8-1-2019

Multi-Polygenic Risk Score Prediction Model for Bipolar Disorder

Travis Mize
mizetrav@isu.edu

MULTI-POLYGENIC RISK SCORE PREDICTION MODEL FOR BIPOLAR DISORDER

By

Travis Mize

Bachelor of Science – Psychology
Idaho State University
2014

A thesis submitted in partial fulfillment
of the requirements for the

Master of Arts – Psychology

Department of Psychology
College of Liberal Arts
The Graduate College

University of Nevada, Las Vegas
August 2019

UNLV | GRADUATE COLLEGE

Thesis Approval

The Graduate College
The University of Nevada, Las Vegas

June 12, 2019

This thesis prepared by

Travis Mize

entitled

Multi-Polygenic Risk Score Prediction Model for Bipolar Disorder

is approved in partial fulfillment of the requirements for the degree of

Master of Arts - Psychology
Department of Psychology

Daniel Allen, Ph.D.
Examination Committee Chair

Kathryn Hausbeck Korgan, Ph.D.
Graduate College Dean

Mira Han, Ph.D.
Examination Committee Co-Chair

Andrew Freeman, Ph.D.
Examination Committee Member

Rochelle Hines, Ph.D.
Examination Committee Member

Kaushik Ghosh, Ph.D.
Graduate College Faculty Representative

Abstract

Bipolar disorder (BP), a severe mental disorder characterized by recurring manic and depressive episodes, has been shown to have a strong genetic underpinning. Current theory suggests that it is the summation of risk alleles, spread across the entirety of the genome, which contributes to the development of BP, as well as other polygenic traits. The comorbid nature of these polygenic traits are often problematic for diagnosticians as the symptomology of the disorders may vary substantially between individuals and can create diagnostic confusion. To alleviate issues such as these, a more objective measure, to be used alongside current diagnostic procedures, is needed. To accomplish this, researchers have begun to turn their attention towards an ever increasing body of publicly available genetic data.

Recently, polygenic risk scores have been implemented in genetic risk prediction. Genome-wide association study (GWAS) summary statistics, derived on a plethora of psychiatric disorders, are readily accessible and provide a cost efficient strategy for generating risk scores. In this study, we attempted to not only predict the diagnosis of bipolar disorder utilizing publicly available genotype information, but to also improve upon current methodology by showing that the inclusion of risk scores calculated on comorbid traits can benefit the accuracy and generalizability of the classification model. While the results reported herein are mixed, this study provides strong support for the feasibility of genetic prediction of psychiatric disorders. This approach was, to our knowledge, entirely novel and the first time it had been implemented in practice.

Table of Contents

Abstract iii

List of Tables v

List of Figures vi

Background and Literature Review 1

Introduction 7

Methods 11

Results 18

Discussion 21

References 43

Curriculum Vitae 52

List of Tables

Table 1. Individual sample information 24

Table 2. GWAS summary statistics information 25

Table 3. AUCs derived on PRSs at multiple different p-value thresholds for multiple traits examined by GWASs 26

Table 4. Rounded AUCs derived on PRSs at multiple different p-value thresholds for multiple traits examined by GWASs 27

Table 5. Regression coefficients for simple logistic regression models 28

Table 6. Regression coefficients for multi-polygenic predictor models 29

Table 7. Classification results for each model examined 30

List of Figures

Figure 1. Probability scores for simple logistic regression using only the BP PGC variable at the non-rounded highest AUC 31

Figure 2. Probability scores for simple logistic regression using only the BP PGC variable at the rounded highest AUC 32

Figure 3. Probability scores for simple logistic regression using only the BP NIMH variable at the non-rounded highest AUC 33

Figure 4. Probability scores for simple logistic regression using only the BP NIMH variable at the rounded highest AUC 34

Figure 5. Probability scores for simple logistic regression using only the SCZ variable at the non-rounded highest AUC 35

Figure 6. Probability scores for simple logistic regression using only the SCZ variable at the rounded highest AUC 36

Figure 7. Probability scores for the multi-polygenic prediction model using all variables at the non-rounded highest AUC 37

Figure 8. Probability scores for the multi-polygenic prediction model using all variables at the rounded highest AUC 38

Figure 9. Probability scores for the multi-polygenic prediction model using all variables, except BP PGC, at the non-rounded highest AUC 39

Figure 10. Probability scores for the multi-polygenic prediction model using all variables, except BP PGC, at the rounded highest AUC 40

Figure 11. Probability scores for the multi-polygenic prediction model using all variables, except BP PGC and SCZ, at the non-rounded highest AUC 41

