CONTEXT MATTERS: USING GENOMIC KNOWLEDGE TO IMPROVE DISORDER CLASSIFICATION MODELS

By: Eric Barnett
A Dissertation in Neuroscience

UPSTATE
MEDICAL UNIVERSITY

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the College of Graduate Studies of State University of New York, Upstate Medical University.

Approved:

Date: January 4, 2023

Table of Contents

Common Abbreviations

Abbreviation	Meaning
ADHD	Attention deficit-hyperactive disorder
AUC	Area under the receiver operating characteristic curve
BPD	Bipolar disorder
CI	Confidence interval
CID	Context informed data matrix
CNN	Convolutional neural network
DBD	Disruptive behavior disorders
gsPRS	Gene set polygenic risk score
GWAS	Genome-wide association study
MDD	Major depressive disorder
ML	Machine learning
NN	Neural network
OR	Odds ratio
PC	Principal component
PCA	Principal components analysis
PRS	Polygenic risk score
RF	Random forest
SNP	Single nucleotide polymorphism
T2D	Type 2 diabetes

Abstract

Context Matters: Using Genomic Knowledge to Improve Disorder Classification Models

Eric J. Barnett

Stephen V. Faraone

Despite heritability estimates that suggest a high ceiling for the classification of many complex genetic disorders, current models have only been moderately successful at accurately classifying cases and controls of these disorders. The knowledge base about the human genome is large and continuously growing, but disorder classification models rarely use any of that information beyond genetic associations. We use three different genomic context data granularities, 4 different machine learning models, and datasets of mood disorders, ADHD, and type 2 diabetes to test hypotheses on whether including genomic context can improve modelling of disorder risk. When predicting whether subjects had been diagnosed with any mood disorder, we found that using polygenic risk scores from other psychiatric disorders in logistic regression models improved classification performance as measured by the area under the receiver operating characteristic curve (AUC). In another study classifying cases of ADHD and controls, we found that the addition of summations of risk based on the genetic variants' inclusion in gene sets associated with ADHD improved AUCs in random forest modelling. The random forest importance scores of those gene set polygenic risk scores showed biological relevance through the correlation of importance scores with relative gene set expression in the brain. In the final study classifying type 2 diabetes cases and controls, for each genetic variant, we attached several types of functional genomic annotations to genotype data. These genomic context informed genotype data were used in convolutional neural networks and significantly improved AUC compared to polygenic risk score models while using a within-model adversarial ancestry task to

adjust for potential confounding due to ancestry. In these models, we found that some risk features developed by context informed data overlapped with features developed with standard genotype input while other risk features were unique to the input type. Together, these studies provide evidence that context matters when looking at the disorder risk conferred by genetic variants in complex genetic disorders.

Introduction

Many psychiatric disorders are highly heritable (Pettersson et al., 2019). While high heritability should make genetic prediction or classification of these disorders effective, performances of models with these tasks have been lower than expected. In studies that build machine learning models for multiple disorders, psychiatric disorders tend to be on the lower end of the classification performance range, even though they have higher heritability than many of the other disorders modeled (Evans et al., 2009; Mittag et al., 2015). One main culprit of model ineffectiveness is the staggering genetic complexity of these disorders, which is in large part due to the disorders' polygenicity, pleiotropy, and heterogeneity.

Psychiatric disorders and other complex genetic disorders like Type 2 Diabetes have been shown to be highly polygenic, meaning the disorders are influenced by many different genomic loci. While researchers have only identified relatively few genetic variants that are associated with most psychiatric disorders at a genome-wide significance threshold, it is thought that the actual number of variants involved in these disorders is in the thousands (Psychiatric Genomics Consortium, 2014; Stahl et al., 2019; Sullivan & Geschwind, 2019; Wray et al., 2018). In most cases, any single common variant changes the risk of a disorder by a small amount. While there is evidence that some rare variants have a large effect on genetic risk, it is hypothesized that in most cases a person's genetic risk of a psychiatric disorder is through the accumulation of these small effects (Visscher et al., 2017; Zoghbi et al., 2021). The polygenicity of these disorders led to the popularity and success of polygenic risk scores (PRS) which sum the risk of a person across many loci across the genome (Visscher et al., 2017). Multiple studies using PRS have shown that the combination of risk at loci at and below the threshold for genome wide significance results in better classification accuracy when modelling the risk of complex genetic disorders in comparison to modelling a single variant or small groups of genome-wide significant variants (Evans et al., 2009; Mittag et al., 2015).

Another factor driving the complexity of these disorders is pleiotropy, which refers to genes affecting two or more phenotypes. This well studied aspect of complex genetic disorders results in genetic overlaps between different traits and disorders (Cross Disorder Group of the Psychiatric Genomic Consortium, 2019). Multiple studies have shown that many common genetic variants are shared among disorders (Anttila et al., 2018; Mullins et al., 2022; Rommelse et al., 2010; Smoller et al., 2013; Tylee et al., 2018). In the clinic, this may be a factor in the widespread comorbidity of psychiatric disorders as pleiotropic variants influence the risk of multiple disorders simultaneously. This pleiotropy and comorbidity along with other evidence suggests that the discrete, categorical disorders in psychiatry may represent overlapping clusters of people at the extremes of multiple continuous traits (Smoller et al., 2019).

Genetic heterogeneity, which refers to the development of risk for some disorder through different genetic mechanisms, is also a critical component of the complexity of psychiatric and other complex genetic disorders. As previously mentioned, individual variants only affect the risk of the disorder by a small amount. This means that when calculating the risk of that variant, the proportion of people that had the variant was only slightly different in people who had the disorder compared to people without the disorder. Therefore, many people are seemingly unimpacted by variants that in other people influence the accumulation of risk that leads to the development of the disorder. Studies comparing distributions of PRS in people with and without some disorder show a similar feature (Hess et al., 2019; Hou et al., 2022). Even when summing risk across the genome, there are many people who have a high combined risk for a disorder but do not have the disorder. On the other hand, there are also people with a low combined risk for a disorder that have the disorder. Multiple factors likely influence this finding, including differing environmental risks, rare variant risk, and genetic resilience. Another explanation is that different sets of genetic risk accumulation have different effects on overall risk. This type of heterogeneity would help explain the low ability of PRS to predict disorders, since finding the meaningful sets of accumulated genetic risk among hundreds of thousands of variants would be a difficult task.

These aspects of the genetic complexity of psychiatric disorders combined with others have limited the performance of models aiming to accurately classify people with and without some psychiatric disorder. In classification models, one common performance metric is area under the receiver operating characteristic curve (AUC). The AUC plots the true positive rate against the false positive rate at many thresholds to separate cases from controls. One then calculates the area under that curve. The AUC is the probability that a random case/control pair are correctly classified. One study derived a mathematical equation relating the maximum AUC for a genetic model given a disorders heritability and prevalence (Wray et al., 2010). Based on the equation, most psychiatric disorders and other complex genetic disorders have a possible maximum AUC above 0.9 when all known genetic variance is explained and many of the disorders have possible AUCs near 0.9 when half of the known genetic variance is explained by the model. Some studies achieve results close and, in some cases, even exceeding these estimated maximum AUCs (Botta et al., 2014; Liu et al., 2021; Sinoquet, 2018). Unfortunately, those studies used methods that are known to result in bias and misleading results that do not generalize to other data sets (Wray et al., 2013). Other studies with less biased methods, resulting in more generalizable models, have reported classification performances far below the estimated limits (Evans et al., 2009; Mittag et al., 2015; Pirooznia et al., 2012). This suggests that current models only capture a small portion of the known genetic variance. Further methodological improvements are likely necessary to capture more of the known genetic variance and achieve higher and more useful AUCs.

Improving the performance of genetic models is useful for a variety of reasons. Most directly, if the classification performance of these models were high enough, they could be useful clinically for screening purposes. One study found that in a suicide risk prediction task, a positive predictive value of only 0.8% for predicting suicide attempts was cost-effective from a health care sector perspective (Ross et al., 2021). However, the performance necessary to be helpful clinically differs between disorders and is dependent on the cost of implementing the model, interventions based on the model, results of interventions based on the model, and ethical considerations of taking potentially harmful actions on false positive predictions (Clayton, 2003; Shickle & Chadwick, 1994). Better classification models also improve the estimation of disorder risk, which would be useful in improving the cost and efficiency of clinical trials. Clinical trials of prevention strategies are often burdened with high costs to have the statistical power necessary to detect improvements. However, a study showed that using polygenic scores to enrich the study sample with people at higher risk for a disorder could increase the proportion of study participants that will experience the onset or progression of the disorder (Fahed et al., 2022). This "prognostic enrichment" improves power and reduces the number of participants necessary to detect a significant change. The strength of the enrichment is tied to how well risk can be estimated, so improvements in risk estimation could further decrease the number of participants necessary for clinical trials. Another benefit of improving classification performance is the potential for improved understanding of the disorder. Given the complexity of these disorders, there is much that we do not understand about the disorders. Interpretable and accurate models could provide insights into the underlying biology of these disorders and indicate areas in which more research should be conducted.

In this dissertation, I present a review paper analyzing the study features that lead to high and often biased AUCs in genomic machine learning literature followed by three primary research studies focused on testing the hypothesis that providing models with more genomic context can improve the models' performance in estimating risk and thereby correctly classifying cases and controls. Across the studies, I analyze different approaches to providing genomic context to answer these questions: (1) Does the inclusion of risk estimates from genetically correlated phenotypes improve risk modeling? (2) Does summation of risk across gene sets associated with a disorder improve risk modeling compared to a genome-wide summation? (3) Does directly providing a model with functional annotation information or more explicit directions on what information to use improve risk modeling? My dissertation includes the development of several novel methods to help answer these questions. I developed the methods using different disorders and data sets that were best suited for the research questions of each study. The results of this work suggest that genomic context can be useful in improving risk modeling for complex genetic disorders and provides multiple pathways to explore in advancing genetic classification models and interpret and validate results of machine learning models.

Genomic Machine Learning Meta-regression: Insights on Associations of Study Features with Reported Model Performance

Eric J Barnett[1], Daniel G Onete[2], Asif Salekin[3], Stephen V Faraone[1,4]

[1]Department of Neuroscience and Physiology, SUNY Upstate Medical University, Syracuse, New York, USA

[2]College of Medicine, MD Program, SUNY Upstate Medical University, Syracuse, New York, USA

[3]Syracuse University, Laboratory for Ubiquitous and Intelligent Sensing, Department of Electrical Engineering and Computer Science, Syracuse, New York USA

[4]Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, New York, USA

Abstract

Background: Many studies have been conducted with the goal of correctly predicting diagnostic status of a disorder using the combination of genomic data and machine learning. The methods of these studies often differ drastically. It is often hard to judge which components of a study led to better results and whether better reported results represent a true improvement or an uncorrected bias inflating performance.

Methods: In this systematic review, we extracted information about the methods used and other differentiating features in genomic machine learning models. We used the extracted features in mixed-effects linear regression models predicting model performance. We tested for univariate and multivariate associations as well as interactions between features.

Results: Of the models reviewed, 71% had some form of data leakage. In univariate models the number of hyperparameter optimizations reported, data leakage due to feature selection, and model type were significantly associated with an increase in reported model performance. In our multivariate model, the number of hyperparameter optimizations, data leakage due to feature selection, and modelling an autoimmune disorder were significantly associated with an increase in reported model performance.

Conclusions: Our results suggest that methods susceptible to data leakage are prevalent among genomic machine learning research, resulting in inflated reported performance. Best practice guidelines that promote the avoidance and recognition of data leakage may help the field avoid biased results.

Introduction

The genetic study of complex disorders has made great strides in the discovery of genome-wide significant genetic loci and substantial evidence for polygenicity (Wray et al., 2014). These discoveries have generated new hypotheses about the etiology of these disorders and have motivated machine learning (ML) efforts to separate cases and controls using genome-wide association study (GWAS) data. While results from early genomic ML research had been promising, the potential pitfalls of such studies have limited their interpretation (Whalen et al., 2022; Wray et al., 2013). Although best practices have previously been described, the methods, reporting, and overall study design for genomic ML studies vary so drastically that it is often difficult to compare and evaluate studies (Libbrecht & Noble, 2015). This between-study heterogeneity may contribute to distrust and underutilization of machine learning results.

To better appreciate the strengths and weaknesses of genomic ML research, one must understand the differences between ML analyses and traditional GWAS. GWAS seeks to determine loci that are statistically significantly different between cases of a disorder and healthy controls and to test for polygenicity (Tam et al., 2019). Since these studies examine hundreds of thousands to millions of loci, researchers apply stringent genome-wide significance thresholds (most commonly $p < 5 \times 10^{-8}$) to reduce reporting false positive results.

In the ML analyses we review here, the primary goal is accurately predicting whether subjects are cases or controls genotype data from each individual. Towards achieving this goal, relying only on loci that meet the threshold for genome-wide significance limits the learning capability of ML models. For example, a schizophrenia GWAS found that while 108 genome wide significant loci were able to explain 3.4% of the variation on the liability scale, including loci that met the nominal significance threshold (0.05) increased the variation explained to 7% (Psychiatric Genomics Consortium, 2014). This effect may be more pronounced in ML models which take advantage of interactions between loci to find patterns that are useful in differentiating cases and controls since more loci give the models more potential to find patterns.

Including additional loci in ML models has some drawbacks. One of the most important aspects of generalizable ML models is avoiding overfitting, which becomes more difficult as the number of loci increases (Ying, 2019). Overfitting occurs when a model learns patterns that are only present within the data used to train the model (See Figure 1). In ML algorithms, the training process learns model parameters that minimize the difference between the predicted and actual case/control labels. Researchers aim to build models that use real, generalizable differences between cases and controls to make each prediction, but in practice, models are free to use whatever patterns best minimize that difference. If a model memorizes the noise specific to only the training data, the model is less motivated to learn patterns that may be more generalizable if using those patterns is less successful than memorizing training data noise. For many ML models, constraints are added to reduce the model’s ability to overfit, but overfitting is rarely completely avoided (Ying, 2019). Each additional locus that a model has access to increases the probability of overfitting but also has the potential to add generalizable information the model can use to increase its ability to separate cases and controls.

Many ML researchers account for overfitting by testing the performance of their models on data that were not seen during training (Cawley & Talbot, 2010). One of two methods is typically used: cross-validation within the training set and validation with data not used at all in the training process. In k-fold cross-validation, researchers randomly split the data into some number of subsets, called folds, and then complete the association analysis and modeling that detect and use loci that are different between cases and controls to best separate the two classes using all but one of the folds. The model is trained using data from k-1 folds and its accuracy is tested in the single fold that was not included in training. This process is repeated until the model has used each fold as the withheld subset. Then the results across all iterations are averaged.

In the hold-out method, researchers randomly split the data into either two (training and test) or three (training, validation, and test) subsets. The training subset is used for the association analysis and modeling. If present, the validation subset is used to tune the model by setting the optimal hyperparameters, which are all the options and configurations that are not trained by the model itself, to best predict the validation subset. Then, the test subset is used to measure and report model performance. Unlike in cross-validation where all folds are the same size, the hold-out method typically uses 60-80% of the data in the training subset, while the remaining data is split evenly between the validation and test subsets. Cross-validation is often used when data is limited since in this method the model has a chance to train using each person in the study. The hold-out method, while not allowing the model to train on each person, is thought to be the more conservative approach and less likely to produce an overfit model.

The value of external testing can easily be lost through methods that leak information about the test subset/fold into the training of a model (Kaufman et al., 2012). When this occurs, the model can use that information to model the specific test data more accurately. Consequently, the test data no longer represent unseen data and no longer account for overfitting to the same degree. This results in a model that is biased in favor of the test data. This problem, called data leakage, is especially detrimental because it often goes unnoticed, leading researchers and their audiences to believe that their model performs exceptionally well even outside the training subset when instead they are observing model bias. The amount of data leakage caused through methodological issues can vary from slight leakages that may result in some overreporting of performance, to major leakages that cause the testing subset/fold to mimic the training performance with near-perfect prediction. One example of minor data leakage could be a cycle of checking the performance of a model in the test subset, then deciding to try more types of models using the same data subsets until the performance meets a threshold deemed worthy of reporting. In this situation, researchers indirectly leak information about the test data into their modeling, allowing them to select the model that performed best in the test data and reporting an inflated result (See Figure 1). An example of major data leakage is using the entire data set to select which of the many different features best separate cases and controls and only splitting the data into subsets during the actual modeling. In this situation, when the features were selected with the entire data set, information about the test data were directly leaked into the process and the features included in the model will perform well in the test data, but in unseen data will either be less predictive or entirely unpredictive if the features selected are specific to the data in the feature selection process.

The choices researchers make regarding which ML models to train, loci to include, and methods to use for measuring performance and optimization are critical decisions that will determine the outcome and validity of their study. But since few studies compare ML variations in the same external data sets, comparisons between studies are difficult even within the same disorder. This leads to a potential dilemma: Is a model with a higher reported performance better than a lower performing model or more overfit to their data?

Here, we report a systematic review that extracted information on model performance, disorder, training size, ML methods, optimization methods, performance measurement methods, and reporting on common issues from all genomic ML papers found in our search. We use this information in a mixed-effects linear regression model predicting model accuracy as measured by the area under the receiver operating characteristic curve (AUC). Within each data set and study involving machine learning models, many factors, such as training size, model type, and disorder, likely have a real, generalizable impact on reported model performance. When looking at a group of studies, those factors may be harder to identify due to the variable effects of data leakage among the studies in the group. We hypothesized that data leakage would be associated with a significant increase in reported AUC. We sought to test this hypothesis and to identify other study features that lead to increased reported AUC. The purpose of our study is not to scrutinize any specific study or model, but to instead look at the current work in the field as a whole to understand what can be strengthened or avoided in future work to improve results.

Methods

Search strategy and selection

To identify studies that used genotype data as input for machine learning models to predict any disorder, we searched PubMed using the key-words '(GWAS OR genotype{ti} OR SNP OR "risk score") AND (classif* OR predict*) AND ("machine learning" OR "data mining"{ti} OR "neural network*" OR "random forest" OR "support vector machine" OR "deep learning")'. The search produced 1435 studies (up to July 6, 2022).

We excluded studies that were not using classifiers to predict a human disorder and studies that did not report results for genotype only models. We excluded studies that did not report testing performance outside the data used to train the model because these models are likely overfit to the data given the ability of machine learning models to learn the random noise within the training data. Since models that use all data on training can achieve near perfect prediction on those data regardless of other methodological features inclusion of these studies could mask other important features, especially when model complexity is high as is often the case in genomic models (Whalen et al., 2022). We also excluded studies that only used variants identified in previous studies, since these studies are generally focused on a small number of specific variants that have been thoroughly studied and validated and therefore do not face the data leakage due to feature selection issue addressed in this paper. We excluded studies that did not report AUC as a performance metric since combining different performance metrics in our analysis would limit interpretation and potentially bias results. After exclusion criteria, 56 studies remained(Abdulaimma et al., 2018; Abraham et al., 2013; Almlöf et al., 2017; An et al., 2017; Antonucci et al., 2020; Badré et al., 2021; Botta et al., 2014; Cánovas et al., 2020; J. Chen et al., 2018; Chen et al., 2021; Chuang & Kuo, 2017; Cope et al., 2021; Evans et al., 2009; Fergus et al., 2020; Garza-Hernandez et al., 2022; Gaudillo et al., 2019; Guo et al., 2016; Hu et al., 2021; Jo et al., 2022; Kang et al., 2011; Kinreich et al., 2021; Kooperberg et al., 2010; Krautenbacher et al., 2021; Kwon et al., 2022; Lauber et al., 2022; Lee et al., 2020; Li et al., 2018; Li et al., 2021; Liu et al., 2021; Mittag et al., 2012; Mittag et al., 2015; Muneeb & Henschel, 2021; Nguyen et al., 2015; Osipowicz et al., 2021; Pal et al., 2017; G. Paré et al., 2017; Pirooznia et al., 2012; Romagnoni et al., 2019; Romero-Rosales et al., 2020; Sinoquet, 2018; Skafidas et al., 2014; Sun et al., 2007; Thomas et al., 2020; Vivian-Griffiths et al., 2019; Wang & Avillach, 2021; Wang et al., 2018; Wang et al., 2016; Wang et al., 2019; Wei et al., 2009; Wei et al., 2013; Widen et al., 2021; Wu et al., 2012; Yan et al., 2021; Zhang et al., 2021; Zhao et al., 2016). These studies provided accuracy statistics for 100 models because some studies modeled multiple disorders. Figure 2 shows the article selection procedure in a PRISMA diagram.

Data extraction

We extracted the following data from each included study: disorder predicted, number of subjects used for training and testing, number of participants the model was trained on, highest AUC, method for testing/reporting model AUC, reporting on optimization, number of hyperparameters optimized, reporting on imputation, reporting on quality control procedures, and type of machine learning method with the highest AUC. We split studies that modeled multiple disorders such that each row of extracted data represented a single disorder from a single study.

Regression analysis

We fit linear mixed-effects models to test the individual and combined contribution of study variables to AUC using STATA17 (StataCorp, 2019). We used this type of model to test the statistical significance of the included study features while accounting for the between-study heterogeneity present in this analysis of many different methods and disorders (Harrer et al., 2021). The standard error of each of our models was estimated using a clustered sandwich estimator clustering on PMID of the included studies, which adjusts standard errors for the lack of statistical independence of results within studies. We fit univariate models with AUC as the dependent variable and used the following as independent variables: data leakage through feature selection, data leakage through hyperparameter optimization, disorder type, reporting of optimization, reporting of quality control procedures, reporting imputation usage, number of hyperparameter optimizations reported, disorder heritability, model type, testing method, and size of training dataset.

“Data leakage due to feature selection” was a binary feature scored as 1 if the data used to test model performance were used to select which features would be included in the model. It was scored zero otherwise. “Data leakage due to optimization” was a binary feature scored 1 if the data used to test model performance was used at all in optimizing the model. It was scored zero otherwise. “Optimization reported” was a binary variable scored 1 if the authors reported any optimization of model hyperparameters. It was scored zero otherwise. “Quality control (QC) reported" was a binary variable scored 1 if authors reported the quality control procedures for their genotype data and zero otherwise. "Imputation reported" was a binary variable scored 1 if authors reported using imputed genotypes in their analysis and zero otherwise. "Number of hyperparameters optimizations" indicated the number of hyperparameters the authors reporting optimizing in the model with the highest AUC. "Disorder heritability" for each disorder included in these studies was gathered from heritability studies of those disorders based on twins or families(Smoller & Finn, 2003) (Almgren et al., 2011; Arkema et al., 2019; Bulik et al., 2006; Cánovas et al., 2020; Christophersen et al., 2009; Elks et al., 2012; Faraone & Larsson, 2018b; Gatz et al., 2006; Gordon et al., 2015; Hamza & Payami, 2010; Hilker et al., 2018; Hottenga et al., 2005; Joergensen et al., 2016; Kuja-Halkola et al., 2016; Kyvik et al., 1995; Lønnberg et al., 2013; MacGregor et al., 2000; Möller et al., 2016; Sandin et al., 2017; Seddon et al., 2005; Svensson et al., 2009; Ullemar et al., 2016; Verhulst et al., 2015; Willemsen et al., 2008; Zdravkovic et al., 2002). Training size was the total number of cases and controls used to train the model. "Model type" was a binary variable scored 1 if the model with the highest AUC was non-linear and 0 if it was linear. "Testing method" was a binary variable scored 1 if cross-validation was used to test model performance and 0 if a hold-out test subset was used to test model performance. Scatter plots of numerical variables and box plots of binary variables are shown in Figures 3 - 6.

We corrected univariate p-values for multiple testing using Bonferroni correction based on the number of features tested across all univariate models. Variables were added to a multivariate model sequentially, ordered based on the p-value of the variables univariate model, and kept if the variable remained significant.

Among the possible interaction terms for the variables used in our models, we identified 6 interactions that could be reasonably hypothesized to impact prediction performance. These 6 terms were: data leakage through feature selection + training size, number of hyperparameter optimizations + training size, testing method + training size, data leakage through hyperparameter optimization + training size, number of hyperparameter optimizations + data leakage through hyperparameter optimization, and number of hyperparameter optimizations + testing method. We tested the potential interaction terms as individual additions to the multivariate model, added the interactions sequentially ordered by the p-value of the initial interaction models, and kept the interactions in the final model if they remained significant after Bonferroni correction based on the number of interactions tested. We also performed sensitivity analyses removing studies with extreme values from models with the number of hyperparameter optimizations and training size variables.

Sensitivity Analyses

For the purpose of sensitivity analyses, we kept all studies that met the condition of:

1.5 => |(value – upper quartile) / IQR|

If the value was above the upper quartile, and:

1.5 => |(value – lower quartile) / IQR|

If the value was below the lower quartile

For the number of hyperparameter optimizations variable, 15 models had extreme values and were excluded. The univariate model modelling AUC with number of hyperparameter optimizations of the remaining models had a coefficient of 0.05, standard error of 0.04, z of 1.39, and p-value of 0.17 before any multiple testing correction. For the training size variable, 11 models had extreme values and were excluded. The univariate model of the remaining models had a coefficient of 5.79 x 10^{-6}, standard error of 1.4 x 10^{-5}, z of 0.41, and p-value of 0.68. For the disorder heritability variable, 2 models had extreme values and were excluded. The resulting univariate model had a coefficient of 9.6 x 10^{-4}, standard error of 0.1, z of 0.01, and p-value of 0.99.

Results

Among the 56 studies and 100 models included, 31 different disorders were modeled. The average AUC among all models was 0.75. Figure 6 shows a histogram of the frequency distribution of reported AUCs. Thirty-two percent (N = 32) of the models modeled eight autoimmune disorders and had a mean AUC of 0.80 ± 0.12. Sixteen percent (N = 16) of the models modeled six psychiatric disorders and had a mean AUC of 0.73 ± 0.15. The remaining 52 models modeled 16 disorders that did not fit into either of these groups and had a mean AUC of 0.73 ± 0.17. The included studies and all data used in our models can be found in Table 1.

Of the models studied, 71% had at least one form of data leakage. Forty-six percent had methods with some degree of data leakage due to feature selection. Sixty-two percent had methods with some degree of data leakage due to optimization. Imputation and quality control procedures were reported in 46% and 91% of the models examined, respectively. Fifty-two percent of the models were non-linear while the remaining 48% were linear models. Cross-validation was used for testing in 65% of the models.

In univariate models the number of hyperparameter optimizations reported, data leakage due to feature selection, and model type were significantly associated with an increase in AUC in the test data after correcting for multiple testing (Table 2). Modelling an autoimmune disorder was nominally associated with an increase in AUC but did not remain significant after correcting for multiple testing. Reporting on optimization, reporting on quality control procedures, reporting imputation usage, training size, testing method, disorder heritability, and data leakage due to optimization were not significant.

When interactions were entered one at a time, the interaction between data leakage due to feature selection and training size, the interaction between number of hyperparameter optimizations and training size, and the interaction between data leakage due to optimization and training size were nominally associated with AUC but did not remain significant after correcting for multiple testing (Table 3). When sequentially adding the interactions, none of the interactions were significant after Bonferroni correction.

Our final multivariate model included the number of hyperparameter optimizations, data leakage due to feature selection, and autoimmune disorder modelling (Table 4). In this model, number of hyperparameter optimizations, data leakage due to feature selection, and modelling an autoimmune disorder were significantly associated with an increase in AUC after correcting for multiple testing. Based on the variable coefficients, our model estimated that the presence of data leakage due to feature selection leads to an overestimation of 0.15 in AUC. Models that classified autoimmune disorders saw an estimated AUC increase of 0.10 compared to all other disorders. The AUC predicted by the multivariate model had an R-squared of 0.53 in a regression model with the actual AUC values of the models examined. A correlation matrix of all variables used in this analysis can be found in Table 5.

Discussion

Our analysis to determine which features of genomic ML studies lead to significant increases in AUC found evidence that, in many studies, methodological issues result in overreporting of model performance. Out of the studies investigated here, 44% had some form of data leakage due to feature selection. The most common form of data leakage due to feature selection was using the entire dataset for the GWAS and later splitting the dataset into multiple subsets or folds. Since the test data were used to determine the loci that best separate cases from controls before ML modeling, even if researchers split the data into separate subsets during ML, information from the test data has already leaked into the process, which would lead to overestimates of accuracy. The coefficient of our multivariate model for this feature indicates that, on average, this form of data leakage increases the AUC by 0.15 after adjusting for other factors. For some applications, an increase in AUC of that magnitude would be enough to move models into a performance range that would falsely suggest clinical utility. Although this is the first demonstration of this problem for a collection of genomic ML studies, others have previously warned of the potential for this type of issue to interfere with results.

Out of the included studies, 56% had some form of data leakage due to optimization. In this form of data leakage, the test data are used multiple times as the researcher is optimizing the hyperparameters of a model to determine which set of those hyperparameters results in the best performance in the test data. A study using random training data showed that improper use of cross-validation, one example of data leakage due to hyperparameter optimization, leads to biased performance in the compromised dataset (Varma & Simon, 2006). This type of data leakage allows the model to overfit the data by giving it many different opportunities to find a model that best separates those specific test data. It is typically used for smaller data sets where it is not feasible to create an independent test set. Instead, a more generalizable practice would be to either use a different subset or cross-validation within the training data to optimize hyperparameters and apply those hyperparameters to a single model in the test data (Cawley & Talbot, 2010).

Although the main effect of data leakage due to optimization was not significant, its interaction with training size was nominally significant. Learning curve analyses have demonstrated that increasing training size results in increased performance in machine learning since the model has more opportunities to learn patterns and features within the data (Banko & Brill, 2001; Halevy et al., 2009). In learning curve analyses, a machine learning model is trained with an increasing number of training examples and prediction performance is measured at each training size (Perlich, 2010). Generally, models will gradually improve with training size until a plateau is reached. The amount of training data necessary to reach the plateau and the prediction performance at the plateau depend on the complexity of the model and the prediction task. Increased training size also makes it more difficult for ML models to find non-generalizable patterns within the noise of large datasets and the correlated increase in test size makes it more difficult to bias towards the test data when data leakage is present (Babyak, 2004). Thus, we thought that the interaction of data leakage due to hyperparameter optimization and training size would be important and included it in our analysis. This finding suggests that optimizing directly on test data should be avoided and verifies that increasing the number of subjects included in machine learning studies could further improve results. Our model also warns that if data leakage due to optimization occurs, studies with this issue may not benefit as much as expected from increasing training size.

The number of hyperparameter optimizations reported by researchers was associated with increased AUC in our univariate and multivariate models. The model type variable was associated with increased AUC in a univariate model but was not associated in multivariate models, most likely due to its correlation with number of hyperparameter optimizations. This correlation was expected since non-linear models have more hyperparameters that can be optimized compared to linear models. Unlike the data leakage features, which are clearly a source of bias, the number of hyperparameters optimized and using non-linear models could lead to genuine improvement, bias, or a balance of the two. The "no free lunch" theorem states that all optimization algorithms have the same performance when averaged over all possible tasks (Wolpert & Macready, 1997). Applied in the context of machine learning hyperparameters, this means that it is impossible to know which type of model or hyperparameters within the model will perform best on a given dataset without prior knowledge. This maxim has led to the best practice of optimizing the hyperparameters of a model by either using a grid search over different values of those hyperparameters or using hyperparameter optimization algorithms and suggests that at least some of the improvement due to the number of hyperparameters optimized is due to genuine prediction improvement.

Conversely, the variable representing the number of hyperparameters optimized may also be a proxy of cryptic data leakage due to optimization. If data leakage due to hyperparameter optimization is present within the study, more optimizations will likely result in a model that is more biased in favor of the test data because the model has more opportunities to use the information it obtains about the test data through data leakage to find the optimal configuration for those data. This is similar to the data leakage illustrated in Figure 1, but instead of being only biased towards the unreported validation subset, the model would also be biased towards the test data. It is possible that even when optimization is reportedly handled appropriately, data leakage due to hyperparameter optimization may still be present. However, these findings should not be interpreted as evidence that hyperparameter optimization should be avoided to reduce data leakage. Hyperparameter optimization is an important component of improving machine learning models and can be done in ways that avoid data leakage. Instead, our results point towards the necessity of establishing and following standard practices that are more careful in avoiding data leakage in the genetics machine learning field.

The type of disorder studied was significantly associated with AUC in our final multivariate model after Bonferroni correction. After adjusting for other significant predictors of AUC, models of autoimmune disorders had significantly higher AUCs compared to both psychiatric disorders and all other disorders. We were surprised by the lack of association between disorder heritability and AUC since disorders with a larger genetic component should theoretically be more predictable in genetic models, but the significance of a disorder type may help explain disorder heritability's lack of significance in our multivariate models. The hypothesis that disorder heritability may be associated with an increase in AUC assumes that the difficulty of extracting and modeling the genetic components is equivalent in all disorders. If the accessibility and ease of modelling genetic risk differ between disorders, we would no longer expect heritability to be associated with AUC. While this conclusion could previously be inferred by comparing the effect sizes of individual loci in GWASs of different disorders, our analysis provides further evidence that the genetic information from some groups of disorders may collectively differ in accessibility and ease of modelling. Differences in accessibility and ease of modelling could be due to differences in genetic architecture or differences in measurement (e.g., differences in misclassification rates). For example, the risk for Celiac disease is largely within a single region, whereas the genetic risk for most psychiatric disorders is spread thinly across the genome (Abraham et al., 2013; Sullivan et al., 2012). Another possibility is that the autoimmune models were consistently different from the other models in some untested way. However, many of the autoimmune models were part of papers that also studied non-autoimmune disorders, which makes this possibility less likely.

