Leveraging multi-modal longitudinal data from the Adolescent Brain Cognitive Development Study to refine obsessive-compulsive disorder phenotype for genetic analyses

Franjo Ivankovic[1,2*†], Aryan Singh[3†], Shashank Kumar[3†], Richa Gupta[3†]

1 Department of Psychiatry, College of Medicine, University of Florida, Gainesville, Florida, USA
2 Genetics Institute, University of Florida, Gainesville, Florida, USA
3 Department of Computer and Information Science and Engineering, University of Florida, Gainesville, Florida, USA

† These authors contributed equally to this work.
* njofrica@ufl.edu

Abstract

In children and adolescents, genetic effects account for most of the obsessive-compulsive disorder (OCD) variance (about 55%), yet exact genetic contributions are still poorly understood. Large initiatives like Adolescent Brain Cognitive Development (ABC) study provide a unique opportunity to study OCD in a population sample of adolescents, however self-report based psychiatric diagnoses might be vulnerable to higher phenotype misclassification rate, ultimately decreasing power of the genome-wide association study. In this study, we leverage multi-modal, longitudinal data from ABCD study and multi-kernel machine learning approaches to accurately classify ambiguous OCD cases and increase power to downstream genetic analyses. We leverage logistic regression model with elastic net penalty to train each kernel independently, followed by metric-weighted case probability score summed across kernels. We evaluated AUC, accuracy, Cohen's kappa, precision, and specificity as weighing metrics for kernel integration. Based on overall performance across various model metrics, Cohen's kappa was chosen. Following the analysis of genetic data and longitudinally averaged diffusion and structural MRI, rest state and task based fMRI, and questionnaire based psychiatric phenotype data, we show this approach has accuracy of 0.90, precision of 0.93, sensitivity of 0.94, and specificity of 0.86, when total model is based on Cohen's kappa weighted integration of individual kernels. Furthermore, we show the Cohen's kappa, a robust measurement for inter-rater reliability for psychiatric disorders in the field of psychiatric epidemiology, outperforms literature-reported Cohen's kappa for OCD, at 0.75 to 0.41. This model also resulted in diagnostic odds ratio of 96.24, indicating its usefulness as a reliable tool for OCD diagnosis. This paper presents a unique and novel way to integrate multi-modal data using machine learning approaches to improve diagnosis of OCD. It is also a proof of concept indicating this approach might work for other psychiatric disorders as well.

Author summary

Large longitudinal initiatives collect abundance of data, and drive sample sizes necessary for genome-wide association studies. However, power of genome-wide

association studies is heavily reliant on accurate phenotype classes, which limits the use of self-report based phenotypes. We developed a multi-kernel logistic regression with elastic net penalty machine learning model, leveraging high-confidence phenotypes and multi-modal longitudinal data to accurately classify ambiguous phenotypes and reduce loss of power due to miss-classification. We integrated individual kernels using Cohen's kappa statistic as a weighing metric. Based on currently available data, we train a model with accuracy of 0.90, precision of 0.93, sensitivity of 0.94, and specificity of 0.86. More importantly, we show this model achieves Cohen's kappa metric of 0.75, which is 1.83-fold increase over literature-reported Cohen's kappa metric for OCD. Thus, we show this model might be able to diagnose OCD more reliably than psychiatrists, by integrating additional modalities such as brain imaging data and genetic data. We believe additional improvements to this model are possible, indicating an immense potential for machine learning assisted diagnosis decision making in mental healthcare.

Introduction

Obsessive-compulsive disorder (OCD) is a neuropsychiatric condition affecting about 1.3% of individuals world-wide, and 2.3% of individuals in the US at some point in their lives [1, 2]. Diagnostic and Statistical Manual of Mental Disorders defines OCD as the presence of obsessions and compulsions that are time consuming, cause clinically significant distress or impairment, and are not attributable to the physiological effects of a substance or another medical condition [3]. OCD is highly heritable with a substantial genetic component, however genetic architecture of OCD is very complex and little is known about exact causative genetic factors and pathways [4].

Large-scale collaborative efforts and advancements in statistical methodology have driven progress in the field of psychiatric genomics, slowly unearthing numerous loci associating with psychiatric disorders and ushering the filed into the era of polygenic and pleiotropic effects, and a paradigm shift towards complex genetic architecture as a probabilistic rather than deterministic factor in disorder etiology [5].

One of the main drivers of the novel discoveries, and power-driving factors in genome-wide association studies in general, is sample size. Collaborative approaches have led to aggregation of large datasets, or summary statistics of smaller datasets, which have been combined by meta analyses to increase power and drive discoveries. Additionally, large coordinated initiatives with wide phenotyping range such as adolescent brain cognitive development (ABCD) study contribute data usable for psychiatric genomics, despite it not being the primary aim thereof [6]. However, simulation studies have shown that phenotype misclassification may present serious challenges to GWAS power, and be as important if not more important than sample size [7]. When it comes to psychiatric disorders like OCD, where phenotype classification agreements between clinicians can be challenging, self-report based data as those in ABCD study can have extensive rates of phenotype misclassification and negatively impact GWAS power [8].