Figure 11. Probability scores for the multi-polygenic prediction model using all variables, except BP PGC and SCZ, at the rounded highest AUC 42

Background and Literature Review

Misdiagnosis, the incorrect diagnosis of an illness or disorder, is an unfortunate reality of the diagnostic process. Misdiagnosis results in the application of ineffective treatment, the need for reassessment, and an increase in total treatment costs, all of which is in addition to the negative impact misdiagnosis can have on a patient's mental and physical well-being. A quick PubMed search will show that the misdiagnosis of psychiatric illnesses has been under intense scrutiny for many years, with bipolar disorder (BP) being of particular concern. In a survey conducted by Hirschfeld, Lewis, and Vornik (2003), the four most frequent misdiagnoses of BP reported were those of unipolar depression, schizophrenia, and borderline or antisocial personality disorder. It is apparent that a more objective measure, to be used alongside current clinical practices, is needed to assist in the diagnosis of not just BP, but many other psychiatric disorders as well.

The wide range of symptomologies of BP are a significant factor as they create confusion during the diagnostic process, even more so when a patient's history is unknown and reliance on self-report data is high. Currently, BP is classified by the Diagnostic and Statistical Manual of Mental Disorders – 5 (American Psychiatric Association, 2013) as having multiple subtypes, with the subtypes reported as BP type I (having experienced one or more manic/mixed episodes), BP type II (never experienced a manic episode, but have experienced one hypomanic and depressive episode), substance/medication-induced bipolar and related disorder (disturbed mood that occurs during substance use or withdrawal), and other specified bipolar and related disorder (atypical BP that does not meet the criteria for BP type I or II). Unfortunately, many patients will only seek help for depressive symptoms and, as such, BP type II is often misdiagnosed and subsequently treated as unipolar depression (Hirschfeld & Vornik, 2004). As it stands, the misdiagnosis of BP, as well as many other psychiatric disorders, is an issue that has yet to be adequately addressed. To improve

upon current methodology, scientists have begun to turn their attention to the biological etiology of psychiatric disorders.

The exact cause of BP is currently unknown, however, many factors are thought to play a significant role. In recent years, the development of BP has been linked to a myriad of biological underpinnings, such as neurochemical imbalances (Manji et al., 2003), structural abnormalities in the brain (Rajkowska, Halaris, & Selemon, 2001), and genetic variations (Craddock & Jones, 1999). While ample progress has been made in the examination of BP in the field of neuroscience and pharmacology, technological improvements, such as DNA microarrays and next generation sequencing (NGS), have opened up many avenues for genetic research. While the price for NGS is steadily decreasing over time (Park & Kim, 2016), its cost is still too high for feasible use on samples of large size. DNA microarrays, on the other hand, remain a simple and extremely cost effective method to examine the genome in a variety of ways. These include the examination of expression levels of various genes, the analyzation of binding sites of transcription factors, or targeted genotyping for regions of particular interest (Bumgarner, 2013). For our purposes, we choose to turn our attention to the broad usage of microarrays to survey single nucleotide polymorphisms (SNPs) spread across the entirety of the genome.

SNPs are single nucleotide changes that occur in abundance within the human genome, yet the impact these small variations have on an organism's development are generally minimal and are most often not selected against, in terms of evolution. With that said, SNPs can have a notable effect on genetic function, such as changes in the amino acid produced by a codon, alterations to expression of a particular gene, and reductions in messenger RNA stability (Shastry, 2009). Current theory, now referred to as the common disease – common variant hypothesis, suggests

that common variants play a prominent role in a diverse array of common diseases and are a prime candidate for targeted study (Cargill et al., 1999).

Microarrays targeting SNPs have been used extensively in genome-wide association studies (GWAS) to identify potential variants involved in the development of various different disorders such as schizophrenia (SCZ; Ripke et al., 2014), BP (Sklar et al., 2011), Crohn's disease (Barret et a., 2008), and many others. A GWAS can be applied to two different types of phenotypes, those that are categorical (i.e. case and control) and those that are quantitative (i.e. height and intelligence). While quantitative GWAS are preferred as they grant more statistical power to detect significant effects (Bush & Moore, 2012), many GWAS conducted on psychiatric disorders are binary (affected or unaffected). For example, an earlier GWAS examining BP conducted by the Psychiatric Genomics Consortium (Sklar et a., 2011), did not attempt to differentiate between different BP subtypes and instead used case control methodology, a limitation that was most likely due to limited sample size. However, as technology improved and the availability of genetic information increased over time, GWAS began to examine differences in subtypes and have identified risk alleles that are associated not only with individual BP subtypes, but other phenotypes as well (Charney, et al., 2017).

