subsets were reported. We also plotted learning curves using the training and test scores from different fractions of the training set to evaluate model overfitting and sample size effect.

Testing ADHD Hypotheses about the Continuity of ADHD from Children to Adults

Our analysis pipeline starts with three base models that classify ADHD in children, adults or combined samples. The base models used data from the corresponding age groups during the model training and validation phase and tested also on data from their corresponding age groups. Therefore, we referred to them as "Child, Child, Child", "Adult, Adult, Adult" and "Both, Both, Both", denoting their corresponding training, validation and test sets.

Next, we tested if the ADHD vs. control sMRI differences seen in adults would be useful in predicting ADHD in children. To do that, we used the model that was trained and validated on the adult data to predict the child

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

test set. We refer this result as "Adult, Adult, Child". If the adult data are irrelevant to the child data, they will result in reduced or non-significant accuracy. We hypothesized that the AUC for this model would be statistically significant.

In like manner, we tested if the ADHD vs. control sMRI differences seen in children are useful in predicting ADHD in adults. To do that, we used the model that was trained and validated on the child data to predict the adult data. We refer to this result as "Child, Child, Adult". If the child data are irrelevant to the adult data, they will result in reduced or non-significant accuracy. We hypothesized that the AUC for this model would be statistically significant.

Model evaluation

We applied a softmax function (42) in our final models to generate a continuous brain risk score (BRS), which assess the probability for each individual to be diagnosed with ADHD. Cohen's *d* effect sizes were computed using the BRSs. In addition, we assessed the clinical utility of the model with sensitivity, specificity, positive predictive power, and negative predictive power using various cut-points to the BRSs.

The overall accuracies (percentage of correct predictions) were also reported and stratified by age, sex, and diagnostic groups. We used logistic regression to determine if prediction errors were significantly influenced by age, sex, diagnostic status, and MRI acquisition site. Pearson's correlation was used to determine if the subgroup accuracies were associated with the sample sizes in their corresponding training sets.

Evaluating model generalizability

We applied our model on 168 samples from four sites that had only provided cases. To prepare these samples for the test, we computed the 16 scaled factor scores based on the PFFA analysis of the training set. AUC statistics could not be computed due to the lack of control subjects. Only prediction accuracies (equivalent to sensitivity) were reported.

Feature importance

Importance scores were computed for 16 MRI factors, age and sex from RF and GB models (the other classifiers do not have a method for computing importance) (43). The importance of age and sex as features was further assessed by comparing the AUCs of models with age and sex excluded. For brain MRI measurement, we computed a composite score by summing the products of their factor loading with their corresponding factor's feature importance scores for both the RF and GB classifiers. We used the composite scores to assess how the different hemispheres and classes of MRI features contributed to predictive accuracy using a linear regression model. The four classes of MRI features were cortical surface areas, cortical thicknesses, subcortical volumes and total intracranial volume (ICV).

Results

Literature Review

Nine studies were retained from the literature search (Table 2), among which three main methods were used to assess their model accuracies: 1) Various k-fold cross-validation (CV) methods: Seven studies used either 10-fold, 70/30, or predefined CV. For the first two types, the training sets were randomly split each time and were trained many times on either 90% or 70% of the data and validated on the remaining portion. Predefined CV used fixed sets of samples as training and validation sets. 2) Leave one out CV (LOOCV): In six studies training was performed many times on the training set excluding one randomly selected sample for validation. Because the k-fold CV and LOOCV studies did not differ from one another in accuracy ($F_{(1, 13)} = 2.69$, $p=0.13$), we combined them into one CV group. Among the above 14 studies, cross-validation accuracies were the only reported results in **nine** studies (64.3%). Only five studies used additional held-out test samples that were not used in the model training and validation. 3) Held-out tests: A total of nine studies reported accuracy evaluation using a held-out test sample that had not been used for either training or validation. Significantly lower

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

accuracies were reported using held-out test samples than those using CV methodology ($F_{(1, 18)}$ = 26.38, p<0.0001, Figure 1). Training sample sizes did not have any significant effect on accuracy ($F_{(1, 18)}$ = 0.33, p=0.57). The correlations between accuracy and training sample size were 0.10 for the cross-validation results (p=0.71) and 0.25 for the held-out tests (p=0.38).

Accuracy of Predictions

The validation AUCs of the ensemble classifiers and their constituent classifiers were listed in Supplementary Table 1 for the best of the three base models. The child data performed better than the adult data for each constituent classifier as well as for the ensemble classifier. The same pattern was also observed for the corresponding test AUCs. Figure 2A shows the test set AUCs (as dots) and their 95% confidence intervals (as horizontal lines) for the three base models, which are denoted by their training, validation and test samples. The vertical line at an AUC of 0.5 indicates a chance level of diagnostic accuracy. The "Child, Child, Child" model had the highest AUC 0.67 with 95%CI (0.60, 0.73) that did not overlap with the 0.5 line, indicating significant predictive accuracy. The combined model ("Both, Both, Both") had a lower AUC (0.61, 95%CI: 0.56, 0.67), which was significantly different from 0.5, but not significantly different from the child model. The "Adult, Adult, Adult" model yielded the lowest test AUC (0.54, 95%CI: 0.45, 0.64), which did not differ significantly from 0.5, and was significantly lower than the Child AUC ($X^2_{(1)}$ = 4.53, p = 0.03).

Tests of Hypotheses

When the Adult model was tested on the child samples (("Adult, Adult, Child"), we obtained a significant AUC(0.59, 95%CI: 0.52, 0.66) that was slightly lower but not significantly different from the "Child, Child, Child", or "Both, Both, Child" AUCs.

When the Child model was tested on the adult samples ("Child, Child, Adult"), we obtained a significant AUC (0.66, 95%CI: 0.57, 0.75) that significantly improved prediction when compared with either the "Adult, Adult, Adult" model ($X^2_{(1)}$=3.88, p =0.048), or the combined model prediction of the adult data ("Both, Both, Adult", $X^2_{(1)}$=5.53, p =0.02).

Model Evaluation

The receiver operating characteristic (ROC) curves (Figure 3) show that the "Child, Child" model predicts child and adult ADHD equally well. The Cohen's d effect sizes comparing the BRS between ADHD and control samples were 0.62 for children (95%CI: 0.37, 0.87) and 0.56 for adults (95%CI: 0.23, 0.90).

Table 3 shows the model accuracies stratified by diagnostic, sex and age subgroups. Logistic regression found that prediction errors were significantly associated with diagnosis ($X^2_{(1)}$ =50.73, p<0.0001), sex ($X^2_{(1)}$ =40.02, p<0.0001), and their interactions ($X^2_{(1)}$ = 84.67, p<0.0001), but not with age ($X^2_{(1)}$ =0, p=0.97) or study site ($X^2_{(18)}$ = 20.43, p=0.31). For both children and adults, the highest accuracies were for female control and male ADHD subjects. When the training sample sizes in subgroups were considered, the accuracies were found significantly associated with sample sizes ($F_{(1, 5)}$ = 51.9, p= 0.0008) and sex ($F_{(1, 5)}$ = 43.28, p= 0.001), but not with diagnosis.

Examining sensitivity and specificity in separate sex groups (Supplementary Figure 1), we found that at default 0.5 BRS cut-point, the female subgroup has low sensitivity and high specificity. By contrast, the male subgroup has high sensitivity and low specificity. We can obtain similar sensitivities of ~20% with high specificities for both sexes (female 83.9% and male 87.4%, Table 4), if we shift the cut-points in opposite directions i.e. female 0.415 and male 0.71.