Initial epidemiological exploration of the ABCD study cohort revealed overrepresentation of OCD in the sample, 13.4% prevalence compared to 1.3% prevalence in early adolescence. Such inflation of OCD prevalence in a sample which was not ascertained for psychiatric disorders or OCD risk factors raises concerns of potentially high phenotype misclassification rates with high potential impact to secondary analyses, like GWAS. In such analyses, high rates of misclassification can be detrimental to test power and p-value inflation (Figure 1). However, the rich data made available in the cohort provides an opportunity to apply machine learning approaches to identify likely positives among the ambiguous cases. In this study, we aim to use neurocognitive, biomedical, sociocultural, genetic, and brain imaging data to determine

likely OCD status among the OCD ambiguous individuals.

The growth in data availability coupled with improvements in computational power has led to a surge in research and application of machine learning to such use cases [9]. The techniques that started out by being used in niche areas like classification, natural language text analysis and basic detection are now showing promising results in understanding large-scale health care data. These methods are not only helping us understand human behavior or predict the outcome, but are also being explored to understand mental health [10, 12]. In recent years, there has been research that have shown how machine learning algorithms can be used to solve various aspects of mental health. Existing research assesses the accuracy, reliability, effectiveness, opportunity and challenges of adoption of such algorithms [13–15]. Others works study the algorithmic effectiveness in predicting the outcomes of clinical interventions (e.g., pharmacological treatments) for specific health conditions.

Consistent longitudinal endorsement in self-report data has been shown to accurately predict clinically accurate psychiatric diagnoses [16]. We utilize similar criteria to classify OCD patients into high-confidence negative cases, high-confidence positive cases, and ambiguous cases. The goal of this analysis is to develop a machine learning model that can estimate true OCD state among ambiguous cases. We use machine learning methods building multiple classifiers on each modality (phenotypes, genotypes, and imaging). We combine these classifiers using weighted probabilities to find the final class of the patient. We used logistic regression model with elastic net penalty to train models. We tested weights using AUC, accuracy (Eq. 1), specificity (Eq. 2), Cohenk's kappa (Eq. 3), and precision (Eq. 4).

$$\text{Accuracy} = \frac{TP + TN}{(TP + FP + TN + FN)} \tag{1}$$

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{2}$$

$$\text{Cohen}'\text{s } \kappa = \frac{2 * (TP * TN + FN * FN)}{(TP + FP) * (FP + TN) + (TP + FN) * (FN + TN)} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

This multi-kernel machine learning model incorporating multi-modal data to estimate classification of a psychiatric disorder such as OCD is a first of its kind. Utility of this model could enable more accurate diagnoses in both research and, potentially, clinical practice.

Materials and methods

P-value simulations

To simulate data for p-value tests, we created a test data with 280 cases and 1120 controls, setting general-population minor allele frequency at 0.1 and case minor allele frequency at $0.175, 0.2, 0.225, 0.25, 0.275, 0.3$. We then simulate number of major and minor alleles among cases and controls, modified by mis-diagnosis rates at $(0.00, 0.15)$. Finally, p-values are calculated using Fisher's exact test on the aforementioned allele counts.

Data Access and Processing

ABCD Study data were accessed through National Institute of Mental Health (NIMH) Data Archives (NDA). Longitudinal data on genetics, neuroimaging, and questionnaires/clinical measure were downloaded using NDA Download Manager (Beta v.0.1.38) and uploaded to University of Florida high-performance computing cluster for further processing. Final sample was composed of 8,661 samples.

Genetic Data

Genetic data for participants were derived from blood and saliva samples and genotyping on National Institute for Drugs and Addiction (NIDA) modified Affymetrix Smokescreen array [17]. Array features 646,247 typed loci, featuring custom smoke-risk loci. Data were pre-processed using Ricopili pipieline [18] and made available in Plink format [19].

Genetic data were further processed to remove problematic IDs (if not fixable), check and impute sex, remove non-autosomes, and exclude samples and probes with low genotyping rate. Subsequently, related individuals were excluded (lower genotyping frequency). Ultimately, probes which were in extreme Hardy-Weinberg disequilibrium ($p < 10^{-7}$) and has a low sample minor-allele frequency ($MAF < 0.05$) were removed, for being at high risk of genotyping errors. Samples with abnormal rates of heterozygosity ($|Het| > mean + 4 * sd$) as well, due to potential inbreeding (low heterozygosity) or cross-contamination (high heterozygosity).

Genetic data were subsequently phased with SHAPEIT4 [20] and imputed with IMPUTE5 [21], and finally filtered for previously associated OCD-loci based on latest OCD GWAS meta-analysis ($p \leq 10^{-2}$) [22]. Final genetic kernel was composed of 26,566 features for 8,713 individuals.

Neuroimaging data

Tabulated summaries of post-processed neuroimaging data obtained via structural magnetic resonance imaging (MRI) and diffusion MRI scans were downloaded from NDA for imaging kernels. As data were pre-processed internally by the ABCD Study group, no additional processing was necessary.

Imaging data were split into four kernels based on the type of imaging experiment. Structural MRI (sMRI), which visualizes brain anatomy, was composed of 5 individual data sets, totalling 2,703 features. Diffusion MRI (dMRI), which visualized water diffusion around the neural tissues, was composed of 8 individual data sets, totalling 7,474 features. Resting state fMRI (rsfMRI), which visualizes brain activity while participant is at rest (i.e. not executing specific tasks) was composed of 7 individual data sets, totalling 776 features. Task based fMRI (tsfMRI), which visualizes brain activity while the participant is executing specific research-related tasks, was composed of 39 individual data sets, totalling 22,339 features. All imaging data was available for 11,841 individuals.