Our study has several limitations that may have limited our ability to detect associations between the included features and AUC. Like all meta-analytic regressions, we were only able to model study characteristics as they were reported within each study. In some studies, information about the test data may have been leaked in inadvertent and unreported ways that we could not examine. The reported AUCs from some models were better than the theoretical maximum AUCs for the disorders and nearly perfect, which could reflect label leakage (Wray et al., 2010). Label leakage is a more direct form of data leakage where the model is given access to class labels, which the model can then use to make a prediction of those same labels. Label leakage is almost always accidental and unreported, so we were unable to study it. Figure 7 shows the peak caused by the near-perfect models in the frequency distribution of reported AUCs. If any confounding study characteristics are present, it is possible that the real effect of included variables could differ from our results. For example, disorder heritability and training size, which we would expect to influence model performance but were not significant in our analysis, indicate the limitations of representing study characteristics in our analysis. Disorder heritability and training size may have had their true effect masked by unreported data leakage, inappropriate method choices within studies, or some other confounding factor. For example, if an inappropriate modeling strategy was selected, leading to a low AUC, that would make it more difficult to detect effects. Our clustered sandwich estimator usage limits but does not eliminate the possibility of an unreported study characteristic in a single study confounding our results. Alternatively, the variables we included could be acting in part as proxies for unreported or untested study characteristics. However, any potential proxy relationships may also have importance for understanding current practices and informing future guidelines.

The studies used in this analysis are heterogeneous due, in part, to the inclusion of any human disorder. These disorders likely have differing genetic complexities and optimal prediction performances, which may have limited our ability to detect differences in AUC based on the features of the study but also highlights the strength of the features that were detected despite this heterogeneity. We were unable to use specific disorders as a feature in our analysis due to the limited number of studies with each disorder. Excluding papers that did not use AUC as a performance metric limited the number of studies we could include in this analysis. Some relevant studies may have been excluded from our analysis because they did not match our search query. Our study was also limited to using what was reported in these studies, which may have limited our ability to fully assess data leakage and optimization. The data leakage features we used in this study were binary, but in practice different methods have different types and severities of data leakage which would be difficult to accurately represent in this type of analysis.

Our analysis of genomic machine learning studies has implications for defining best practices for genomic ML studies. Although such studies may eventually lead to clinically actionable risk calculators, publications that overestimate results will not be replicated, which could lead the field to prematurely abandon ML. We found data leakage due to both feature selection and optimization to be prevalent. The former leads to increased AUCs that likely overestimate the models' true performance outside the training data while the latter limits or hides the effect of increasing training size. Perhaps our most important finding is that 71% of studies showed some form of data leakage, which suggests a need to promote better machine learning practices in the field as is being done by the MLPsych consortium (Quinn et al., 2022). We also found evidence that suggests data leakage due to optimization may occur even when studies report methods that should minimize the effects of data leakage. If genomic machine learning methods and results are to be improved and trusted, researchers must recognize and avoid these issues. Thorough best practice guidelines that promote the avoidance of data leakage and other common issues will be critical as the field grows and advances.

Table 1: Features of the Analyzed Genetic Machine Learning Classifier Models

pmid	AUC	psych	autoimmune	training size	disorder h2	model type	dl opt	optimization	number of opt	dl fs	testing method	imputation reported	qc reported
18133515	0.5784	0	0	4314	0.69	1	1	1	1	1	0	0	1
18466563	0.94	0	1	825	0.6	1	1	1	1	1	1	1	1
19553258	0.668	1	0	3348	0.725	0	1	1	1	0	1	1	1
19553258	0.6	0	0	3406	0.48	0	1	1	1	0	1	1	1
19553258	0.627	0	1	3228	0.75	0	1	1	1	0	1	1	1
19553258	0.61	0	0	3432	0.54	0	1	1	1	0	1	1	1
19553258	0.666	0	1	3340	0.6	0	1	1	1	0	1	1	1
19553258	0.749	0	1	3443	0.72	0	1	1	1	0	1	1	1
19553258	0.601	0	0	3404	0.69	0	1	1	1	0	1	1	1
19816555	0.84	0	1	3443	0.72	1	0	1	6	0	0	1	1
20842684	0.637	0	1	2808	0.75	0	0	1	1	0	0	1	1
21427131	0.72	0	1	2223	0.75	0	1	1	1	0	0	1	1
22081063	0.576	1	0	3625	0.725	0	1	1	1	0	0	1	1
22777693	0.61	1	0	3502	0.725	1	1	1	2	1	1	1	1
22777693	0.56	0	0	13243	0.41	1	1	1	2	1	1	1	1
22777693	0.88	0	1	3504	0.72	1	1	1	2	1	1	1	1
22777693	0.62	0	0	3503	0.69	1	1	1	2	1	1	1	1
22965006	0.749	0	0	855	0.83	0	0	0	0	1	0	0	0
22987127	0.9271	0	0	541	0.41	1	1	1	2	1	1	0	1
23203348	0.734	1	0	4806	0.725	0	1	1	1	1	1	0	1
23203348	0.877	0	1	6785	0.75	1	0	1	1	0	0	0	1
23203348	0.683	0	0	4864	0.48	0	1	1	1	1	1	0	1
23203348	0.768	0	1	4686	0.75	0	1	1	1	1	1	0	1
23203348	0.7	0	0	4890	0.54	0	1	1	1	1	1	0	1
23203348	0.766	0	1	4798	0.6	0	1	1	1	1	1	0	1
23203348	0.895	0	1	4901	0.72	0	1	1	1	1	1	0	1
23203348	0.713	0	0	4862	0.69	0	1	1	1	1	1	0	1
23731541	0.864	0	1	15173.6667	0.75	0	0	1	1	0	0	1	1
23731541	0.826	0	1	13584	0.67	0	0	1	1	0	0	1	1
24695491	0.959	1	0	5000	0.725	1	1	1	2	1	1	1	1
24695491	0.999	0	0	5000	0.48	1	1	1	2	1	1	1	1
24695491	0.955	0	1	5000	0.75	1	1	1	2	1	1	1	1
24695491	0.969	0	0	5000	0.54	1	1	1	2	1	1	1	1
24695491	0.996	0	1	5000	0.6	1	1	1	2	1	1	1	1
24695491	0.94	0	1	5000	0.72	1	1	1	2	1	1	1	1
24695491	0.979	0	0	5000	0.69	1	1	1	2	1	1	1	1
25708662	0.975	0	0	364	0.58	1	1	1	3	1	1	0	0
25708662	0.959	0	0	541	0.41	1	1	1	3	1	1	0	0
26285210	0.5834	1	0	3500	0.725	1	1	1	3	0	1	0	1
26285210	0.5843	0	0	3500	0.48	1	1	1	2	0	1	0	1
26285210	0.6178	0	1	3500	0.75	1	1	1	2	0	1	0	1
26285210	0.5617	0	0	3500	0.54	0	1	1	1	0	1	0	1
26285210	0.7276	0	1	3500	0.6	1	1	1	2	0	1	0	1
26285210	0.8682	0	1	3500	0.72	1	1	1	2	0	1	0	1
26285210	0.5861	0	0	3500	0.69	1	1	1	2	0	1	0	1
26792494	0.693	1	0	4402	0.56	0	0	1	1	0	0	1	1
27080919	0.89	0	1	705	0.72	0	0	1	1	0	0	0	1
27444562	0.7239	0	1	1590	0.68	1	0	1	.	0	0	0	1
28045094	0.702	1	0	2035	0.725	0	0	1	1	0	0	1	1
28358032	0.855	0	0	737	0.58	0	1	1	2	1	1	1	1
28512778	0.72	0	1	5000	0.75	1	0	0	.	0	0	1	1
28740209	0.78	0	1	3871	0.61	1	1	1	2	0	1	0	1
28979001	0.602	0	0	2000	0.69	1	1	1	4	0	0	1	1
29587628	0.958	1	0	4806	0.725	1	1	1	13	1	1	0	1
29587628	0.968	0	0	4864	0.48	1	1	1	13	1	1	0	1
29587628	0.952	0	1	4686	0.75	1	1	1	13	1	1	0	1
29587628	0.94	0	0	4890	0.54	1	1	1	13	1	1	0	1
29587628	0.962	0	1	4798	0.6	1	1	1	13	1	1	0	1
29587628	0.957	0	1	4901	0.72	1	1	1	13	1	1	0	1
29587628	0.961	0	0	4862	0.69	1	1	1	13	1	1	0	1
30183645	0.9998	0	0	1221.6	0.25	1	1	1	4	1	0	0	1
30193110	0.69	0	0	401	0.77	0	1	1	1	1	1	0	1
30204480	0.7	0	0	139	0.74	1	1	1	2	0	1	0	0
30276764	0.905	1	0	10858.5	0.79	0	1	1	9	1	0	1	0
30516002	0.697	1	0	11853	0.79	0	0	0	0	0	1	1	1
31316157	0.802	0	1	34634	0.75	1	0	1	.	1	0	1	1
31564248	0.69	0	1	111	0.75	1	0	0	0	0	0	0	1
31595034	0.75	1	0	140	0.49	0	1	1	.	0	1	1	1
31800601	0.64	0	0	128	0.75	1	1	1	3	1	1	0	1
31948640	0.48	1	0	440	0.79	0	0	1	2	0	0	0	0
32106268	0.62	0	0	137	0.31	1	0	1	1	0	0	1	1
32324812	0.844	0	0	1382	0.58	0	0	1	1	1	0	1	1
32758450	0.654	0	0	72791	0.16	0	1	1	2	0	0	1	1
32887683	0.671	0	0	7505	0.54	0	0	1	1	0	0	1	1

32997854	0.54	0	1	1410	0.82	0	0	1	1	0	0	1	1
33009504	0.674	0	0	39288.8	0.31	1	0	1	3	0	0	0	0
33714937	0.955	0	0	3144	0.83	1	0	1	12	1	0	0	1
33866309	0.703	0	0	7240	0.58	0	0	0	0	0	1	0	0
33874881	0.97	0	0	69	0.69	1	1	1	6	1	0	0	1
34003914	0.78	0	0	27215	0.46	1	0	1	2	0	0	1	1
34109382	0.957	1	0	1486	0.74	1	0	1	.	1	0	0	1
34209487	0.493	0	0	2791	0.48	0	0	1	1	0	1	0	1
34209487	0.505	0	0	2791	.	0	0	1	1	0	1	0	1
34209487	0.65	0	1	2910	0.72	0	0	1	1	0	1	0	1
34209487	0.637	0	0	2910	0.69	0	0	1	1	0	1	0	1
34209487	0.558	0	0	2791	0.54	0	0	1	1	0	1	0	1
34209487	0.527	0	0	2791	.	0	0	1	1	0	1	0	1
34209487	0.551	0	0	2910	.	0	0	1	1	0	1	0	1
34209487	0.574	0	0	2791	0.74	0	0	1	1	0	1	0	1
34327330	0.57	0	0	1519	0.58	1	0	0	0	0	1	0	0
34336000	0.981	0	0	1316	0.58	1	1	1	.	0	0	0	1
34430925	0.738	1	0	12065	0.79	1	0	1	7	1	0	1	1
34745218	0.85	0	0	367	0.41	0	0	1	1	1	0	0	1
34749286	0.598	1	0	1010	0.79	0	0	1	3	0	1	1	1
34892435	0.999	0	0	988	0.58	1	1	1	.	0	0	0	1
35086918	0.82	0	0	4099	0.62	1	1	1	4	0	0	0	1
35183061	0.82	0	0	981	0.58	1	1	1	.	1	1	1	1
35239643	0.609	0	0	3688	0.69	0	0	1	1	1	1	1	1
35239643	0.507	0	0	3951	0.48	0	0	1	1	1	1	1	1
35306380	0.667	0	1	8421	0.75	0	1	1	1	0	1	1	1

Table 2: Univariate mixed-effects linear model results

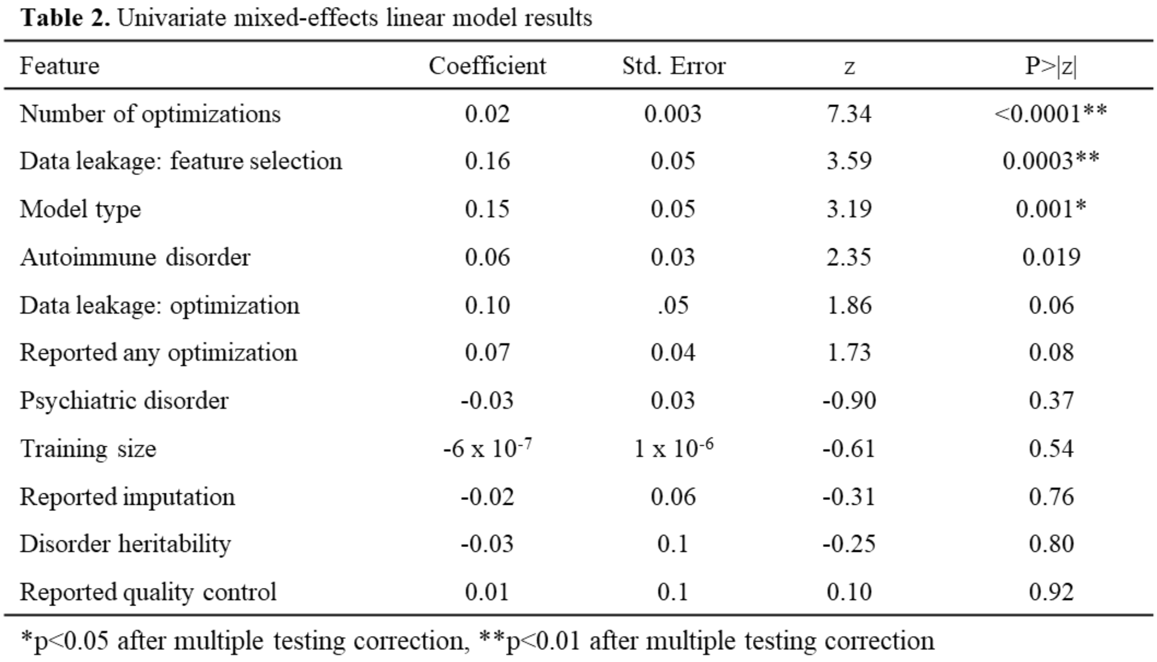


Table 2. Univariate mixed-effects linear model results

Feature	Coefficient	Std. Error	z	P>|z|
Number of optimizations	0.02	0.003	7.34	<0.0001**
Data leakage: feature selection	0.16	0.05	3.59	0.0003**
Model type	0.15	0.05	3.19	0.001*
Autoimmune disorder	0.06	0.03	2.35	0.019
Data leakage: optimization	0.10	.05	1.86	0.06
Reported any optimization	0.07	0.04	1.73	0.08
Psychiatric disorder	-0.03	0.03	-0.90	0.37
Training size	-6 x 10^{-7}	1 x 10^{-6}	-0.61	0.54
Reported imputation	-0.02	0.06	-0.31	0.76
Disorder heritability	-0.03	0.1	-0.25	0.80
Reported quality control	0.01	0.1	0.10	0.92

*p<0.05 after multiple testing correction, **p<0.01 after multiple testing correction

Table 3: Multivariate mixed-effects linear model interaction results

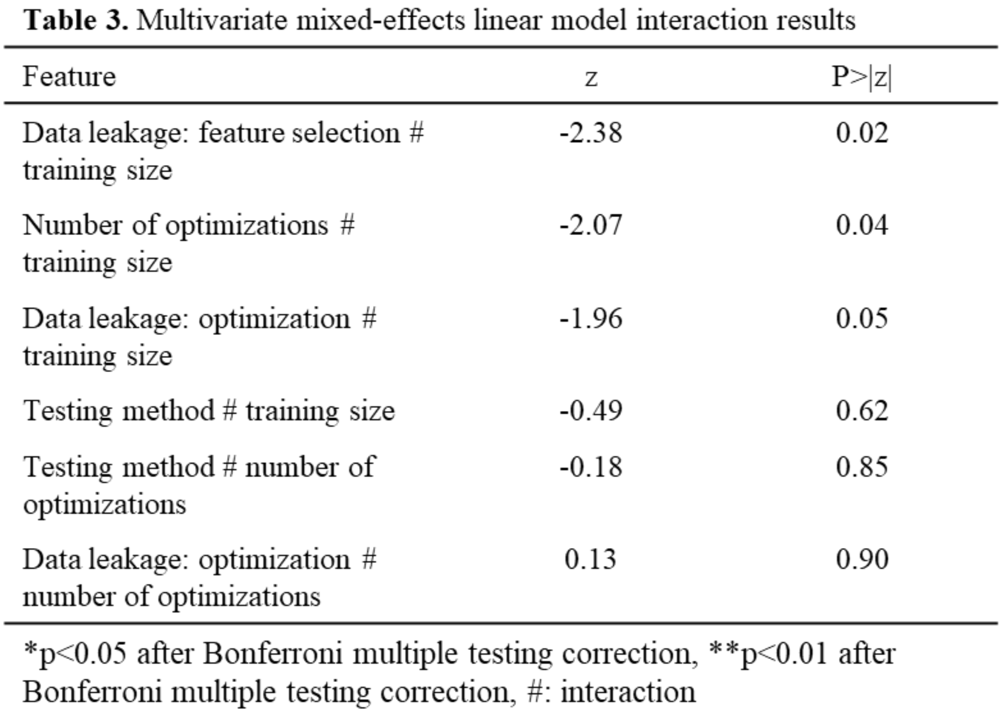


Table 3. Multivariate mixed-effects linear model interaction results

Feature	z	P>|z|
Data leakage: feature selection # training size	-2.38	0.02
Number of optimizations # training size	-2.07	0.04
Data leakage: optimization # training size	-1.96	0.05
Testing method # training size	-0.49	0.62
Testing method # number of optimizations	-0.18	0.85
Data leakage: optimization # number of optimizations	0.13	0.90

*p<0.05 after Bonferroni multiple testing correction, **p<0.01 after Bonferroni multiple testing correction, #: interaction

Table 4: Multivariate mixed-effects linear model results

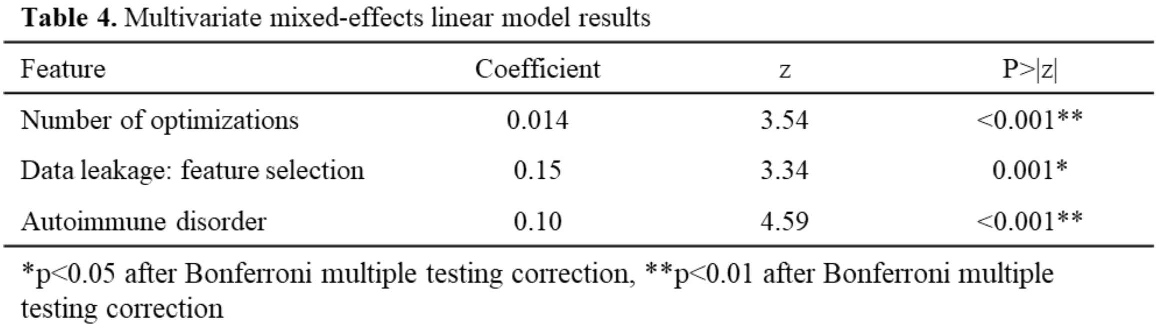
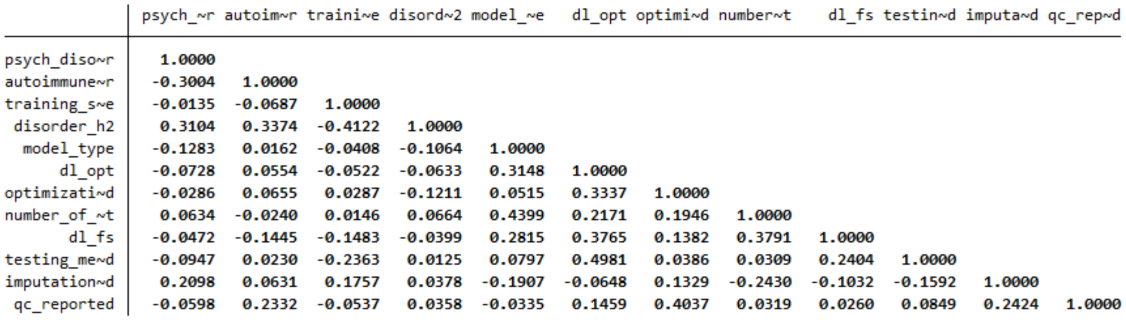


Table 4. Multivariate mixed-effects linear model results

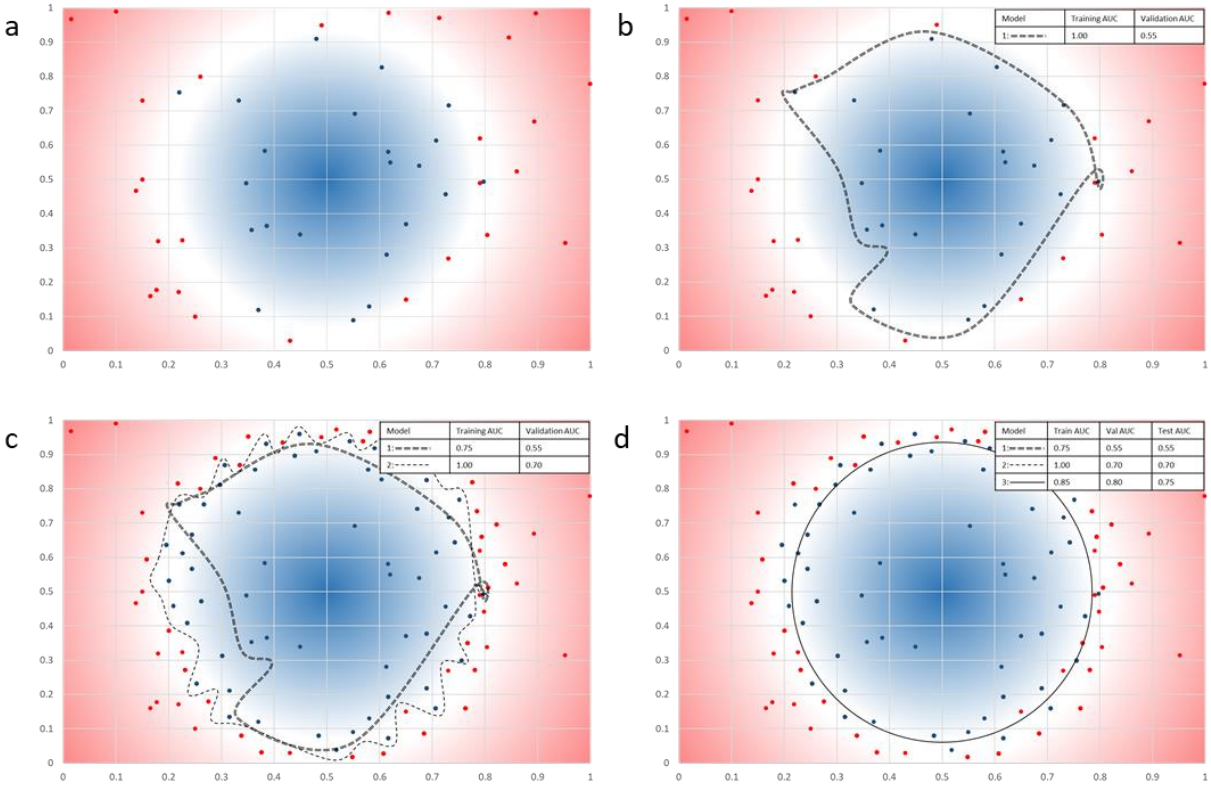
Feature	Coefficient	z	P>|z|
Number of optimizations	0.014	3.54	<0.001**
Data leakage: feature selection	0.15	3.34	0.001*
Autoimmune disorder	0.10	4.59	<0.001**

*p<0.05 after Bonferroni multiple testing correction, **p<0.01 after Bonferroni multiple testing correction

Table 5: Correlation matrix of variables used in the analysis.



```
             | psych_~r autoim~r traini~e disord~2 model_~e   dl_opt optimi~d number~t    dl_fs testin~d imputa~d qc_rep~d
-------------+------------------------------------------------------------------------------------------------------------
psych_diso~r |   1.0000
autoimmune~r |  -0.3004   1.0000
training_s~e |  -0.0135  -0.0687   1.0000
 disorder_h2 |   0.3104   0.3374  -0.4122   1.0000
  model_type |  -0.1283   0.0162  -0.0408  -0.1064   1.0000
      dl_opt |  -0.0728   0.0554  -0.0522  -0.0633   0.3148   1.0000
optimizati~d |  -0.0286   0.0655   0.0287  -0.1211   0.0515   0.3337   1.0000
number_of_~t |   0.0634  -0.0240   0.0146   0.0664   0.4399   0.2171   0.1946   1.0000
       dl_fs |  -0.0472  -0.1445  -0.1483  -0.0399   0.2815   0.3765   0.1382   0.3791   1.0000
testing_me~d |  -0.0947   0.0230  -0.2363   0.0125   0.0797   0.4981   0.0386   0.0309   0.2404   1.0000
imputation~d |   0.2098   0.0631   0.1757   0.0378  -0.1907  -0.0648   0.1329  -0.2430  -0.1032  -0.1592   1.0000
 qc_reported |  -0.0598   0.2332  -0.0537   0.0358  -0.0335   0.1459   0.4037   0.0319   0.0260   0.0849   0.2424   1.0000
```



Model	Training AUC	Validation AUC
1: — — — —	1.00	0.55

Model	Training AUC	Validation AUC
1: — — — —	0.75	0.55
2: - - - - - -	1.00	0.70

Model	Train AUC	Val AUC	Test AUC
1: — — — —	0.75	0.55	0.55
2: - - - - - -	1.00	0.70	0.70
3: ———	0.85	0.80	0.75

Figure 1: Overfitting, sample size, and data leakage

Plot Overview: In this illustrative example, we have condensed most of the known genetic variance for a hypothetical disorder into 2 continuous features scaled from 0 to 1. In these graphs the axes represent those 2 continuous features, and each point represents a person in the training subset based on their values for each feature. People who are cases of the hypothetical disorder are shown in red while people who are controls are shown in blue. The true population distribution of cases and controls, which would be unknown to us when modeling since we can only sample from the population, is represented by the background, with the intensity of the gradient representing how likely a person is to be a case (red) or control (blue) given the 2 features. The white space represents a region in which the likelihood of being a case or control is similar and perfect separation of cases and controls is impossible with only genetic data. Our task is to build a model that predicts whether each person is a case or control. To do this task in this example, we will build models that enclose all predicted controls. The optimal solution in this case would be a circle splitting the white region, but since our data are not a perfect representation of the population distribution, the models we make will inevitably be less optimal. We will discuss several factors that commonly impact how similar a model can be to the optimal model.

Overfitting: Many models are easily capable of achieving perfect prediction on the data that are used to build or train the model. However, that performance may be specific to the training data, as illustrated by the overfit Model 1. To get a more generalizable measure of model performance, we should test performance in data that are not used to train the model. If we were to test Model 1 in a different sample of unseen data (validation subset) we may find that it is not nearly as predictive in that subset since the model was trained using some patterns that generalize to the population and some patterns in the training data that do not match the population distribution.

Advantages of Increasing Training Size: If we increased the training size (only case/control boundary examples illustrated) it would become clear that Model 1 was overfit to the original data and the model does not generalize to the population. At this point it may still be possible to get perfect prediction by overfitting the model to the data with more complex models, as illustrated by Model 2, but overfitting becomes increasingly difficult as more data are added to the training subset.

Bias due to Data Leakage: To find the model with the best prediction in unseen data, we can use a validation subset (not shown) to optimize which model type we use and which hyperparameters or options to use within a model. During this process, we indirectly learn information about the validation subset as we find the model that best predicts that specific set of unseen data. Therefore, our model is biased towards correctly predicting the validation subset due to indirectly “leaking” information from the validation subset during model optimization. Each additional model trained during optimization increases the potential for data leakage, but also increases the potential of finding the optimal model. The best model among many trials optimizing towards the validation subset is illustrated by Model 3. Since the data in the validation subset are no longer truly unseen data, we should test the best model in a new, completely unseen subset to avoid data leakage from inflating our reported model performance. We can use the test subset (not shown) for this purpose to report the performance of the optimized model on unseen data. In this example, the validation data contained more cases than controls in the unshaded area wherein the population is equally split between cases and controls. This resulted in choosing a model that performs better in the validation subset compared to the test subset because the test subset and population do not have all the same patterns seen in the validation data. The imperfect generalization of models to unseen data is expected, which is why a final test on truly unseen data is important for accurate performance reporting. Testing more models on the test subset may lead to the same indirect data leakage occurring in the test subset, resulting in inflated reported performance, so use of the test subset should be reserved for the final model.

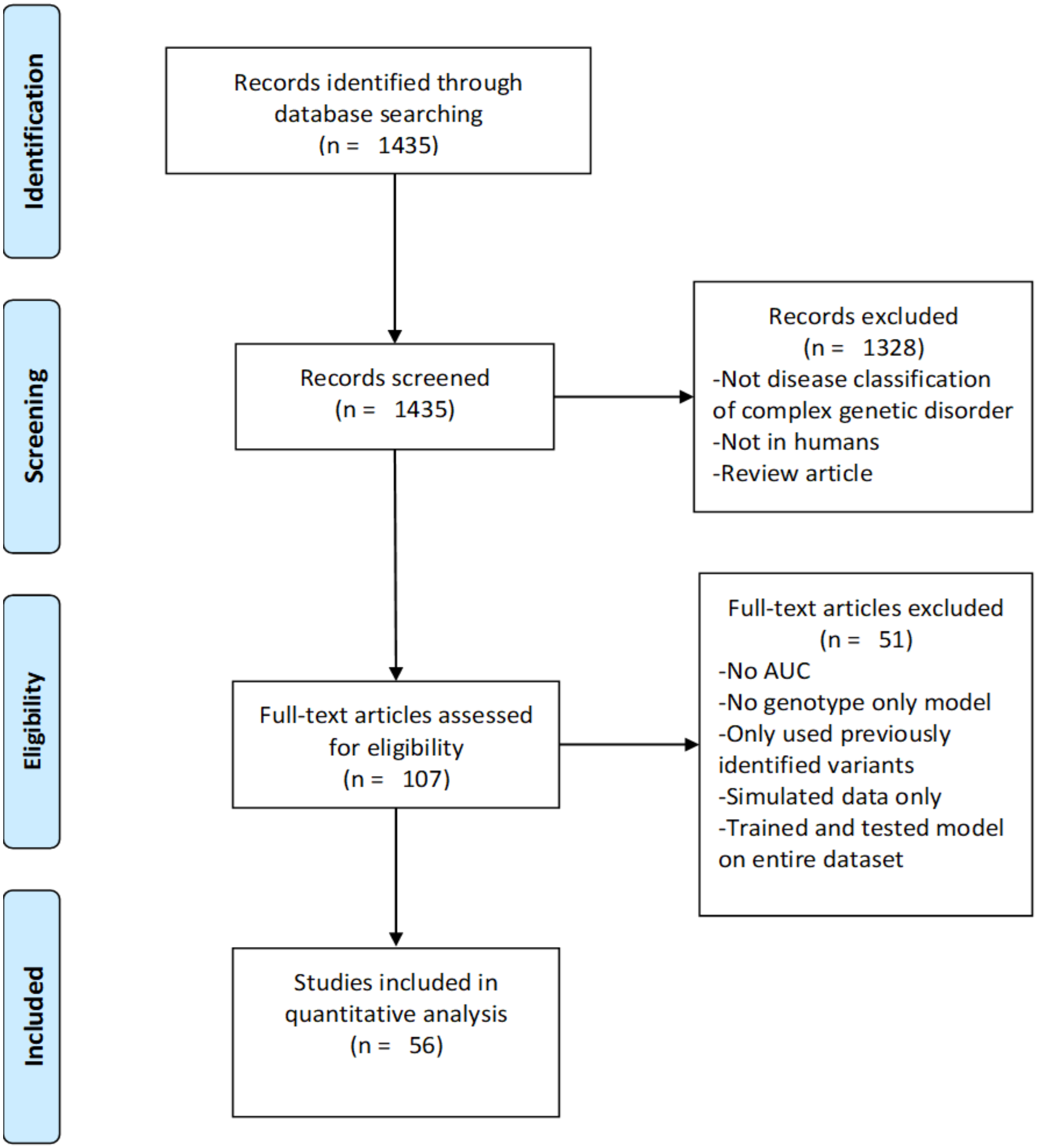

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 1435)
Records screened (n = 1435)
Records excluded (n = 1328)
-Not disease classification of complex genetic disorder
-Not in humans
-Review article
Full-text articles assessed for eligibility (n = 107)
Full-text articles excluded (n = 51)
-No AUC
-No genotype only model
-Only used previously identified variants
-Simulated data only
-Trained and tested model on entire dataset
Studies included in quantitative analysis (n = 56)

Figure 2: The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) diagram of article selection procedure.