Feature importance

Feature importance scores from RF and GB models were significantly correlated (Pearson's r=0.66, p=0.003, Supplementary Figure 2). The ranked features using the average scores between both models were listed in Supplementary File 1. Factor 1 ranked the first and sex ranked last. Age ranked higher than sex but was still among the lowest. When we excluded age and sex from the model, we obtained an AUC 0.65 for the child

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

samples (95%CI: 0.58, 0.71), which was similar to the AUC of the child set with age and sex included as features. Excluding age and sex resulted a lower AUC 0.59 (95%CI: 0.49, 0.68) for the adult samples, which also was not significantly different from that of the adult model with age and sex included. This suggests that the predictive information afforded by age and sex is redundant with the predictive information afforded by the sMRI features.

The ranked brain features and their corresponding importance scores are in Supplementary File 1. The importance scores were similar between left and right brain hemispheres but differed significantly across feature types ($F_{[3,153]} = 134$, $p<0.0001$). ICV had the highest score (Figure 4B). The mean scores for surface area measures (0.125, 95%CI: 0.122, 0.127) and subcortical volumes (0.125, 95%CI: 0.117, 0.134) were similar and both were significantly higher than those for cortical thickness (0.102, 95%CI: 0.099, 0.104).

Learning curves and effects of sample size on model performance

Figure 5 plots the learning curves for the prediction in children (Left) and adults (Right), showing similar converging trends for training and testing AUCs as the sample size increases. Final AUCs did not reach the accuracy achieved during training when all samples were used. The characteristics of the learning curves suggest some degrees of overfitting and that increasing sample size should improve performance and reduce overfitting. By extrapolating the training and test accuracies, we would predict that collecting more data would improve accuracy to about 0.75. Further improvement would likely require additional predictive features, such as functional MRI data.

Model generalizability

Testing on the four excluded sites showed an overall 79.8% sensitivity (95%CI: 73.6%, 85.9%). Sensitivity did not differ between the child (82%, 50 male subjects, 95%CI: 71.0%, 93.0%) and adult samples (78.8%, 18 females and 100 males, 95%CI: 71.3%, 82.3%). There were no differences in sensitivities across the four sites. However, sensitivity was significantly lower for females (11.1%, 95%CI: -0.05%, 27.2%) than for males (88%, 95%CI: 82.7%, 93.3%) ($X^2_{(1)} = 9.39$, $p=0.002$), consistent with the test results in Table 3.

Discussion

We achieved three main goals. First, our review found that many prior studies seeking to develop clinically useful classifiers for ADHD based on MRI data did not use a held-out test set and reported overly optimistic assessments of classification accuracy. Second, our results from ENIGMA ADHD data suggest that clinically useful classification may be possible, although achieving that will still require larger samples and, perhaps, additional predictive features. Third, we used ML in an innovative manner to provide supporting evidence for the continuity of ADHD's pathophysiology from childhood to adulthood.

What Makes Our Study Different From Prior Studies

Almost 40% prior studies of ADHD MRI classifiers only reported cross-validation (CV) accuracies (23, 24, 44-50)). Samples used in iterations of cross-validation influence the hyperparameter estimation. Therefore, CV results may overestimate actual accuracy, as our results showed. The sample size of ENIGMA ADHD data allowed us to properly estimate model accuracies from true test sets that were not involved in hyperparameter tuning.

Other concerns regarding many previous studies were confounding factors such as sites and case-control imbalance. For example, ADHD-200 dataset comprised of data from many different sites and has more controls (63%) than ADHD (51). Learning algorithms can be confounded by base rate of the disorder and difference among sites. In our study, we applied oversampling of under-represented groups in each age subgroup and site. In addition, we used AUC statistic for model evaluation, instead of commonly used accuracy (percentage correct), which can be influenced by case-control imbalances in data sets. Our results showed indeed that we removed the confounding effects of different acquisition sites and age groups. Although our accuracies were modest (61.2% for children and 62.1% for adults), they were at the high end of prior

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

results with the held-out test sets (Figure 1). Two studies reported higher test accuracies than ours (52, 53). Both studies utilized functional MRI in addition to sMRI data as features. More importantly, both reported significant site variations. One estimated that the site information alone resulted in 66% prediction accuracy(52). Our results, in contrast, generalized well to different sites.

Another factor that has contributed to our models' generalizability was the rigorous regularization. Hyperparameters that favor less overfitting were selected so that models had training AUCs close to the validation AUCs. Learning curve analysis also helped us to assess how well our models learned with increasing sample sizes, and whether our models were overfit. No previous studies of ADHD have implemented and reported these measures.

Clinical Utility

Examining different accuracies in subgroups of sex and diagnosis, we found that male ADHD and female control groups had the highest accuracies for both children and adults (73.3% ~ 89.2%). Same pattern was observed in samples from the excluded sites. Conditional probability analysis suggests that sex-specific BRS thresholds should be considered. For example, shifting BRS cut-points in opposite directions for male and female groups, we achieve ~20% sensitivity and specificity >84% for both sexes. In this case, males also had a high PPP (71.1%), meaning that our model was correct most of the time when predicting a male as having ADHD, but it is wrong ~58% most of the time when classifying someone as non-ADHD. The lower PPP (40%) for girls means that we are often wrong (60% of time) when classifying a girl as having ADHD, although we are often correct (65.6%) when classifying a girl as non-ADHD. Although the current levels of accuracy do not suggest our models for clinical practice, our learning curve analysis indicate that increasing sample sizes could improve the model performance (54, 55), particularly more samples from the underrepresented sex and diagnostic groups, i.e. female ADHD and male control samples.

Machine Learning Tests of Hypotheses

Our results support the hypotheses about the continuity of child and adult ADHD pathophysiology and extend the results of prior ENIGMA ADHD studies (28, 29). Firstly, consistent with prior ENIGMA reports, we found that adult ADHD could not be successfully discriminated from the controls when using only adult data. It could be due to the smaller sample size, or larger variations in brain differences in adults rendering it more difficult to discriminate. However, we show that using child data to train and validate the model, we can significantly improve the adult ADHD prediction, which suggests that the ADHD vs. control differences observed in children provide information relevant to adult ADHD and argues against recent hypothesis that adult ADHD is etiologically distinct from childhood ADHD (33). Indeed, our BRS estimated similar case-control effect sizes (Cohen's *d*) for children and adults. Both were two to three times greater than those of the individual regions reported in prior ENIGMA studies (Cohen's d 0.09 ~0.25) (28, 29). Secondly, some main features in our prediction models were consistent with preceding ENIGMA reports, for example, our most important feature ICV (28). Previously, total surface area was identified as the most significant measure with the highest Cohen's d effect size (28). We also found high scores for many surface areas in our model. One caution in interpreting importance scores is that one feature may mask the importance of the others because of high correlations, particularly for interconnected brain structures.

Limitations

First, although we eliminated the confounding effects of age and sites, we still observed sex and sample size differences. Future studies with more samples from under-represented groups will help improve model performance and generalizability. Second, we only used sMRI data. Incorporating other imaging modalities could help improve classification accuracy. Finally, we used pre-defined structures from ENIGMA standard image processing pipeline as features. It is possible that other methods such as one using 3D images as input

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

features, in a convolutional neural network would uncover useful features leading to increased classification accuracy.

In conclusion, our application of ML to the ENIGMA ADHD data suggests that clinically useful classification may be possible, although achieving that will require larger samples. ML can uncover ADHD vs. control structure differences in adults that were not detected in prior ENIGMA ADHD reports using standard statistical methods. These analyses show that sMRI differences associated with ADHD are similar for adults and youth, which supports the continuity of ADHD's pathophysiology from childhood to adulthood.

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

List of Tables:

Table 1. Sample characteristics
Table 2. Machine learning literature on ADHD neuroimaging data.
Table 3. Test accuracies (and subject numbers) by sex, diagnosis and age groups
Table 4. Conditional probability analysis of different sex groups

Supplementary Table 1. Highest Validation Score (AUC) for The Individual Classifiers and the Ensemble Classifiers

Supplementary Table 2. Ranked feature importance scores for factors and brain MRI measurements.