Questionnaire and phenotype data

Tabulated questionnaire and phenotype data were downloaded from NDA for questionnaire kernel and label definition. Based on the longitudinal OCD diagnosis values, the samples were classified into OCD positive (diagnosed at every assessment, N = 274), OCD negative (never diagnosed, N = 7,307), and OCD ambiguous (diagnosed at a single assessment, N = 1,080). OCD positive and OCD negative cases were subset and used in this study.

Questionnaire kernel was constructed by pulling numeric variables from questionnaire data sets, as well as numeric-transforming string variables, into a final kernel with a total of 1,317 features for 11,876 individuals. Data were pulled for all psychiatric, cognitive, and behavioral questionnaires and tests - from 60 data sets. OCD diagnostic measures were removed.

Data processing

For each kernel except the genetic, all data sets have been loaded in and joined. Subsequently, all features with $> 10\%$ missingness rate were removed from the analysis. The remaining features were mean-imputed. Feature with the range of 0 were removed and remaining data were range normalized to range $\langle 0, 1 \rangle$. Finally, features with low variance (< 0.001) were removed.

Machine learning model - Method

A total of four processed data sets were defined for all samples: phenotypes, genotypes, diffusion MRI, structural MRI, resting state fMRI, and task based fMRI. We built 6 different kernels based on these categories: phenotype, genotype, sMRI, dMRI, rsfMRI, and tsfMRI (Figure 2).

Learning Model

All kernels were individually trained over multiple epochs by under-sampling to remove effects of label imbalance. We calculated the mean of the parameters from all the batches to get the parameter of the final model. Each of the kernels uses Logistic Regression model with elastic net as the penalty for every batch. For every kernel, test accuracy and various metrics based on confusion-matrix were calculated.

Hyper-parameters:

- Penalty: Elastic Net
- Solver: Saga
- L1 Ratio: 0.9

Kernel Combination

We have tested various metrics using metric-weighted additive model (2), including: AUC, accuracy (Eq. 1), specificity (Eq. 2), Cohen's kappa (Eq. 3), and precision (Eq. 4). Additionally, we have conducted a leave-one-out analysis to determine the influence of lowest metric kernel, models and their weights are shown in supplemental table S1.

Support Methods

We compared model performance using Cohen's kappa score to Cohen's kappa scores reported in psychiatry and psychiatric epidemiology literature, reference Cohen's kappa of 0.41 [23].

Issues

During the analysis, two noticeable issues arose: there was substantially large number of features compared to observations, and data sets were substantially imbalanced in terms of classes (i.e. OCD diagnosis status).

While penalized elastic net regression does reduce the number of features (weights set to 0) to mitigate over-fitting, closer inspection has revealed that a large number of features were not informative enough to warrant computational resources. In order to accomplish this in genetic data, we have only selected those variants which were previously associated with OCD in latest OCD GWAS meta-analysis ($p \leq 10^{-2}$) [22].

In order to mitigate effects of imbalanced data, we have split the training sample in epochs which were composed of balanced number of sub-sampled cases from the training set and unique controls from the training set, iterated until training set controls were exhausted.

Success Measurement

The aim of this project is to develop a machine learning model to classify individuals with OCD status. In order to measure our success, we have utilized confusion-matrix based assessments like accuracy (Eq. 1), specificity and sensitivity. At the baseline, we expect accuracy of 0.5, sensitivity of 0.5, and specificity of 0.5, assuming prevalence of 0.5.

Code availability

- Github repository: `https://github.com/FranjoIM/CIS3930_MLP`.
- Data: `https://abcdstudy.org` (public facing URL, restricted access data).

Results

Summary

Performance of each kernel in terms of accuracy has been visually represented in figure 3. Phenotype kernel vastly outperformed all other kernels in terms of accuracy. Additionally, plots of receiver operative curves and precision-recall curves are shown for each individual kernel in figure 4. Based on AUC metrics and F1 scores, we can see that phenotype kernel outperforms all other kernels. The kernels based on imaging and genetic data performed similarly, very close to expected baseline performance.

The kernel-wise training results have been also summarised in figure 1. For each kernel and total mode, we calculated Cohen Kappa, Accuracy, positive predicted value, false omission rate, false discovery rate, negative predictive value, sensitivity, fall-out, specificity, miss rate, positive likelihood ratio, negative likelihood ratio, markedness, informedness, prevalence threshold, diagnostic odds ratio, balanced accuracy, F1 score, Fowikes-Mallows index, Matthews correlation coefficient, and threat score. While, in terms of most metrics, imaging and genetic data remained unremarkable, Cohen's kappa metric for phenotype kernel has substantially outperformed psychiatrists (0.60 to 0.41).

Evaluation of weight-metrics and kernel inclusions has shown that Cohen's kappa weighing generally yields the best outcome in terms of accuracy, precision, specificity, and sensitivity. Eliminations of specific modules did not appear to significantly change those metrics - as majority of weight distribution has been allocated to best-performing kernel: phenotype. These results are graphically summarized in figure 5.

Based on these data, we have chosen to use Cohen's kappa as a weighing metric for final model, with formula show in in equation 5. Based on this metric, dMRI kernel was completely reduced and not considered in the total model. The final model had Cohen's kappa of 0.75, accuracy of 0.90, precision of 0.93, sensitivity of 0.86, and specificity of 0.94. These metrics are graphically summarized in figure 6.

Table 1. Tabulated summary of kernel perfomance metrics.