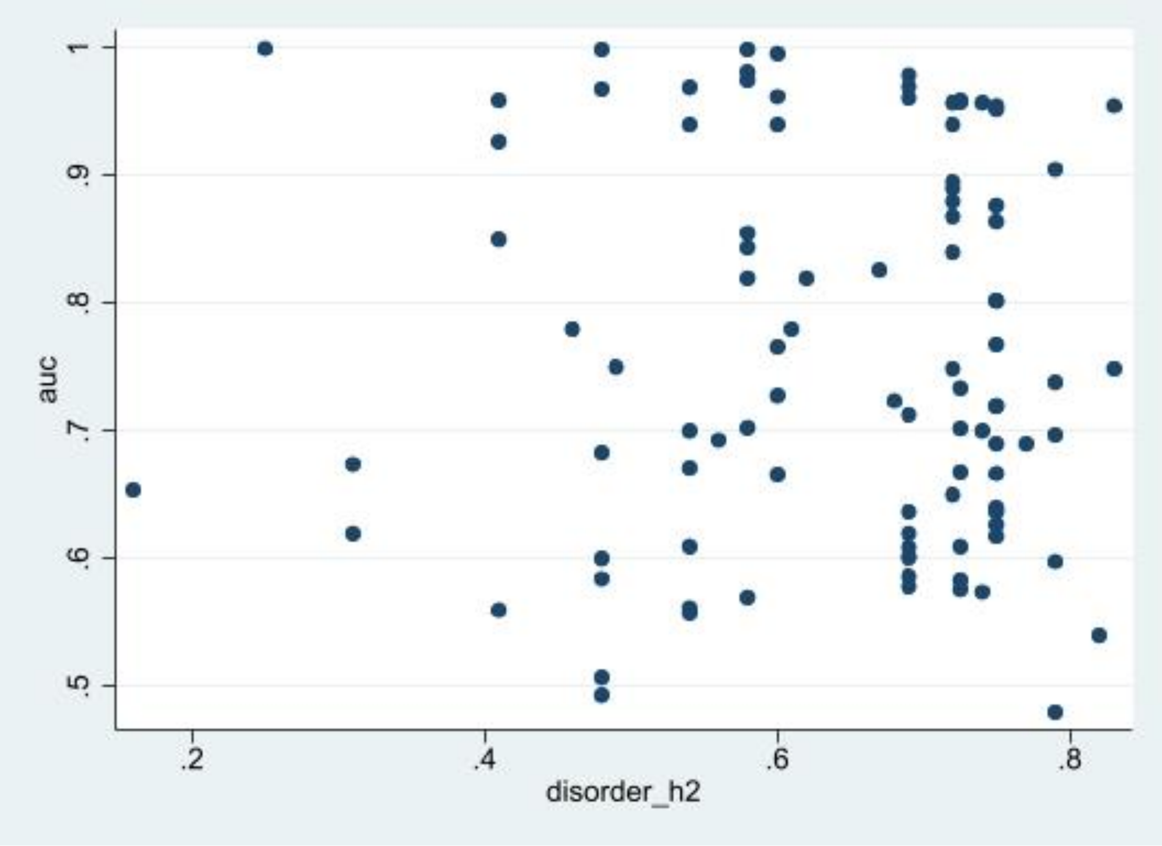


Figure 3: Scatterplot of AUC and disorder heritability. AUC represents the area under the receiver operating characteristic curve reported by each study and disorder_h2 represents an estimation of disorder heritability based on twin-studies.

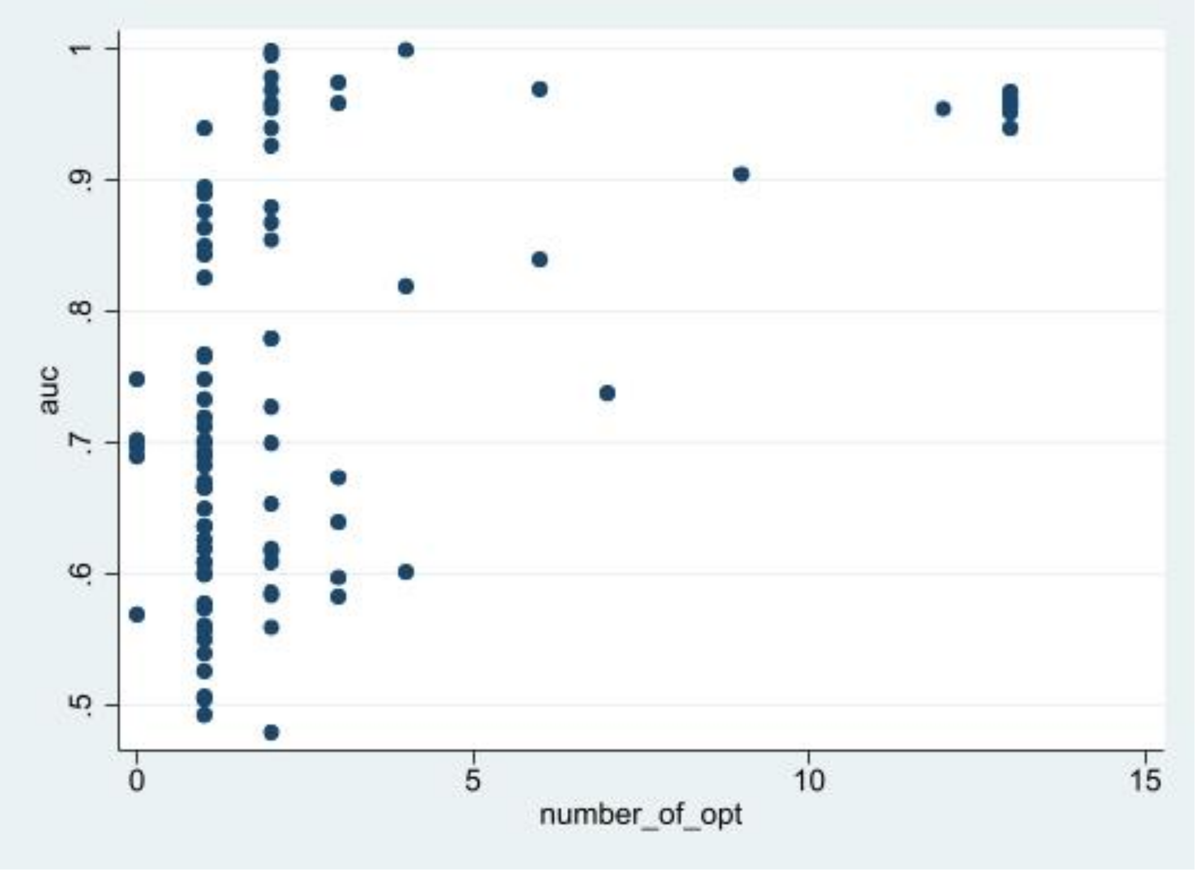


Figure 4: Scatterplot of AUC and number of hyperparameter optimizations reported. AUC represents the area under the receiver operating characteristic curve reported by each study and number_of_opt represents the number of hyperparameter optimizations that were reported by the study.

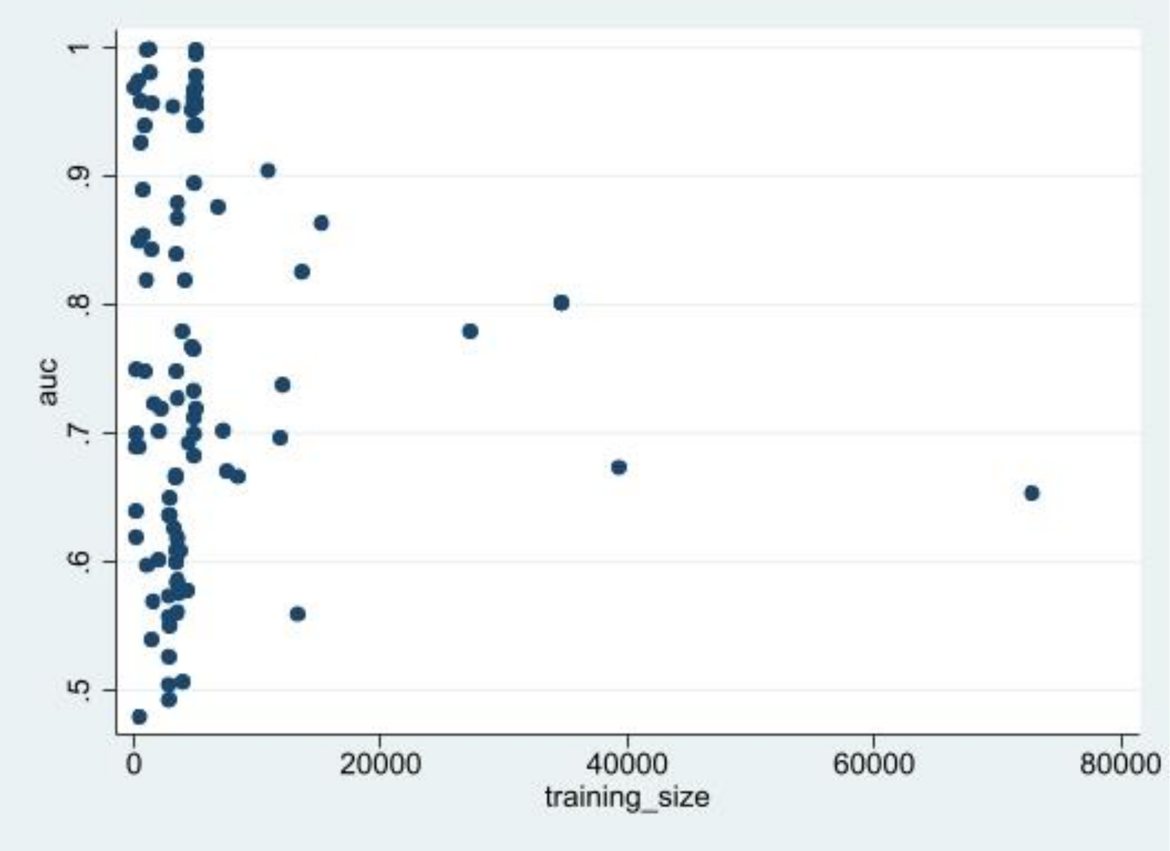


Figure 5: Scatterplot of AUC and training size. AUC represents the area under the receiver operating characteristic curve reported by each study and training size represents the total number of cases and controls used to train the model.

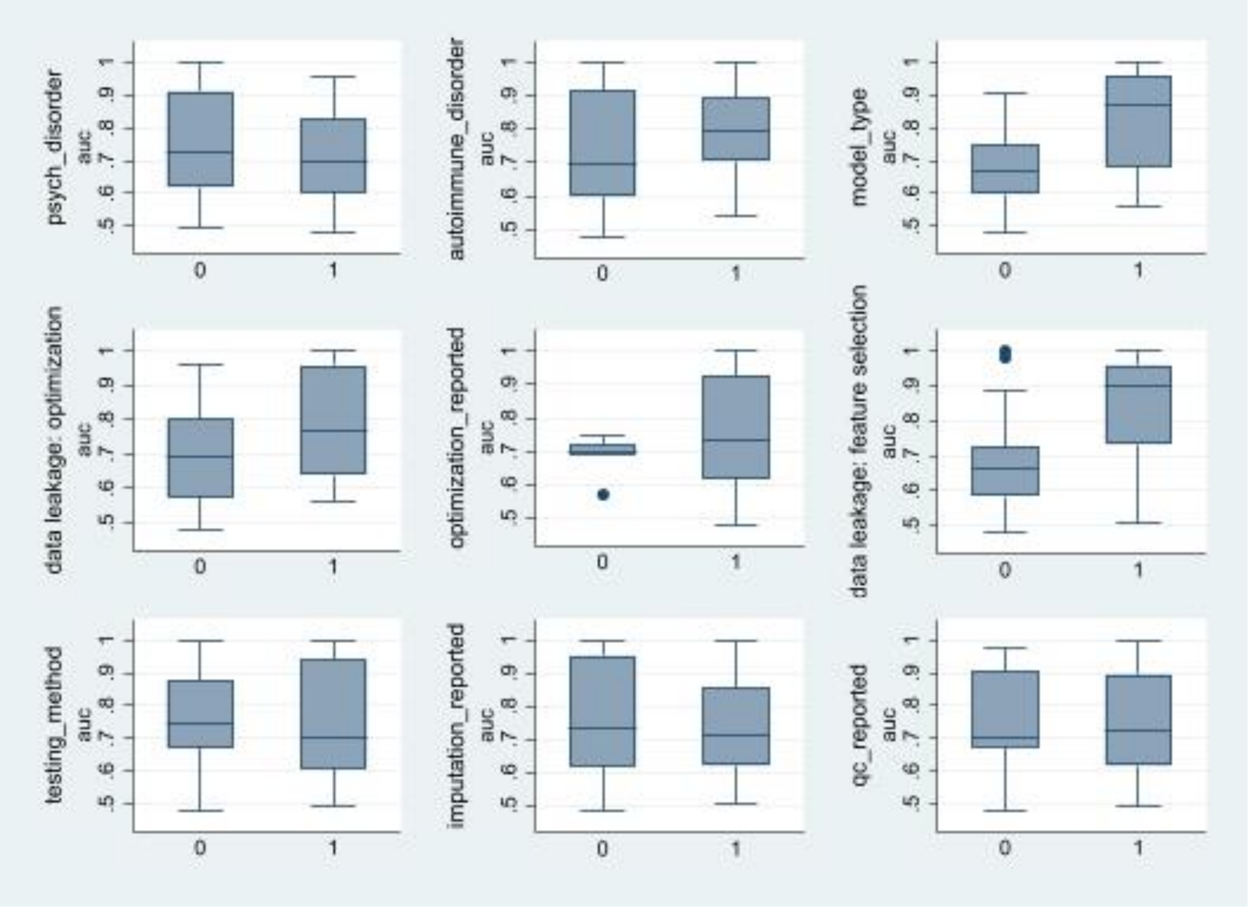


Figure 6. Box plots of all binary features used in our analyses. The "psych_disorder" variable represented whether the disorder studied was a psychiatric disorder. The "autoimmune_disorder" variable represented whether the disorder studied was an autoimmune disorder. The "model_type" variable represented whether the model used was linear (0) or non-linear (1). The "optimization_reported" variable represented whether the study reported any optimization of model hyperparameters. The "data leakage: feature selection variable represented whether the subset used to test the performance of the model was used to select the features used within the model". The "data leakage: optimization" variable represented whether the subset used to test the performance of the model was used to optimize the hyperparameters of the model. The "testing_method" variable represented whether model

performance was tested using cross-validation (0) or hold-out (1). The “imputation_reported” variable represented whether the study reported using imputed genotypes. The “qc_reported” variable represented whether the study reported any quality control procedures.

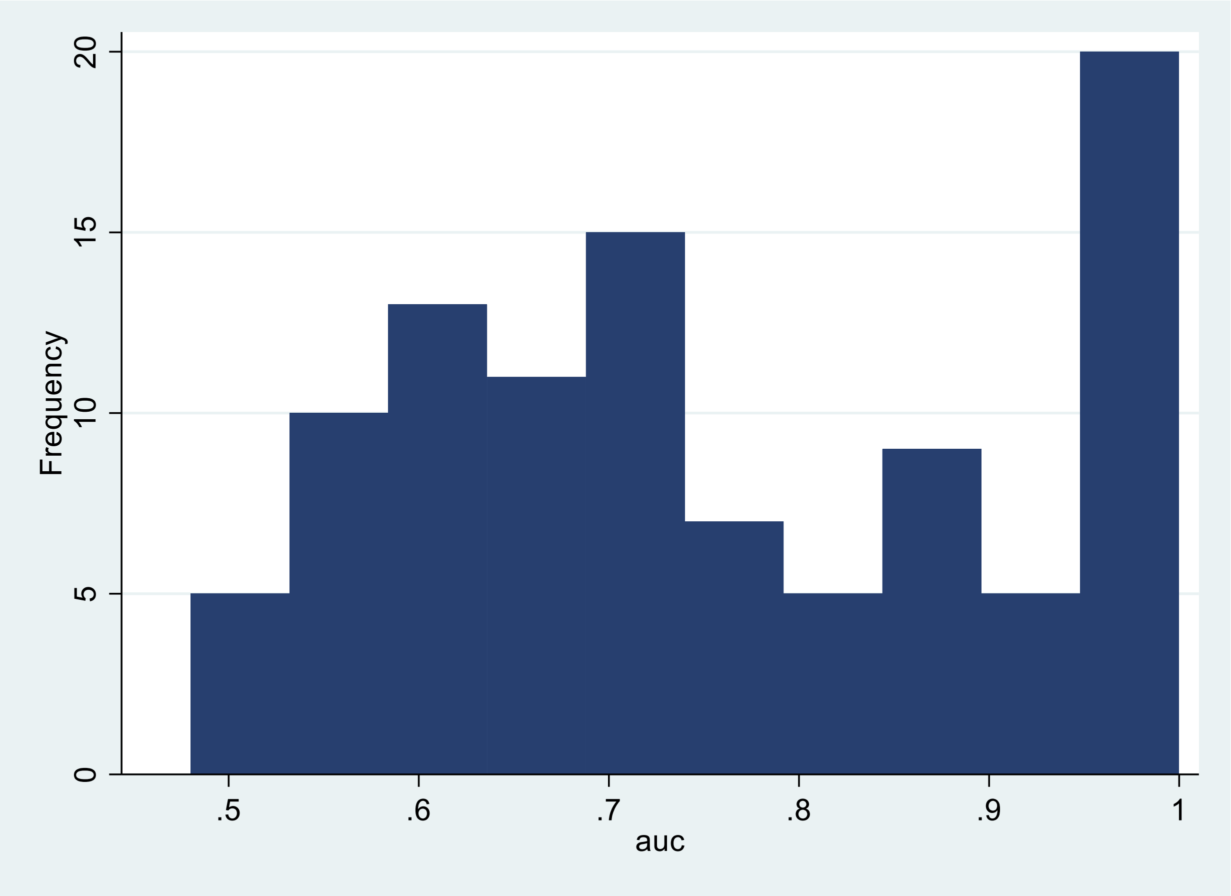


Figure 7: A histogram showing frequency distribution of reported AUCs. AUC represents the area under the receiver operating characteristic curve reported by each study and frequency represents how often an AUC within each histogram bin is reported within our study sample.

Identifying Pediatric Mood Disorders From Transdiagnostic Polygenic Risk Scores: A Study of Children and Adolescents

Eric J. Barnett[1], Joseph Biederman[2,3], Alysa Doyle[3,4], Jonathan Hess, PhD[5],

Maura DiSalvo[2], and Stephen V. Faraone[,5]

[1] Department of Neuroscience and Physiology, SUNY Upstate Medical University, Syracuse, NY, USA;

[2] Clinical and Research Programs in Pediatric Psychopharmacology and Adult ADHD, Massachusetts General Hospital, Boston, MA, USA;

[3] Department of Psychiatry, Harvard Medical School, Boston, MA, USA;

[4]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA.

[5]Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY, USA

Abstract

Objective: Mood disorders often co-occur with attention deficit-hyperactive disorder (ADHD), disruptive behavior disorders (DBD), and aggression. We aimed to determine if polygenic risk scores (PRSs) based on external genome-wide association studies (GWASs) of these disorders would be associated with mood disorders.

Methods: We combined six independent family studies that had genetic data and diagnoses for mood disorders that were made using different editions of the DSM. We predict mood disorders, either concurrently or in the future, in participants between 6 and 17 years of age using PRSs calculated using summary statistics of GWASs for ADHD, ADHD with DBD, major depressive disorder (MDD), bipolar disorder (BPD), and aggression to compute PRSs.

Results: In our sample of 485 youths, 356 (73%) developed a subthreshold or full mood disorder and 129 (27%) did not. The cross-validated mean areas under the receiver operating characteristic curve (AUCs) for the seven models predicting development of any mood disorder ranged from 0.552 in the base model of age and sex to 0.648 in the base + all five PRSs model. When included in the base model individually, the ADHD PRS (OR=1.65, p<0.001), aggression PRS (OR=1.27, p=0.02), and MDD PRS (OR=1.23, p=0.047) were significantly associated with the development of any mood disorder.

Conclusions: PRSs ADHD, MDD, BPD, DBDs and aggression were modestly associated with the development of mood disorders. These findings extend evidence for transdiagnostic genetic components of psychiatric illness and demonstrate that PRSs calculated using traditional diagnostic boundaries can be useful within a transdiagnostic framework.

Introduction

Mental health problems, which have been shown to account for 45% of the global burden of disease in people between 10 and 25 years old, are complicated by low healthcare utilization (Gore et al., 2011). One study found that 66.9% of adolescents in need of healthcare services for psychiatric disorders received none (Costello et al., 2007). Many studies have found that shorter duration of untreated disease is correlated with improved treatment response in mood disorders (Ghio et al., 2014; Hung et al., 2017; Kraus et al., 2020) (Berk et al., 2011; Bukh et al., 2013). Thus, better screening for these disorders would help get patients the care they need in the developmental period that gives the largest opportunity to improve outcomes.

In contrast to the Diagnostic and Statistical Manual of Mental Disorders (DSM) view of psychiatric disorders as distinct entities, several lines of evidence suggest that a transdiagnostic paradigm may be more appropriate. Several decades of research has shown that psychiatric comorbidity is pervasive for childhood disorders (Biederman et al., 1990; Neuman et al., 2001). More recently, multiple groups have shown that many common genetic variants are shared among psychiatric disorders (Anttila et al., 2018; Smoller et al., 2019) and that these disorders share some brain variations documented by neuroimaging (Radonjić et al., 2021). In parallel, many studies indicate that most psychiatric disorders fall along a continuum that is not discrete from the rest of the population (Smoller et al., 2019). These findings suggest that the discrete, categorical disorders of the DSM may be better represented as overlapping clusters of people expressing the extremes of multiple continuous traits. This evidence suggests that transdiagnostic evidence may be useful when screening for disorders because, for example, subthreshold or biomarker manifestations of one disorder might be useful when diagnosing other disorder. Polygenic risk scores (PRSs) estimate risk by summing the effects of common genetic variants across the genome. For psychiatric conditions, such scores have not yet proven useful in clinical settings. Nevertheless, given the genetic correlations among many psychiatric disorders, a PRS from one disorder may be predictive of other genetically correlated disorders (Smoller et al., 2019). There are two reasons that PRS from other disorders may be useful. First, it is well known that the reliability of a measure comprising correlated predictors is higher than the reliability of each individual predictor (Nunnally, 1978). Thus, adding risk estimates calculated based on other disorders may result in better estimates of the true risk of psychopathology. A second point to consider is that the risk estimate derived from a comorbid disorder may be much more reliable than the risk estimate derived from the disorder that is the target of the predictive model. This occurs when the sample used to derive the PRS for the comorbid disorder is much larger than the sample used to derive PRS for the target disorder.

Here, we use PRSs to predict the onset of mood disorders in children and adolescents. We know from prior clinical and epidemiologic studies that mood disorders frequently co-occur with attention deficit-hyperactive disorder (ADHD) (Demontis et al., 2019; Faraone & Biederman, 1997; Nigg et al., 2019), disruptive behavior disorders (Anney et al., 2008; Demontis et al., 2021) and aggression (Zhang-James & Faraone, 2016). Moreover, GWAS studies have shown that these disorders share common genetic variants (Demontis et al., 2019; Demontis et al., 2021; Schiweck et al., 2021; van Hulzen et al., 2017).

Given these associations, and the lack evidence for prediction of mood disorders in children using PRS derived from mood disorders alone (Biederman et al., 2021), we sought to capitalize on correlated associations to predict the emergence of any mood disorders in a sample of well-characterized youth. Our analytic strategy aimed to determine if PRSs based on external genome-wide association studies (GWASs) of bipolar disorder (BPD), major depressive disorder (MDD), ADHD, ADHD with disruptive behavior disorders (DBD), and aggression could improve upon the predictive accuracy afforded by a mood disorder PRS alone.

Methods

Sample

The sample was derived from six independent studies using identical assessment methodology: 1) and 2) were prospectively controlled family studies of boys and girls 6 to 17 years of age with and without DSM-III-R ADHD and their first-degree relatives (Boys Study: N=140 ADHD probands with N=454 first-degree relatives and N=120 control probands with N=368 first-degree relatives; Girls Study: N=140 ADHD and N=122 Controls) (Biederman et al., 1996) (Biederman et al., 2006); 3) was a prospective controlled family study of youth 10 to 18 years of age with and without DSM-IV pediatric BP-I disorder and their first degree relatives (N=105 BP-I probands with N=320 first-degree relatives and N=98 control probands with N=288 first-degree relatives) (Wilens et al., 2008); 4) was a prospective family study of youth 6 to 17 years of age of both sexes with active symptoms of DSM-IV BP-I Disorder and their first degree relatives (N=239

BP-I probands with N=687 first-degree relatives) (Wozniak et al., 2012); 5) was a cross-sectional family study of men and women 18 to 55 years of age with and without DSM-IV ADHD and their first degree relatives (N=224 ADHD probands with N=300 first-degree relatives and N=146 control probands with 118 first-degree relatives) (Faraone et al., 2006); and 6) was a cross-sectional linkage study of families with two or more full biological siblings with a lifetime diagnosis of DSM-IV ADHD (N=271 families, N=1,170 genotyped individuals) (Faraone et al., 2008). Subjects from the boys ADHD study were followed-up after one, four, 10, and 16 years; subjects from the girls ADHD study were followed-up after five and 11 years; and subjects from the BPD family study were followed-up after four, five, and six years. Subjects from the other three studies had cross-sectional data only.

The pediatric ADHD studies (1 and 2) recruited subjects from pediatric and psychiatric clinics. The BPD studies and ADHD linkage study (3, 4, and 6) recruited subjects from referrals to the Clinical and Research Programs in Pediatric Psychopharmacology at the Massachusetts General Hospital and through advertisements in the community. The ADHD linkage study (6) also recruited subjects from pediatric clinics and private child psychiatry practices. The adult ADHD study (5) recruited subjects from psychiatric clinics and advertisements in the community. Controls were recruited from pediatric clinics, advertisements to hospital personnel and community newspapers, and Internet postings. Potential subjects were excluded from all six studies if they had major sensorimotor handicaps, inadequate command of the English language, or a Full-Scale IQ <70 (<80 for the pediatric and adult ADHD studies), and from all studies except the adult ADHD study (1,2,3,4,6) if they were adopted or if their nuclear family was not available for study. Potential subjects were also excluded from all four ADHD studies (1,2, 5, 6) if they had psychosis, from the pediatric ADHD and BPD studies if they had autism, from the BPD studies if their BP-I disorder was due solely to a medication reaction, and from the ADHD linkage study if they did not want to provide a blood sample. For all six studies, every subject 18 and older provided written informed consent. Children and adolescents provided written assent to participate and the parents provided written informed consent. The Partners Human Research Committee approved these studies.

For the current study, we restricted our sample to subjects who were 6 to 17 years of age, had genetic data available, and had diagnosis information for bipolar disorder and major depressive disorder. Based on these criteria, our sample consisted of 485 subjects including 112 subjects from the boys ADHD study (1), 144 subjects from the girls ADHD study (2), 21 subjects from the controlled BPD study (3), 80 subjects from the BPD family study (4), 10 subjects from the adult ADHD study (5), and 117 subjects from the ADHD linkage study (6). There was also one subject with genetic data included from an "unselected" clinic population referred for psychiatric care at MGH for which there were no exclusion criteria and for whom we received approval from the Partners Human Research Committee to review, analyze, and report on anonymously.

Diagnostic Assessments

In all six studies, psychiatric assessments of subjects younger than 18 were made with the Kiddie Schedule for Affective Disorders – Epidemiologic Version (K-SADS-E) (Orvaschel & Puig-Antich, 1987). For subjects 12 and younger, diagnoses were based on independent interviews with parents. For subjects 13 to 17, diagnoses were based on independent interviews with parents and direct interviews with children and adolescents. Data were combined such that endorsement of a diagnosis by either reporter resulted in a positive diagnosis.

Extensively trained and supervised psychometricians with undergraduate degrees (or graduate degrees for the ADHD linkage study) in psychology or a related field conducted all interviews. For the pediatric ADHD studies, ADHD linkage study, and the controlled BPD study, raters were blind to the ascertainment status of the families. For the BPD family study, raters were blind to the study assignment and whether the subject was a proband or sibling. For the adult ADHD study, raters were blind to the subject's ascertainment status, ascertainment site, and all prior assessments. To assess the reliability of our overall diagnostic procedures, we computed kappa coefficients of agreement by having experienced, blinded, board-certified child and adult psychiatrists and licensed experienced clinical psychologists diagnose subjects from audiotaped interviews made by the assessment staff. Based on 500 assessments from interviews of children and adults, the median kappa coefficient was 0.98 for the pediatric ADHD studies, adult ADHD study, and the controlled BPD study, and 0.99 for the BPD family study. Based on 173 assessments from interviews of children and adults, the median kappa coefficient was 0.99 for the ADHD linkage study.

Socioeconomic status (SES) was measured using the 5-point Hollingshead scale (Hollingshead, 1975). A higher score indicates being of a lower socioeconomic status.

Polygenic Risk Scores

All participants provided blood for DNA extraction and genome wide genotyping of 585,979 SNPs on the Illumina PsychArray. A minimum call rate of 98% was set to exclude variants and individuals with missing data. In addition, we removed variants that showed significant departure from Hardy-Weinberg equilibrium (p < 1×10-6) and variants with a minor allele frequency (MAF) less than 1%. Following these steps, data 504,432 variants were retained. The Michigan Imputation Server was used to perform automated haplotype phasing with Eagle v.2.4 and imputation of missing genotypes with Minimac4 based on the Haplotype Reference Consortium (version r1.1 2016), a reference panel of 64,940 haplotypes from individuals of predominantly European ancestry. After genotype imputation, quality control steps were performed to exclude variants with a MAF less than 1%, variants with a call rate under 98%, and variants that were not robustly imputed ($R^2 < 90\%$). In order to detect variation between patients due to ancestry, we performed a principal component analysis (PCA) on directly genotyped variants that exhibited a minimum MAF of 10% and approximate linkage equilibrium (Plink command: --indep-pairwise

100 10 0.2). Variants found in the extended major histocompatibility (MHC) locus of chromosome 6 (24mb – 35mb) were excluded to avoid biasing our PCA due to extensive linkage disequilibrium (LD). Top principal components (PCs) were included in initial analyses to check and adjust for potential confounding due to ancestry.

At the time of writing, we used published genome-wide summary statistics from the largest available genome-wide association meta-analyses of ADHD, ADHD with DBD, MDD, BPD, and aggression to compute PRSs (Demontis et al., 2019; Demontis et al., 2021; Stahl et al., 2019; Wray et al., 2018). We used imputed genome-wide SNP genotypes (n SNPs = 8,063,863) to calculate PRSs for three neuropsychiatric disorders (ADHD, BPD, and MDD). All PRSs were computed using the conventional LD-pruning and p-value thresholding (P+T) method (Purcell et al., 2009). Pre-processing steps were followed to exclude uncommon SNPs (MAF < 10%), insertions and deletions, variants in the extended MHC locus, variants with an imputation quality score less than 90%, strand-ambiguous variants (i.e., CG, AT), and variants not included in our target dataset from the GWAS summary statistics. We then used Plink v.1.9 to perform a greedy pruning of SNP associations (or "clumping") such that the resultant SNP set was largely LD-independent. The parameters used for the clumping algorithm were as follows: --clump-p1 1.0 –clump-p2 1.0 –clump-kb 250 –clump-r2 0.1. When computing PRSs in our dataset, we chose the p-value threshold that was reported to have maximized the phenotype variance explained (R2) in a sample that was independent from the initial training sample that was used to derive the PRS formula (ADHD: $p \leq 0.2$; BPD: $p \leq 0.01$; MDD: $p \leq 0.05$; aggression: $p < 0.1$ (Elam et al., 2018). ADHD with DBD was an exception to this criteria; in absence of a reported best p-value threshold, we computed PRS for ADHD with DBD using a threshold of $p < 0.5$. PRSs were standardized to a mean of zero and unit variance for downstream statistical analyses.

Statistical Analysis

First, we stratified patients by lifetime development of any subthreshold or full mood disorder (MDD or BPD) and compared them on sociodemographic characteristics using t-tests, ordered logistic regression, and Pearson's chi-square tests. We included subthreshold cases based on a meta-analysis that showed evidence for the validity of subthreshold cases (Vaudreuil et al., 2019). Participants were classified as having a subthreshold mood disorder if they did not meet full criteria, had three or more symptoms, and had a duration of symptoms that lasted at least one week to qualify as an episode. For subjects from the pediatric ADHD studies and the BPD family study, we defined lifetime history of any mood disorder as positive if the subject met subthreshold or full diagnostic criteria for MDD or BPD at any assessment (baseline or follow-up visits). For subjects with cross-sectional data from the adult ADHD study, ADHD linkage study, and controlled BPD study, we defined lifetime history of any mood disorder as positive if the subject met subthreshold or full diagnostic criteria for MDD or BPD at the time of assessment. Next, we examined the predictive utility of seven models to identify any mood disorder versus no mood disorder. We started by using multiple logistic regression to test a model that predicted any mood disorder from age, sex, and the first 10 principal components from a principal components analysis, which reduces the dimensionality of the genetic data to explain as much variance as possible. If the ancestry of our subjects differed between cases and controls, the principal components would be predictive and control for differing ancestries between samples (Price et al., 2006). However, none of the principal components were significantly predictive and were therefore excluded from other models to minimize overfitting and overestimation of performance that could be caused by overparameterization. We used a logistic regression model using age and sex to predict any mood disorder as the base model. We then added each PRS to the base model individually (i.e., base model + BPD PRS; base model + MDD PRS; base model + ADHD PRS; base model + ADHD with DBD PRS; base model + Aggression PRS) to test the predictive utility of each PRS. Finally, we tested a model that included all five PRSs plus the base model. We assessed the predictive utility of the models using receiver operating characteristic (ROC) curve analysis with 10x cross-validation and summarized the results using mean area under the curve (AUC) statistics across the ten folds. In our 10x cross-validation protocol, we randomly split subjects into 10 folds and each fold is iteratively held out of model fitting to measure the prediction performance in that fold on a model fit using the other 9 folds. All AUCs we report are based only on the prediction performance in the withheld folds during cross-validation. AUC represents the probability that a randomly selected case/control pair are accurately classified. AUCs from different models were compared using the DeLong test (DeLong et al., 1988) for comparing AUCs. Our test for equality of AUC statistics used a single AUC based on cross-validated probabilities for each model. We also performed two sensitivity analyses; the first restricted the sample to Caucasian patients and the second predicted full mood disorders only. All analyses were two-tailed and performed at the 0.05 alpha level using Stata (Version 16.1) (StataCorp, 2019).