Figure Legends

Figure 1. Reported accuracies from previously published studies plotted against their training sample sizes.

Blue dots are results from held-out test set methodology and red triangles are results from cross-validation methodology. Higher accuracy scores were found for cross-validation compared with held-out test methodology.

Figure 2. Test set performance of the best ensemble models.

Area under the receiver operating characteristic curve (AUC) accuracy statistics for the held-out test results were plotted (as dots) with their 95% confidence intervals (as horizontal lines). The models were defined based on what samples were used for training, validation, and testing (Train, Valid, Test). The vertical line at an AUC of 0.5 indicates a chance level of diagnostic accuracy. If the 95%CI does not overlap with the 0.5 vertical line, it indicates significant predictive accuracy.

Three base models are plotted on the top group: the “Adult, Adult, Adult”, the “Both, Both, Both”, and the “Child, Child, Child” models. These used either only the adult, or the child or both samples respectively to train, validate, and test. Middle portion plots the adult and child subset AUCs from the “Both, Both, Both” models. Bottom portion showed the test AUC for adult samples by using the child model (“Child, Child, Adult”, and the test AUC for the child samples by using the adult model (“Adult, Adult, Child”).

Figure 3. ROC curves for ADHD prediction in adults and children.

Receiver operating characteristic (ROC) curves for our best model were compared for the test set prediction results in adults (red ROC) and children (blue ROC).

Figure 4. Feature Importance.

A. The importance scores were derived from the two models that provide such scores: Random Forests (RF) and Gradient Boosting (GB). The features used are 16 brain MRI factors , age, and sex. These scores indicate the degree to which each feature contributed in predicting ADHD diagnostic status. The scores from two models were significantly correlated. Factor 1 ranked highest and sex ranked lowest.

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

B. The composite importance scores for MRI brain features were plotted to show mean differences across four main classes. The scores were generated by summing the products of the importance scores of the 16 factors in RF and GB models and the factor loading of the individual brain regions in each factor.

Figure 5. Learning curves for model prediction of the child ADHD (Left) and adult ADHD (Right). The learning curves plot the training (red line) and test (green line) AUCs achieved for increasing training sample sizes. The whole training data were randomly split into eight parts. We started the training with 1/8th of the total data, and repeated the process at an increment of 1/8th each time. For both graphs, a converging trend of training and testing AUCs was observed, although the final test AUCs did not reach the training AUCs when all the samples were used. The converging pattern and the gap indicate the presence of overfitting and suggest more samples are needed to improve model performance.

Supplementary Figure 1. Sensitivity and Specificity Analysis of Different Sex groups. Classification sensitivity and specificity were computed and plotted for males and females separately at various probability cut-offs, which we referred as the brain risk score (BRS) that dichotomizes the case and control.

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. Faraone SV (2005): The scientific foundation for understanding attention-deficit/hyperactivity disorder as a valid psychiatric disorder. *Eur Child Adolesc Psychiatry*. 14:1-10.
2. Faraone SV, Biederman J, Spencer T, Wilens T, Seidman LJ, Mick E, et al. (2000): Attention deficit hyperactivity disorder in adults: an overview. *Biological psychiatry*. 48:9-20.
3. Bruchmuller K, Margraf J, Schneider S (2012): Is ADHD diagnosed in accord with diagnostic criteria? Overdiagnosis and influence of client gender on diagnosis. *J Consult Clin Psychol*. 80:128-138.
4. Visser SN, Danielson ML, Bitsko RH, Holbrook JR, Kogan MD, Ghandour RM, et al. (2014): Trends in the parent-report of health care provider-diagnosed and medicated attention-deficit/hyperactivity disorder: United States, 2003-2011. *J Am Acad Child Adolesc Psychiatry*. 53:34-46 e32.
5. Ginsberg Y, Quintero J, Anand E, Casillas M, Upadhyaya HP (2014): Underdiagnosis of attention-deficit/hyperactivity disorder in adult patients: a review of the literature. *Prim Care Companion CNS Disord*. 16.
6. The Express Scripts Lab (2014): Turning Attention to ADHD: U.S. Medication Trends for Attention Deficit Hyperactivity Disorder.
7. Elder TE (2010): The importance of relative standards in ADHD diagnoses: evidence based on exact birth dates. *J Health Econ*. 29:641-656.
8. Ljung T, Chen Q, Lichtenstein P, Larsson H (2014): Common etiological factors of attention-deficit/hyperactivity disorder and suicidal behavior: a population-based study in Sweden. *JAMA Psychiatry*. 71:958-964.
9. Faraone SV, Bonvicini C, Scassellati C (2014): Biomarkers in the diagnosis of ADHD--promising directions. *Curr Psychiatry Rep*. 16:497.
10. Scassellati C, Bonvicini C, Faraone SV, Gennarelli M (2012): Biomarkers and attention-deficit/hyperactivity disorder: a systematic review and meta-analyses. *J Am Acad Child Adolesc Psychiatry*. 51:1003-1019 e1020.
11. Joseph N, Zhang-James Y, Perl A, Faraone SV (2015): Oxidative Stress and Attention Deficit Hyperactivity Disorder: A Meta-Analysis *J Atten Disord*. 19:915-924.
12. Ritsner MS (2009): Neuropsychological Endophenotypes and Biomarkers. *The Handbook of Neuropsychiatric Biomarkers, Endophenotypes and Genes*: Springer Netherlands.
13. Snyder SM, Rugino TA, Hornig M, Stein MA (2015): Integration of an EEG biomarker with a clinician's ADHD evaluation. *Brain and Behavior*.
14. Dane AV, Schachar RJ, Tannock R (2000): Does actigraphy differentiate ADHD subtypes in a clinical research setting? *Journal of the American Academy of Child and Adolescent Psychiatry*. 39:752-760.
15. Sole Puig M, Perez Zapata L, Puigcerver L, Esperalba Iglesias N, Sanchez Garcia C, Romeo A, et al. (2015): Attention-Related Eye Vergence Measured in Children with Attention Deficit Hyperactivity Disorder. *PLoS One*. 10:e0145281.
16. Faraone SV, Newcorn JH, Antshel KM, Adler L, Roots K, Heller M (2016): The Groundskeeper Gaming Platform as a Diagnostic Tool for Attention-Deficit/Hyperactivity Disorder: Sensitivity, Specificity, and Relation to Other Measures. *J Child Adolesc Psychopharmacol*. 26:672-685.
17. Homack S, Riccio CA (2006): Conners' Continuous Performance Test (2nd ed.; CCPT-II). *J Atten Disord*. 9:556-558.
18. Riccio CA, Reynolds CR (2001): Continuous performance tests are sensitive to ADHD in adults but lack specificity. A review and critique for differential diagnosis. *Ann N Y Acad Sci*. 931:113-139.
19. Corkum PV, Siegel LS (1993): Is the Continuous Performance Task a valuable research tool for use with children with Attention-Deficit-Hyperactivity Disorder? *J Child Psychol Psychiatry*. 34:1217-1239.
20. Thome J, Ehlis AC, Fallgatter AJ, Krauel K, Lange KW, Riederer P, et al. (2012): Biomarkers for attention-deficit/hyperactivity disorder (ADHD). A consensus report of the WFSBP task force on biological markers and the World Federation of ADHD. *World J Biol Psychiatry*. 13:379-400.
21. Valera EM, Faraone SV, Murray KE, Seidman LJ (2007): Meta-analysis of structural imaging findings in attention-deficit/hyperactivity disorder. *Biol Psychiatry*. 61:1361-1369.
22. Consortium A- (2012): The ADHD-200 Consortium: A Model to Advance the Translational Potential of Neuroimaging in Clinical Neuroscience. *Frontiers in systems neuroscience*. 6:62.