	Phenotype	Genotype	sMRI	dMRI	rsfMRI	tsfMRI
Type	Questionnaire	Genotypes	Imaging	Imaging	Imaging	Imaging
AUC	0.93	0.56	0.59	0.57	0.60	0.56
Cohen's Kappa	0.60	0.02	0.04	0.00	0.08	0.10
Accuracy	0.85	0.53	0.53	0.50	0.56	0.56
Precision	0.86	0.52	0.55	0.50	0.58	0.63
False Omission Rate	0.44	0.44	0.48	0.50	0.45	0.46
False Discovery Rate	0.14	0.48	0.45	0.50	0.42	0.38
Negative Predictive Value	0.84	0.56	0.52	0.50	0.55	0.54
Sensitivity	0.84	0.76	0.32	0.18	0.42	0.30
Fall-out	0.14	0.70	0.26	0.18	0.30	0.18
Specificity	0.86	0.30	0.74	0.82	0.58	0.82
Miss Rate	0.16	0.24	0.68	0.82	0.58	0.70
Positive Likelihood Ratio	6.00	1.09	1.23	1.00	1.40	1.67
Negative Likelihood Ratio	0.19	0.80	0.92	1.00	0.83	0.85
Markedness	0.70	0.08	0.07	0.00	0.13	0.16
Informedness	0.70	0.06	0.06	0.00	0.12	0.12
Prevalence Threshold	1.21	8.47	8.36	∞	4.57	4.27
Diagnostic Odds Ratio	32.25	1.36	1.34	1.00	1.69	1.95
Balanced Accuracy	0.85	0.53	0.53	0.50	0.56	0.56
F1 Score	0.85	0.62	0.41	0.26	0.49	0.41
Fowlkes-Mallows Index	0.85	0.63	0.42	0.30	0.49	0.43
Matthews Correlation Coefficient	0.70	0.07	0.07	0.00	0.13	0.14
Threat Score	0.74	0.45	0.25	0.15	0.32	0.25

Table shows summary of 22 confusion matrix-derived metrics used to evaluate performance of each kernel in the study.

$$\begin{aligned}\text{Total model} = {} & 0.7143 * \text{Phenotype}+ \\ & 0.1190 * \text{tsfMRI}+ \\ & 0.0952 * \text{rsfMRI}+ \\ & 0.0476 * \text{sMRI}+ \\ & 0.0238 * \text{Genotype}+ \\ & 0.0000 * \text{dMRI}\end{aligned} \tag{5}$$

Successes

At the baseline, our total model has substantially outperformed all baseline metrics, as shown in figure 6, while individual kernels had variable performance (with phenotype kernel consistently performing the best).

In terms of specific metrics, the total model has Cohen's kappa of 0.75, accuracy of 0.90, precision of 0.93, false omission rate of 0.13, false discovery rate of 0.07, negative predictive value of 0.87, sensitivity of 0.86, specificity of 0.94, positive likelihood ratio of 14.40, negative likelihood ratio of 0.15, informedness of 0.80, prevalence threshold of 1.12, and diagnostic odds ratio of 96.24.

These metrics, particularly Cohen's kappa and diagnostic odds ratio appear to show our model to be a reliable method for prediction of OCD case status.

Issues

As previously described, due to a large number of features and imbalanced classes, feature selection steps in terms of elimination of features with high missingness, 0 range, low variance, or previously reported associations, and subsampling, iterative approach to training were used to successfully ameliorate these effects. This is best evident in apparent lack of over-fitting, as can be inferred form model's performance in the test set.

Discussion

Results

Our total model has substantially outperformed both individual kernels (including Phenotype kernel) and psychiatry field trials in terms of Cohen's kappa, with values of 0.75, 0.60, and 0.41 for total model, phenotype (highest performing) kernel, and psychiatric field-trials, respectively. An important aspect of Cohen's kappa metric is that in addition to measuring simple agreement, Cohen's kappa is more robust by the virtue of taking into account the possibility of agreement by chance. While exact interpretation of this metric is difficult, it is a preferred measure in the field of psychiatric epidemiology, and thus useful for this analysis. Out total multi-kernel model outperformed psychiatric field-trials in terms of Cohen's kappa by nearly 1.83-fold as compared to the literature OCD kappa [23].

Furthermore, positive and negative likelihood ratios, as well as diagnostic odds ratio are commonly used metrics in evidence-based medicine to determine the effectiveness and value of performing a diagnostic test. Based on our total model, individuals who are assigned a label of OCD positive have 14.40-fold increase in odds of having OCD. Conversely, individuals who are assigned a label of OCD negative have 6.67-fold increase in odds of not having OCD. Our total model also has a remarkably high diagnostic odds ratio of 96.24, making it a reliable tool for OCD diagnosis.

Based on our test, the prevalence threshold is 1.12. Prevalence threshold indicates a point below which tests' positive predictive value (i.e. precision) begins to sharply decrease. As OCD prevalence in our sample is 2.61, and prevalence rate of OCD in adult and adolescent population in the US is 2.3 and 1.6, respectively, our model can be applied reliably to OCD [2, 24].

Informedness test estimates the quality of a test - giving a value of 0 when test is not useful and randomly assigns class and value of 1 when test is perfect, without false positives or false negatives. Our total model has informedness of 0.80, making it a very useful and reliable test for OCD. Ultimately, this is reflected in test sensitivity and specificity, which are calculated at 0.94 and 0.86, respectively.