The PCs and PRSs were standardized based on the means and standard deviations of the current sample. For each model, we assessed goodness of fit using Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), and Nagelkerke's pseudo-R2. When comparing AIC and BIC values across models, lower values indicate a better fit model. When comparing Nagelkerke's pseudo-R2 values across models, higher values indicate a greater percentage of variance explained by the model. The amount of variance explained by the PRS variables is calculated as the difference of Nagelkerke's pseudo-R2 in the model including the PRS compared with the base model.

Results

Sociodemographic Characteristics

In our sample of 485 youths, 356 (73%) developed a subthreshold or full mood disorder and 129 (27%) did not. As shown in Table 1, there were significant differences between those who did and did not develop a mood disorder in socioeconomic status (SES) and sex. Those who developed a mood disorder were of lower SES and had a greater percentage of males compared to those who did not. There were no significant differences between the groups in age or race.

Prediction of Any Mood Disorder by Polygenic Risk Scores

As shown in Table 2, the fit statistics were the best for the base model + all five PRSs with the lowest AIC (AIC=1.110) and second lowest BIC (BIC=-2427.326). The BIC, which penalizes more heavily for complex models, was slightly lower in the base model + ADHD PRS (BIC=-2440.554) compared to the base model + all five PRSs. The base model + all five PRSs had the highest Nagelkerke R^2 (R^2=0.114), with the five PRSs explaining 9.8% of the variance when comparing this model to the base model (R^2=0.016). The base model + ADHD PRS had the next highest Nagelkerke R^2 (R^2=0.081), with the ADHD PRS explaining 6.5% of the variance when comparing this model to the base model. The base model + BPD PRS, base model + MDD PRS, base model + ADHD with DBD PRS, and base model + Aggression PRS performed no better than the base model itself, explaining only 0.05% to 1.5% of the variance. All comparisons marked significant at $p < 0.01$ were statistically significant after correction for multiple comparisons.

The cross-validated mean AUC statistics for the seven models ranged from 0.552 for the base model to 0.648 for the base model + all five PRSs (Table 2, Figure 1). Pairwise comparisons revealed that the base model + ADHD PRS performed significantly better at identifying youths with any mood disorder than all the other models except for the base model + Aggression PRS and the base model + all five PRSs.

The base model + all five PRSs performed significantly better at identifying youths with any mood disorder than all the other models except for the base model + ADHD PRS.

When included in the base model individually, the BPD PRS (OR=1.14, p=0.20, 95% CI: 0.93 – 1.40) and ADHD with DBD PRS (OR=1.17, p=0.13, 95% CI: 0.95 – 1.45) were not significantly associated with any mood disorder, but the MDD PRS (OR=1.23, p=0.047, 95% CI: 1.00 – 1.52), ADHD PRS (OR=1.65, p<0.001, 95% CI: 1.33 – 2.05), and Aggression PRS (OR=1.27, p=0.02, 95% CI: 1.03 – 1.56) were significant before correction for multiple testing. Higher MDD, ADHD and Aggression PRSs were associated with increased odds of having a mood disorder. When all five PRSs were included in the base model at the same time, only the ADHD PRS (OR=1.68, p<0.001, 95% CI: 1.34 – 2.10) and Aggression PRS (OR=1.33, p=0.01, 95% CI: 1.07 – 1.66) remained significant.

Discussion

Although genetic associations among psychiatric disorders have been well documented (Smoller et al., 2019), this study is the first to use PRSs for several psychiatric disorders to predict the emergence of mood disorders in youth. Using PRSs for ADHD, MDD, BPD, DBDs and aggression, we could modestly predict the lifetime development of mood disorders (operationalized as subthreshold or full presentation) in a set of independent family studies. These models extend evidence for transdiagnostic components of psychiatric illness using genetic data and demonstrate that PRSs computed using traditional diagnostic boundaries can be leveraged within a transdiagnostic approach to child psychopathology.

Several factors might explain the failure of the BPD and MDD PRSs to predict new onsets of these disorders. We studied youth who were 6 to 17 years of age but the samples that generated the PRSs were mostly ascertained as adults. One interpretation of our findings is that the genomic etiology of the early onset mood disorders associated with ADHD differs from the genomic etiology of adult-onset mood disorders. If so, there may be a neurodevelopmental mood disorder associated with ADHD and aggression. The ADHD PRS may have been significantly predictive in this sample due to its ability to predict an ADHD specific depression, which could also explain the overrepresentation of males in the group with any mood disorder. Consistent with this idea, Levitan et al. (Levitan et al., 2020) proposed a neurodevelopmental theory of depression and inflammation associated with obesity and metabolic dysfunction, which are also seen in ADHD (Q. Chen et al., 2018; Chen et al., 2019; Muntaner-Mas et al., 2020).

It would be reasonable to suspect that our inclusion of subthreshold cases of BPD and MDD may have limited the success of those PRSs in predicting any mood disorder since both GWASs only include cases meeting full criteria. If the genetic risk architecture of subthreshold disorders differs from that of the corresponding full threshold disorders, the former would be less accurately predicted by PRS generated by the latter type of sample. However, results were similar when we predicted only full mood disorders, which is consistent with the continuum theory of psychiatric disorders (Smoller et al., 2019).

We followed up the single PRS models by testing a model that included age, sex, and all 5 PRSs. This model had the highest AUC among the models we tested of 0.65. This was a significant improvement from all models except the base + ADHD PRS model. Compared with the ADHD PRS model, the 5 PRSs model also had a lower variance among cross-validation folds. In the base + 5 PRSs model, only the ADHD and aggression PRSs were significantly associated with any mood disorder after controlling for age, sex, and the other PRSs. This suggests that, while the inclusion of multiple PRSs was not enough to significantly improve prediction in our sample, it is noteworthy that the aggression PRS was associated with the development any mood disorder even after controlling for the ADHD PRS.

This study has several limitations. The study sample used here differs from the population that would be screened as part of a transdiagnostic clinical staging paradigm. We are also limited by using PRSs as the only genetic sources of information as that may have limited the flexibility of the models to represent the genetic architecture of the complex disorders we are attempting to predict. Because allele frequencies differ across races and ethnicities, more work is needed to collect data from underrepresented groups. The data sets we used in these analyses were also included as part of larger GWAS studies that were used to estimate PRSs. This could lead to overestimation of the prediction performance of overlapping PRSs but given the large size of the GWAS relative to the overlapping data sets used here the impact is likely minimal. In our analysis, 8 out of 485 children were related, which could also lead to overestimating prediction performance. Given the lack of significance of any principal components and the minimal use of related individuals, the impact of this limitation is likely small.

While our models predicting any mood disorder using genetic data show the potential of using genetic information alone, further improvements might be made by using genetic risk profiles alongside clinical interviews and other biomarkers in screening for these disorders. While the complementarity of genetic data with other sources of data in mood disorders still needs to be investigated, studies in other disorders have found that combining genetic data with other data sources leads to improved prediction (Li et al., 2018; Xu et al., 2011). The low predictive accuracy of the models presented here makes it unlikely that they would be clinically useful individually and therefore do not warrant reporting conditional probability or other more clinically relevant metrics nor are they relevant to any specific clinical setting. Instead, our study shows that in this opportunistic sample gathered from data available to us, the genetics a child is born with are modestly predictive of that child developing a mood disorder with simple models. If further genetic risk modelling improvements are made and used alongside the clinical interviews currently implemented in screening, we may eventually improve detection of patients at risk for mood disorders.

Table 1. Sociodemographic characteristics

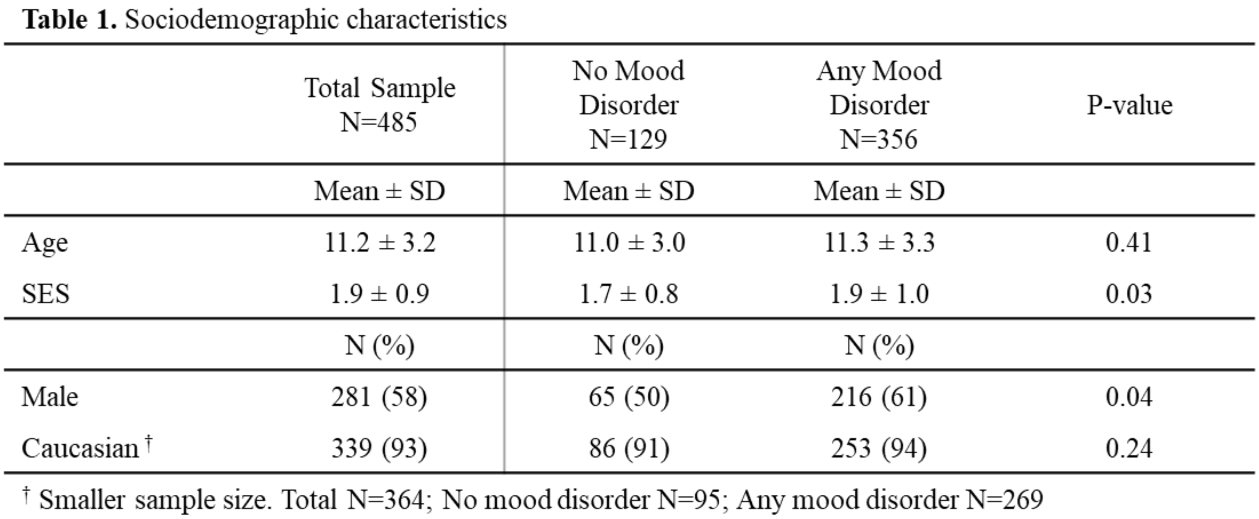


Table 1. Sociodemographic characteristics

	Total Sample N=485	No Mood Disorder N=129	Any Mood Disorder N=356	P-value
	Mean ± SD	Mean ± SD	Mean ± SD	
Age	11.2 ± 3.2	11.0 ± 3.0	11.3 ± 3.3	0.41
SES	1.9 ± 0.9	1.7 ± 0.8	1.9 ± 1.0	0.03
	N (%)	N (%)	N (%)	
Male	281 (58)	65 (50)	216 (61)	0.04
Caucasian [†]	339 (93)	86 (91)	253 (94)	0.24

[†] Smaller sample size. Total N=364; No mood disorder N=95; Any mood disorder N=269

Table 2: Comparison of fit statistics and predictive utility

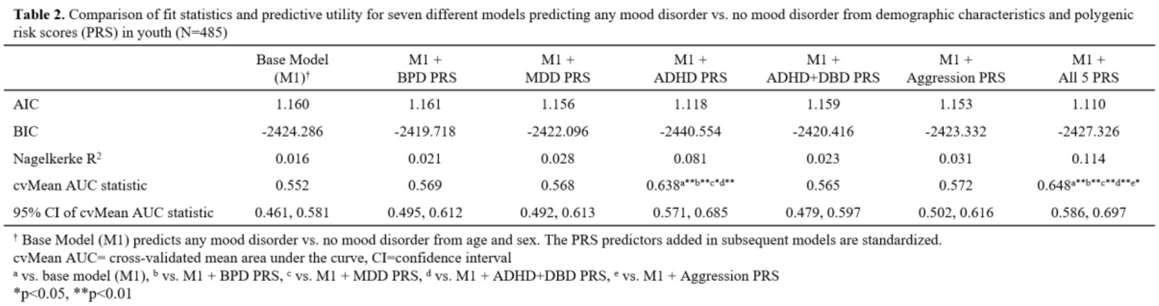


Table 2. Comparison of fit statistics and predictive utility for seven different models predicting any mood disorder vs. no mood disorder from demographic characteristics and polygenic risk scores (PRS) in youth (N=485)

	Base Model (M1)†	M1 + BPD PRS	M1 + MDD PRS	M1 + ADHD PRS	M1 + ADHD+DBD PRS	M1 + Aggression PRS	M1 + All 5 PRS
AIC	1.160	1.161	1.156	1.118	1.159	1.153	1.110
BIC	-2424.286	-2419.718	-2422.096	-2440.554	-2420.416	-2423.332	-2427.326
Nagelkerke R^2	0.016	0.021	0.028	0.081	0.023	0.031	0.114
cvMean AUC statistic	0.552	0.569	0.568	$0.638^{a**b**c*d**}$	0.565	0.572	$0.648^{a**b**c**d**e*}$
95% CI of cvMean AUC statistic	0.461, 0.581	0.495, 0.612	0.492, 0.613	0.571, 0.685	0.479, 0.597	0.502, 0.616	0.586, 0.697

† Base Model (M1) predicts any mood disorder vs. no mood disorder from age and sex. The PRS predictors added in subsequent models are standardized.

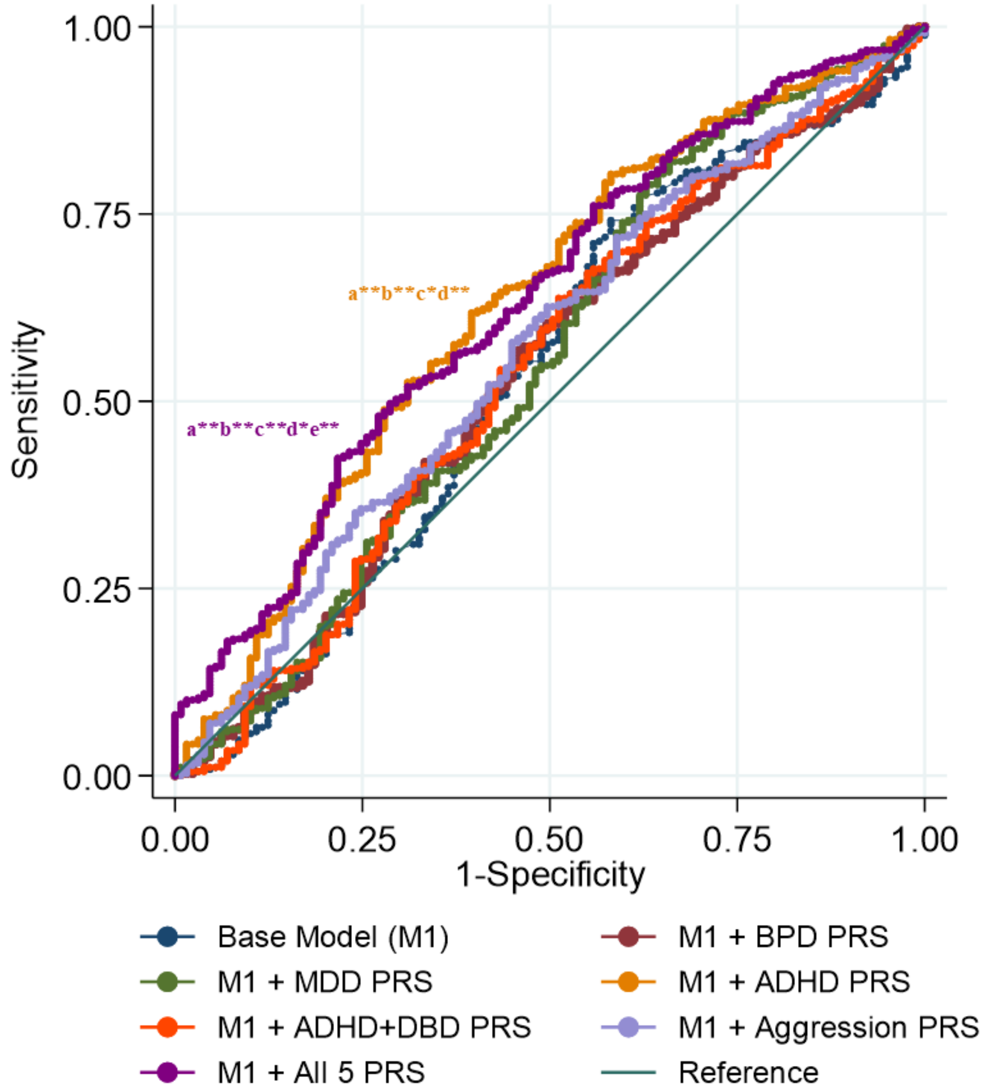
cvMean AUC= cross-validated mean area under the curve, CI=confidence interval

[a] vs. base model (M1), [b] vs. M1 + BPD PRS, [c] vs. M1 + MDD PRS, [d] vs. M1 + ADHD+DBD PRS, [e] vs. M1 + Aggression PRS

*p<0.05, **p<0.01

Figure 1. Receiver operating characteristic (ROC) curves

Figure 1. Receiver operating characteristic (ROC) curves for the seven models tested to predict any mood disorder vs. no mood disorder in youth



[a] vs. base model (M1), [b] vs. M1 + BPD PRS, [c] vs. M1 + MDD PRS, [d] vs. M1 + ADHD+DBD PRS, [e] vs. M1 + Aggression PRS; * p<0.05, ** p<0.01

Improving Machine Learning Prediction of ADHD Using Gene Set Polygenic Risk Scores and Risk Scores from Genetically Correlated Phenotypes

Eric Barnett[1], Yanli Zhang-James[2], Stephen V Faraone[1,2]

[1]Department of Neuroscience and Physiology, SUNY Upstate Medical University, Syracuse, New York, USA

[2]Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, New York, USA

Abstract

Background: Polygenic risk scores (PRSs), which sum the effects of SNPs throughout the genome to measure risk afforded by common genetic variants, have improved our ability to estimate disorder risk for Attention-Deficit/Hyperactivity Disorder (ADHD) but the accuracy of risk prediction is rarely investigated.

Methods: In a study of 10,887 participants across 9 cohorts, we performed gene set analysis of GWAS data to select gene sets associated with ADHD within a training subset. For each gene set, we generated gene set polygenic risk scores (gsPRSs), which sum the effects of SNPs for each selected gene set. We created gsPRS for ADHD and for phenotypes having a high genetic correlation with ADHD. These gsPRS were added to the standard PRS as input to machine learning models predicting ADHD.

Results: On the test subset, a random forest (RF) model using PRSs from ADHD and genetically correlated phenotypes and an optimized group of 20 gsPRS had an area under the receiving operating characteristic curve (AUC) of 0.72 (95% CI: 0.70 – 0.74). This AUC was a statistically significant improvement over logistic regression models and RF models using only PRS from ADHD and genetically correlated phenotypes.

Conclusions: Summing risk at the gene set level and incorporating genetic risk from disorders with high genetic correlations with ADHD improved the accuracy of predicting ADHD. Learning curves suggest that additional improvements would be expected with larger study sizes. Our study suggests that better accounting of genetic risk and the genetic context of allelic differences results in more predictive models.

Introduction

The field of psychiatric genomics has made great strides discovering genetic loci that are significantly associated with psychiatric disorders (Demontis et al., 2019; Psychiatric Genomics Consortium, 2014; Stahl et al., 2019). These discoveries have generated new hypotheses about the genomic architecture complex pathogenesis of many of these disorders. The combination of risk conferring alleles has improved the prediction of psychopathology (Smoller et al., 2019).

A multi-site ADHD GWAS found that 12 genome-wide-significant loci captured a small amount of the heritability of ADHD while risk profiles using all loci captured a significantly larger amount of heritability, which proved the usefulness of loci that are, individually, are not significantly different between cases and controls (Demontis et al., 2019). Even in this study of over 20,000 people with ADHD, the complex genetic architecture of the disorder makes predicting generalizable risk and establishing significance at each common variant difficult.

Previous work has shown that ADHD has significant genetic overlaps with other psychiatric and non-psychiatric disorders (Brikell et al., 2018; Chen et al., 2017; Cole et al., 2009; Faraone et al., 2017; Rommelse et al., 2010; Skoglund et al., 2015). This supports the theory that ADHD risk comprises traits that are also present in the phenotypes with which it is genetically correlated. The risk estimation of SNPs in genetically correlated disorders could be more predictive in ADHD relative to the risk estimation of SNPs in ADHD GWASs due to larger sample sizes being better for estimating risk. In addition, when dealing with disorders with high heterogeneity like ADHD it is possible that other less heterogeneous phenotypes better estimate risk for genetic loci for some clusters of patients. Therefore, using the genetic overlap with other disorders could be useful in improving the predictive modeling of ADHD.

A review of twin-studies of ADHD found that the mean heritability of ADHD across 37 studies was 74% (Faraone & Larsson, 2018a). The high heritability of ADHD suggests that predicting ADHD using genetic and environmental data is achievable. However, reports on predictive models of ADHD using genetic data are limited. Significant improvements in prediction and our understanding of the disorder must be made before genetic information can be used in the clinic as part of future objective diagnoses and personalized medicine plans that aim to improve outcomes in ADHD.

One potential area of improvement is balancing the flexibility of models to detect robust risk patterns with complexity and generalizability. Combining the risk at SNPs across the genome into a single polygenic risk score (PRS) has proven to be a successful way to create a more useful and generalizable feature than any individual SNP (Wray et al., 2020). However, summing all SNPs into a single value per individual limits any modelling method's capacity to learn more complicated patterns and interactions. On the other end of the complexity spectrum, using individual SNPs as input into machine learning models of complex and heterogenous disorders like ADHD leads to concerns of overfitting and lack of generalizability (Ying, 2019). Combining risk at the gene set level could be an effective middle ground between these two extremes. While research using features combining risk at the gene set level to predict a disorder is limited, gene set association analyses have shown that this middle ground can be useful.

While machine learning classification models of ADHD using genomic data have not been reported, many researchers have used such models to predict diagnoses for other heritable complex disorders (Almlöf et al., 2017; Evans et al., 2009; Mittag et al., 2015; Wang et al., 2016; Wei et al., 2013). Collectively, these studies have shown the potential of machine learning to predict many disorders but concerns of how well these models would perform on unseen external data sets remain. In addition, many machine learning methods generate "black-box" models that are uninterpretable. Since most models lack the performance necessary for clinical application, interpretable models may provide additional useful results apart from the model that would otherwise only be an intermediate to eventual models that will be useful clinically. Interpretable genomic models could yield biological insights by finding new loci of interest or new groups of loci that together improve models. These models also could incorporate further model validation by relating the output to our understanding of the biology behind the disorder.

Here, we balance these issues by summing risk across gene sets to create gene set polygenic risk scores (gsPRSs) that may be used alongside PRS to improve predictive accuracy by providing the model with information about gene sets associated with ADHD. We hypothesized that including gsPRSs as input into machine learning models would improve prediction performance compared to models that use only traditional PRS. We also supplemented the model with summary statistics from phenotypes with high genetic correlations with ADHD as additional features to test if these data are useful to improve ADHD prediction.

Methods

Data Preprocessing and Splitting

Quality control and imputation were done using the RICOPILI pipeline (Lam et al., 2020). After quality control, 2455 ADHD cases and 8432 controls across 9 cohorts with European ancestry aggregated by the PGC were available for analysis as part of the fast-track data analysis pipeline (Demontis et al., 2019). We excluded SNPs with a minor allele frequency < 0.01, missing genotype rate > 0.05, and deviating from Hardy-Weinberg equilibrium in controls at $p < 1 \times 10^{-5}$.

The participants were randomly split into a training subset containing 1673 cases and 5818 controls, a validation subset containing 406 cases and 1329 controls, and a test subset containing 376 cases and 1285 controls. The training subset was used to teach the model to differentiate different cases and controls by optimizing the parameters within the model. The validation subset was used to estimate the model performance outside examples used to train the model and to optimize model hyper-parameters. The test subset was used for reporting the results of our final models on an unseen sample.

Gene Set Association Analysis

Using the SNP association p-values generated in the SNP association analysis, we used MAGMA to compare allele frequencies between cases and controls at the gene and gene sets level (de Leeuw et al., 2015). Both analyses used an extended gene window starting 35 kilobases upstream and ending 10 kilobases downstream of each gene to account for cis regulatory elements. The complete MsigDB gene ontology gene sets collection was used as input into the analysis. The gene sets most associated with this study sample have been previously reported (Demontis et al., 2019).

Polygenic Risk Scoring

From the associations collected from gene set analysis, we selected the most associated gene sets based on their p-values. To avoid including the same risk signal multiple times within a score, we adjusted SNPs tagging each gene set for linkage disequilibrium using PRS-CS, a tool that infers posterior effect sizes of each SNP after removing overlaps due to linkage disequilibrium. This method was used instead of removing SNPs by clumping and thresholding so that we could retain all SNPs in the analysis without inflating results due to linkage disequilibrium. From these linkage disequilibrium-adjusted SNPs effect sizes, we used polygenic weighted scoring of all SNPs with an association p-value less than or equal to 0.5 in the training subset to generate a risk profile for each gene set in each subject using Plink. We calculated genome-wide polygenic risk profiles using the same combination of PRS-CS and Plink scoring. For comparison, we also generated PRS using the clumping and thresholding method. The Plink parameters used for the clumping algorithm were: --clump-p1 1.0 –clump-p2 1.0 –clump-kb 250 –clump-r2 0.1.

Correlated Trait/Disorder Polygenic Risk Scoring

We calculated additional risk profiles using SNP effects estimated from GWASs of disorders and traits with the highest genetic correlation with ADHD and heritability over 0.1 found using GWAS Atlas (Watanabe et al., 2019). After excluding similar phenotypes based on study size, the included phenotypes were age at first sexual intercourse(Watanabe et al., 2019), opioid use(Wu et al., 2019), college completion(Rietveld et al., 2013), childhood IQ(Benyamin et al., 2014), childhood extreme obesity(Riveros-McKay et al., 2019), autism spectrum disorder(Grove et al., 2019), time spent watching television(Watanabe et al., 2019), psychiatric cross-disorder risk(Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013), intracranial volume(Adams et al., 2016), and myopia(Watanabe et al., 2019). We calculated the gsPRS for genetically correlated disorders on the gene sets most associated with ADHD diagnosis by using the SNP effects from the summary statistics for each trait in additional MAGMA gene set analyses. We included PRS and 100 gsPRS for each of the 10 phenotypes in machine learning feature selection.

Machine Learning Preprocessing and Feature Selection

We adjusted each polygenic risk score for ancestry by extracting the top 5 principal components from a principal components analysis (PCA) of the training subset and using those 5 principal components in a generalized linear model predicting each polygenic risk score. We replaced the unadjusted polygenic risk score with the residual of each prediction using the 5 principal components. We normalized each score between 0 and 1 using min-max normalization and balanced cases and controls in each subset by random case up-sampling with replacement.

For gsPRS only models, we started by selecting gsPRS from the 40 gene sets most associated with ADHD within the training subset. The 40 gene set inclusion value was selected based on manual tuning in which we looked at ranges of starting gene set values between 10 and 400. We optimized the hyperparameters of a random forest based on this initial set of features. Then, we performed a random iterative feature selection process in which we kept and recorded the most important features, based on the permutation feature importance calculated from the mean difference in Gini impurity, and randomly replaced the less important features with a different gsPRS feature until the model found a set of gsPRS that outperformed the previous best set. We reoptimized the random forest hyperparameters after 3300 intervals, a number such that each gene set would likely be used in three models. We repeated this reoptimization strategy until the set of gsPRS was stable. At the end of this process, we selected the best group of 20 gsPRS based on random forest importance score for model performance evaluation.

In the models that included gsPRS and PRS-CS, we used the same random iterative feature selection approach used in the gsPRS only model, but also included the genome-wide PRS-CS scores calculated from the training subset and summary statistics from GWAS of related disorders in every model.

Machine Learning Model Optimization

Within Scikit-learn, we used grid search optimization to select the best hyperparameters for all models using the AUC in the validation subset (Pedregosa et al., 2011). We optimized multiple types of models to better compare the performance of different methods within the validation subset and select the best model for this application. Exploring multiple models is essential given that for any given problem, one algorithm may be ideal, but it is not possible to know in advance what algorithm will be best (Wolpert & Macready, 1997).

Model Performance Evaluation and Feature Importance Tracking

To measure the performance of the models selected with grid search optimization we used area under the receiver operating characteristic curve (AUC) in the test subset. Data leakage is a common issue in machine learning research normally caused by inadvertently learning information about the test data that improves performance in those specific data. One way data leakage can occur is through testing many models on the test data, which increases the chance of selecting a model that is randomly configured in a way that is more optimal for the test data but not generalizable. With the goal of minimizing data leakage that might bias our results towards the test data, we tested model performance in the test subset only on the model with the highest AUC in the validation subset for each analysis. We estimated the known genetic variance explained by each of the models using a formula developed for the genetic interpretation of AUCs using 0.75 as the heritability estimate and 0.05 as the prevalence estimate (Wray et al., 2010). We compared AUCs from different models using DeLong's test for two correlated ROC curves. We also tested the probability of achieving the AUC in the best gsPRS grouping by comparing the AUC in the test subset with the distribution of AUCs from 10,000 models with random gsPRS groups of the same size. All models included the PRS for all correlated phenotype summary statistics. We used learning curves to model whether additional training examples would improve model performance and to compare models.

To calculate a more generalizable importance score for each gsPRS outside of the best group of gsPRS, we estimated feature importance for each gsPRS and PRS-CS feature in RF models with a random group of gsPRS calculated from gene sets associated with ADHD and tracked the permutation feature importance that measures the decrease in model performance when a single feature value is randomly shuffled. We calculated the mean feature importance of each gsPRS across 10,000 models that used 40 random gsPRSs each. We did not use feature importance scores calculated from the test subset for feature selection or any optimization.

Testing Biological Relevance of gsPRS Feature Importance

To further validate our methods by testing for correlations with the known neurobiology of ADHD, we computed correlations between tissue-specific gene expression and feature importance (Faraone, 2018; Faraone & Biederman, 1998). For ADHD, we would expect that most gene sets truly associated with the disorder would be more relevant to the brain and less relevant to other tissues. Therefore, if the importance of the gsPRS generated in our analysis are correlated with brain expression relative to all other tissues, we can be more confident that gsPRSs are collectively picking up a real generalizable risk feature instead of modelling random noise. We used a dataset containing gene expression data for 54 tissue types from the genotype-tissue expression (GTEx) project. We combined this gene expression data into gene set expression data for the same gene sets used in the gsPRSs. We estimated relative gene set expression in the brain using the Preferential Expression Measure formula which estimates how different the expression of a gene is relative to the expected expression level. We fit a linear model predicting gene set expression in brain tissues relative to non-brain tissues using the MAGMA gene set association p-value to establish a baseline. We fit a linear model using the base model with gsPRS feature importance as a second predictor to test for association of gene set expression with gsPRS importance score after controlling for MAGMA gene set associations. We fit linear models with the same dependent and independent variables using only gene sets calculated from the ADHD training subset or from the group of correlated phenotypes to test whether each group was independently associated with gene set expression. To test whether the association between mean importance score and relative gene set brain expression in the brain was dependent on whether the gsPRS was calculated from the ADHD training subset, we estimated predictive margins using STATA16's margins command, which computes the average probability for each observation at a fixed level of a selected variable. In our analyses, these predictive margins estimate the average relative gene set expression in the brain for each gsPRS while fixing the ADHD vs non-ADHD variable to each value. A meta-analysis on subcortical brain volume differences in ADHD found that the volumes of the accumbens, amygdala, caudate, hippocampus, and putamen were smaller in participants with ADHD (Hoogman et al., 2017). We fit linear models predicting gene expression in these brain regions implicated in ADHD relative to all other brain regions with gene set expression as the dependent variable and MAGMA gene set association p-value and gsPRS feature importance.

Results

Model performance

To establish baseline performance, we measured the prediction performance in the test subset of a logistic regression with the PRS calculated from the training subset. This PRS only logistic regression had an AUC of 0.62 (95% CI: 0.60 – 0.64) in the test subset and explained 5.0% of the known genetic variance. Replacing PRS with PRS-CS in another logistic regression model led to an AUC of 0.66 (Figure 1; 95% CI: 0.64 – 0.68) and explained 9.0% of the known genetic variance. We then measured the performance of logistic regression and random forest models containing the PRS-CS from the training subset and PRS-CS calculated from summary statistics from phenotypes with a heritability above 0.1 with the highest genetic correlation to ADHD (Table 1). The logistic regression model had an AUC of 0.66 (95% CI: 0.64 – 0.68) while a random forest model using the same input had an AUC of 0.69 (Figure 1; 95% CI: 0.67 – 0.71) in the test subset and explained 12.8% of the known genetic variance.

After using our feature selection method to select the best group of 20 gsPRS, we trained a random forest model using the selected group and all PRS-CS. In the test subset, this model had an AUC of 0.72 (Figure 1; 95% CI: 0.70 – 0.74) and explained 17.4% of the known genetic variance. This was a significant improvement in comparison to the RF that included only the PRS-CS from each trait (p = 0.0057, DeLong's test for two correlated ROC curves). The RF model with all PRS-CS and the best group of 20 gsPRS also had a significantly higher AUC (p = 1.2 x 10^{-6}, Delong's test for two correlated ROC curves). compared to a lasso model fit with all PRS-CS and gsPRS as input, which had an AUC of 0.65 (95% CI: 0.63 – 0.67). The AUC of the best group model was greater than 99.6% of the 10,000 random group models. The mean AUC of the random group models was 0.69. All the gene sets used in the random groups were associated with ADHD in the training subset with a p-value of less than 0.05 without correction for multiple testing.