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

23. Iannaccone R, Hauser TU, Ball J, Brandeis D, Walitza S, Brem S (2015): Classifying adolescent attention-deficit/hyperactivity disorder (ADHD) based on functional and structural imaging. *Eur Child Adolesc Psychiatry*. 24:1279-1289.
24. Hart H, Chantiluke K, Cubillo AI, Smith AB, Simmons A, Brammer MJ, et al. (2014): Pattern classification of response inhibition in ADHD: toward the development of neurobiological markers for ADHD. *Hum Brain Mapp*. 35:3083-3094.
25. Johnston BA, Mwangi B, Matthews K, Coghill D, Konrad K, Steele JD (2014): Brainstem abnormalities in attention deficit hyperactivity disorder support high accuracy individual diagnostic classification. *Hum Brain Mapp*. 35:5179-5189.
26. Wolfers T, Buitelaar JK, Beckmann CF, Franke B, Marquand AF (2015): From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci Biobehav Rev*. 57:328-349.
27. Uddin LQ, Dajani DR, Voorhies W, Bednarz H, Kana RK (2017): Progress and roadblocks in the search for brain-based biomarkers of autism and attention-deficit/hyperactivity disorder. *Translational psychiatry*. 7:e1218.
28. Hoogman M, Bralten J, Hibar DP, Mennes M, Zwiers MP, Schweren LSJ, et al. (2017): Subcortical brain volume differences in participants with attention deficit hyperactivity disorder in children and adults: a cross-sectional mega-analysis. *Lancet Psychiatry*. 4:310-319.
29. Hoogman M, Muetzel R, Guimaraes J, Shumskaya E, Mennes M, Zwiers M, et al. (2019(Submitted)): Brain imaging of the cortex in ADHD: alterations observed in large-scale clinical and population-based samples. *JAMA Psychiatry*.
30. Franke B, Michelini G, Asherson P, Banaschewski T, Bilbow A, Buitelaar JK, et al. (2018): Live fast, die young? A review on the developmental trajectories of ADHD across the lifespan. *European neuropsychopharmacology : the journal of the European College of Neuropsychopharmacology*. 28:1059-1088.
31. Franke B, Faraone SV, Asherson P, Buitelaar J, Bau CH, Ramos-Quiroga JA, et al. (2012): The genetics of attention deficit/hyperactivity disorder in adults, a review. *Mol Psychiatry*. 17:960-987.
32. Faraone SV, Biederman J, Mick E (2006): The age-dependent decline of attention deficit hyperactivity disorder: a meta-analysis of follow-up studies. *Psychol Med*. 36:159-165.
33. Faraone SV, Biederman J (2016): Can Attention-Deficit/Hyperactivity Disorder Onset Occur in Adulthood? *JAMA Psychiatry*. 73:655-656.
34. Shaw P, Eckstrand K, Sharp W, Blumenthal J, Lerch JP, Greenstein D, et al. (2007): Attention-deficit/hyperactivity disorder is characterized by a delay in cortical maturation. *Proc Natl Acad Sci U S A*. 104:19649-19654.
35. Onnink AM, Zwiers MP, Hoogman M, Mostert JC, Kan CC, Buitelaar J, et al. (2014): Brain alterations in adult ADHD: effects of gender, treatment and comorbid depression. *European neuropsychopharmacology : the journal of the European College of Neuropsychopharmacology*. 24:397-409.
36. Sibley MH, Swanson JM, Arnold LE, Hechtman LT, Owens EB, Stehli A, et al. (2017): Defining ADHD symptom persistence in adulthood: optimizing sensitivity and specificity. *J Child Psychol Psychiatry*. 58:655-662.
37. Pedregosa F, Gaël Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. (2011): Scikit-learn: Machine Learning in Python *Journal of Machine Learning Research*. 12:2825–2830.
38. Platt J (2000): Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *Advances in large margin classifiers*: Cambridge: MIT Press., pp 61-74.
39. Breiman L (2001): Random Forests. *Machine Learning*. 45:5-32.
40. Altman NS (1992): An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*. 46:175-185.
41. Friedman JH (2001): machine. *The Annals of Statistics*. 29:1189-1232.
42. Bishop C (2006): *Pattern Recognition and Machine Learning*. Springer-Verlag New York.
43. Breiman L (1984): *Classification and regression trees*. Belmont, Calif.: Wadsworth International Group.
44. Du J, Wang L, Jie B, Zhang D (2016): Network-based classification of ADHD patients using discriminative subnetwork selection and graph kernel PCA. *Comput Med Imaging Graph*. 52:82-88.
45. Qureshi MN, Min B, Jo HJ, Lee B (2016): Multiclass classification for the differential diagnosis on the ADHD subtypes using recursive feature elimination and hierarchical extreme learning machine: Structural MRI study. *PLoS One*. 11:e0160697.

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

46. Dai D, Wang J, Hua J, He H (2012): Classification of ADHD children through multimodal magnetic resonance imaging. *Front Syst Neurosci*. 6:63.
47. Brown MR, Sidhu GS, Greiner R, Asgarian N, Bastani M, Silverstone PH, et al. (2012): ADHD-200 Global Competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. *Front Syst Neurosci*. 6:69.
48. Deshpande G, Wang P, Rangaprakash D, Wilamowski B (2015): Disorder Classification From Functional Magnetic Resonance Imaging Data. *IEEE Transactions on Cybernetics*. 45:2668-2679.
49. Peng X, Lin P, Zhang T, Wang J (2013): Extreme learning machine-based classification of ADHD using brain structural MRI data. *PLoS One*. 8:e79476.
50. Fair DA, Nigg JT, Iyer S, Bathula D, Mills KL, Dosenbach NU, et al. (2012): Distinct neural signatures detected for ADHD subtypes after controlling for micro-movements in resting state functional connectivity MRI data. *Front Syst Neurosci*. 6:80.
51. Consortium HD (2012): The ADHD-200 Consortium: A Model to Advance the Translational Potential of Neuroimaging in Clinical Neuroscience. *Front Syst Neurosci*. 6:62.
52. Bernhardt BC, Sen B, Borle NC, Greiner R, Brown MRG (2018): A general prediction model for the detection of ADHD and Autism using structural and functional MRI. *PloS one*. 13:e0194856.
53. Zou L, Zheng J, Miao C, McKeown MJ, Wang ZJ (2017): 3D CNN Based Automatic Diagnosis of Attention Deficit Hyperactivity Disorder Using Functional and Structural MRI. *IEEE Access*. 5:23626-23636.
54. Goodfellow I, Bengio Y, Courville A (2016): *Deep Learning (Adaptive Computation and Machine Learning series)*. MIT Press.
55. Bengio Y, Courville A, Vincent P (2013): Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell*. 35:1798-1828.

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Tables

Table 1. Sample Characteristics

Diagnosis		Child (Age <18)		Adult (Age ≥18)		Total
		Female	Male	Female	Male	
Control	N of Subjects	309	548	233	230	1320
	Mean Age	10.7	11.7	34.0	30.1	18.6
	SD of Age	2.8	2.9	12.0	9.9	12.1
ADHD	N of Subjects	179	667	217	330	1393
	Mean Age	11.2	12.0	34.2	29.6	19.5
	SD of Age	3.1	2.9	11.6	11.1	12.2
Total	N of Subjects	488	1215	450	560	2713
	Mean Age	10.9	11.9	34.1	29.8	19.1
	SD of Age	2.9	2.9	11.8	10.6	12.2

*Note: SD, standard deviation; N, total numbers.

Table 2. Machine Learning Literature on Classification of ADHD Using Neuroimaging Data.