While these results are promising, further improvements are possible. For example, for imaging kernels, instead of using tabulated imaging data, actual minimally-processed brain images could be put through a deep learning model to potentially improve the performance of imaging kernels. Furthermore, the genetic kernel is likely susceptible to ancestry effects, so additional work could help reduce such effects and improve genetic kernel performance.

Novelty

Literature on machine learning applications in the field of psychiatry remains relatively sparse. Multi-modal based machine learning approaches even more so - our paper appears to be the first to integrate multi-disciplinary approaches (traditional psychiatry and psychology in terms of phenotypic measures, genetics in terms of genotypes, and neurobiology in terms of neuroanatomical imaging) to determine a mental health

diagnosis in an individual. Here, we show a promising potential to such models, and basis on which, with further improvements, diagnosing mental illnesses could become easier and more reliable.

Conclusion

In this paper, we aimed to develop a machine learning model that can utilize diverse sources of data (phenotype, genotype, brain imaging) to diagnose individuals with OCD. While each modality varies in terms of performance given a specific metric (e.g. accuracy of phenotype modality being at 0.85 and genotype at 0.53), integration of these methods ultimately refines final predictions and yields optimal results. This is evident in Cohen's kappa metric, commonly utilized in the field of psychiatric epidemiology. On an individual level, Cohen's kappa is highest for Phenotype kernel at 0.60, whereas in total model, Cohen's kappa is 0.75. This shows utility of including non-phenotype modalities, even when those kernels have less than optimal performance. Ultimately, our total model performs much better than literature-reported Cohen's kappa for OCD of 0.41, further showing evidece for its utility. This novel approach will allow integration of data and potentially improve accuracy of OCD diagnosis, thus help faster delivery of appropriate treatments.

Figures

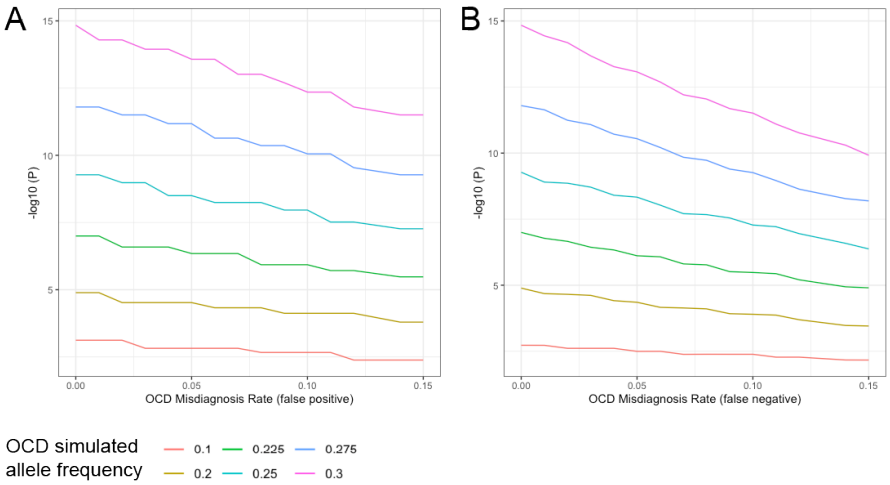

Fig 1. GWAS p-value simulation P-value deterioration curves per minor-allele frequency bins (0.1, 0.2, 0.225, 0.25, 0.275, 0.3), plotted against rates of phenotype misclassifications. A: P-value deterioration as a function of misclassified case status. B: P-value deterioration as a function of misclassified control status.

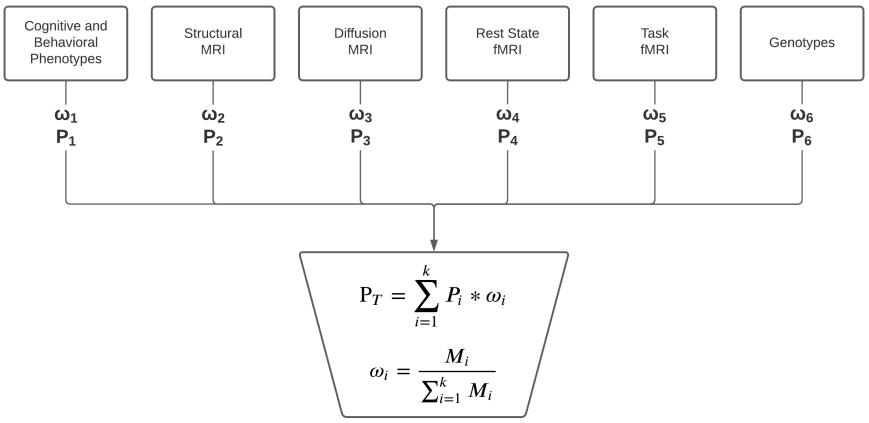

Fig 2. Machine learning model. A diagram showing how data are kernelized and processed, and class probabilities combined into a final class estimation.