Over time, researchers developed statistical methods to estimate the shared genetic effects of risk alleles on multiple binary traits (Lee et al., 2012), an approach referred to as genetic correlation. Recent studies report compelling evidence for genetic correlations between BP type I and SCZ, as well as larger genetic correlations between BP type II and major depressive disorder (Stahl et al., 2019). The examination of these correlations, found among many different psychiatric disorders, has been under investigation for more than a decade. Significant effort has been made to not only understand the etiology of these disorders, but the underlying pleiotropy as well. It is

now widely held that a large number of the genetic variants being discovered today contribute to the development of multiple different disorders, and that this genetic overlap is linked to the comorbid nature of many of these disorders (Rzhetsky et al., 2007), however, the effect these variants have, in terms of their impact on development, vary widely from disorder to disorder, as reported by recent GWAS. A consistent, yet unfortunate, result outlined by these experiments is that even when SNPs are found to be statistically significant, the effect sizes reported tend to be small with the total variance accounted for being less than expected. Findings such as these eventually led to the development of polygenic theories, the idea that a large number of these disorders are due to an accumulation of many small genetic factors (Abdolmaleky, Thiagalingam, & Wilcox, 2005). A prominent technique that has been implemented extensively in recent years that makes use of the effect sizes reported by GWAS is that of polygenic risk scores (PRS).

A PRS is a single value estimate sum of risk alleles for a given phenotype, spread across an individual's entire genome, which can be used to determine an individual's risk for that particular trait. The use of PRSs has increased extensively over the last decade as the field moves away from traditional monogenic theories and evidence for polygenic traits continues to grow. Interestingly, one of the first studies to incorporate the use of PRSs examined schizophrenia and bipolar disorder (Purcell et al., 2009). In this study, the International Schizophrenia Consortium concluded that thousands of SNPs explained a "third of the total variation in liability" and that the genetic risk of these SNPs was shared more with BP than any of the other disorders examined.

A current limitation of PRS methodology is that in order to calculate a risk score, a significance threshold for SNP inclusion must be set. If a GWAS significance threshold of $5x10^{-8}$ were to be implemented in practice, it is likely that too few SNPs would be included, and the overall detection of signal would be drastically limited. Yet, if a threshold is too relaxed (i.e. p-

value = 1), then the signal might be masked as there could potentially be too much noise incorporated into the score. To date, many studies looking to apply PRS methodology decide on a set of arbitrary thresholds and compare across the thresholds post hoc to determine those which worked best for the targeted population.

There are two commonly implemented practices for risk score calculation, that of determining risk stratification, where the goal is to order individuals by their level of risk, essentially creating a spectrum, or that of genetic prediction, where the goal is to accurately differentiate between an affected individual and an unaffected control (Chatterjee, Shi, & Garcia-Closas, 2016). The majority of studies conducted thus far have attempted to perform the latter, generally by establishing the efficacy of the approach by calculating an AUC for the final classification model. A limitation of this approach is that in order to calculate a PRS for a given individual, a SNP must be denoted not only in the GWAS, but the sample of interest as well.

A wide variety of microarray platforms used by independent groups made this necessity difficult as they examined distinct genotypes by implementing different microarray platforms that were produced by separate companies. This eventually led to complications in the inference of probable risk markers when scientists attempted to combine samples for performing meta-analyses to increase power for detection. The development of statistical techniques to impute the unavailable genetic information thus became necessary. It was found that the alleles between two SNPs could be estimated quite successfully by using haplotype information derived from the population. With the efforts of the International HapMap Project (Gibbs et al., 2003), the available haplotype information of the population expanded and was eventually condensed into large files referred to as reference panels. With more complex methods available in today's world, reference

panels are now generally created by subjecting many human genomes to whole-genome and exome sequencing, then combining this information to create said panels (McCarthy et al., 2016).

As the knowledge and understanding of genetic imputation developed over time, scientists began to incorporate data from multiple ethnic backgrounds into a single reference panel, as ancestral human migration and demographic histories were known to be complex. Howie, Marchini, and Stephens (2011) provided evidence that these diverse reference panels not only provided a higher accuracy in imputation, but also had generalizability to a wider variety of ethnic populations. This discovery not only increased the total number of genotypes available for examination, it also reduced the amount of error during inference and led to a substantial improvement in the quality of imputed calls. In order to assess the overall accuracy of imputation, a method was developed where imputed SNPs were masked then reimputed using nearby variants. The reimputed SNPs could then be compared with the masked SNPs to evaluate the amount of concordance (Howie, Donnelley, & Marchini, 2009).