Author, Year	Training Set		Test Set		Sample Source	Best ML Methods/ Models	Features	Type of Test	Highest Percent Correct
	Sample Size	ADHD%	Sample Size	ADHD%					
Brown MR 2012	668	36%	171	45%	ADHD-200	SVM	rs-fMRI	10-fold CV	71%
								test	55%
Colby JB, 2012	776	37%	197		ADHD-200	SVM	sMRI and rs-fMRI	test	55%
Dai D, 2012	624	36%	165		ADHD-200	MKL	sMRI and rs-fMRI	10-fold CV	68%
								test	62%
Deshpande G, 2015	1177	37%			ADHD-200	FCCANN	rs-fMRI	LOOCV	90%
Du J, 2016	216	55%			ADHD-200	SVM	rs-fMRI	10-fold CV	95%
Eloyan A, 2012	572	37%	193	34%	ADHD-200	Voting	sMRI, rs-fMRI and demographics	test	61%
								~1/4 randomly chosen validation sample	78%
Fair DA 2013	104	50%			ADHD-200	SVM	rs-fMRI	LOOCV	83%
Ghiassian S, 2016	769	36%	171	45%	ADHD-200	MHPC	sMRI and rs-fMRI and demographics	test	70%
Hao A, 2015	216	55%	41	71%	ADHD-200 (NYU subset)	DBAY	rs-fMRI	test	49%
	85	28%	50	46%	ADHD-200 (Peking subset)	DBAY	rs-fMRI	test	54%
	83	27%	11	27%	ADHD-200 (KKI subset)	DBAY	rs-fMRI	test	72%
Hart H, 2014	60	50%			clinic and local community	GPC	task-fMRI	LOOCV	77%

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Iannaccone R, 2015	36	50%			Outpatient clinic and local schools	SVM	task-fMRI	LOOCV	78%
Igual L, 2012	78	50%			URNC database	SVM	sMRI of the caudate nucleus	5-fold CV	72%
Johnston BA, 2014	68	50%			clinic and local schools	SVM	sMRI	LOOCV	93%
Kuang D, 2014	83		11		ADHD-200 (KKI subset)	DBN	rs-fMRI	test	73%
	85		50		ADHD-200 (Peking subset)	DBN	rs-fMRI	test	54%
	222		41		ADHD-200 (NYU subset)	DBN	rs-fMRI	test	37%
Olivetti E, 2015	923	38%			ADHD-200	extremely randomized tree	rs-fMRI	10-fold CV	66%
Peng X, 2013	110	50%			ADHD-200 (Peking subset)	ELM	sMRI	LOOCV	90%
Qureshi MN, 2016	106	50%			ADHD-200	H-ELM	sMRI	10-fold CV	80%
								70/30 CV	85%
Sen B, 2018	558	50%	171	45%	ADHD-200	SVM	sMRI + rs-fMRI	5-fold CV	68.9%
								test	67.3%
Zou L, 2017	559	35%	171	45%	ADHD-200	CNN	sMRI	test	65.9%
							rs-fMRI	test	66.0%
							sMRI + rs-fMRI	test	69.2%

*Note: ML, Machine learning;

fMRI, funcitonal MRI; rs-fMRI, resting state- functional MRI; sMRI, structure MRI;

CV, cross-validation; LOOCV, leave-one-out cross validation;

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SVM, support vector machine; MKL, multi-kernel learning; FCCANN, fully connected cascade artificial neural network; GBM, a gradient boosting method; MHPC, the histogram of oriented gradients (HOG)-feature-based patient classification; GPC, Gaussian process classifiers; DBAY, Deep Bayesian Network; DBN, Deep Belief Network; ELM, extreme learning machine; H-ELM, hierarchical extreme learning machine; CNN, convolutional neural network

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 3. Test accuracies (and subject numbers) by sex, diagnosis and age groups								
		Adult (age>18)				Child (age≤18)		
		Female	Male	Total		Female	Male	Total
Control	Accuracy	89.2%	27.6%	62.1%		80.4%	43.2%	59.2%
	95% CI	(78.7%, 99.7%)	(10.3%, 44.9%)	(50.1%, 74.1%)		(9.6%, 91.1%)	(31.7%, 54.8%)	(50.7%, 67.8%)
	Sample Size	37	29	66		56	74	130
ADHD	Accuracy	15.4%	84.9%	62.0%		20.0%	73.3%	63.1%
	95% CI	(0.5%, 30.2%)	(74.9%, 94.9%)	(51.1%, 73.0%)		(3.1%, 36.9%)	(64.7%, 81.9%)	(54.7%, 71.5%)
	Sample Size	26	53	79		25	105	130
Total	Accuracy	58.7%	64.6%	62.1%		61.7%	60.9%	61.2%
	95% CI	(46.2%, 71.2%)	(54.1%, 75.2%)	(54.1%, 70.1%)		(50.9%, 72.5%)	(53.7%, 68.1%)	(55.2%, 67.1%)
	Sample Size	63	82	145		81	179	260

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 4. Conditional Probability Analysis of Subgroups

	Female Subjects		Male Subjects	
Cut point	0.5	0.415	0.5	0.71
Sensitivity	3.92%	19.6%	88.6%	20.25%
Specificity	96.8%	83.9%	24.27%	87.4%
Positive Predictive Power (PPP)	40%	40%	64.22%	71.1%
Negative Predictive Power (NPP)	64.75%	65.55%	58.14%	41.67%

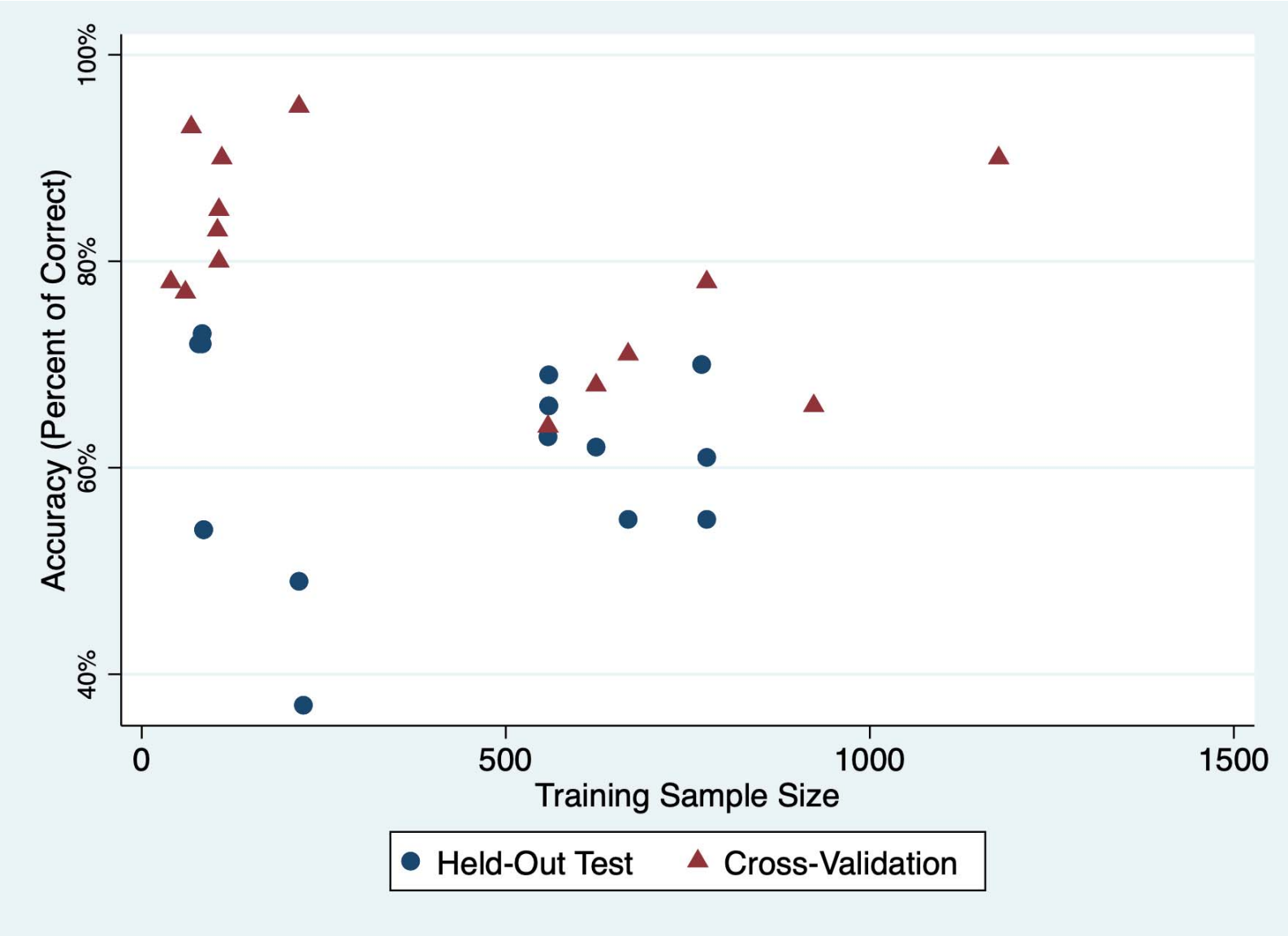
bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 1. Highest Validation Score (AUC) of The Individual Classifiers and the Ensemble Classifiers					
Training and Validation Datasets	**k_Nearest Neighbors**	**Support Vector**	**Random Forest**	**Gradient Boosting**	**Ensemble**
Both, Both	0.57	0.58	0.58	0.58	0.60
Child, Child	0.59	0.66	0.63	0.64	0.68
Adult, Adult	0.55	0.54	0.53	0.6	0.56