We trained and optimized another random forest model using only gsPRS. The model had an AUC of 0.61 (95% CI: 0.59 – 0.63) in the test subset. The AUC of the best group model was greater than 99.1% of the 10,000 random group models.

Random Forest Learning Curve and Feature Importance Analyses

For the best random forest model, we generated a learning curve (Figure 2) that plots the AUC against the number of training examples (Perlich, 2010). We also optimized a random forest model using only PRS-CS and generated a learning curve (Figure 3) for comparison.

Using the optimized random forest model, we generated feature importance scores in the test subset for all the features used in the model. The most important features and their importance scores are listed in Table 2.

Testing Biological Relevance of gsPRS Feature Importance

The base linear model we fit with relative gene set expression as the dependent variable and MAGMA gene set association p-value as the independent variable showed a significant negative correlation between the two variables (p = 1 x 10^{-5}). The model adding mean gsPRS importance score as an independent variable showed a significant positive correlation between mean gsPRS importance score and relative gene set expression after controlling for MAGMA gene set association p-value ($p = 2 \times 10^{-4}$). We found no significant differences in gene expression between brain regions implicated in ADHD and other brain regions.

The base + mean gsPRS feature importance model we fit using only gsPRS calculated from the ADHD training subset showed a significant positive correlation between mean gsPRS importance score and relative gene set expression in the brain ($p = 0.008$). The same model fit using only gsPRS calculated from the correlated phenotypes also showed a significant positive correlation between mean gsPRS importance score and relative gene set expression in the brain ($p = 0.003$). An additional linear model we fit adding an independent variable specifying whether the gsPRS was calculated in the ADHD training subset or a correlated phenotype and that variable's interaction with importance score showed that the correlation of gene set expression in the brain with mean gsPRS importance was negatively dependent on whether the gsPRS was calculated in the ADHD training subset ($p = 5 \times 10^{-4}$). As illustrated in Figure 4, our predictive margins analysis of this interaction estimated a significant positive association with a slope of 4.7 ($p < 0.001$) when the variable indicating development in the ADHD training subset was fixed to 0, meaning the gsPRS was developed using one of the correlated phenotypes, and a significant positive association with a slope of 0.60 ($p = 0.015$) when the same variable was fixed to 1, meaning the gsPRS was developed using the ADHD training subset.

Discussion

This study is the first to produce gene set specific risk profiles predicting the presence/absence of a psychiatric disorder with machine learning. The addition of optimized groups of gsPRS to genome wide PRS-CS significantly improves prediction performance compared to both models without gsPRS and models with random groups of gsPRS. We further validated these results by testing for biological correlation of the random forest importance scores, which showed that importance scores were significantly positively associated with increased relative gene set expression in the brain.

Compared to simpler models that rely on a single PRS value per individual and more complex models that rely on “black-box” dimension reduction methods, gsPRS models have the potential of offering more interpretability and have the possibility to shed light on mechanisms involved in risk prediction and test specific gene set hypotheses. To improve interpretation of our models, we generated two sets of feature importance measurements that capture similar, but distinct information regarding the predictiveness of gsPRS. The feature importance measurements from the best group of gsPRS (Table 2) show how useful each gsPRS and PRS-CS were in that specific model. Unsurprisingly, the ranking is led by the PRS-CS from a cross-disorder GWAS that studied the shared risk across multiple psychiatric disorders including ADHD and the PRS-CS calculated from the training subset. Those PRS-CS are followed by a group of gsPRS that collectively led to significant improvements in prediction. It is likely that this group contains less overlapping risk information relative to other groupings since such overlaps would increase model complexity without adding value for prediction. However, overlapping gsPRS could still be important individually or in different groupings. Therefore, we calculated average gsPRS feature importance in 10,000 models that each used 40 gsPRS as input. This average represents how often and how strongly each gsPRS was able to improve prediction.

With this list of gsPRSs and their feature importance, we sought to further validate our methods by testing for correlations with what is known about the neurobiology of ADHD (Faraone, 2018; Faraone & Biederman, 1998). Our baseline regression analysis found a significant negative correlation between relative gene set expression in the brain and MAGMA gene set association p-value. This met our expectation since MAGMA is a widely used tool and we would expect that gene sets more associated with ADHD and correlated phenotypes would be correlated with increased relative expression of that gene set in the brain, which is consistent with the report of

Demontis et al (2019). Our analysis adding mean gsPRS importance score to the baseline regression analysis found that, even after correcting for MAGMA gene set association, mean gsPRS importance score was significantly positively correlated with relative gene set expression in the brain. This suggests that the mean gsPRS importance scores can be used to select biologically relevant gene sets beyond their association with ADHD as calculated using MAGMA. This finding suggests that combining MAGMA and mean gsPRS importance scores could provide a better way to prioritize gene sets for future study compared with using MAGMA alone.

We were also interested to test whether the correlations between relative gene set expression in the brain and mean gsPRS importance scores were dependent on whether the gsPRS was calculated using the ADHD training subset or from summary statistics of the correlated phenotypes. In both groups (ADHD and correlated phenotypes), the correlation between gene set expression and importance scores remained significant but the correlation of relative gene set expression in the brain and gsPRS importance score was stronger when the gsPRS was developed using summary statistics from correlated phenotypes (see Figure 4). This finding may seem counterintuitive, considering that most of the gsPRS from other phenotypes had low gsPRS importance scores relative to the gsPRS calculated using the ADHD training subset. However, when a gsPRS from a correlated phenotype is predictive in ADHD that gene set has shown an association and importance in its initial study, the ADHD training subset in our study, and the ADHD test subset in our study. We find it unsurprising that gsPRSs calculated from such generalizable gene sets would be more likely to represent true risk signals and therefore be more likely to have increased relative gene set expression values in the brain.

The learning curves suggest that the performance improvements from gsPRS should increase with increasing sample size. More complex models generally require more data to train, as demonstrated by the early stages of the learning curve that show perfect training subset performance and no predictability in the validation subset as the model is complex enough and sample size is low enough to memorize the training data instead of learning patterns among those data. In both learning curves, it is evident that the model is better at predicting the training data compared to the validation data even after selecting hyperparameters that specifically maximize prediction in the validation subset. This further illustrates the importance of testing performance on data the model does not learn from during training to get an accurate representation of model performance and generalization. As training size increases, the model can no longer rely on memorization and starts to learn patterns that generalize to the validation subset. The continued validation subset prediction improvements at the highest training sizes suggest that the model could still benefit from more training data. In comparison, the learning curve of a random forest model using only PRS-CS shows a quick plateau to optimal performance and additional training size does not further improve performance.

Our study has several limitations that could limit performance. To best estimate model performance and reduce overfitting, we split our data into several subsets, thereby limiting the number of study participants available to train the models. We also adjusted for the effects of the top 5 principal components in a PCA of the training subset to control for ancestry. This adjustment could inadvertently remove non-confounding information that might have improved performance and likely does not remove all ancestry information. A better method of selectively removing known confounders like ancestry would likely further improve both the performance and generalizability of these models. The gene sets we used to sum sets of SNPs into gsPRS, although capture the biological functions and pathways, may not be ideally suited for prediction tasks. A more data-driven approach to develop sets of SNPs that best collectively predict diagnosis may be necessary to maximize prediction performance.

More advanced machine learning methods and architectures may also lead to more predictive models. Including data beyond genotype information like clinical data and data that captures at least a portion of the environmental component of ADHD pathology could help machine learning models better estimate ADHD risk and better separate ADHD cases and controls. With the right set of interpretation tools, models that can accurately discriminate ADHD cases and controls would be useful in improving our understanding of the disorder and allow for testing specific hypotheses.

Figure 1. Model ROC Comparison. The logistic regression model using traditional PRS methods had an AUC of 0.62 (95% CI: 0.60 – 0.64). The logistic regression model using PRS-CS methods had an AUC of 0.66 (95% CI: 0.64 – 0.68). The random forest model using PRS-CS and an optimized group of 20 gsPRS had an AUC of 0.72 (95% CI: 0.70 – 0.74).

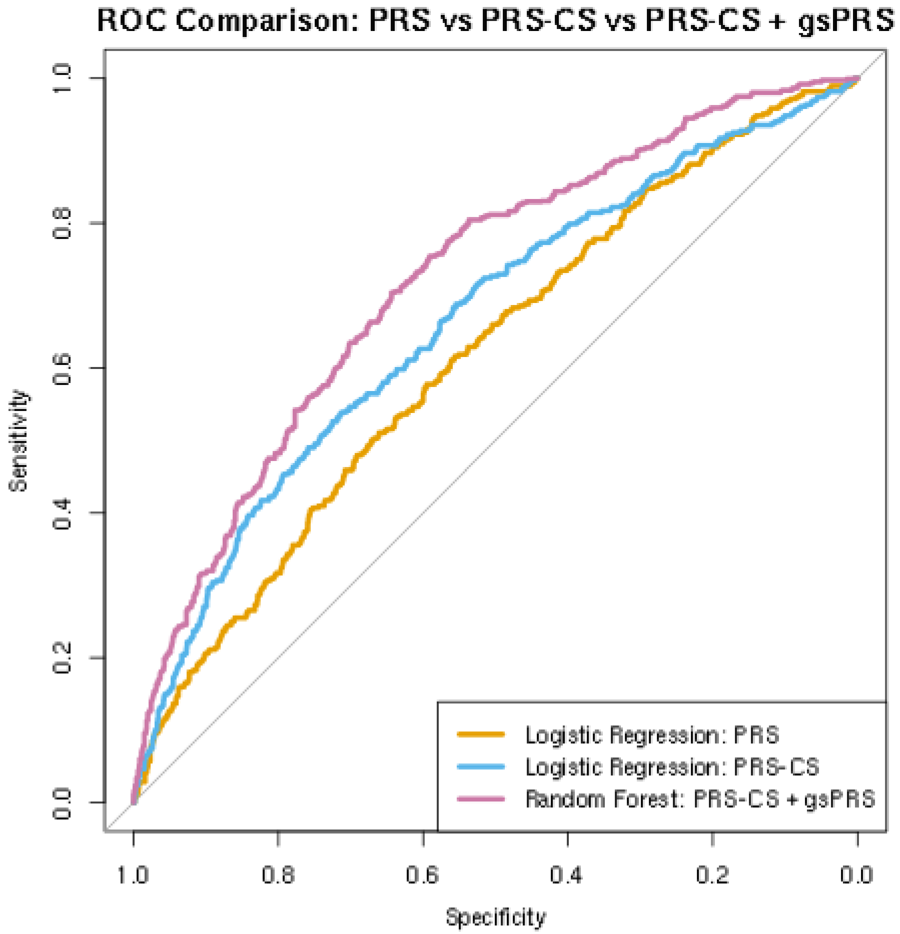


Figure 2. Learning Curves for gsPRS + PRS-CS Random Forest Model. The learning curve analysis of the random forest model containing all PRS-CS and the best group of 20 gsPRS. Each point represents the accuracy of the model when trained with a set number of training examples.

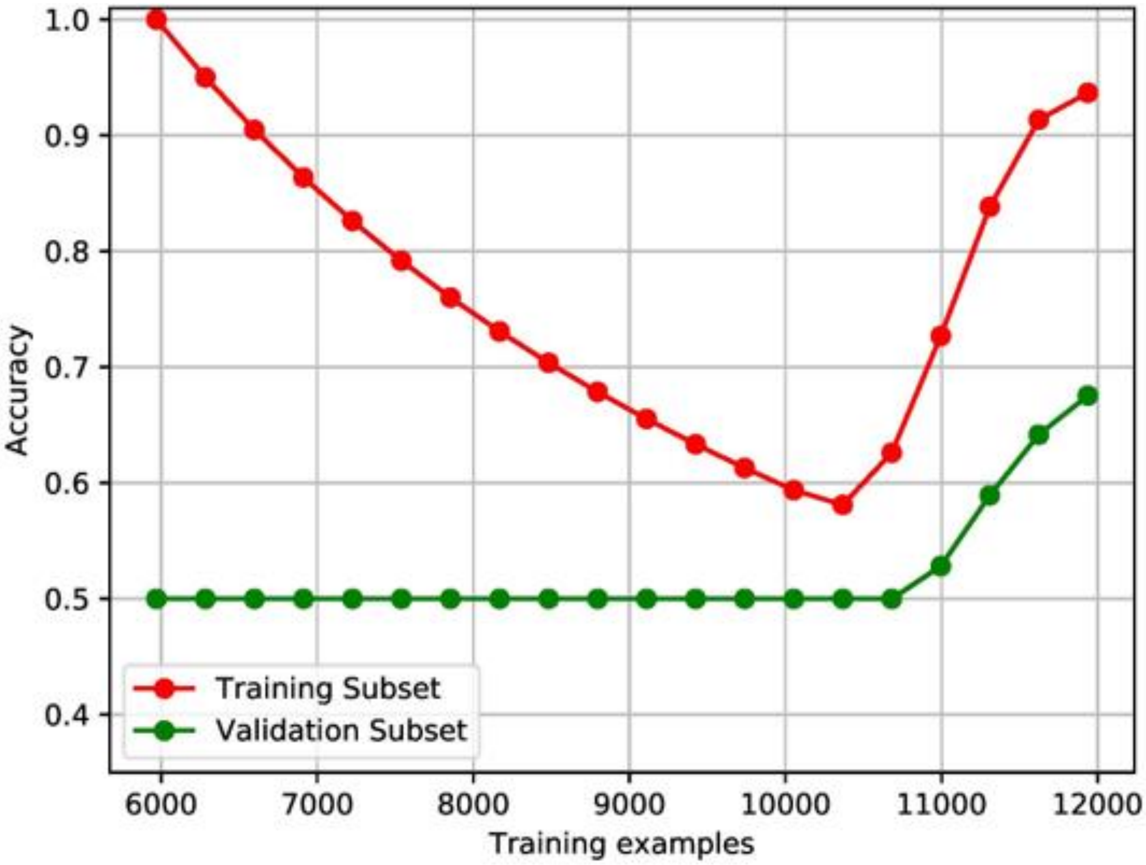


Figure 3. Learning Curves for PRS-CS only Random Forest Model.

The learning curve analysis of the random forest model containing all PRS-CS. Each point represents the accuracy of the model when trained with a set number of training examples.

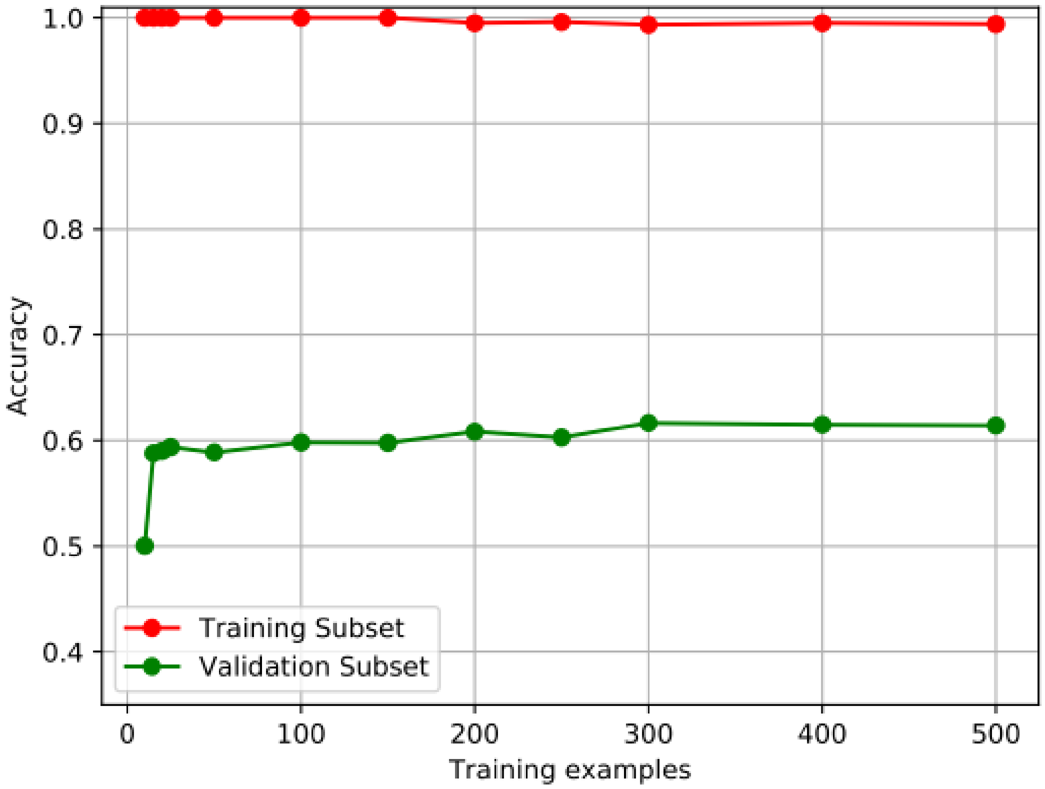


Figure 4. Predictive Margins Analysis of Interaction between Importance Score and Relative Gene Set Expression. When the variable indicating gsPRS development in the ADHD training subset was fixed to 0 (developed in a correlated disorder) there was a significant positive association with a slope of 4.7 (p < 0.001). When the variable was fixed to 1 (developed in the ADHD training subset) there was a significant positive association with a slope of 0.60 (p = 0.015).

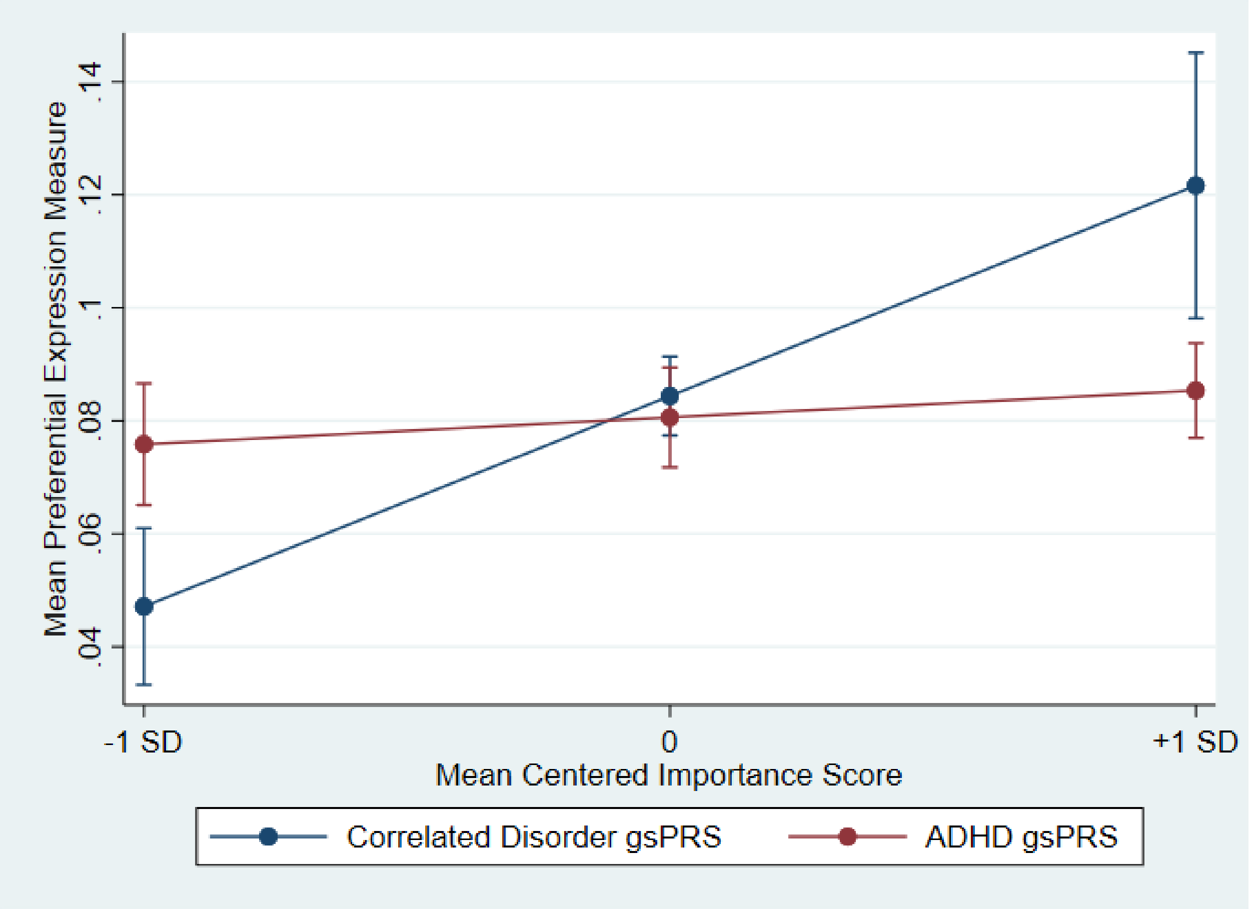


Table 1. Genetically correlated phenotypes for which external GWAS summary statistics were used to generate additional genetic risk features.

Phenotype	genetic correlation	N	SNP.h2
Age at first sexual intercourse	-0.584	339614	0.1132
Opioid use	0.565	78808	0.146
College completion	-0.524	95427	0.105
Childhood IQ	-0.461	12441	0.2744
Extreme obesity (childhood)	0.436	7916	0.5078
Autism spectrum disorder	0.384	46350	0.1944
Time spent watching television (TV)	0.372	365236	0.1023
PGC cross disorder	0.262	61220	0.1715
Intracranial Volume	-0.248	26577	0.2467
Myopia	-0.217	78647	0.1532

Table 2. Top feature importance scores for the best group of gsPRS and PRS random forest model.

Feature	Phenotype	Importance
Genome Wide Polygenic Risk Score	pgc cross disorder	0.0324
Genome Wide Polygenic Risk Score	training subset	0.0250
GO_CELLULAR_RESPONSE_TO_ENDOGENOUS_STIMULUS	training subset	0.0055
GO_POSITIVE_REGULATION_OF_PROTEIN_MODIFICATION_PROCESS	training subset	0.0044
GO_RESPONSE_TO_TOXIC_SUBSTANCE	training subset	0.0032
GO_REGULATION_OF_PRESYNAPSE_ORGANIZATION	training subset	0.0027
GO_REGULATION_OF_PLASMA_LIPOPROTEIN_PARTICLE_LEVELS	training subset	0.0026
GO_POSITIVE_REGULATION_OF_TRANSFERASE_ACTIVITY	training subset	0.0026
GO_PRESYNAPSE_ORGANIZATION	training subset	0.0025
GO_SYNAPTIC_SIGNALING	pgc cross disorder	0.0024
Genome Wide Polygenic Risk Score	age first had sexual intercourse	0.0022
GO_REGULATION_OF_VASCULAR_PERMEABILITY	training subset	0.0019
GO_PRIMARY_ALCOHOL_BIOSYNTHETIC_PROCESS	training subset	0.0018
GO_HEAD_DEVELOPMENT	pgc cross disorder	0.0017
GO_LIPASE_ACTIVITY	training subset	0.0017
GO_REGULATION_OF_NEURON_DIFFERENTIATION	pgc cross disorder	0.0014
GO_MICROTUBULE_PLUS_END_BINDING	training subset	0.0013
GO_NEURON_DIFFERENTIATION	pgc cross disorder	0.0012
GO_CEREBELLAR_GRANULAR_LAYER_FORMATION	training subset	0.0011
GO_SOMATODENDRITIC_COMPARTMENT	pgc cross disorder	0.0009
GO_POSITIVE_REGULATION_OF_EXCITATORY_POSTSYNAPTIC_POTENTIAL	training subset	0.0008
GO_REGULATION_OF_MEMBRANE_POTENTIAL	pgc cross disorder	0.0008

Classifying Type 2 Diabetes Mellitus using Genomic Context Informed Genotype Data and Within-model Ancestry Adjustment

Eric Barnett[1], Stephen V Faraone[1,2]

Affiliations

[1]Department of Neuroscience and Physiology, SUNY Upstate Medical University, Syracuse, New York, USA

[2]Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, New York, USA

Abstract

Background: The development of a high-risk prediabetes group has led to better prevention of type 2 diabetes mellitus (T2D). Despite high heritability estimates, no genetic data are used in establishing this high-risk group. This is in part because accuracy of current genetic classification models for T2D is lower than expected given the disorder's heritability.

Methods: In a matched pairs study of 78,528 participants from the UK Biobank, we built a context informed data matrix that included genomic annotation information and other genomic context for a set of genetic variants. In a novel analysis, we used this data with genomic context as input to convolutional neural networks and compared several model architectures with adversarial tasks to test hypotheses on the usefulness of genomic context and adjustment for confounding due to ancestry.

Results: An ancestry adjusted neural network using genotype data (AUC: 0.66) and an ancestry adjusted convolutional neural network using context informed genotype data (AUC: 0.65) both outperformed an ancestry adjusted polygenic risk score approach (AUC: 0.57) in classifying type 2 diabetes. Adversarial ancestry tasks eliminated the predictability of ancestry without changing model performance.

Conclusions: Our results suggest that context informed genotype and standard genotype data input can both be useful in classifying T2D and can find genomic risk features specific to each input type. Within-model adjustment for ancestry shows promise in eliminating confounding while retaining genetic risk information.

Introduction

It is estimated that 6.3% of the world’s population has type 2 diabetes mellitus (T2D), a prevalence that has steadily increased since the center for disease control and prevention (CDC) began tracking in 1958 (CDC Division of Diabetes Translation, 2017; Centers for Disease Control and Prevention, 2022). To prevent diabetes, a high-risk group called prediabetes was developed, which includes people that do not meet the criteria for diabetes but have the presence of IFG and/or IGT and/or A1C 5.7-6.4%(American Diabetes Association Professional Practice Committee, 2021). Multiple studies have shown that interventions in this high-risk group are effective in decreasing the incidence of diabetes in the group (Knowler et al., 2002; Knowler et al., 2009; Ramachandran et al., 2006; Tuomilehto et al., 2001). One risk factor that is regularly not included in establishing the high-risk group is the genetic risk of an individual, despite heritability estimates that range from 25% in shorter term follow-up studies to 80% in longer term follow-up studies (Prasad & Groop, 2015).

Genome-wide association studies of T2D have been successful in identifying hundreds of genetic variants that are significantly associated with the disorder (Mahajan et al., 2022; Xue et al., 2018). Studies using machine learning models to classify people with and without T2D have been comparatively less successful, with reported model performances that are lower than expected relative to the heritability of T2D and the classification performance of diseases with similar heritability (Abraham et al., 2013; Evans et al., 2009; Mittag et al., 2012). This lower-than-expected performance could be attributed a variety of factors. One possible factor is that the genetic risk for T2D is more complex, with small amounts of risk spread throughout the genome, compared to disorders like type 1 diabetes mellitus where the risk is largely contained in a single region (Noble & Valdes, 2011). If this were the case, more complex models or more useful input may be necessary to better model the genetic risk.

Increasing machine learning model complexity comes with its own disadvantages, mainly in terms of model generalizability(Whalen et al., 2022). A complex model is more capable of modeling patterns of input features specific to the data that the model is trained on. If those patterns are more effective at classifying correctly compared to the real effects, those real effects may be avoided in favor of the patterns specific to the training data. When this occurs, a model will “overfit” to the training data. Meaning the model will perform well on the data it was trained on, but the performance will drop when testing on unseen data that do not share the same patterns present in the training data.

One concerning example of learning patterns specific to the training data occurs due to population stratification (Freedman et al., 2004). In population stratification, the cases and controls within the training data have systematic ancestry differences. A machine learning model could use those different frequencies of genetic variants between ancestry groups to accurately classify cases and controls, but in reality, the model is only classifying the ancestry groups. When testing such a model on data that do not have the same unbalanced ancestry between cases and controls, the model’s classification performance will decline.

In genome-wide association studies (GWAS) and in models that use polygenic risk scores, population stratification is commonly accounted for by including principal components as covariates in the analyses determining associations and calculating risk scores (Price et al., 2006).

It is thought that most of the ancestry information is contained in the top principal components, so removing those effects from the model allows for the true disease risk to be detected and used. However, it is difficult to know how many principal components are necessary to remove ancestry. In more complex non-linear models, it is unclear whether current linear adjustments are sufficient to stop the model from using ancestry information. Within-model adjustment of ancestry may be more effective at eliminating confounding by ancestry.

One cause of overfitting complex models to training data is the number of parameters within a model (Ying, 2019). One study found that many of the most popular machine learning models can almost perfectly classify data with random labels within training data due to the huge number of parameters used within the models compared to the size of the training data (Zhang et al., 2017). This is of particular importance in genomic machine learning models, where the number of genetic variants available within many data sets can be in the millions while the number of study participants are normally orders of magnitude lower. One common approach to this issue is limiting the number of genetic variants used in the model(Ying, 2019). This can be effective in reducing overfitting, but in complex genetic disorders may limit the maximum model performance if genetic variants that influence risk of the disorder are removed.

In this article, we use convolutional neural networks (CNNs) to limit the number of parameters without losing information from removing genetic variants. The CNN’s capacity to learn local patterns and reduce dimensions has made it popular in the image recognition and classification field. Several studies have had success using CNNs on genetic data, largely focusing on disease classification or annotation prediction based on genotype values (Waldmann et al., 2020; Zhou & Troyanskaya, 2015). One area that has not been investigated is whether CNNs, which specialize at finding local patterns, can be used alongside genetic annotations to create a representation of the genomic context surrounding genetic variants in disease classification models.

Given the strengths of CNNs and availability of information about the genome, we sought to incorporate the two in models that search for T2D genetic risk patterns based on the combination of genotype data and the nearby and correlated genetic annotations. Our analytic strategy aimed to determine if training a novel CNN informed by genomic context could improve classification performance and if the patterns learned by such a model were different than those obtained in models without genomic context. We also sought to test if using gradient reversal layers to establish adversarial multi-task networks could control for confounding by ancestry within the models.

Methods

Data Acquisition and Preprocessing

We obtained genotype and phenotype data for UK Biobank (Sudlow et al., 2015), a prospective cohort study of over 500,000 individuals in the United Kingdom. The genotype data were generated by a combination of the UK Biobank Axiom array and UK BiLEVE Axiom array. In the resulting genotype data, 1,037 sample outliers, multi-allelic single nucleotide polymorphisms (SNPs), and SNPs with a minor allele frequency (MAF) < 1% were removed, resulting in 641,018 SNPs (Marchini, 2015). These SNPs were used to impute ungenotyped SNPs, leading to a dataset of 73,355,667 variants. For our analyses, we removed all SNPs with MAF < 1%, all SNPs with > 5% missingness, all individuals with > 5% missingness, and one member of any estimated kinship equal to or closer than second-degree relatives using Plink. We randomly split the data into a training subset (70%), validation subset (15%), and testing subset (15%). In each subset, we performed 1:1 case-control matched pairing based on age and sex using the MatchIt package in R (Ho et al., 2011; R Core Team, 2014). The resulting training subset had 55,168 subjects, the validation subset had 11,712 subjects, and the test subset had 11,648 subjects.

Context Informed Data Matrix Construction

Model Overview and Input: We trained several models with differing architectures and input to test hypotheses. Simplified diagrams of model architectures, inputs, and outputs are shown in Figure 1. An overview of the models used in our analysis is shown in Table 1. The genotype input used in our NN models is the number of alternate alleles at each SNP for each person. The additional input used in our CNN models is a context informed data matrix (CID) that contains each SNP along with information about that SNP; columns in a CID represent each SNP and rows contain genomic annotations and risk values at each SNP. Figure 2 shows a simplified example of a CID. For each individual, the CID, which largely contains information about each SNP that does not change between individuals, is multiplied by the genotype values of the individual to get an individualized-CID for model training. For all T2D classification models, we used a binary variable representing the presence of the ICD-10 code for T2D, E11, in each individual as the label. Our models output values from 0 to 1 with the goal of minimizing the error in predicting the label for each individual.

Genomic Annotation: The genomic annotations provided information on whether the SNP occurs at a location known to be an miRNA, DNASE hypersensitivity site, CPG island, gene, intron, 5’ UTR, 3’ UTR, splice site, promotor, transcription factor binding site using the AnnotationHub (Morgan & Cauce, 1999) and VariantAnnotation (Obenchain et al., 2014) packages in R. We did this by loading the SNP locations and annotation ranges, and then finding the overlaps between the two. Any SNP that had an overlapping location with an annotation was coded as a 1 for that annotation or was otherwise coded as a zero. We included whether the SNP was a coding variant using the same overlap method.

We also added annotations indicating whether the SNP was within the range of each of the 20 gene sets most associated with T2D among gene sets in the lowest 10% standard error in MAGMA gene set analysis (de Leeuw et al., 2015). We used the lowest 10% standard error threshold to ensure that the resulting gene set annotations were relatively denser due to using larger gene sets and gene set association was consistent in the training subset as determined by MAGMA gene set analysis. The results of MAGMA gene set analysis on the training subset can be found in the supplementary results.