The implementation of these diverse theories and methodologies is how we believe an objective measure can be procured to assist in the diagnosis of psychiatric disorders. Many studies have already been conducted in an attempt to use PRSs in classification models to detect differences between affected and unaffected samples, however, they have been met with limited success and apply a wide array of techniques. This study attempted to build on current understanding and application by employing some techniques discussed herein, as well as newer procedures such as including PRSs derived from multiple different traits. This approach has been shown to be feasible in our previous work (Chen et al., 2018).

Introduction

Bipolar disorder, a severe mental disorder characterized by recurring manic and depressive episodes, has been shown to have a strong genetic underpinning with heritability rates as high as 85% (Smoller & Finn, 2003). A substantial amount of research has been conducted to discover candidate genes, such as ANK3 (Sklar et al., 2011) and CACNA1C (Ferreira et al., 2008), however, these genes collectively account for only a small portion of the overall heritability and there has yet to be a major finding that implicates any single gene as a major contributor to the development of BP. Current theory suggests that it is the summation of risk alleles, spread across the entirety of the genome, that lead to the development of BP, as well as other polygenic traits, such as SCZ (Gejman, Sanders, & Duan, 2010).

The polygenic nature of these traits may explain why recent studies have found an overlap in genetic risk loci between BP and SCZ (Purcell et al., 2009). It is probable that some genetic variants involved in the development of BP may also be implicated in the development of other, possibly comorbid, traits (Lydall et al., 2011; Lee et al., 2013). This shared genetic underpinning may explain why disorders such as anxiety, depression, substance abuse, and many other psychiatric illnesses, often accompany the diagnosis of BP. The comorbid nature of these traits are often problematic for diagnosticians as the symptomology of the disorders can vary substantially between individuals and are prone to creating diagnostic confusion that is ultimately dependent upon the expertise of the clinician, leading to diagnoses with large variance. To alleviate issues such as these, a more objective measure, to be used alongside current diagnostic procedures, is needed.

Recently, PRSs have been implemented in the use of genetic risk prediction. A PRS for an individual is the summation of the number of risk alleles the individual carries, weighted by the

effect sizes these risk alleles carry for a given phenotype. As a PRS is an aggregated estimate of genetic risk for a particular trait, exactly how these genetic risks are defined will have a direct impact on the estimate. Traditionally, researchers have used only well replicated markers, such as GWAS findings, to define genetic risks, however, due to the small effects of individual alleles, GWAS validated markers only account for a small fraction of phenotype variation. This generates a PRS, calculated from the effect sizes of these GWAS variants, with an overall limited application. To ensure the practicality and usefulness of a PRS, researchers are required to optimize the significance threshold for marker selection to calculate a precise PRS. One approach is to select a series of thresholds to compute the PRSs and evaluate the predictive effect of these PRSs on a trait of interest (Euesden, Lewis, & O'Reilly, 2014). Generally speaking, as the significance threshold becomes more stringent (i.e. association p-value decreases), fewer markers would be included, and the selected markers would also be more specific to the trait of interest, leading to an increase in specificity. As the significance threshold is relaxed (i.e. association p-value increases), sensitivity increases, as well as the overall amount of noise. To combat the sensitivity versus specificity issue, researchers have begun to test the predictive power of a PRS at a series of different p-value thresholds (PT), while comparing the accuracy of prediction across all thresholds implemented (So & Sham, 2016). With this approach, a more accurate prediction is obtained, as well as a set of genetic variants that are, in theory, most likely to contribute to the development of the trait.

While individual genotype data can be difficult to obtain, GWAS summary statistics are readily accessible and cover a wide variety of traits. Due to their ease of accessibility and large power, GWAS summary statistics are a prime candidate for use in genetic prediction. GWAS summary data provides population information for SNPs of a given trait and can be used in the calculation of a PRS to predict the genetic risk of a trait of interest for an individual, given that an

individual's genotype information is available. When risk scores are compared across a large sample size, it becomes feasible to separate the sample into affected and unaffected groups, based on their predicted risk for a given trait. Interestingly, recent studies have shown that the addition of a comorbid trait into the prediction model may increase the accuracy of prediction for a trait, as compared to the use of a single trait predicting itself (Krapohl et al., 2017). While there has yet to be any conclusive results in the utilization of genetic data for the singular prediction of BP, there has been substantial effort from the field to predict other, highly heritable, traits. These include the prediction of SCZ using blood-based biomarkers (Chan et al., 2015), the prediction of educational achievement, body mass index, and general cognitive ability (Krapohl et al., 2017), as well as the prediction of ten complex traits ranging from mental health disorders, such as major depressive disorder, to cardiometabolic traits, such as total level of cholesterol (So & Sham, 2016).