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures

Figure 1. Reported accuracies from previously published studies plotted against their training sample sizes



bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2. Test set performance of the ensemble models

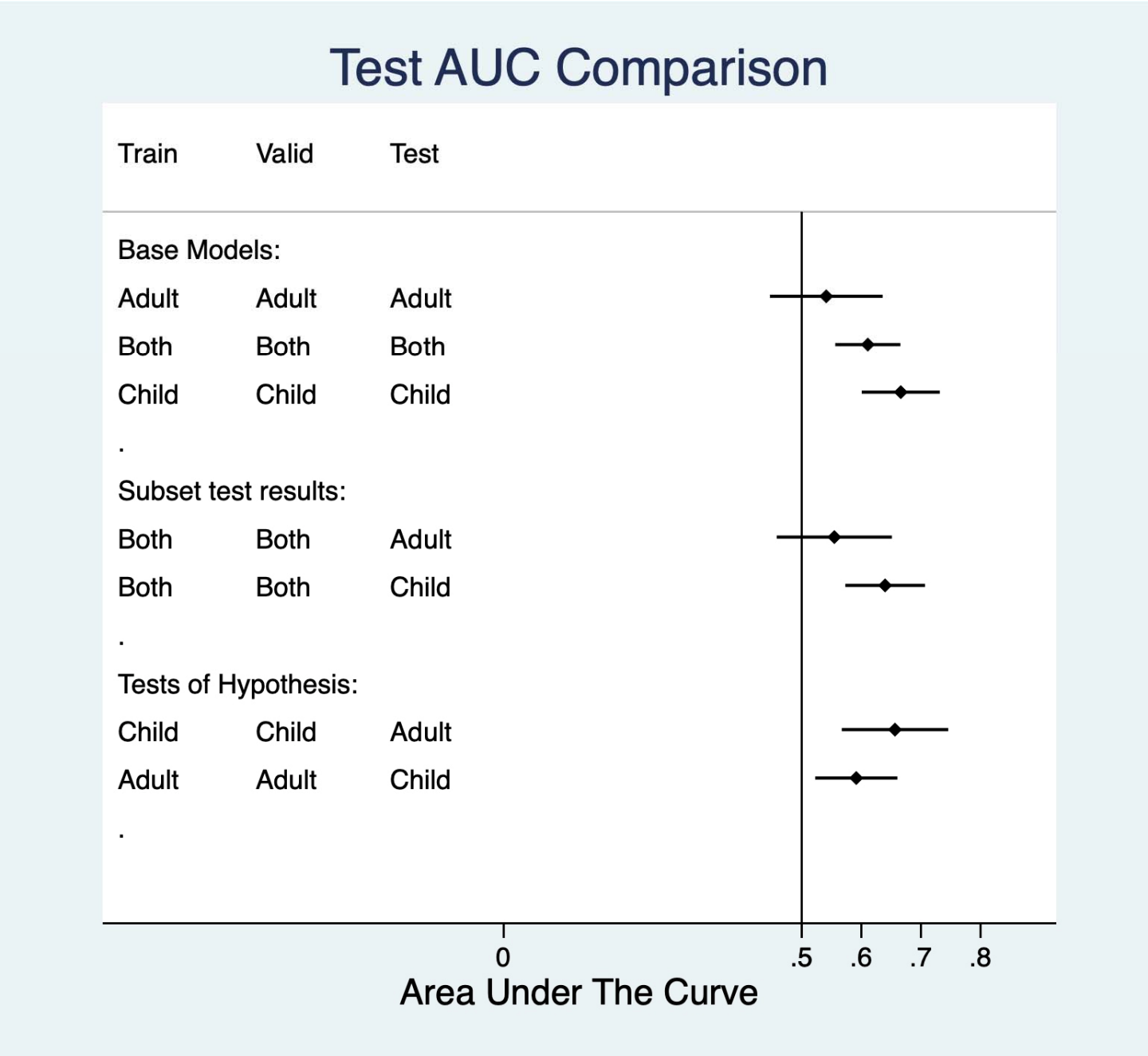
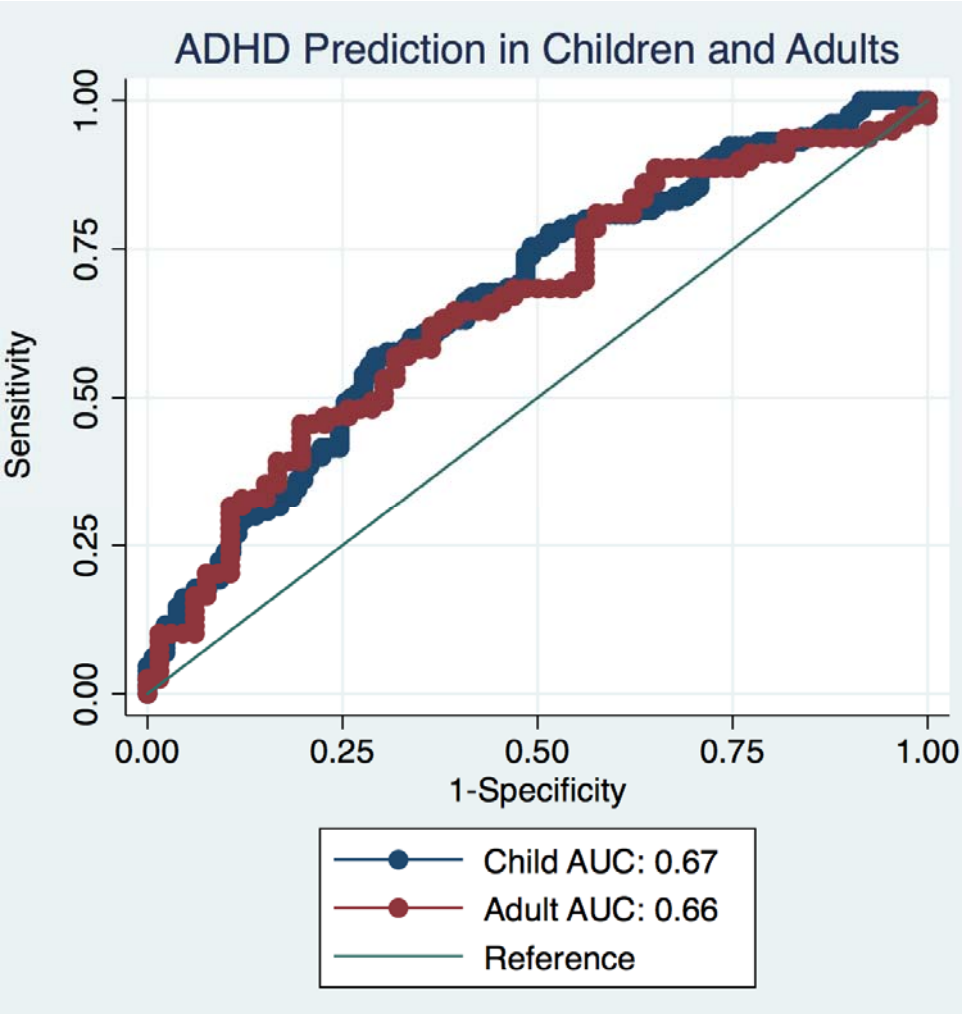