Risk Values: To give the model information about documented disease risk associations, we added the log odds ratios for each SNP from an external GWAS on T2D to the CID (Mahajan et al., 2022). Only SNPs present in both the UK Biobank data and the external GWAS were included in the model. To reduce computational cost of this analysis, we further reduced the number of SNPs for ML models by only included SNPs that were associated with T2D in the external GWAS at $p \leq 0.01$. This reduction left us with 11,730 SNPs for machine learning models. We also included odds ratios from correlated disorders based on genetic correlations calculated by GWAS Atlas (Watanabe et al., 2019). From the list of the 100 most correlated traits with T2D, we selected traits that had ICD-10 codes available in the UK Biobank data. The included traits were overweight and obesity, disorders of lipoprotein metabolism and other lipidemias, essential hypertension, chronic ischemic heart disease, cholelithiasis, angina pectoris, other disorders of the urinary system, and pain in throat and chest. We performed GWASs on these traits using genotyped UK Biobank individuals that were not included in our machine learning models. The top 10 principal components (PCS) were used as covariates in the model to adjust for population stratification. We included the log odds ratios calculated from these GWASs in the CID.

Correlation to Annotation: It is likely that, for most SNPs, the risk associated with the SNP is due to another genetic change that is associated with the SNP through linkage disequilibrium. To account for this possibility, we added an annotation for each previously described binary annotation that indicates the maximum correlation each SNP has to a SNP with the binary annotation. Correlations were calculated using Plink (Purcell et al., 2007). In cases where the SNP has the annotation, the correlation is 1 and the resulting annotation value is identical to the original annotation value. These annotations representing how correlated each SNP was to a SNP with each genomic annotation replaced the original genomic annotations in the CID.

Model Hyperparameter Optimization

Within TensorFlow (Abadi et al., 2016), we used the KerasTuner (O'Malley, 2019) framework to optimize the hyperparameters of our models. Using the Hyperband search algorithm, we searched for the optimal number of genomic convolutional blocks, number of filters within each block, filter and pool width within each block, number of dense layers, number of nodes and dropout rate within each layer, gradient reversal weights, L2 regularization presence and factor, number of epochs, and learning rates. The objective of the search algorithm was to maximize the area under the receiver operating characteristic curve (AUC) in the validation subset for classification tasks and maximize validation subset R^2 for regression tasks.

Detecting and Adjusting Ancestry Confounding with Principal Components and Adversarial Learning

ML models use all information available to them to produce the best performance possible. This can result in models that perform well mainly due to a confounding variable. As previously described, one common confounder is ancestry. To reduce our model's capability to use ancestry to predict T2D status, we ran a principal components analysis using the SNPs that were not included in our ML models nor correlated with the SNPs ($r^2 < 0.2$) that were used in the ML models. We extracted the top 10 PCs, as is frequently used for ancestry inference (Price et al., 2006), and used them as labels in several models.

To establish whether population stratification was an issue in our data, we built a model that used the ancestral principal components to predict T2D status (Model 2a). To determine if the subset of SNPs used as predictors in our model could recreate the principal components within the model, we used the genotype data as input in a neural network model that predicted the PCs

(Model 2b). We determined the effectiveness of all model architectures in estimating the PCs with mean squared error (MSE) and R^2. In models that had a positive R^2 value, we also tracked T2D classification AUC from the PC estimates.

To test if a neural network could classify T2D and estimate PCs using the same layers, we designed a multi-task model with both tasks (Model c, Figure 1c). All layers besides output layers were shared between the tasks. An additional task classifying T2D from the PC estimates was used to compare classification performance between the PC estimates and the true PCs.

To test whether, in classifying T2D from genotype data, our model inadvertently uses PC-like ancestral information, we built a multi-task model where one task classified T2D from the genotype input and another task estimated PCs (Model d, Figure 1d) from the output of the dense layers. In this model, we used a stop gradient layer between the PC estimation task and the dense layers, which stop backpropagation. By using this layer, we stopped training such that the task estimating PCs did not influence the dense layers upstream from the layer unique to their tasks. This strategy allows us to track the performance of a task, in this case PC estimation, without interfering with shared layer training. If the PC estimation task was predictive, this would suggest the model classifying T2D from genotype data used ancestral information that could be directly transformed into PC estimations.

To test whether we could teach the network not to use ancestry information when predicting T2D, we built a multi-task model (Model e, Figure 1e) with the same two tasks as the previous model but replaced the stop gradient layer with a gradient reversal layer, a technique initially developed in the domain adaptation field of machine learning (Ganin et al., 2016). This layer reverses the direction of gradients in gradient descent, thereby directing the weights in the layers upstream of the gradient reversal layer to adjust in a way that maximizes the error in the task, instead of the typical minimization. By using this gradient reversal layer, we create an “adversarial task”, which directs the model to find patterns that are ancestry invariant by penalizing classification strategies that lead towards more accurate estimation of ancestral PCs. Layers that are downstream of the gradient reversal layer still adjust weights in a way that minimizes error in the task. This means the adversarial task still attempts to minimize error and accurately estimate PCs in the layers that are unique to the task, thereby leaving adjustment of shared layers as the only option in maximizing error. We used this architecture to remove any PC-like ancestry information from the shared layers. If the model is unable to estimate PCs, it suggests the ancestry information present within the PCs is not being used in the shared layers.

Convolutional Neural Network Model Architecture for Genomic Data

In all our CNN models, we used the matrices as input into genomic convolutional blocks, which are repeating units within the model that contain a combination of convolutional layers with ReLU activation functions, pooling layers, and batch normalization layers. Within convolutional layers, the convolutional filters require spatial invariance, meaning a signal on one part of the two-dimensional data structure means the same thing as the same signal anywhere else in the data structure This is true for images, for which convolutional layers were originally developed. In contrast, the height dimension of the context informed data matrix used as input in our model, which is the annotation information at each SNP location, has no spatial meaning and filters at different heights could mean vastly different things, which violates spatial invariance. To address this issue, we set the height of each convolutional filter equal to the total number of rows present in the matrix to assure that output signals from one part of the matrix were equivalent to those from another part. After each convolutional layer, the output was fed into a batch normalization layer. After the genomic convolutional blocks, the output was used as input into one or more dense layers with ReLU activation functions, depending on hyperparameter optimization. Finally, we used a dense layer with a sigmoid activation function to produce an output prediction of the T2D status for each person. We used the Adam optimizer and binary cross-entropy loss function to train our models.

To test whether genotype only models and context informed genotype models can share the same features, we trained a CNN model that had two input sources and a T2D classification task specific to each input, but shared intermediate layers (Model f, Figure 1f). One input was the CID as previously described. The second input was the genotype data portion of the CID replicated such that it had the same dimensions as the CID. Both inputs shared all layers except the output layer specific to each input/task. If both tasks can classify T2D, it would suggest that there are similarities in the patterns used in the two input types. A PC estimation task with a stop gradient layer was used to track whether ancestry information was used within the model.

To test whether overlaps in the patterns found and used in both tasks are present when only training the CNN with CID input, we trained another CNN model (Model g, Figure 1g). In this model, we used the same base structure as model f, but used a stop gradient layer prior to the output layer of the genotype only model, thereby preventing the task from training layers besides the output layer unique to the task. If the task is still able to classify T2D diagnosis with these constraints, it suggests that at least some of the similarities in the patterns used for both tasks are used even when not directed to find shared patterns.

To test whether the CNN with CID input can find patterns that the genotype only model cannot use to classify T2D, we built a third CNN model (Model h, figure 2h). In this model, the stop gradient layers in model g were replaced with gradient reversal layers. This effectively forces the shared layers away from any patterns that could be used by the genotype only input to classify T2D. Likewise, the shared layers are forced away from using ancestry information that could be used to estimate PCs. If the CID input can classify T2D using the same layers as the genotype only input, it would suggest that the CID input can model patterns that the genotype only input is not able to model. To test whether the CNN with genotype input can find patterns the model with CID input cannot, we built a similar model (Model i, Figure 2i) that switched the main and adversarial tasks of model h.

To test whether models h and i contained non-overlapping T2D risk information we built a model that used a combination of the risk features generated by models h and I to predict T2D. To do this, we concatenated the output from the final intermediate layer in both models. This combined output was used as input into a neural network model (Model h+i) with the task of classifying T2D diagnosis. We compared the performance of this combined model to the performance of models h and i individually.

Logistic Regression Models and Model Comparison

We fit logistic regressions using the PRS from all SNPs and from the same set of SNPs used in the genomic CNN for comparisons. For both sets of SNPs, we used Plink to prune correlated SNPs and calculate polygenic risk scores (PRSs). We adjusted both PRSs for ancestry by regressing the PCs on the PRSs and keeping the residual. We used this residual in the glm function within the base stats package in R to fit a logistic regression using only the training subset. We predicted the test subset T2D status and the performance within the test subset to compare to other models. We used AUC to measure performance in the test subset. AUC confidence intervals were computed with 2,000 stratified bootstrap replicates using the pROC R package (Robin et al., 2011). We also used the pROC package to compare the performance between models with DeLong’s test for two correlated ROC curves.

Results

In the following, "genotype data" refers to the set of genotypes used as input to the models, "GWAS PCs" refers to ancestrally informative PCs estimated from the full set of GWAS SNPs and "ML PC estimates" refers to PCs estimated in machine learning models using genotype data as input.

Table 2 shows the performance of all models on the test subset. The hyperparameter optimization ranges and values are in Table 3. Model a (Figure 1a) had a small but significant level of predictive accuracy, with an AUC of 0.55 (95% CI: 0.54 – 0.56). Model b, which used genotype input to estimate PCs (Figure 2b), which were calculated in a GWAS that did not include the genotype data used in the NN model, had an R^2 of 0.73. The additional task that predicted T2D diagnosis from ML PC estimates in this model had a significant AUC of 0.55 (95% CI: 0.54 – 0.56). The prediction of T2D using GWAS PCs and the prediction of T2D using ML PCs estimates were not significantly different ($p = 0.65$, Delong's test for two correlated ROC curves). The PRS and PC adjusted PRS logistic regressions had AUCs of 0.60 (95% CI: 0.59 – 0.61) and 0.57 (95% CI: 0.56 – 0.58), respectively. These two models were significantly different ($p = 2e\text{-}13$).

Model c (Figure 1c) was significantly predictive in all tasks. The T2D classification task had an AUC of 0.66 (95% CI: 0.65 – 0.67, the estimation of PCs task had an R^2 of 0.62 and MSE of 0.38, and the T2D from PC estimates task had an AUC of 0.56 (95% CI: 0.55 – 0.57). The prediction of T2D from GWAS PCs and ML PC estimates based on genotype input was not significantly different ($p = 0.35$). Model d (Figure 1d) had an AUC of 0.65 (95% CI: 0.64 – 0.66) for T2D classification and an $R^2 < 0$, indicating that the model's predictions are worse predictions than using the mean, and MSE of 1.70 for the PC estimation task. 0Model e (Figure 1e) had an AUC of 0.66 (95% CI: 0.65 – 0.67) for the T2D classification task, while the estimation of PCs task had an $R^2 < 0$ and MSE of 2.32.

Model f (Figure 1f) had an AUC of 0.62 (95% CI: 0.61 – 0.63) for the T2D from CID input task and 0.63 (95% CI: 0.62 – 0.64) for the T2D from genotype only input task. The PCs estimation task with stop gradient layer designed to track but not influence PC usage in the shared layers had a $R^2 < 0$ and MSE of 0.69.

Model g (Figure 1g) had an AUC of 0.65 (95% CI: 0.64 – 0.66) for the classification of T2D from CID input task. The AUC for classifying T2D from genotype only was 0.54 (95% CI: 0.53 – 0.55), which was significantly lower than the T2D from CID input task ($p = 2e\text{-}16$).

Model h (Figure 1h) had an AUC of 0.59 (95% CI: 0.58 – 0.60) for the T2D from CID input task. The classification of T2D from genotype only input task was not significantly predictive (AUC = 0.50). The adversarial PCs estimation task had a $R^2 < 0$ and MSE of 3.30. Model I (Figure 1i) had an AUC of 0.57 (95% CI: 0.56 – 0.58) in its main task of classifying T2D from genotype input. This was a small but significant decrease in AUC compared to the model h (Figure 1h) (p = 0.0002). Model h+i had an AUC of 0.61 (95% CI: 0.60 – 0.62). This was a significant increase compared to both model h (p = 8.4e-4) and model I (p = 1.8e-11).

Discussion

Our results show that using genetic annotations to provide context for common genetic variants provides information not found in prior genomic machine learning models. Our methods using convolutional layers with genomic context are the first of their kind and results suggest that unique risk patterns can be learned through these methods. We report the first use of gradient reversal layers in genomics machine learning research, which we found to be useful in both ancestry adjustment and hypothesis testing. The series of models we tested in our study also provided evidence of confounding due to ancestry and suggest within-model control methods could be useful for directly excluding and tracking ancestry information.

In a large dataset of over 70,000 subjects, a NN model (Model a, Figure 1a) significantly predicted T2D using the top 10 PCs from a PCA in which we used only SNPs that were not included or in linkage disequilibrium with SNPs that were used in estimating our models. We also show that the genotype information that was included in the model accurately estimated the PCs in another NN (Model b, Figure 1b). Taken together, these two results suggest that machine learning models can recreate ancestry information and use those data to classify T2D diagnosis and, likely, other disorders. Model b (Figure 1b) tested this ancestry inference to disorder diagnosis pathway directly and resulted in T2D classification performance not significantly different from classifying directly from GWAS PCs. Model e (Figure 1e) did not perform any worse than Model c (Figure 1c), despite eliminating its ability to estimate PCs and use those ancestry data. It is likely that since the layers within the adversarial model were forced away from using ancestry information, the nodes of the model that would have otherwise been occupied with ancestry inference were instead used to represent additional real risk features of T2D.

It's possible that some models use ancestry information while others do not, depending on model architecture, data structure, or even initialization. However, it is hard to know in traditional approaches when ancestry is influencing a model. Our study provides a pair of solutions for this dilemma. First, we use a subtask that estimates ancestry-adjusted PCs from the output of layers used in the main classification task. This subtask cannot influence the weights of the shared layers due to a stop gradient layer. Without the stop gradient layer, this subtask would have the opposite of our desired effect, as backpropagation would encourage the weights in the layers of the main classification task to change in a way that improves the model’s ability to estimate PCs and thereby use ancestry information. Instead, the task with stop gradient essentially monitors whether ancestry information is used without changing the main network. If ML estimation of PCs is significant with the stop gradient layer, it indicates that ancestry information may be present within the shared layers. In this case, the stop gradient layer can be replaced with an adversarial layer which reverses the direction of the weight changes suggested by gradient descent when entering the upstream layers used by the main task. This would direct the model to maximize errors in PC estimation, thereby avoiding ancestry information. This capability can be seen by comparing the MSE between the models in Figure 1d and 1e. When the stop gradient layer is replaced with a gradient reversal layer it creates an adversarial task that increases the MSE of the PC estimation task from 1.70 to 2.32. Weights of the strength of the adversarial task can be adjusted as needed to eliminate ancestry information. If ancestry estimates beyond PCs are available, those data can be used in the same way to guide models away from using ancestry information. Studies have shown that genetic risk models, which are largely developed on people with European ancestry, do not generalize well to other ancestry groups (Duncan et al., 2019; Martin et al., 2019). Adversarial ancestry tasks may help minimize this discrepancy by directing models towards finding ancestry invariant patterns that should be more consistent across different ancestry groups.

In model f (Figure 1f), we trained a model with genomic context information (CID Input) and without genomic context information (genotype input) to classify T2D. The input types trained and used the same shared layers, apart from output layers that were unique to each input. The classification performance resulting from the two input types were similar. This suggests that the input types can find and use the same patterns within the data to classify T2D. This result was expected since the two input types share the same genotype information. Likewise, models c and g were not significantly different, suggesting that enough information can be learned from genotype only data to match the performance of CID models with the genomic context information included in our analyses.

However, in a different model (Model g, Figure 1g) that only allowed the CID input to train the shared layers, we found that the CID input was significantly better at classifying T2D compared to genotype input that did not train the shared layers (p = 2e-16, Delong's test for two correlated ROC curves). This suggests that while the two input types share similarities in their ability to represent the risk of T2D, CID input can build models that have unique risk representations that cannot estimate risk when using genotype data alone.

Our use of gradient reversal layers to create adversarial tasks in model h and model i further separated out input dependent features of CNN models. While both models' performance declined relative to other CNN models, the main inputs still had AUCs that were statistically significant. In comparison, the alternate input (genotype input for model h and CID input for model i) was unable to significantly classify T2D using the same shared layers used by the main input. This suggests that the risk representations within the models leading to T2D classification were unique to the main input. We found that model h, which was trained to classify T2D from CID input with an adversarial task for genotype input was significantly better at classifying T2D compared to Model i, which trained to classify T2D from genotype input with an adversarial task for CID input. This could mean that fewer or less useful risk representations are unique to genotype input or could represent a failure of hyperparameter optimization to find the best models in the more complex hyperparameter search spaces of these two models. In addition, the optimized gradient reversal weight for the adversarial CID input task (0.0012) was very small compared to the optimized gradient reversal weight for the adversarial genotype input task (0.18). This suggests that the risk representations created by the CID input are harder to find and therefore easier to avoid. One possible explanation of the unique risk representations is that they represent the same underlying risk features but are constructed in a way that can only be used with one input type. To test this theory, we trained model h+i, which combines the features of the final intermediate layers of model h and model i. If the features from models h and i were overlapping, one would expect the model using the features from both models to have the same classification performance. We found that model h+i was significantly better at classifying T2D compared to both individual models. This suggests that the risk representations found in models h and i are at least in part unique to those input types. This could mean that some of the genetic risk for T2D can be transformed into higher order risk features using genomic context (features from model h) while other risk cannot be transformed with the set of genomic context information used in our analyses. Further investigation of the genetic variants and higher order risk features used and developed in these models could reveal insights for future genetic risk modelling efforts.

Type 2 Diabetes has a history of low classification performance in genetic models. The performance of our NN and CNN models improve on prior results. One study reported an AUC of 0.60 when using a gradient boosted and LD adjusted heuristic polygenic score and an AUC of 0.61 when using LDpred polygenic scoring (Paré et al., 2017; Vilhjálmsson et al., 2015). The same study reported an AUC of 0.58 in an LD unadjusted polygenic risk score and an AUC of 0.58 when using traditional pruning and thresholding polygenic scoring methods, which is similar to our polygenic risk score methods and results. Another study that investigated many different models and feature encodings reported a maximum AUC in T2D across all models/encodings of 0.59 (Mittag et al., 2015). Our best genotype (model e, AUC: 0.66) and CID (model g, AUC: 0.65) models are both improvements over these prior studies. While these are significant improvements, they are small; further advances will be necessary to reach the goal of clinical utility.

Our study had several limitations that may have limited our models' ability to classify T2D. Due to computational limitations, we were only able to use a small subset of the genetic variants. It is possible that using more genetic variants would increase the CNNs ability to detect local patterns since the genetic variants would be spaced closer together and the distance between variants would be more uniform. Increasing the number of genetic variants would likely impact the number of weights necessary in a CNN less than a NN because the weights in convolutional layers are not determined by the number of inputs and the pooling layers following convolutional layers reduce the dimensionality of the data. In each optimized CNN model, the best model had at least some degree of pooling and convolutional filter width, suggesting that local feature summation is useful to the model and reducing dimensions through pooling does not reduce classification performance. Insufficient optimization of hyperparameters may have also limited performance of our models. In more complex models, the hyperparameter space becomes too large to efficiently explore all possible solutions. Therefore, it is likely that our model architectures are not the optimal solution. Further exploration of the hyperparameter space could improve results. The annotations used to create the CID may not be the best combination of information. Further study on the best genomic context information to use in classification tasks may improve upon our results. There has been some evidence that ancestry information is also present outside of the top 10 principal components (Privé et al., 2020). Our models are only able to judge the ancestry component based on the PC labels provided, so it is possible that other ancestry data are still included in the model and undetected. Including more PCs as labels in adversarial tasks may further reduce the models' ability to use ancestry information.

In summary, we have described novel CNN/NN architectures that combine genomic context informed genotype data, and within model ancestry detection/adjustment. Our results indicate that this may be a useful direction for improving our ability to classify complex genetic disorders. While classification performance remains too low for clinical utility and earlier detection of T2D risk, incremental improvements such as those reported here may get to the point of clinical utility in the future.

Figure 1: Simplified model architecture diagrams. The Black arrows represent the layers that connect each input to each output. The green arrows represent the positive feedback from backpropagation that aims to minimize error. The unfilled/white arrows represent stop gradient layers, which prevent the task from changing the weights in all layers upstream from the stop gradient layer. Red arrows represent gradient reversal layers of adversarial tasks, which reverse the direction of the weight changes and maximize loss for the task in any layer upstream from the gradient reversal layer.

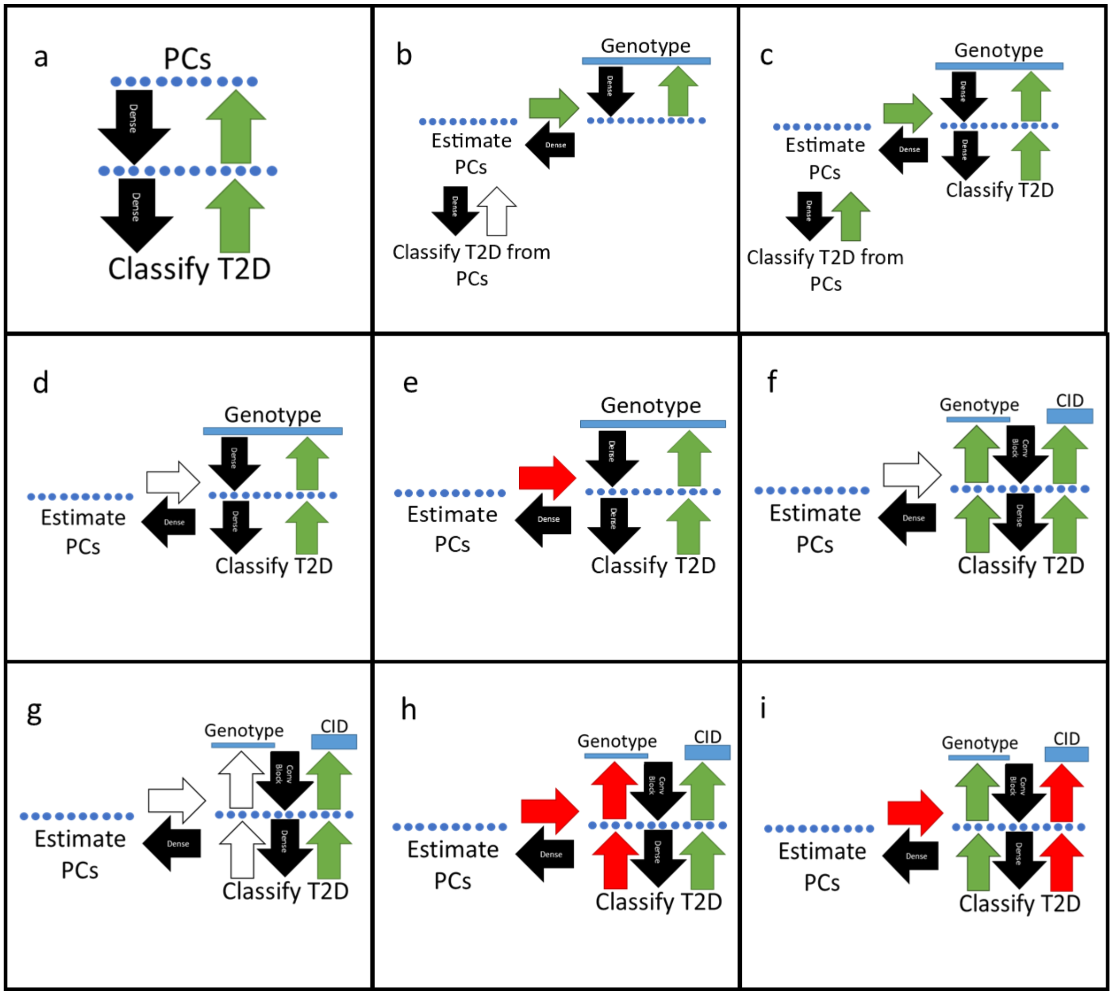
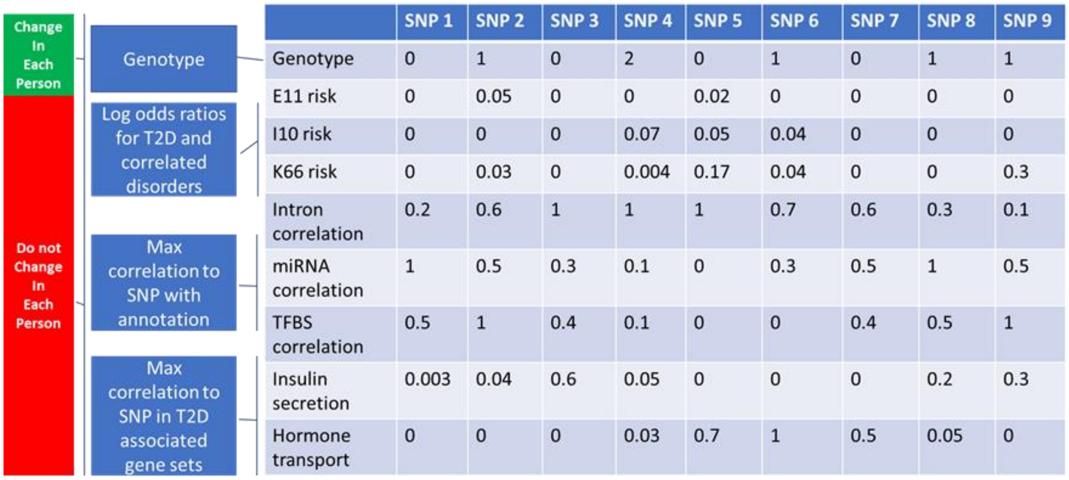


Figure 2: An illustrative example of a portion of a context informed data matrix (CID). A CID is constructed for each person within the study. For each person, the annotation values are multiplied by the allele count at each SNP and the resulting individualized annotation matrix is used as input into machine learning models.


Change In Each Person
Do not Change In Each Person
Genotype
Log odds ratios for T2D and correlated disorders
Max correlation to SNP with annotation
Max correlation to SNP in T2D associated gene sets

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7	SNP 8	SNP 9
Genotype	0	1	0	2	0	1	0	1	1
E11 risk	0	0.05	0	0	0.02	0	0	0	0
I10 risk	0	0	0	0.07	0.05	0.04	0	0	0
K66 risk	0	0.03	0	0.004	0.17	0.04	0	0	0.3
Intron correlation	0.2	0.6	1	1	1	0.7	0.6	0.3	0.1
miRNA correlation	1	0.5	0.3	0.1	0	0.3	0.5	1	0.5
TFBS correlation	0.5	1	0.4	0.1	0	0	0.4	0.5	1
Insulin secretion	0.003	0.04	0.6	0.05	0	0	0	0.2	0.3
Hormone transport	0	0	0	0.03	0.7	1	0.5	0.05	0

Table 1. Overview of Machine Learning Models

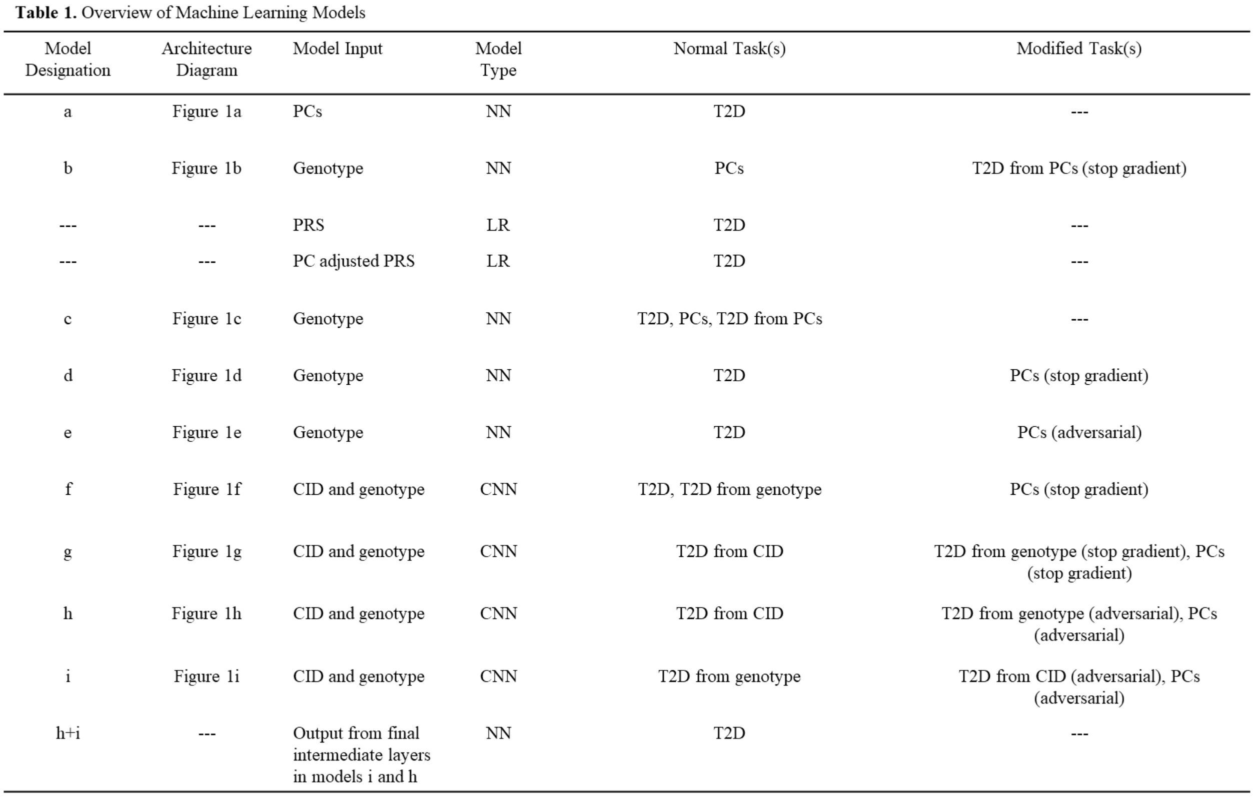


Table 1. Overview of Machine Learning Models

Model Designation	Architecture Diagram	Model Input	Model Type	Normal Task(s)	Modified Task(s)
a	Figure 1a	PCs	NN	T2D	---
b	Figure 1b	Genotype	NN	PCs	T2D from PCs (stop gradient)
---	---	PRS	LR	T2D	---
---	---	PC adjusted PRS	LR	T2D	---
c	Figure 1c	Genotype	NN	T2D, PCs, T2D from PCs	---
d	Figure 1d	Genotype	NN	T2D	PCs (stop gradient)
e	Figure 1e	Genotype	NN	T2D	PCs (adversarial)
f	Figure 1f	CID and genotype	CNN	T2D, T2D from genotype	PCs (stop gradient)
g	Figure 1g	CID and genotype	CNN	T2D from CID	T2D from genotype (stop gradient), PCs (stop gradient)
h	Figure 1h	CID and genotype	CNN	T2D from CID	T2D from genotype (adversarial), PCs (adversarial)
i	Figure 1i	CID and genotype	CNN	T2D from genotype	T2D from CID (adversarial), PCs (adversarial)
h+i	---	Output from final intermediate layers in models i and h	NN	T2D	---