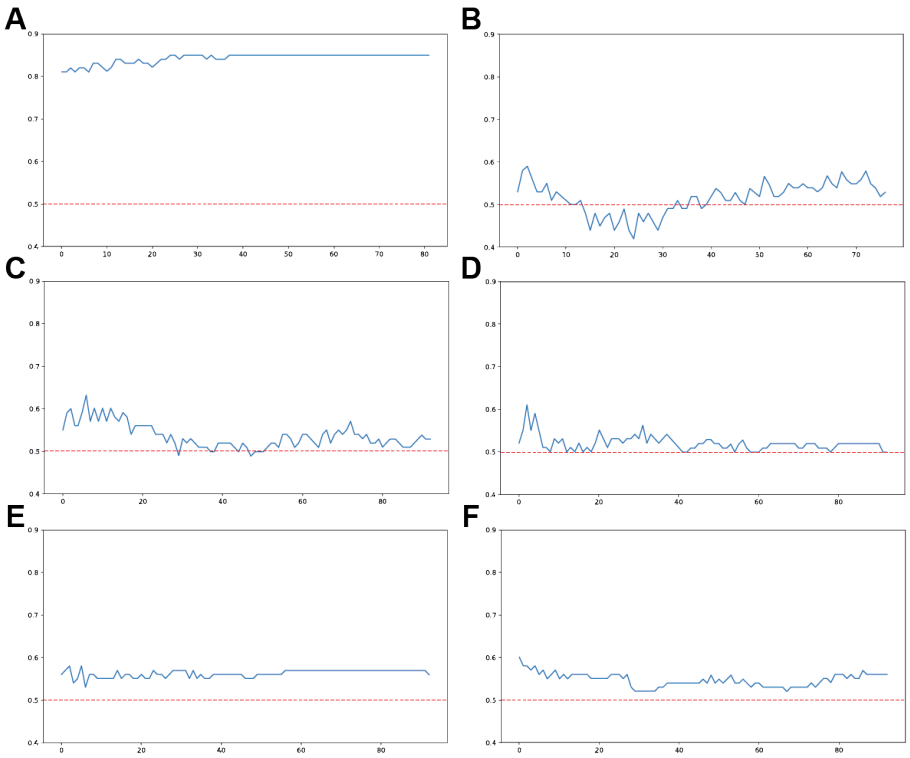

Fig 3. Kernel performace over epochs. Figure shows performance in terms of accuracy over epochs for each of the six kernels: phenotype (A), genotype (B), structural MRI (C), diffusion MRI (D), resting state fMRI (E), and task fMRI (F). Red line at 0.5 accuracy is associated with random case-control allocation.

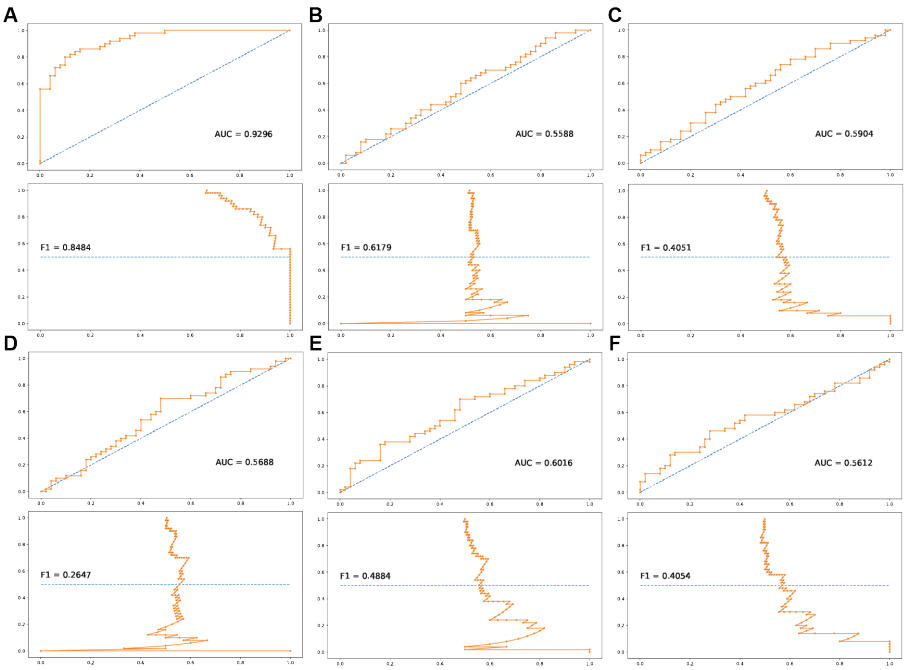

Fig 4. Receiver operating and precision-recall curves. Kernel-wise ROC curve and precision-recall curve for each of the six kernels: phenotype (A), genotype (B), structural MRI (C), diffusion MRI (D), resting state fMRI (E), and task fMRI (F). Each ROC curve includes area under the curve calculation, each precision-recall curve includes F1 score.

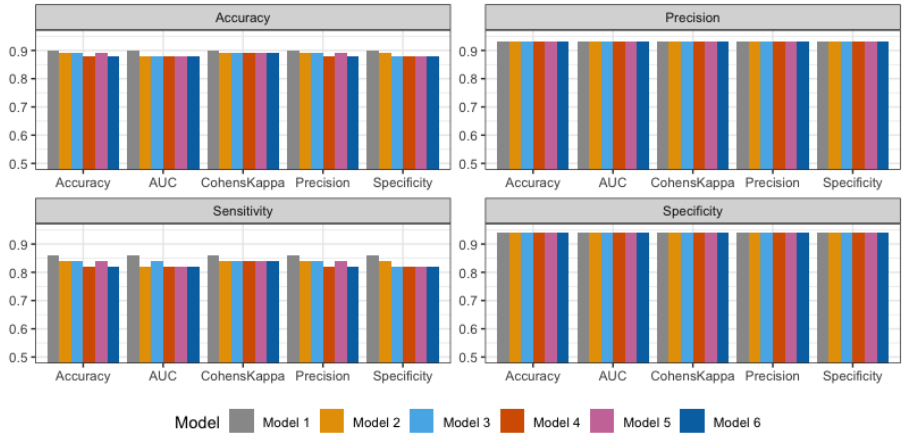

Fig 5. Chart summary of kernel and total performance. Accuracy, precision, sensitivity, and specificity metrics for model selection analysis. Models were based on weighing metric (accuracy, AUC, Cohen's kappa, precision, specificity), where analyses were done eliminating one kernel at a time, based on performance. Model 1 has only the highest performing kernel in the specific weighing metric, model 6 includes all kernels. Full table with models, included kernels, and their weights, are shown in table S1.