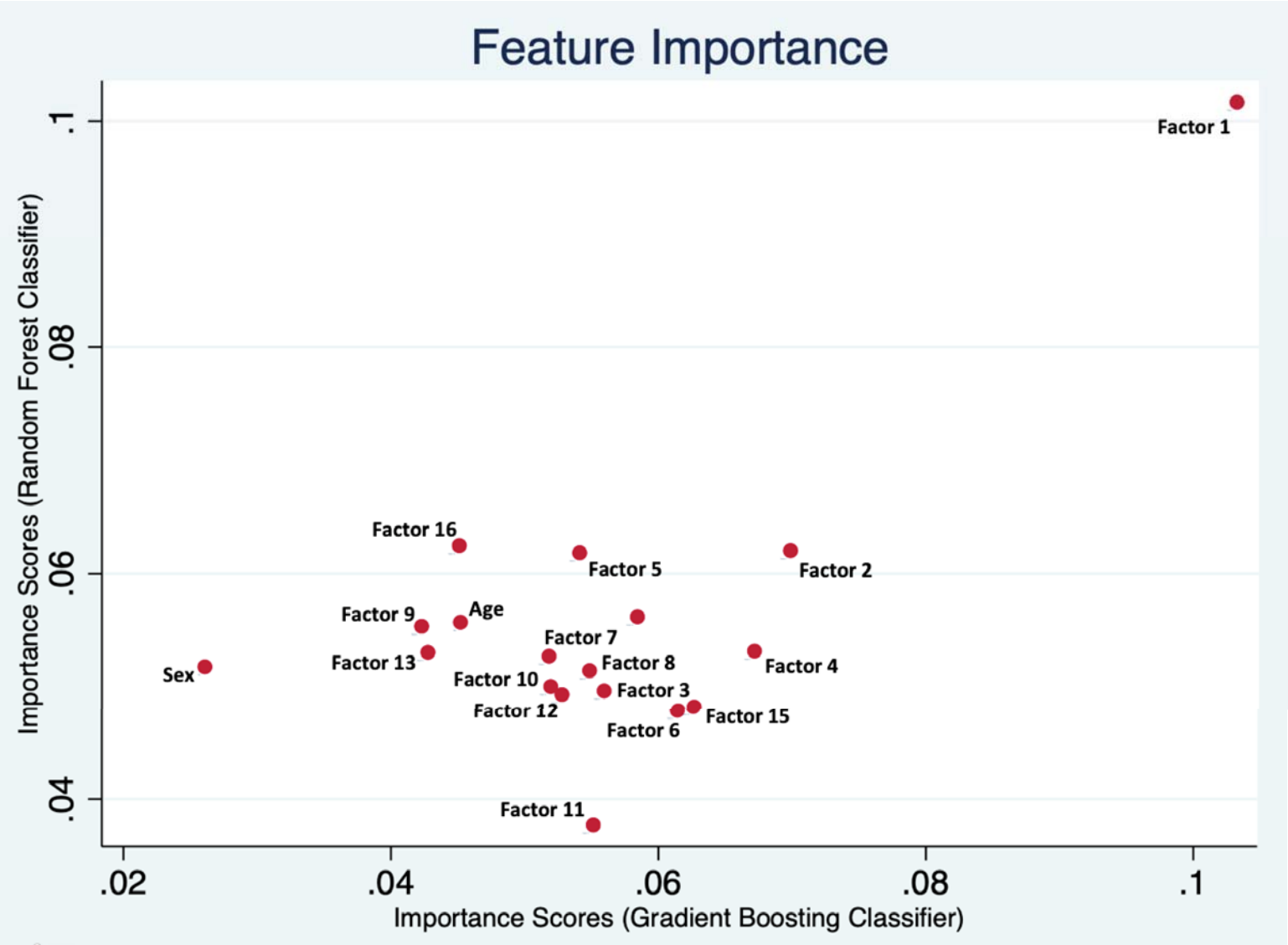
Figure 3. ROC Curves: Predicting ADHD in Children and Adults.



bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

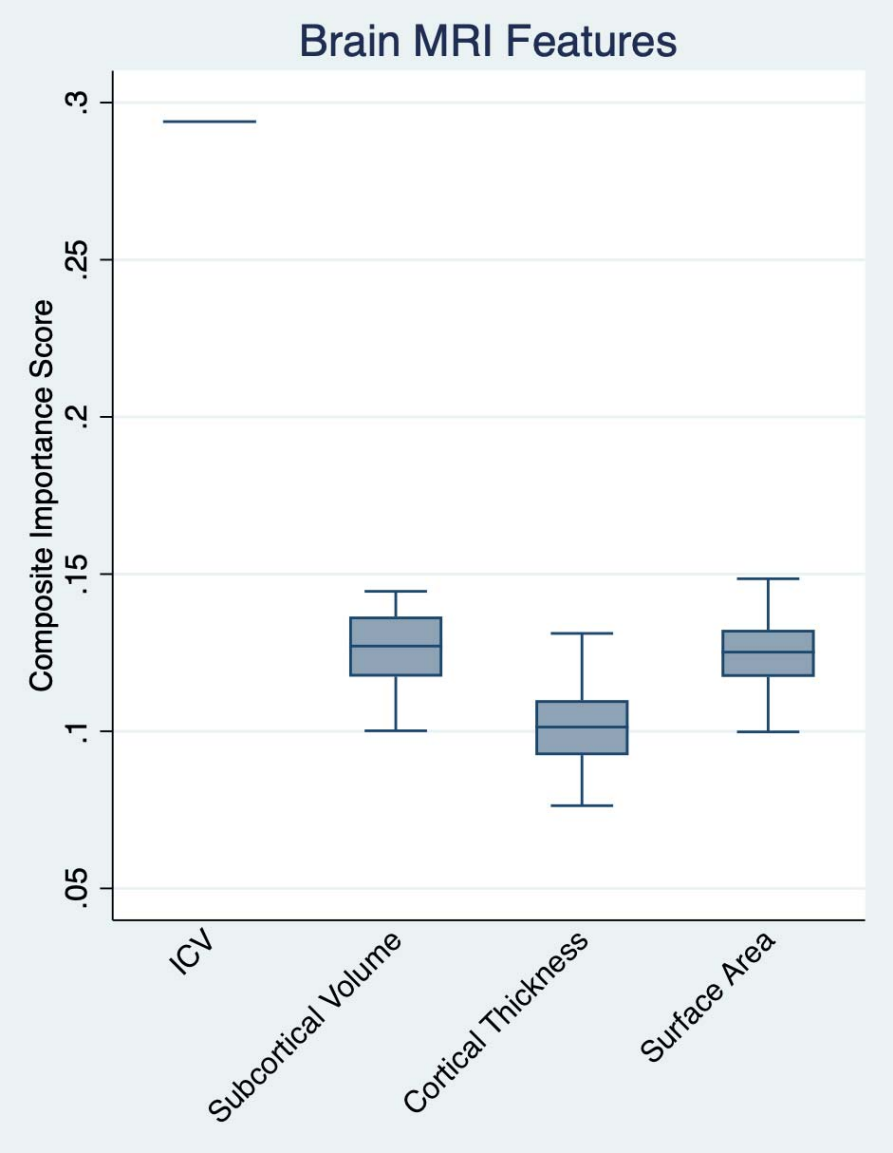
Figure 4. Feature Importance.