Table 2. Machine Learning Model Results

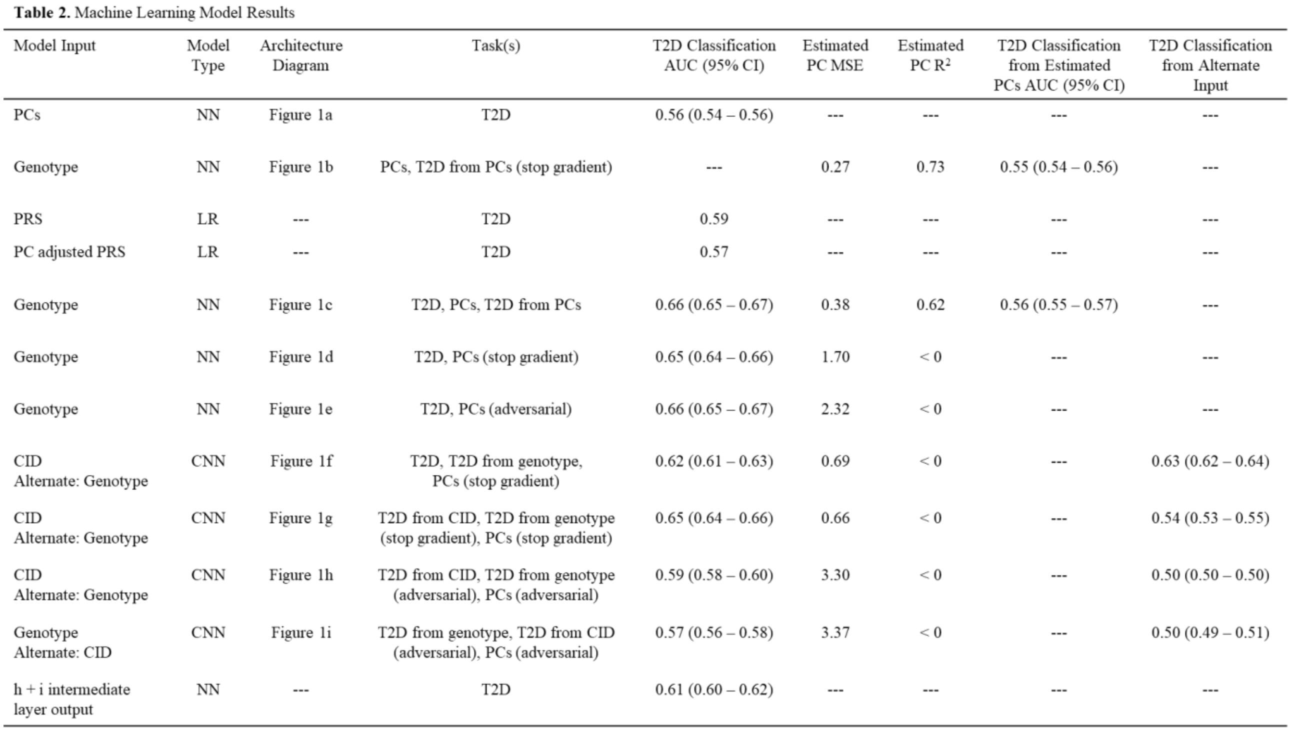


Table 2. Machine Learning Model Results

Model Input	Model Type	Architecture Diagram	Task(s)	T2D Classification AUC (95% CI)	Estimated PC MSE	Estimated PC R^2	T2D Classification from Estimated PCs AUC (95% CI)	T2D Classification from Alternate Input
PCs	NN	Figure 1a	T2D	0.56 (0.54 – 0.56)	---	---	---	---
Genotype	NN	Figure 1b	PCs, T2D from PCs (stop gradient)	---	0.27	0.73	0.55 (0.54 – 0.56)	---
PRS	LR	---	T2D	0.59	---	---	---	---
PC adjusted PRS	LR	---	T2D	0.57	---	---	---	---
Genotype	NN	Figure 1c	T2D, PCs, T2D from PCs	0.66 (0.65 – 0.67)	0.38	0.62	0.56 (0.55 – 0.57)	---
Genotype	NN	Figure 1d	T2D, PCs (stop gradient)	0.65 (0.64 – 0.66)	1.70	< 0	---	---
Genotype	NN	Figure 1e	T2D, PCs (adversarial)	0.66 (0.65 – 0.67)	2.32	< 0	---	---
CID Alternate: Genotype	CNN	Figure 1f	T2D, T2D from genotype, PCs (stop gradient)	0.62 (0.61 – 0.63)	0.69	< 0	---	0.63 (0.62 – 0.64)
CID Alternate: Genotype	CNN	Figure 1g	T2D from CID, T2D from genotype (stop gradient), PCs (stop gradient)	0.65 (0.64 – 0.66)	0.66	< 0	---	0.54 (0.53 – 0.55)
CID Alternate: Genotype	CNN	Figure 1h	T2D from CID, T2D from genotype (adversarial), PCs (adversarial)	0.59 (0.58 – 0.60)	3.30	< 0	---	0.50 (0.50 – 0.50)
Genotype Alternate: CID	CNN	Figure 1i	T2D from genotype, T2D from CID (adversarial), PCs (adversarial)	0.57 (0.56 – 0.58)	3.37	< 0	---	0.50 (0.49 – 0.51)
h + i intermediate layer output	NN	---	T2D	0.61 (0.60 – 0.62)	---	---	---	---

Table 3. Hyperparameter Optimization Ranges

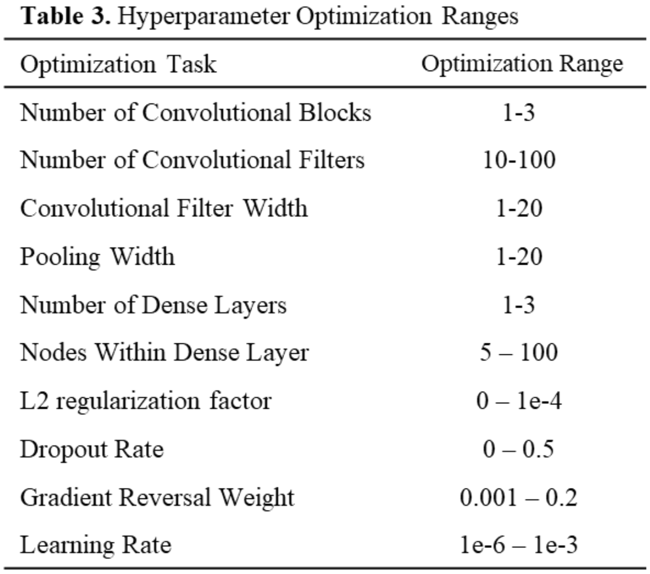


Table 3. Hyperparameter Optimization Ranges

Optimization Task	Optimization Range
Number of Convolutional Blocks	1-3
Number of Convolutional Filters	10-100
Convolutional Filter Width	1-20
Pooling Width	1-20
Number of Dense Layers	1-3
Nodes Within Dense Layer	5 – 100
L2 regularization factor	0 – 1e-4
Dropout Rate	0 – 0.5
Gradient Reversal Weight	0.001 – 0.2
Learning Rate	1e-6 – 1e-3

Discussion and Final Remarks

I presented a review paper and three primary research studies in this dissertation that all help to answer the question: Can we use genomic context to improve complex genetic disorder classification models? The most obvious way to measure improvement in classification models is to focus solely on AUC or some other performance metric. However, much of the work presented in this dissertation suggests that model improvements can be found in other ways. Improvements can come in the form of interesting or useful model outputs or byproducts, better validation techniques for testing generalizability, and better methods and tools for hypothesis testing. While we did find evidence that improvements can be made in terms of model AUC, we also found that focusing solely on increasing a metric can inadvertently lead to worse models.

The genomic machine learning meta-regression review paper I presented here showed that a staggering 71% of models published in peer-reviewed journals had some form of data leakage. At least in part due to this data leakage, many papers report terrific and often unbelievable model classification performances. However, if applied to a new dataset, most of the models with the highest reported AUCs would perform worse than the models with low AUCs that used methods that don’t artificially inflate AUC. Ungeneralizable, inflated-AUC-models are destructive to machine learning work in several ways. First, the models don’t offer any useful information about the disorder, because they are too overfit to the data to separate real genetic risk signals from noise. Similarly, they offer no useful applications for the disorder because they are specific to the dataset used to train the model. In addition to the lack of useful outputs, these models unintentionally negatively affect generalizable models. For the audience that does not understand the pitfalls of machine learning, generalizable models with lower AUCs will be seen as less impressive compared to the study with near-perfect classification performance. For the audience that does understand the pitfalls, every result is suspicious due to the prevalence of inappropriate methods. Reduction or elimination of publishing models that use faulty methods would be a huge boon for the field, which makes educating interested audiences about the issues and promoting the avoidance of data leakage and other machine learning pitfalls critical for the future of genomic machine learning.

The first question I sought to answer with this dissertation was: Does the inclusion of risk estimates from genetically correlated phenotypes improve risk modeling? Our study aiming to classify the diagnosis of any mood disorder based on logistic regression models with and without polygenic risk scores of psychiatric disorders found that the inclusion of additional risk estimates from correlated phenotypes did significantly improve model AUC. We also found this to be true in our study of gsPRS, which found that many of the most important gsPRS in the random forest model were from disorders besides ADHD. In addition, when investigating the correlations between the random forest importance scores and relative gene set expression, we found that the correlation was higher in correlated disorders, suggesting that the gsPRS that had high importance scores tended to be real risk effects since we would expect gene sets that are informative of ADHD risk to be expressed more in the brain relative to other tissue types. Risk scores for genetically correlated phenotypes were also included in the context informed data matrix (CID) used as input into CNNs in our third study, which found that the CID was able to find patterns that genotype only input could not use. However, since this study only looked at the overall addition of genomic context, further investigation is necessary to determine if the inclusion of the correlated risk scores had an impact on establishing the unique risk representation of the CID model. These studies individually and collectively strengthen the evidence that risk estimates from genetically correlated phenotypes are useful by showing model improvements in different model types (logistic regression, random forest, and CNN), disorders (any mood disorder, ADHD, T2D), and granularities (whole genome summation, gene set summation, individual genetic variants).

The second question I aimed to answer in this work was: Does summation of risk across gene sets associated with a disorder improve risk modeling compared to genome-wide summations? In the gsPRS analyses, we found that the inclusion of gene set summations of risk did significantly improve the AUC of ADHD classification models compared to models that only used genome wide polygenic risk scores. In addition to AUC improvements, using gsPRS also let us generate potentially informative output in the form of the average feature importance of each gene set. Finally, gsPRS inclusion made it possible to further validate the results based on their biological relevance, which helps improve confidence that the results were generalizable. Gene sets associated with T2D were used as input into the CID but were not tested specifically to determine if they had a significant effect on the results of our analyses. However, it is notable that the convolutional filter widths and pooling widths selected by the hyperparameter optimization algorithm we used were almost all greater than 1, suggesting that the CNN also found some level of local risk summation beneficial.

The final question my work sought to answer was: Does directly providing a model with functional annotation information or more explicit directions on what information to use improve risk modeling? Judging based on AUC, the answer to this question might be no, given the current computational limitations of our CNN model using functional annotation information as part of the input. Our analysis showed that while the CNN was an improvement over logistic regression, it was not significantly different from the performance of our neural network models that only used genotypes as input. However, using adversarial tasks, we were able to demonstrate that the CNN using CID input created a representation of the genetic risk of T2D that the genotype input could not replicate using the same shared layers. This suggests that genomic context based on functional annotation information could play a role in the overall genetic risk of the disorder. We also used a series of multi-task neural networks with specialized layers that either permitted, stopped, or reversed training to show that ancestry information can be reconstructed in machine learning models but can also be avoided by giving the model explicit directions on which information to avoid. The avoidance of ancestry confounds and ability to test for unique risk representations are both interesting tools for strengthening generalizability and testing hypotheses, and therefore improve our risk modeling even when AUC results remain similar.

Together, the work presented in this dissertation provides evidence in several ways that genomic context in many forms, model types, and disorders can lead to improvements in risk modelling. Incremental, generalizable, and verifiable improvements like those described in these studies will be necessary to continue to push the genetic machine learning field towards the goals of clinical utility and a better understanding of complex genetic disorders.

Acknowledgements

These projects have received funding from the European Union’s Horizon 2020 research and innovation programme grant agreement No 965381.

The data used for the gene expression analyses described in this manuscript were obtained from the Genotype-Tissue Expression (GTEx) Portal. The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., & Devin, M. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Abdulaimma, B., Hussain, A., Fergus, P., Al-Jumeily, D., Lisboa, P., Huang, D.-S., & Radi, N. (2018). Improving Type 2 Diabetes Phenotypic Classification by Combining Genetics and Conventional Risk Factors. *IEEE Congress on Evolutionary Computation (CEC)*.

Abraham, G., Kowalczyk, A., Zobel, J., & Inouye, M. (2013). Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genet Epidemiol*, *37*(2), 184-195. https://doi.org/10.1002/gepi.21698

Almgren, P., Lehtovirta, M., Isomaa, B., Sarelin, L., Taskinen, M. R., Lyssenko, V., Tuomi, T., & Groop, L. (2011). Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia*, *54*(11), 2811-2819. https://doi.org/10.1007/s00125-011-2267-5

Almlöf, J. C., Alexsson, A., Imgenberg-Kreuz, J., Sylwan, L., Bäcklin, C., Leonard, D., Nordmark, G., Tandre, K., Eloranta, M. L., Padyukov, L., Bengtsson, C., Jönsen, A., Dahlqvist, S. R., Sjöwall, C., Bengtsson, A. A., Gunnarsson, I., Svenungsson, E., Rönnblom, L., Sandling, J. K., & Syvänen, A. C. (2017). Novel risk genes for systemic lupus erythematosus predicted by random forest classification. *Sci Rep*, *7*(1), 6236. https://doi.org/10.1038/s41598-017-06516-1

An, L., Adeli, E., Liu, M., Zhang, J., Lee, S. W., & Shen, D. (2017). A Hierarchical Feature and Sample Selection Framework and Its Application for Alzheimer's Disease Diagnosis. *Sci Rep*, *7*, 45269. https://doi.org/10.1038/srep45269

Anney, R. J., Lasky-Su, J., O'Dushlaine, C., Kenny, E., Neale, B. M., Mulligan, A., Franke, B., Zhou, K., Chen, W., Christiansen, H., Arias-Vasquez, A., Banaschewski, T., Buitelaar, J., Ebstein, R., Miranda, A., Mulas, F., Oades, R. D., Roeyers, H., Rothenberger, A., . . . Gill, M. (2008). Conduct disorder and ADHD: Evaluation of conduct problems as a categorical and quantitative trait in the international multicentre ADHD genetics study. *Am J Med Genet B Neuropsychiatr Genet*, *147B*(8), 1369-1378. https://doi.org/10.1002/ajmg.b.30871

Antonucci, L. A., Pergola, G., Pigoni, A., Dwyer, D., Kambeitz-Ilankovic, L., Penzel, N., Romano, R., Gelao, B., Torretta, S., Rampino, A., Trojano, M., Caforio, G., Falkai, P., Blasi, G., Koutsouleris, N., & Bertolino, A. (2020). A Pattern of Cognitive Deficits Stratified for Genetic and Environmental Risk Reliably Classifies Patients With Schizophrenia From Healthy Control Subjects. *Biol Psychiatry*, *87*(8), 697-707. https://doi.org/10.1016/j.biopsych.2019.11.007

Anttila, V., Bulik-Sullivan, B., Finucane, H. K., Walters, R. K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G. J., Gormley, P., Malik, R., Patsopoulos, N. A., Ripke, S., Wei, Z., Yu, D., Lee, P. H., Turley, P., Grenier-Boley, B., Chouraki, V., Kamatani, Y., . . . Neale, B. M. (2018). Analysis of shared heritability in common disorders of the brain. *Science*, *360*(6395), eaap8757. https://doi.org/10.1126/science.aap8757

Arkema, E., Rosside, s. M., Sjöwall, C., Svenungsson, E., & Simard, J. (2019). Heritability and Familial Risk of Systemic Lupus Erythematosus in Sweden: A Population-based Case-control Study [abstract]. *Arthritis Rheumatol, 71 (suppl 10)*. https://acrabstracts.org/abstract/heritability-and-familial-risk-of-systemic-lupus-erythematosus-in-sweden-a-population-based-case-control-study

Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med*, *66*(3), 411-421. https://doi.org/10.1097/01.psy.0000127692.23278.a9

Badré, A., Zhang, L., Muchero, W., Reynolds, J. C., & Pan, C. (2021). Deep neural network improves the estimation of polygenic risk scores for breast cancer. *J Hum Genet*, *66*(4), 359-369. https://doi.org/10.1038/s10038-020-00832-7

Banko, M., & Brill, E. (2001). Scaling to Very Very Large Corpora for Natural Language Disambiguation. *In Proc. of ACL-2001*.

Berk, M., Brnabic, A., Dodd, S., Kelin, K., Tohen, M., Malhi, G. S., Berk, L., Conus, P., & McGorry, P. D. (2011). Does stage of illness impact treatment response in bipolar disorder? Empirical treatment data and their implication for the staging model and early intervention. *Bipolar Disord*, *13*(1), 87-98. https://doi.org/10.1111/j.1399-5618.2011.00889.x

Biederman, J., Faraone, S., Milberger, S., Guite, J., Mick, E., Chen, L., Mennin, D., Marrs, A., Ouellette, C., Moore, P., Spencer, T., Norman, D., Wilens, T., Kraus, I., & Perrin, J. (1996). A prospective 4-year follow-up study of attention-deficit hyperactivity and related disorders. *Arch Gen Psychiatry*, *53*(5), 437-446. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8624187

Biederman, J., Green, A., DiSalvo, M., & Faraone, S. V. (2021). Can polygenic risk scores help identify pediatric bipolar spectrum and related disorders?: A systematic review. *Psychiatry Res 299*, 113843. https://doi.org/10.1016/j.psychres.2021.113843

Biederman, J., Monuteaux, M., Mick, E., Spencer, T., Wilens, T., Klein, K., Price, J. E., & Faraone, S. V. (2006). Psychopathology in females with attention-deficit/hyperactivity disorder: A controlled, five-year prospective study. *Biol Psychiatry*, *60*(10), 1098-1105.

Biederman, J., Newcorn, J., & Sprich, S. (1990). Comorbidity in attention deficit hyperactivity disorder. In Task Force on DSM-IV (Ed.), *Source Book for DSM-IV* (pp. 145-162). American Psychiatric Association.

Botta, V., Louppe, G., Geurts, P., & Wehenkel, L. (2014). Exploiting SNP correlations within random forest for genome-wide association studies. *PLoS One*, *9*(4), e93379. https://doi.org/10.1371/journal.pone.0093379

Bukh, J. D., Bock, C., Vinberg, M., & Kessing, L. V. (2013). The effect of prolonged duration of untreated depression on antidepressant treatment outcome. *J Affect Disord*, *145*(1), 42-48. https://doi.org/10.1016/j.jad.2012.07.008

Bulik, C. M., Sullivan, P. F., Tozzi, F., Furberg, H., Lichtenstein, P., & Pedersen, N. L. (2006). Prevalence, Heritability, and Prospective Risk Factors for Anorexia Nervosa. *Archives of General Psychiatry, 63*(3), 305-312. https://doi.org/10.1001/archpsyc.63.3.305

Cánovas, R., Cobb, J., Brozynska, M., Bowes, J., Li, Y. R., Smith, S. L., Hakonarson, H., Thomson, W., Ellis, J. A., Abraham, G., Munro, J. E., & Inouye, M. (2020). Genomic risk scores for juvenile idiopathic arthritis and its subtypes. *Ann Rheum Dis*, *79*(12), 1572-1579. https://doi.org/10.1136/annrheumdis-2020-217421

Cawley, G. C., & Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, *11*, 2079-2107.

CDC Division of Diabetes Translation. (2017). Long-term Trends in Diabetes. *Diabetes Surveillance System*

http://www.cdc.gov/diabetes/data

Centers for Disease Control and Prevention. (2022). Estimates of Diabetes and Its Burden in the United States. *National Diabetes Statistics Report* https://www.cdc.gov/diabetes/data/statistics-report/index.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fdiabetes%2Fdata%2Fstatistics%2Fstatistics-report.html

Chen, J., Wu, J. S., Mize, T., Shui, D., & Chen, X. (2018). Prediction of Schizophrenia Diagnosis by Integration of Genetically Correlated Conditions and Traits. *J Neuroimmune Pharmacol*, *13*(4), 532-540. https://doi.org/10.1007/s11481-018-9811-8

Chen, Q., Hartman, C. A., Haavik, J., Harro, J., Klungsoyr, K., Hegvik, T. A., Wanders, R., Ottosen, C., Dalsgaard, S., Faraone, S. V., & Larsson, H. (2018). Common psychiatric and metabolic comorbidity of adult attention-deficit/hyperactivity disorder: A population-based cross-sectional study. *PLoS One*, *13*(9), e0204516. https://doi.org/10.1371/journal.pone.0204516

Chen, Q., Hartman, C. A., Kuja-Halkola, R., Faraone, S. V., Almqvist, C., & Larsson, H. (2019). Attention-deficit/hyperactivity disorder and clinically diagnosed obesity in adolescence and young adulthood: a register-based study in Sweden. *Psychol Med*, *49*(11), 1841-1849. https://doi.org/10.1017/s0033291718002532

Chen, X., Chen, D. G., Zhao, Z., Zhan, J., Ji, C., & Chen, J. (2021). Artificial image objects for classification of schizophrenia with GWAS-selected SNVs and convolutional neural network. *Patterns (N Y)*, *2*(8), 100303. https://doi.org/10.1016/j.patter.2021.100303

Christophersen, I. E., Ravn, L. S., Budtz-Joergensen, E., Skytthe, A., Haunsoe, S., Svendsen, J. H., & Christensen, K. (2009). Familial aggregation of atrial fibrillation: a study in Danish twins. *Circ Arrhythm Electrophysiol*, *2*(4), 378-383. https://doi.org/10.1161/circep.108.786665

Chuang, L. C., & Kuo, P. H. (2017). Building a genetic risk model for bipolar disorder from genome-wide association data with random forest algorithm. *Sci Rep*, *7*, 39943. https://doi.org/10.1038/srep39943

Clayton, E. W. (2003). Ethical, legal, and social implications of genomic medicine. *N Engl J Med*, *349*(6), 562-569.

Cope, J. L., Baukmann, H. A., Klinger, J. E., Ravarani, C. N. J., Böttinger, E. P., Konigorski, S., & Schmidt, M. F. (2021). Interaction-Based Feature Selection Algorithm Outperforms Polygenic Risk Score in Predicting Parkinson's Disease Status. *Front Genet*, *12*, 744557. https://doi.org/10.3389/fgene.2021.744557

Costello, E. J., Copeland, W., Cowell, A., & Keeler, G. (2007). Service costs of caring for adolescents with mental illness in a rural community, 1993-2000. *Am J Psychiatry*, *164*(1), 36-42. https://doi.org/10.1176/appi.ajp.2007.164.9.A36

Cross Disorder Group of the Psychiatric Genomic Consortium. (2019). Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell*, *179*(7), 1469-1482.e1411. https://doi.org/10.1016/j.cell.2019.11.020

de Leeuw, C. A., Mooij, J. M., Heskes, T., & Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol*, *11*(4), e1004219. https://doi.org/10.1371/journal.pcbi.1004219

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, *44*(3), 837-845.

Demontis, D., Walters, R. K., Martin, J., Mattheisen, M., Als, T. D., Agerbo, E., Baldursson, G., Belliveau, R., Bybjerg-Grauholm, J., Baekvad-Hansen, M., Cerrato, F., Chambert, K., Churchhouse, C., Dumont, A., Eriksson, N., Gandal, M., Goldstein, J. I., Grasby, K. L., Grove, J., . . . Neale, B. M. (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat Genet*, *51*(1), 63-75. https://doi.org/10.1038/s41588-018-0269-7

Demontis, D., Walters, R. K., Rajagopal, V. M., Waldman, I. D., Grove, J., Als, T. D., Dalsgaard, S., Ribasas, M., Bybjerg-Grauholm, J., Bækvad-Hansen, M., Werge, T., Nordentoft, M., Mors, O., Mortensen, P. B., Cormand, B., Hougaard, D. M., Neale, B. M., Franke, B., Faraone, S. V., & Børglum, A. D. (2021). Risk variants and polygenic architecture of disruptive behavior disorders in the context of attention-deficit/hyperactivity disorder. *Nat Commun*, *12*(1), 576. https://doi.org/10.1038/s41467-020-20443-2

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., & Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nat Commun*, *10*(1), 3328. https://doi.org/10.1038/s41467-019-11112-0

Elam, K. K., Chassin, L., & Pandika, D. (2018). Polygenic risk, family cohesion, and adolescent aggression in Mexican American and European American families: Developmental pathways to alcohol use. *Dev Psychopathol*, *30*(5), 1715-1728. https://doi.org/10.1017/s0954579418000901

Elks, C. E., den Hoed, M., Zhao, J. H., Sharp, S. J., Wareham, N. J., Loos, R. J., & Ong, K. K. (2012). Variability in the heritability of body mass index: a systematic review and meta-regression. *Front Endocrinol (Lausanne)*, *3*, 29. https://doi.org/10.3389/fendo.2012.00029

Evans, D. M., Visscher, P. M., & Wray, N. R. (2009). Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum Mol Genet*, *18*(18), 3525-3531. https://doi.org/10.1093/hmg/ddp295

Fahed, A. C., Philippakis, A. A., & Khera, A. V. (2022). The potential of polygenic scores to improve cost and efficiency of clinical trials. *Nat Commun*, *13*(1), 2922. https://doi.org/10.1038/s41467-022-30675-z

Faraone, S. V., & Biederman, J. (1997). Do attention deficit hyperactivity disorder and major depression share familial risk factors? *Journal of Nervous and Mental Disease, 185*(9), 533-541.

Faraone, S. V., Biederman, J., Doyle, A. E., Murray, K., Petty, C., Adamson, J., & Seidman, L. (2006). Neuropsychological studies of late onset and subthreshold diagnoses of adult attention-deficit/hyperactivity disorder. *Biol Psychiatry*, *60*(10), 1081-1087. https://doi.org/10.1016/j.biopsych.2006.03.060

Faraone, S. V., Doyle, A. E., Lasky-Su, J., Sklar, P. B., D'Angelo, E., Gonzalez-Heydrich, J., Kratochvil, C., Mick, E., Klein, K., Rezac, A. J., & Biederman, J. (2008). Linkage analysis of attention deficit hyperactivity disorder. *Am J Med Genet B Neuropsychiatr Genet*, *147B*(8), 1387-1391. https://doi.org/10.1002/ajmg.b.30631

Faraone, S. V., & Larsson, H. (2018). Genetics of attention deficit hyperactivity disorder. *Mol Psychiatry*, *24*(4), 562-575. https://doi.org/10.1038/s41380-018-0070-0

Fergus, P., Montanez, C. C., Abdulaimma, B., Lisboa, P., Chalmers, C., & Pineles, B. (2020). Utilizing Deep Learning and Genome Wide Association Studies for Epistatic-Driven Preterm Birth Classification in African-American Women. *IEEE/ACM Trans Comput Biol Bioinform*, *17*(2), 668-678. https://doi.org/10.1109/tcbb.2018.2868667

Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. A., Patterson, N., Gabriel, S. B., Topol, E. J., Smoller, J. W., Pato, C. N., Pato, M. T., Petryshen, T. L., Kolonel, L. N., Lander, E. S., Sklar, P., Henderson, B., Hirschhorn, J. N., & Altshuler, D. (2004). Assessing the impact of population stratification on genetic association studies. *Nat Genet*, *36*(4), 388-393. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15052270

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The journal of machine learning research*, *17*(1), 2096-2030.

Garza-Hernandez, D., Estrada, K., & Trevino, V. (2022). Multivariate genome-wide association study models to improve prediction of Crohn's disease risk and identification of potential novel variants. *Comput Biol Med*, *145*, 105398. https://doi.org/10.1016/j.compbiomed.2022.105398

Gatz, M., Reynolds, C. A., Fratiglioni, L., Johansson, B., Mortimer, J. A., Berg, S., Fiske, A., & Pedersen, N. L. (2006). Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry*, *63*(2), 168-174. https://doi.org/10.1001/archpsyc.63.2.168

Gaudillo, J., Rodriguez, J. J. R., Nazareno, A., Baltazar, L. R., Vilela, J., Bulalacao, R., Domingo, M., & Albia, J. (2019). Machine learning approach to single nucleotide polymorphism-based asthma prediction. *PLoS One*, *14*(12), e0225574. https://doi.org/10.1371/journal.pone.0225574

Ghio, L., Gotelli, S., Marcenaro, M., Amore, M., & Natta, W. (2014). Duration of untreated illness and outcomes in unipolar depression: a systematic review and meta-analysis. *J Affect Disord*, *152-154*, 45-51. https://doi.org/10.1016/j.jad.2013.10.002

Gordon, H., Trier Moller, F., Andersen, V., & Harbord, M. (2015). Heritability in inflammatory bowel disease: from the first twin study to genome-wide association studies. *Inflamm Bowel Dis*, *21*(6), 1428-1434. https://doi.org/10.1097/mib.0000000000000393

Gore, F. M., Bloem, P. J., Patton, G. C., Ferguson, J., Joseph, V., Coffey, C., Sawyer, S. M., & Mathers, C. D. (2011). Global burden of disease in young people aged 10-24 years: a systematic analysis. *Lancet*, *377*(9783), 2093-2102. https://doi.org/10.1016/s0140-6736(11)60512-6

Guo, Y., Wei, Z., Keating, B. J., & Hakonarson, H. (2016). Machine learning derived risk prediction of anorexia nervosa. *BMC Med Genomics*, *9*, 4. https://doi.org/10.1186/s12920-016-0165-x

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, *24*(2), 8-12. https://doi.org/doi: 10.1109/MIS.2009.36.

Hamza, T. H., & Payami, H. (2010). The heritability of risk and age at onset of Parkinson's disease after accounting for known genetic risk factors. *J Hum Genet*, *55*(4), 241-243. https://doi.org/10.1038/jhg.2010.13

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2021). *Doing Meta-Analysis With R: A Hands-On Guide* (1st ed.). Chapman & Hall/CRC Press. https://www.routledge.com/Doing-Meta-Analysis-with-R-A-Hands-On-Guide/Harrer-Cuijpers-Furukawa-Ebert/p/book/9780367610074

Hess, J. L., Tylee, D. S., Mattheisen, M., Consortium, S. W. G. o. t. P. G., (iPSYCH), L. F. I. f. I. P. P., Borglum, A. D., Als, T. D., Grove, J., Werge, T., Mortensen, P. B., Mors, O., Nordentoft, M., Hougaard, D. M., Byberg-Grauholm, J., Baekvad-Hansen, M., Greenwood, T. A., Tsuang, M. T., Curtis, D., Steinberg, S., . . . Glatt, S. J. (2019). A polygenic resilience score moderates the genetic risk for schizophrenia. *Mol Psychiatry*, *1038*, s41380-41019-40463-41388. https://doi.org/10.1038/s41380-019-0463-8

Hilker, R., Helenius, D., Fagerlund, B., Skytthe, A., Christensen, K., Werge, T. M., Nordentoft, M., & Glenthøj, B. (2018). Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register. *Biol Psychiatry*, *83*(6), 492-498. https://doi.org/10.1016/j.biopsych.2017.08.017

Ho, D., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, *42*(8), 1 - 28. https://doi.org/10.18637/jss.v042.i08

Hollingshead, A. B. (1975a). *Four Factor Index of Social Status*. Yale Press.