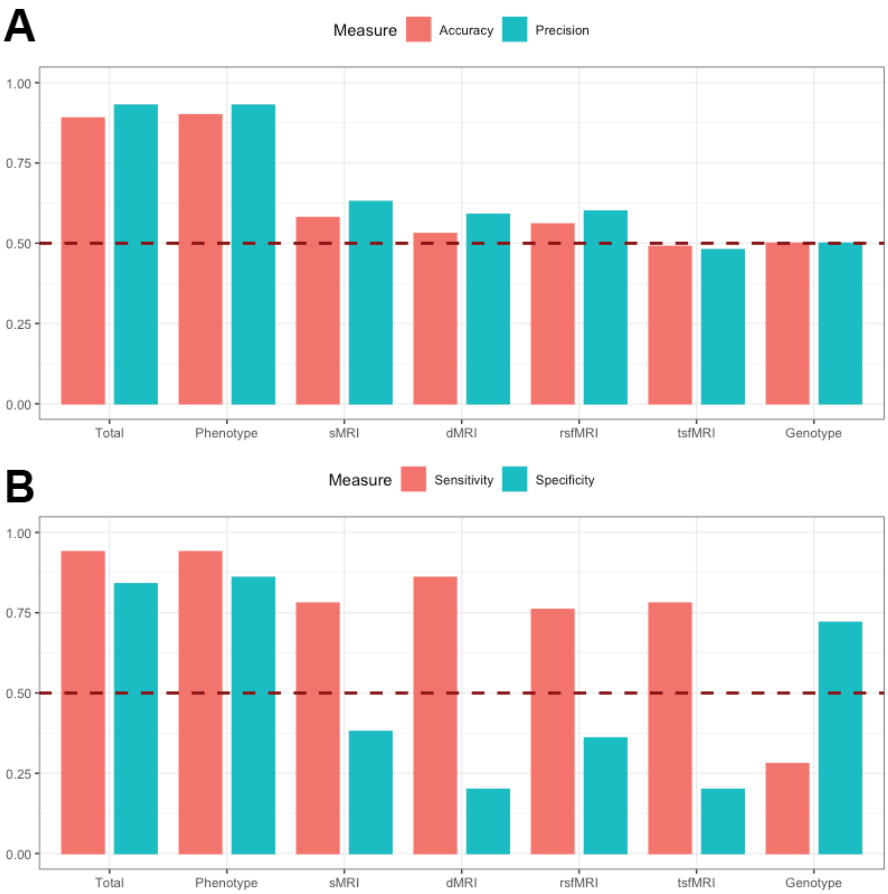

Fig 6. Test set analysis of kernels and final model. Accuracy and precision (A) and sensitivity and specificity (B) analysis of individual kernels and total model using test-set data.

Supporting information

Table S1. Kernel weights based on 5 selective metrics and model reduction. Table shows weights for each kernel based on 5 different weighing metrics, as well as kernel-elimination adjusted weights.

Table S1.

Kernel (Metric)	Model 6	Modeel 5	Model 4	Model 3	Model 2	Model 1
AUC						
Phenotype	0.2441	0.2862	0.3457	0.4387	0.6078	1.0000
rsfMRI	0.1575	0.1846	0.2230	0.1830	0.3922	-
sMRI	0.1549	0.1815	0.2193	0.2783	-	-
dMRI	0.1496	0.1754	0.2119	-	-	-
tsfMRI	0.1470	0.1723	-	-	-	-
Genetic	0.1470	-	-	-	-	-
Accuracy						
Phenotype	0.2408	0.2805	0.3400	0.4315	0.6028	1.0000
tsfMRI	0.1586	0.1848	0.2249	0.2843	0.3972	-
rsfMRI	0.1586	0.1848	0.2249	0.2843	-	-
sMRI	0.1501	0.1749	0.2120	-	-	-
Genetic	0.1501	0.1749	-	-	-	-
dMRI	0.1416	-	-	-	-	-
Specificity						
Phenotype	0.2087	0.2251	0.2654	0.3440	0.5119	1.0000
tsfMRI	0.1990	0.2147	0.2531	0.3280	0.4881	-
dMRI	0.1990	0.2147	0.2531	0.3280	-	-
sMRI	0.1796	0.1937	0.2248	-	-	-
rsfMRI	0.1408	0.1518	-	-	-	-
Genetic	0.0728	-	-	-	-	-
Cohen's Kappa						
Phenotype	0.7143	0.7143	0.7317	0.7692	0.8571	1.0000
tsfMRI	0.1190	0.1190	0.1220	0.1282	0.1429	-
rsfMRI	0.0952	0.0952	0.0976	0.1026	-	-
sMRI	0.0476	0.0476	0.0488	-	-	-
Genetic	0.0238	0.0238	-	-	-	-
dMRI	0.0000	-	-	-	-	-
Precision						
Phenotype	0.2363	0.2739	0.3282	0.4155	0.5772	1.0000
tsfMRI	0.1731	0.2006	0.2405	0.3043	0.4228	-
rsfMRI	0.1593	0.1847	0.2214	0.2802	-	-
sMRI	0.1511	0.1752	0.2099	-	-	-
Genetic	0.1429	0.1656	-	-	-	-
dMRI	0.1374	-	-	-	-	-