A.



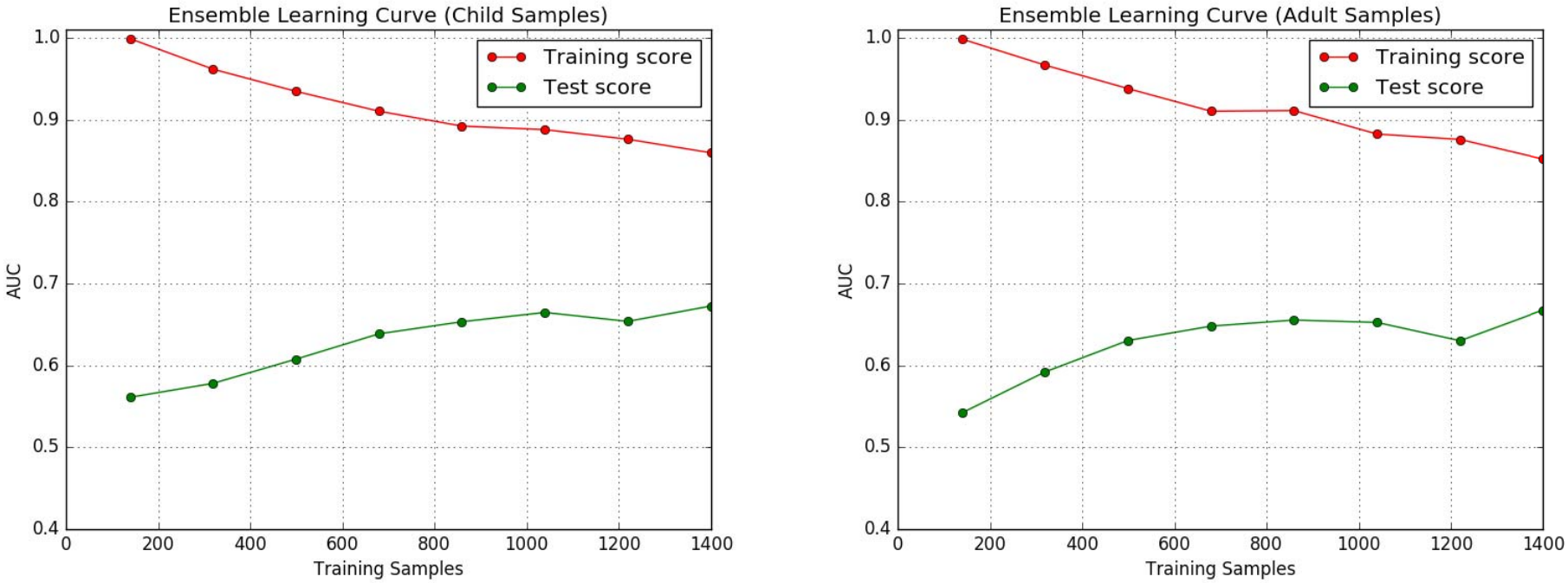
bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

B.



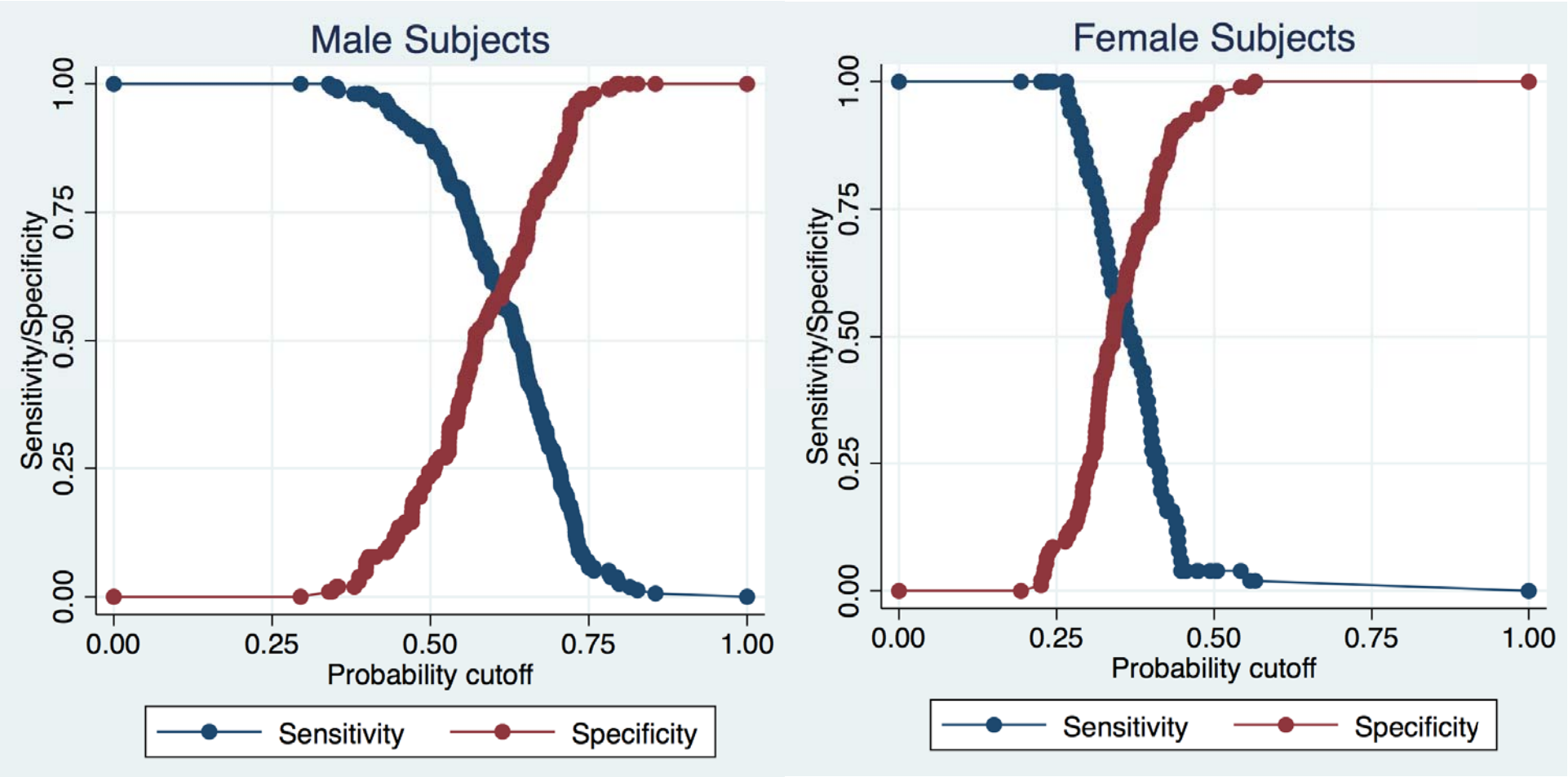
bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 5. Learning Curves for the best ensemble model for the adult model (Left) and child model (Right).



bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figure 1. Sensitivity and Specificity Analysis of Different Sex Groups



bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/546671; this version posted February 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.