Hottenga, J.-J., Boomsma, D. I., Kupper, N., Posthuma, D., Snieder, H., Willemsen, G., & de Geus, E. J. C. (2005). Heritability and Stability of Resting Blood Pressure. *Twin Research and Human Genetics*, *8*(5), 499-508. https://doi.org/10.1375/twin.8.5.499

Hou, J., Hess, J. L., Armstrong, N., Bis, J. C., Grenier-Boley, B., Karlsson, I. K., Leonenko, G., Numbers, K., O'Brien, E. K., Shadrin, A., Thalamuthu, A., Yang, Q., Andreassen, O. A., Brodaty, H., Gatz, M., Kochan, N. A., Lambert, J. C., Laws, S. M., Masters, C. L., . . . Glatt, S. J. (2022). Polygenic resilience scores capture protective genetic effects for Alzheimer's disease. *Transl Psychiatry*, *12*(1), 296. https://doi.org/10.1038/s41398-022-02055-0

Hu, K., Wang, M., Liu, Y., Yan, H., Song, M., Chen, J., Chen, Y., Wang, H., Guo, H., Wan, P., Lv, L., Yang, Y., Li, P., Lu, L., Yan, J., Wang, H., Zhang, H., Zhang, D., Wu, H., . . . Liu, B. (2021). Multisite schizophrenia classification by integrating structural magnetic resonance imaging data with polygenic risk score. *Neuroimage Clin*, *32*, 102860. https://doi.org/10.1016/j.nicl.2021.102860

Hung, C. I., Liu, C. Y., & Yang, C. H. (2017). Untreated duration predicted the severity of depression at the two-year follow-up point. *PLoS One*, *12*(9), e0185119. https://doi.org/10.1371/journal.pone.0185119

Jo, T., Nho, K., Bice, P., & Saykin, A. J. (2022). Deep learning-based identification of genetic variants: application to Alzheimer's disease classification. *Brief Bioinform*, *23*(2). https://doi.org/10.1093/bib/bbac022

Joergensen, T. M., Christensen, K., Lindholt, J. S., Larsen, L. A., Green, A., & Houlind, K. (2016). Editor's Choice - High Heritability of Liability to Abdominal Aortic Aneurysms: A Population Based Twin Study. *Eur J Vasc Endovasc Surg*, *52*(1), 41-46. https://doi.org/10.1016/j.ejvs.2016.03.012

Kang, J., Kugathasan, S., Georges, M., Zhao, H., & Cho, J. H. (2011). Improved risk prediction for Crohn's disease with a multi-locus approach. *Hum Mol Genet*, *20*(12), 2435-2442. https://doi.org/10.1093/hmg/ddr116

Kaufman, S., Rossett, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance [Article No. 15]. *ACM Tranactions on Knowledge Discovery from Data*, *6*(4). https://doi.org/https://doi.org/10.1145/2382577.2382579

Kinreich, S., Meyers, J. L., Maron-Katz, A., Kamarajan, C., Pandey, A. K., Chorlian, D. B., Zhang, J., Pandey, G., Subbie-Saenz de Viteri, S., Pitti, D., Anokhin, A. P., Bauer, L., Hesselbrock, V., Schuckit, M. A., Edenberg, H. J., & Porjesz, B. (2021). Predicting risk for Alcohol Use Disorder using longitudinal data with multimodal biomarkers and family history: a machine learning study. *Mol Psychiatry*, *26*(4), 1133-1141. https://doi.org/10.1038/s41380-019-0534-x

Knowler, W. C., Barrett-Connor, E., Fowler, S. E., Hamman, R. F., Lachin, J. M., Walker, E. A., & Nathan, D. M. (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med*, *346*(6), 393-403. https://doi.org/10.1056/NEJMoa012512

Knowler, W. C., Fowler, S. E., Hamman, R. F., Christophi, C. A., Hoffman, H. J., Brenneman, A. T., Brown-Friday, J. O., Goldberg, R., Venditti, E., & Nathan, D. M. (2009). 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. *Lancet*, *374*(9702), 1677-1686. https://doi.org/10.1016/s0140-6736(09)61457-4

Kooperberg, C., LeBlanc, M., & Obenchain, V. (2010). Risk prediction using genome-wide association studies. *Genet Epidemiol*, *34*(7), 643-652. https://doi.org/10.1002/gepi.20509

Kraus, C., Kadriu, B., Lanzenberger, R., Zarate, C. A., Jr., & Kasper, S. (2020). Prognosis and Improved Outcomes in Major Depression: A Review. *Focus (Am Psychiatr Publ)*, *18*(2), 220-235. https://doi.org/10.1176/appi.focus.18205

Krautenbacher, N., Kabesch, M., Horak, E., Braun-Fahrländer, C., Genuneit, J., Boznanski, A., von Mutius, E., Theis, F., Fuchs, C., & Ege, M. J. (2021). Asthma in farm children is more determined by genetic polymorphisms and in non-farm children by environmental factors. *Pediatr Allergy Immunol*, *32*(2), 295-304. https://doi.org/10.1111/pai.13385

Kuja-Halkola, R., Lebwohl, B., Halfvarson, J., Wijmenga, C., Magnusson, P. K., & Ludvigsson, J. F. (2016). Heritability of non-HLA genetics in coeliac disease: a population-based study in 107 000 twins. *Gut*, *65*(11), 1793-1798. https://doi.org/10.1136/gutjnl-2016-311713

Kwon, O. S., Hong, M., Kim, T. H., Hwang, I., Shim, J., Choi, E. K., Lim, H. E., Yu, H. T., Uhm, J. S., Joung, B., Oh, S., Lee, M. H., Kim, Y. H., & Pak, H. N. (2022). Genome-wide association study-based prediction of atrial fibrillation using artificial intelligence. *Open Heart*, *9*(1). https://doi.org/10.1136/openhrt-2021-001898

Kyvik, K. O., Green, A., & Beck-Nielsen, H. (1995). Concordance rates of insulin dependent diabetes mellitus: a population based study of young Danish twins. *Bmj*, *311*(7010), 913-917. https://doi.org/10.1136/bmj.311.7010.913

Lauber, C., Gerl, M. J., Klose, C., Ottosson, F., Melander, O., & Simons, K. (2022). Lipidomic risk scores are independent of polygenic risk scores and can predict incidence of diabetes and cardiovascular disease in a large population cohort. *PLoS Biol*, *20*(3), e3001561. https://doi.org/10.1371/journal.pbio.3001561

Lee, S., Liang, X., Woods, M., Reiner, A. S., Concannon, P., Bernstein, L., Lynch, C. F., Boice, J. D., Deasy, J. O., Bernstein, J. L., & Oh, J. H. (2020). Machine learning on genome-wide association studies to predict the risk of radiation-associated contralateral breast cancer in the WECARE Study. *PLoS One*, *15*(2), e0226157. https://doi.org/10.1371/journal.pone.0226157

Levitan, R. D., Zhang, C. X. W., Knight, J. A., Hung, R., Lye, J., Murphy, K., Atkinson, L., Bocking, A., Lye, S., & Matthews, S. G. (2020). Using Precision Medicine with a Neurodevelopmental Perspective to Study Inflammation and Depression. *Curr Psychiatry Rep*, *22*(12), 87. https://doi.org/10.1007/s11920-020-01206-8

Li, J., Pan, C., Zhang, S., Spin, J. M., Deng, A., Leung, L. L. K., Dalman, R. L., Tsao, P. S., & Snyder, M. (2018). Decoding the Genomics of Abdominal Aortic Aneurysm. *Cell*, *174*(6), 1361-1372.e1310. https://doi.org/10.1016/j.cell.2018.07.021

Li, L., Huang, Y., Han, Y., & Jiang, J. (2021). Use of deep learning genomics to discriminate Alzheimer's disease and healthy controls. *Annu Int Conf IEEE Eng Med Biol Soc*, *2021*, 5788-5791. https://doi.org/10.1109/embc46164.2021.9629983

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nat Rev Genet*, *16*(6), 321-332. https://doi.org/10.1038/nrg3920

nrg3920 (Youngstrom et al.)

Liu, L., Feng, X., Li, H., Cheng Li, S., Qian, Q., & Wang, Y. (2021). Deep learning model reveals potential risk genes for ADHD, especially Ephrin receptor gene EPHA5. *Brief Bioinform*. https://doi.org/10.1093/bib/bbab207

Lønnberg, A. S., Skov, L., Skytthe, A., Kyvik, K. O., Pedersen, O. B., & Thomsen, S. F. (2013). Heritability of psoriasis in a large twin sample. *Br J Dermatol*, *169*(2), 412-416. https://doi.org/10.1111/bjd.12375

MacGregor, A. J., Snieder, H., Rigby, A. S., Koskenvuo, M., Kaprio, J., Aho, K., & Silman, A. J. (2000). Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum*, *43*(1), 30-37. https://doi.org/10.1002/1529-0131(200001)43:1<30::Aid-anr5>3.0.Co;2-b

Mahajan, A., Spracklen, C. N., Zhang, W., Ng, M. C. Y., Petty, L. E., Kitajima, H., Yu, G. Z., Rüeger, S., Speidel, L., Kim, Y. J., Horikoshi, M., Mercader, J. M., Taliun, D., Moon, S., Kwak, S. H., Robertson, N. R., Rayner, N. W., Loh, M., Kim, B. J., . . . Morris, A. P. (2022). Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nat Genet*, *54*(5), 560-572. https://doi.org/10.1038/s41588-022-01058-3

Marchini, J. (2015). UK Biobank Phasing and Imputation Documentation. *Department of Statistics, University of Oxford, On behalf of UK Biobank, Version 1.2*.

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*, *51*(4), 584-591. https://doi.org/10.1038/s41588-019-0379-x

Mittag, F., Büchel, F., Saad, M., Jahn, A., Schulte, C., Bochdanovits, Z., Simón-Sánchez, J., Nalls, M. A., Keller, M., Hernandez, D. G., Gibbs, J. R., Lesage, S., Brice, A., Heutink, P., Martinez, M., Wood, N. W., Hardy, J., Singleton, A. B., Zell, A., . . . Sharma, M. (2012). Use of support vector machines for disease risk prediction in genome-wide association studies: concerns and opportunities. *Hum Mutat*, *33*(12), 1708-1718. https://doi.org/10.1002/humu.22161

Mittag, F., Römer, M., & Zell, A. (2015). Influence of Feature Encoding and Choice of Classifier on Disease Risk Prediction in Genome-Wide Association Studies. *PLoS One*, *10*(8), e0135832. https://doi.org/10.1371/journal.pone.0135832

Möller, S., Mucci, L. A., Harris, J. R., Scheike, T., Holst, K., Halekoh, U., Adami, H. O., Czene, K., Christensen, K., Holm, N. V., Pukkala, E., Skytthe, A., Kaprio, J., & Hjelmborg, J. B. (2016). The Heritability of Breast Cancer among Women in the Nordic Twin Study of Cancer. *Cancer Epidemiol Biomarkers Prev*, *25*(1), 145-150. https://doi.org/10.1158/1055-9965.Epi-15-0913

Morgan, C. J., & Cauce, A. M. (1999). Predicting DSM-III-R disorders from the Youth Self-Report: analysis of data from a field study. *Journal of the American Academy of Child and Adolescent Psychiatry*, *38*(10), 1237-1245.

Mullins, N., Kang, J., Campos, A. I., Coleman, J. R. I., Edwards, A. C., Galfalvy, H., Levey, D. F., Lori, A., Shabalin, A., Starnawska, A., Su, M. H., Watson, H. J., Adams, M., Awasthi, S., Gandal, M., Hafferty, J. D., Hishimoto, A., Kim, M., Okazaki, S., . . . Ruderfer, D. M. (2022). Dissecting the Shared Genetic Architecture of Suicide Attempt, Psychiatric Disorders, and Known Risk Factors. *Biol Psychiatry*, *91*(3), 313-327. https://doi.org/10.1016/j.biopsych.2021.05.029

Muneeb, M., & Henschel, A. (2021). Eye-color and Type-2 diabetes phenotype prediction from genotype data using deep learning methods. *BMC Bioinformatics*, *22*(1), 198. https://doi.org/10.1186/s12859-021-04077-9

Muntaner-Mas, A., Ortega, F. B., Femia, P., Kiive, E., Eensoo, D., Mäestu, J., Franke, B., Reif, A., Faraone, S. V., & Harro, J. (2020). Low cardiorespiratory fitness and obesity for ADHD in childhood and adolescence: A 6-year cohort study. *Scand J Med Sci Sports*, *31*(4), 903-913. https://doi.org/10.1111/sms.13905

Neuman, R. J., Heath, A., Reich, W., Bucholz, K. K., Madden, P. A. F., Sun, L., Todd, R. D., & Hudziak, J. J. (2001). Latent class analysis of ADHD and comorbid symptoms in a population sample of adolescent female twins. *J Child Psychol Psychiatry*, *42*(7), 933-942. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11693588

Nguyen, T. T., Huang, J., Wu, Q., Nguyen, T., & Li, M. (2015). Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. *BMC Genomics*, *16 Suppl 2*(Suppl 2), S5. https://doi.org/10.1186/1471-2164-16-s2-s5

Nigg, J. T., Karalunas, S. L., Gustafsson, H. C., Bhatt, P., Ryabinin, P., Mooney, M. A., Faraone, S. V., Fair, D. A., & Wilmot, B. (2019). Evaluating chronic emotional dysregulation and irritability in relation to ADHD and depression genetic risk in children with ADHD. *J Child Psychol Psychiatry*, *61*(2), 205-214. https://doi.org/10.1111/jcpp.13132

Noble, J. A., & Valdes, A. M. (2011). Genetics of the HLA region in the prediction of type 1 diabetes. *Curr Diab Rep*, *11*(6), 533-542. https://doi.org/10.1007/s11892-011-0223-x

Nunnally, J. C. (1978). *Psychometric Theory*. McGraw Hill.

O'Malley, T. a. B., Elie and Long, James and Chollet, Fran\c{c}ois and Jin, Haifeng and Invernizzi, Luca and others. (2019). KerasTuner. *\url{https://github.com/keras-team/keras-tuner*.

Obenchain, V., Lawrence, M., Carey, V., Gogarten, S., Shannon, P., & Morgan, M. (2014). VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*, *30*(14), 2076-2078. https://doi.org/10.1093/bioinformatics/btu168

Orvaschel, H., & Puig-Antich, J. (1987). *Schedule for Affective Disorders and Schizophrenia for School-Age Children: Epidemiologic Version*. Nova University.

Osipowicz, M., Wilczynski, B., & Machnicka, M. A. (2021). Careful feature selection is key in classification of Alzheimer's disease patients based on whole-genome sequencing data. *NAR Genom Bioinform*, *3*(3), lqab069. https://doi.org/10.1093/nargab/lqab069

Pal, L. R., Kundu, K., Yin, Y., & Moult, J. (2017). CAGI4 Crohn's exome challenge: Marker SNP versus exome variant models for assigning risk of Crohn disease. *Hum Mutat*, *38*(9), 1225-1234. https://doi.org/10.1002/humu.23256

Paré, G., Mao, S., & Deng, W. Q. (2017). A machine-learning heuristic to improve gene score prediction of polygenic traits. *Scientific Reports*, *7*(1), 12665. https://doi.org/10.1038/s41598-017-13056-1

Perlich, C. (2010). Learning Curves in Machine Learning. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 577-580). Springer US. https://doi.org/10.1007/978-0-387-30164-8_452

Pettersson, E., Lichtenstein, P., Larsson, H., Song, J., Agrawal, A., Borglum, A. D., Bulik, C. M., Daly, M. J., Davis, L. K., Demontis, D., Edenberg, H. J., Grove, J., Gelernter, J., Neale, B. M., Pardinas, A. F., Stahl, E., Walters, J. T. R., Walters, R., Sullivan, P. F., . . . Polderman, T. J. C. (2019). Genetic influences on eight psychiatric disorders based on family data of 4 408 646 full and half-siblings, and genetic data of 333 748 cases and controls. *Psychol Med*, *49*(7), 1166-1173. https://doi.org/10.1017/s0033291718002039

Pirooznia, M., Seifuddin, F., Judy, J., Mahon, P. B., Potash, J. B., & Zandi, P. P. (2012). Data mining approaches for genome-wide association of mood disorders. *Psychiatr Genet*, *22*(2), 55-61. https://doi.org/10.1097/YPG.0b013e32834dc40d

Prasad, R. B., & Groop, L. (2015). Genetics of type 2 diabetes-pitfalls and possibilities. *Genes (Basel)*, *6*(1), 87-123. https://doi.org/10.3390/genes6010087

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, *38*(8), 904-909. https://doi.org/ng1847 (Youngstrom et al.)

10.1038/ng1847

Privé, F., Luu, K., Blum, M. G. B., McGrath, J. J., & Vilhjálmsson, B. J. (2020). Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics*, *36*(16), 4449-4457. https://doi.org/10.1093/bioinformatics/btaa520

Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci [Research Support, N.I.H., Extramural

Research Support, Non-U.S. Gov't]. *Nature*, *511*(7510), 421-427. https://doi.org/10.1038/nature13595

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, *81*(3), 559-575. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17701901

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., & Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), 748-752. https://doi.org/nature08185 (Youngstrom et al.)

10.1038/nature08185

Quinn, T. P., Hess, J. L., Marshe, V. S., Barnett, M. M., Hauschild, A. C., Maciukiewicz, M., Elsheikh, S. M., Emanuel, S., Trakadis, Y. J., Breen, M. S., Barnett, E. J., Zhang James, Y., Ahsen, M. E., Cao, H., Chen, J., Hou, J., Salekin, A., Lin, P. I., Nicodemus, K. K., . . . Glatt, S. J. (2022). Signal from Noise: Using Machine Learning to Distil Knowledge from Data in Biological Psychiatry. *PsyArXiv*, 1-54.

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved January 21, 2015 from http://www.r-project.org

Radonjić, N. V., Hess, J. L., Rovira, P., Andreassen, O., Buitelaar, J. K., Ching, C. R. K., Franke, B., Hoogman, M., Jahanshad, N., McDonald, C., Schmaal, L., Sisodiya, S. M., Stein, D. J., van den Heuvel, O. A., van Erp, T. G. M., van Rooij, D., Veltman, D. J., Thompson, P., & Faraone, S. V. (2021). Structural brain imaging studies offer clues about the effects of the shared genetic etiology among neuropsychiatric disorders. *Mol Psychiatry*, *26*(6), 2101-2110. https://doi.org/10.1038/s41380-020-01002-z

Ramachandran, A., Snehalatha, C., Mary, S., Mukesh, B., Bhaskar, A. D., & Vijay, V. (2006). The Indian Diabetes Prevention Programme shows that lifestyle modification and metformin prevent type 2 diabetes in Asian Indian subjects with impaired glucose tolerance (IDPP-1). *Diabetologia*, *49*(2), 289-297. https://doi.org/10.1007/s00125-005-0097-z

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Romagnoni, A., Jégou, S., Van Steen, K., Wainrib, G., & Hugot, J. P. (2019). Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. *Sci Rep*, *9*(1), 10351. https://doi.org/10.1038/s41598-019-46649-z

Romero-Rosales, B. L., Tamez-Pena, J. G., Nicolini, H., Moreno-Treviño, M. G., & Trevino, V. (2020). Improving predictive models for Alzheimer's disease using GWAS data by incorporating misclassified samples modeling. *PLoS One*, *15*(4), e0232103. https://doi.org/10.1371/journal.pone.0232103

Rommelse, N. N., Franke, B., Geurts, H. M., Hartman, C. A., & Buitelaar, J. K. (2010). Shared heritability of attention-deficit/hyperactivity disorder and autism spectrum disorder. *Eur Child Adolesc Psychiatry*, *19*(3), 281-295. https://doi.org/10.1007/s00787-010-0092-x

Ross, E. L., Zuromski, K. L., Reis, B. Y., Nock, M. K., Kessler, R. C., & Smoller, J. W. (2021). Accuracy Requirements for Cost-effective Suicide Risk Prediction Among Primary Care Patients in the US. *JAMA Psychiatry*, *78*(6), 642-650. https://doi.org/10.1001/jamapsychiatry.2021.0089

Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Hultman, C., Larsson, H., & Reichenberg, A. (2017). The Heritability of Autism Spectrum Disorder. *Jama*, *318*(12), 1182-1184. https://doi.org/10.1001/jama.2017.12141

Schiweck, C., Arteaga-Henriquez, G., Aichholzer, M., Thanarajah, S. E., Vargas-Cáceres, S., Matura, S., Grimm, O., Haavik, J., Kittel-Schneider, S., Ramos-Quiroga, J. A., Faraone, S. V., & Reif, A. (2021). Comorbidity of ADHD and adult bipolar disorder: A systematic review and meta-analysis *Neurosci Biobehav Rev*, *124*, 100-123. https://doi.org/10.1016/j.neubiorev.2021.01.017

Seddon, J. M., Cote, J., Page, W. F., Aggen, S. H., & Neale, M. C. (2005). The US twin study of age-related macular degeneration: relative roles of genetic and environmental influences. *Arch Ophthalmol*, *123*(3), 321-327. https://doi.org/10.1001/archopht.123.3.321

Shickle, D., & Chadwick, R. (1994). The ethics of screening: is 'screeningitis' an incurable disease? *J Med Ethics*, *20*(1), 12-18. https://doi.org/10.1136/jme.20.1.12

Sinoquet, C. (2018). A method combining a random forest-based technique with the modeling of linkage disequilibrium through latent variables, to run multilocus genome-wide association studies. *BMC Bioinformatics*, *19*(1), 106. https://doi.org/10.1186/s12859-018-2054-0

Skafidas, E., Testa, R., Zantomio, D., Chana, G., Everall, I. P., & Pantelis, C. (2014). Predicting the diagnosis of autism spectrum disorder using gene pathway analysis. *Mol Psychiatry*, *19*(4), 504-510. https://doi.org/10.1038/mp.2012.126

Smoller, J. W., Andreassen, O. A., Edenberg, H. J., Faraone, S. V., Glatt, S. J., & Kendler, K. S. (2019). Psychiatric genetics and the structure of psychopathology. *Mol Psychiatry*, *24*(3), 409-420. https://doi.org/10.1038/s41380-017-0010-4

Smoller, J. W., Craddock, N., Kendler, K., Lee, P. H., Neale, B. M., Nurnberger, J. I., Ripke, S., Santangelo, S., & Sullivan, P. F. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis [Research Support, N.I.H., Extramural]. *Lancet*, *381*(9875), 1371-1379. https://doi.org/10.1016/S0140-6736(12)62129-1

Smoller, J. W., & Finn, C. T. (2003). Family, twin, and adoption studies of bipolar disorder. *American Journal of Medical Genetics*, *123C*(1), 48.

Stahl, E. A., Breen, G., Forstner, A. J., McQuillin, A., Ripke, S., Trubetskoy, V., Mattheisen, M., Wang, Y., Coleman, J. R. I., Gaspar, H. A., de Leeuw, C. A., Steinberg, S., Pavlides, J. M. W., Trzaskowski, M., Byrne, E. M., Pers, T. H., Holmans, P. A., Richards, A. L., Abbott, L., . . . Sklar, P. (2019). Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat Genet*, *51*(5), 793-803. https://doi.org/10.1038/s41588-019-0397-8

StataCorp. (2019). *Stata Statistical Software: Release 16*. In StataCorp LLC.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med*, *12*(3), e1001779. https://doi.org/https://doi.org/10.1371/journal.pmed.1001779

Sullivan, P. F., Daly, M. J., & O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet*, *13*(8), 537-551. https://doi.org/10.1038/nrg3240

Sullivan, P. F., & Geschwind, D. H. (2019). Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders. *Cell*, *177*(1), 162-183. https://doi.org/10.1016/j.cell.2019.01.015

Sun, Y. V., Cai, Z., Desai, K., Lawrance, R., Leff, R., Jawaid, A., Kardia, S. L., & Yang, H. (2007). Classification of rheumatoid arthritis status with candidate gene and genome-wide single-nucleotide polymorphisms using random forests. *BMC Proc*, *1 Suppl 1*(Suppl 1), S62. https://doi.org/10.1186/1753-6561-1-s1-s62

Svensson, A. C., Sandin, S., Cnattingius, S., Reilly, M., Pawitan, Y., Hultman, C. M., & Lichtenstein, P. (2009). Maternal effects for preterm birth: a genetic epidemiologic study of 630,000 families. *Am J Epidemiol*, *170*(11), 1365-1372. https://doi.org/10.1093/aje/kwp328

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nat Rev Genet*, *20*(8), 467-484. https://doi.org/10.1038/s41576-019-0127-1

Thomas, M., Sakoda, L. C., Hoffmeister, M., Rosenthal, E. A., Lee, J. K., van Duijnhoven, F. J. B., Platz, E. A., Wu, A. H., Dampier, C. H., de la Chapelle, A., Wolk, A., Joshi, A. D., Burnett-Hartman, A., Gsur, A., Lindblom, A., Castells, A., Win, A. K., Namjou, B., Van Guelpen, B., . . . Hsu, L. (2020). Genome-wide Modeling of Polygenic Risk Score in Colorectal Cancer Risk. *Am J Hum Genet*, *107*(3), 432-444. https://doi.org/10.1016/j.ajhg.2020.07.006

Tuomilehto, J., Lindström, J., Eriksson, J. G., Valle, T. T., Hämäläinen, H., Ilanne-Parikka, P., Keinänen-Kiukaanniemi, S., Laakso, M., Louheranta, A., Rastas, M., Salminen, V., & Uusitupa, M. (2001). Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *N Engl J Med*, *344*(18), 1343-1350. https://doi.org/10.1056/nejm200105033441801

Tylee, D. S., Sun, J., Hess, J. L., Tahir, M. A., Sharma, E., Malik, R., Worrall, B. B., Levine, A. J., Martinson, J. J., Nejentsev, S., Speed, D., Fischer, A., Mick, E., Walker, B. R., Crawford, A., Grant, S. F. A., Polychronakos, C., Bradfield, J. P., Sleiman, P. M. A., . . . Glatt, S. J. (2018). Genetic correlations among psychiatric and immune-related phenotypes based on genome-wide association data. *Am J Med Genet B Neuropsychiatr Genet*, *177*(7), 641-657. https://doi.org/10.1002/ajmg.b.32652

Ullemar, V., Magnusson, P. K., Lundholm, C., Zettergren, A., Melén, E., Lichtenstein, P., & Almqvist, C. (2016). Heritability and confirmation of genetic association studies for childhood asthma in twins. *Allergy*, *71*(2), 230-238. https://doi.org/10.1111/all.12783

van Hulzen, K. J. E., Scholz, C. J., Franke, B., Ripke, S., Klein, M., McQuillin, A., Sonuga-Barke, E. J., Group, P. A. W., Kelsoe, J. R., Landen, M., Andreassen, O. A., Group, P. G. C. B. D. W., Lesch, K. P., Weber, H., Faraone, S. V., Arias-Vasquez, A., & Reif, A. (2017). Genetic Overlap Between Attention-Deficit/Hyperactivity Disorder and Bipolar Disorder: Evidence From Genome-wide Association Study Meta-analysis. *Biol Psychiatry*, *82*(9), 634-641. https://doi.org/10.1016/j.biopsych.2016.08.040

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*, 91. https://doi.org/10.1186/1471-2105-7-91

Vaudreuil, C. A. H., Faraone, S. V., Di Salvo, M., Wozniak, J. R., Wolenski, R. A., Carrellas, N. W., & Biederman, J. (2019). The morbidity of subthreshold pediatric bipolar disorder: A systematic literature review and meta-analysis (Alcohol Research: Current Reviews Editorial). *Bipolar Disord*, *21*(1), 16-27. https://doi.org/10.1111/bdi.12734

Verhulst, B., Neale, M. C., & Kendler, K. S. (2015). The heritability of alcohol use disorders: a meta-analysis of twin and adoption studies. *Psychol Med*, *45*(5), 1061-1072. https://doi.org/10.1017/s0033291714002165

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P. R., Bhatia, G., Do, R., Hayeck, T., Won, H. H., Kathiresan, S., Pato, M., Pato, C., Tamimi, R., Stahl, E., Zaitlen, N., Pasaniuc, B., . . . Price, A. L. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet*, *97*(4), 576-592. https://doi.org/10.1016/j.ajhg.2015.09.001

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet*, *101*(1), 5-22. https://doi.org/10.1016/j.ajhg.2017.06.005

Vivian-Griffiths, T., Baker, E., Schmidt, K. M., Bracher-Smith, M., Walters, J., Artemiou, A., Holmans, P., O'Donovan, M. C., Owen, M. J., Pocklington, A., & Escott-Price, V. (2019). Predictive modeling of schizophrenia from genomic data: Comparison of polygenic risk score with kernel support vector machines approach. *Am J Med Genet B Neuropsychiatr Genet*, *180*(1), 80-85. https://doi.org/10.1002/ajmg.b.32705

Waldmann, P., Pfeiffer, C., & Mészáros, G. (2020). Sparse Convolutional Neural Networks for Genome-Wide Prediction [Methods]. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00025

Wang, H., & Avillach, P. (2021). Diagnostic Classification and Prognostic Prediction Using Common Genetic Variants in Autism Spectrum Disorder: Genotype-Based Deep Learning. *JMIR Med Inform*, *9*(4), e24754. https://doi.org/10.2196/24754

Wang, H. Y., Chang, S. C., Lin, W. Y., Chen, C. H., Chiang, S. H., Huang, K. Y., Chu, B. Y., Lu, J. J., & Lee, T. Y. (2018). Machine Learning-Based Method for Obesity Risk Evaluation Using Single-Nucleotide Polymorphisms Derived from Next-Generation Sequencing. *J Comput Biol*, *25*(12), 1347-1360. https://doi.org/10.1089/cmb.2018.0002

Wang, Y., Li, Y., Pu, W., Wen, K., Shugart, Y. Y., Xiong, M., & Jin, L. (2016). Random Bits Forest: a Strong Classifier/Regressor for Big Data. *Sci Rep*, *6*, 30086. https://doi.org/10.1038/srep30086

Wang, Y., Miller, M., Astrakhan, Y., Petersen, B. S., Schreiber, S., Franke, A., & Bromberg, Y. (2019). Identifying Crohn's disease signal from variome analysis. *Genome Med*, *11*(1), 59. https://doi.org/10.1186/s13073-019-0670-6

Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T. J. C., van der Sluis, S., Andreassen, O. A., Neale, B. M., & Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet*, *51*(9), 1339-1348. https://doi.org/10.1038/s41588-019-0481-0

Wei, Z., Wang, K., Qu, H. Q., Zhang, H., Bradfield, J., Kim, C., Frackleton, E., Hou, C., Glessner, J. T., Chiavacci, R., Stanley, C., Monos, D., Grant, S. F., Polychronakos, C., & Hakonarson, H. (2009). From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet*, *5*(10), e1000678. https://doi.org/10.1371/journal.pgen.1000678

Wei, Z., Wang, W., Bradfield, J., Li, J., Cardinale, C., Frackelton, E., Kim, C., Mentch, F., Van Steen, K., Visscher, P. M., Baldassano, R. N., & Hakonarson, H. (2013). Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am J Hum Genet*, *92*(6), 1008-1012. https://doi.org/10.1016/j.ajhg.2013.05.002

Whalen, S., Schreiber, J., Noble, W. S., & Pollard, K. S. (2022). Navigating the pitfalls of applying machine learning in genomics. *Nat Rev Genet*, *23*(3), 169-181. https://doi.org/10.1038/s41576-021-00434-9

Widen, E., Raben, T. G., Lello, L., & Hsu, S. D. H. (2021). Machine Learning Prediction of Biomarkers from SNPs and of Disease Risk from Biomarkers in the UK Biobank. *Genes (Basel)*, *12*(7). https://doi.org/10.3390/genes12070991

Wilens, T. E., Biederman, J., Adamson, J. J., Henin, A., Sgambati, S., Gignac, M., Sawtelle, R., Santry, A., & Monuteaux, M. C. (2008). Further evidence of an association between adolescent bipolar disorder with smoking and substance use disorders: A controlled study. *Drug Alcohol Depend*, *95*(3), 188-198. https://doi.org/S0376-8716(08)00050-1 (Youngstrom et al.)

10.1016/j.drugalcdep.2007.12.016

Willemsen, G., van Beijsterveldt, T. C. E. M., van Baal, C. G. C. M., Postma, D., & Boomsma, D. I. (2008). Heritability of Self-Reported Asthma and Allergy: A Study in Adult Dutch Twins, Siblings and Parents. *Twin Research and Human Genetics*, *11*(2), 132-142. https://doi.org/10.1375/twin.11.2.132

Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems for Optimization. *IEEE Transactions on Evolutionary Computation*

1(1), 1-32.

Wozniak, J., Faraone, S. V., Martelon, M., McKillop, H., & Biederman, J. (2012). Further evidence for robust familiality of pediatric bipolar I disorder: results from a very large controlled family study of pediatric bipolar I disorder and a meta-analysis. *J Clin Psychiatry*, *73*(10), 1328-1334. https://doi.org/10.4088/JCP.12m07770

Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A., Dudbridge, F., & Middeldorp, C. M. (2014). Research review: Polygenic methods and their application to psychiatric traits. *J Child Psychol Psychiatry*, *55*(10), 1068-1087. https://doi.org/10.1111/jcpp.12295

Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., Adams, M. J., Agerbo, E., Air, T. M., Andlauer, T. M. F., Bacanu, S. A., Baekvad-Hansen, M., Beekman, A. F. T., Bigdeli, T. B., Binder, E. B., Blackwood, D. R. H., Bryois, J., Buttenschon, H. N., Bybjerg-Grauholm, J., . . . Sullivan, P. F. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet*, *50*(5), 668-681. https://doi.org/10.1038/s41588-018-0090-3

Wray, N. R., Yang, J., Goddard, M. E., & Visscher, P. M. (2010). The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet*, *6*(2), e1000864. https://doi.org/10.1371/journal.pgen.1000864

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., & Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet*, *14*(7), 507-515. https://doi.org/10.1038/nrg3457

nrg3457 (Youngstrom et al.)

Wu, Q., Ye, Y., Liu, Y., & Ng, M. K. (2012). SNP selection and classification of genome-wide SNP data using stratified sampling random forests. *IEEE Trans Nanobioscience*, *11*(3), 216-227. https://doi.org/10.1109/tnb.2012.2214232

Xu, M., Tantisira, K. G., Wu, A., Litonjua, A. A., Chu, J. H., Himes, B. E., Damask, A., & Weiss, S. T. (2011). Genome Wide Association Study to predict severe asthma exacerbations in children using random forests classifiers. *BMC Med Genet*, *12*, 90. https://doi.org/10.1186/1471-2350-12-90

Xue, A., Wu, Y., Zhu, Z., Zhang, F., Kemper, K. E., Zheng, Z., Yengo, L., Lloyd-Jones, L. R., Sidorenko, J., Wu, Y., McRae, A. F., Visscher, P. M., Zeng, J., & Yang, J. (2018). Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat Commun*, *9*(1), 2941. https://doi.org/10.1038/s41467-018-04951-w

Yan, Q., Jiang, Y., Huang, H., Swaroop, A., Chew, E. Y., Weeks, D. E., Chen, W., & Ding, Y. (2021). Genome-Wide Association Studies-Based Machine Learning for Prediction of Age-Related Macular Degeneration Risk. *Transl Vis Sci Technol*, *10*(2), 29. https://doi.org/10.1167/tvst.10.2.29

Ying, X. (2019). *An Overview of Overfitting and its Solutions* IOP Conf. Series: Journal of Physics: Conf. Series 1168 (2019) 022022,

Zdravkovic, S., Wienke, A., Pedersen, N. L., Marenberg, M. E., Yashin, A. I., & De Faire, U. (2002). Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. *J Intern Med*, *252*(3), 247-254. https://doi.org/10.1046/j.1365-2796.2002.01029.x

Zhang-James, Y., & Faraone, S. V. (2016). Genetic architecture for human aggression: A study of gene-phenotype relationship in OMIM. *Am J Med Genet B Neuropsychiatr Genet*, *171*(5), 641-649. https://doi.org/10.1002/ajmg.b.32363

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization (2016). *arXiv preprint arXiv:1611.03530*.

Zhang, W., Dong, Y., Sartor, O., & Zhang, K. (2021). Comprehensive Analysis of Multiple Cohort Datasets Deciphers the Utility of Germline Single-Nucleotide Polymorphisms in Prostate Cancer Diagnosis. *Cancer Prev Res (Phila)*, *14*(7), 741-752. https://doi.org/10.1158/1940-6207.Capr-20-0534

Zhao, L. P., Bolouri, H., Zhao, M., Geraghty, D. E., & Lernmark, Å. (2016). An Object-Oriented Regression for Building Disease Predictive Models with Multiallelic HLA Genes. *Genet Epidemiol*, *40*(4), 315-332. https://doi.org/10.1002/gepi.21968

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods*, *12*(10), 931-934. https://doi.org/10.1038/nmeth.3547

Zoghbi, A. W., Dhindsa, R. S., Goldberg, T. E., Mehralizade, A., Motelow, J. E., Wang, X., Alkelai, A., Harms, M. B., Lieberman, J. A., Markx, S., & Goldstein, D. B. (2021). High-impact rare genetic variants in severe schizophrenia. *Proc Natl Acad Sci U S A*, *118*(51). https://doi.org/10.1073/pnas.2112560118