Acknowledgments

We would like to acknowledge Kiley Graim, PhD for feedback and assistance with machine learning methods, and Carol Mathews MD, for feedback and assistance with psychiatric and genetics methods. We would also like to acknowledge their home

departments for funding, resources, and infrastructure that made this project available. We would especially like to acknowledge National Institute for Mental Health and the ABCD Study initiative for making data available for analysis.

References

1. Fawcett, EJ, Power H, Fawcett JM. Women are at greater risk of OCD than men: A meta-analytic review of OCD prevalence worldwide. J. Clin. Psychiatry. 2020; 81(4).

2. Ruscio AM, Stein DJ, Chiu WT, Kessler RC. The epidemiology of obsessive-compulsive disorder in the National Comorbidity Survey Replication. Mol. Psychiatry. 2010; 15(1), 53–63.

3. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5th ed. American Psychiatric Association. 2013.

4. Strom NI, Soda T, Mathews CA, Davis LK. A dimensional perspective on the genetics of obsessive-compulsive disorder. Transl. Psychiatry. 2021; 11(401).

5. Reynolds T, Johnson EC, Huggett SB, Bubier JA, Palmer RHC, Agrawal A, et al. Interpretation of psychiatric genome-wide association studies with multispecies heterogeneous functional genomic data integration. Neuropsychopharmacology. 2021; 46(1):86-97.

6. Bjork JM, Straub LK, Provost RG, Neale MC. The ABCD study of neurodevelopment: Identifying neurocircuit targets for prevention and treatment of adolescent substance abuse. Curr. Treat. Options Psychiatry. 2017; 4(2):196-209.

7. Manchia M, Cullis J, Turecki G, Rouleau GA, Uher R, Alda M. The Impact of Phenotypic and Genetic Heterogeneity on Results of Genome Wide Association Studies of Complex Diseases. PLoS ONE. 2013; 8(10):e76295.

8. Freedman R, Lewis DA, Michels R, Pine DS, Schultz SK, Tamminga CA, et al. The Initial Field Trials of DSM-5: New Blooms and Old Thorns. Am. J. Psychiatry. 2013; 170(1):1-5.

9. Thieme A, Belgrave D, Doherty G. Machine Learning in Mental Health. ACM Trans. Comput. Hum. Interact. 2020 Oct 31;27(5):1-53.

10. Shatte ABR, Hutchinson DM, Teague SJ.

11. Psychol. Med. 2019 Jul;49(09):1426-48.

12. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015 Jul 17;349(6245):255-60.

13. Lee Y, Ragguett R, Mansur RB, Boutilier JJ, Rosenblat JD, Trevizol A, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. J. Affect. Disord. 2018 Dec;241:519-32.

14. Mohr DC, Zhang M, Schueller SM. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. Annu. Rev. Clin. Psychol. 2017 May 8;13(1):23-47.

15. Rohani DA, Faurholt-Jepsen M, Kessing LV, Bardram JE. Correlations Between Objective Behavioral Features Collected From Mobile and Wearable Devices and Depressive Mood Symptoms in Patients With Affective Disorders: Systematic Review. JMIR mhealth uhealth. 2018 Aug 13;6(8):e165.

16. Sordo Vieira L, Nguyen B, Nutley SK, Bertolace L, Ordway A, Simpson H, et al. Self-endorsement of psychiatric illness in an online patient registry is a good indicator of the existence of psychiatric illness Submitted, unpublished.

17. Baurley JW, Edlund CK, Pardamean CI, Conti DV, Bergen AW. Smokescreen: a targeted genotyping array for addiction research BMC Genomics. 2016 Feb 27;17(145).

18. Lam M, Awasthi S, Watson HJ, Goldstein J, Panagiotaropoulou G, Trubetskoy V, et al. RRICOPILI: Rapid Imputation for COnsortias PIpeLIne Bioinformatics. 2020 Feb 1;36(3):930-3.

19. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets GigaSci. 2015 Dec;4(1).

20. Delaneau O, Zagury J, Robinson MR, Marchini JL, Dermitzakis ET. Accurate, scalable and integrative haplotype estimation Nat Commun. 2019 Dec;10(1).

21. Rubinacci S, Delaneau O, Marchini J. Genotype imputation using the Positional Burrows Wheeler Transform PLoS Genet. 2020 Nov 16;16(11):e1009049.

22. IOCDF-GC, OCGAS. Revealing the complex genetic architecture of obsessive–compulsive disorder using meta-analysis Mol Psychiatry. 2018 May;23(5):1181-8.

23. Chmielewski M, Clark LA, Bagby RM, Watson D. Method matters: Understanding diagnostic reliability in DSM-IV and DSM-5 Journal of Abnormal Psychology. 2015 Aug;124(3):764-9.

24. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication Arch Gen Psychiatry. 2005 Jun;62(6):593-602.