Due to the sheer complexity of these polygenic traits, there has yet to be an established approach for model selection. In this study, we aimed to create a multi-polygenic prediction model (MPM) that could accurately categorize individuals into one of two categories, affected with bipolar disorder or unaffected. Similar to the approach used by Krapohl et al. (2017), risk scores generated from the use of comorbid traits via GWAS summary statistics were used as multiple predictors in the formation of a single model, rather than implementing BP in a model by itself. We hypothesized that the inclusion of comorbid traits would improve the accuracy and of the final model. In addition, we also implemented a PRS calculation method similar to that of So and Sham (2016) where arbitrary PTs were set to ascertain the best threshold for which SNPs should be chosen. It was believed that this approach would be more likely to circumvent overfitting of the training data and would lead to a model with greater generalizability. The combination of these two approaches was hypothesized to create a model with enough sensitivity to include crucial

SNPs, yet have high enough specificity to reduce noise and combat issues of overfitting. This approach was, to our knowledge, entirely novel and the first time it had been implemented in practice.

Methods

TARGET SAMPLES

In predictive analytics, a training and test sample, both with the same trait of interest, are required to generate and assess the efficacy of a given model. The training sample is used for feature selection, the process of selecting the necessary features that allude to the highest achievable signal while reducing the number of features as much as possible, as well as the formation of the actual model (i.e. hyperparameters and regression coefficients). The testing sample is then used to evaluate the application of the model on a group of individuals that are independent of the training sample. Successful implementation of the model on an unrelated sample provides backing to the hypotheses and reduces the risk of overfitting. When predicting a binary variable, such as bipolar disorder, the goal is to generate a model that is able to separate the entirety of the sample into two groups, in this case, affected and unaffected, with the final assessment being that of the accuracy of separation.

Target samples were downloaded from NIMH Genetics, a repository for genetic samples and data funded by the National Institute of Mental Health. The samples used in this study are a subset of a genome-wide association study conducted by the Psychiatric Genome Wide Association Study Consortium (Sklar et al., 2011). The Psychiatric Genomics Consortium (PGC) Bipolar Disorder Working Group association study was comprised of eleven different samples, however, only seven of those samples were made available for public access and thus, only seven of the eleven samples were included in this study. A description of the individual samples, such as their ancestry, sample size, and microarray platform used for genotyping are shown in Table 1.

Target samples were combined if the same microarray was used for genotyping, producing a total of four different samples to undergo imputation. The samples were named based on

abbreviations of their microarray platform, Affy 6, Affy 5, Affy 500k, and i550, and are referred to as such from here on. Each of the four target samples was imputed using the software IMPUTE2 (Howie, Donnelly, & Marchini, 2009) and the 1000 Genomes Phase 3, build 37 reference panel (The 1000 Genomes Project Consortium, 2015). Refer to the Imputation subsection for a more in depth explanation of imputation procedures.

After imputation, the Affy 6, Affy 5, and i550 data sets were combined into one data set (N = 4754) to serve as the sample for training and validation of the model, whereas the Affy 500k data set was left out for independent testing. The Affy 500k data set was kept separate for independent testing as it was the largest individual data set (N = 2519) of the four combined samples.

GENOME-WIDE ASSOCIATION STUDY SUMMARY STATISTICS

Effect sizes provided via GWAS study summary statistics were used for PRS calculation. All GWAS data sets used in this study were acquired from public databases, such as LD Hub and IBDGenetics. Table 2 provides detailed information about which GWAS data sets were included in this study, as well as where they were obtained. A total of 29 data sets were used for PRS calculation based on prior association with BP and public availability. Each of the 29 predictor data sets contained information such as, but not limited to, SNP ID, chromosome number, base-pair position, effect size (beta or odds ratio), and p-value. The majority of these studies examined individuals of European descent and had sample sizes that ranged from 4,596 (Preschool Internalizing Problems) to 307,354 (Major Depressive Disorder 2018).

IMPUTATION AND QUALITY CONTROL

For a SNP to be included in PRS calculation, it must be present in both the samples and the GWAS from which the effect size is being estimated. Therefore, imputation was performed to increase the total amount of SNPs available for PRS calculation. Pre-imputation quality control was performed on the Affy 6, Affy 5, Affy 500k, and i550 data sets utilizing PLINK software. As per the recommendations of Anderson et al. (2010) to ensure higher quality calls, SNPs were removed if they met any of the following criteria: genotype call rate less than 95%, minor allele frequency less than 1%, and/or Hardy-Weinberg Equilibrium less than $5x10^{-6}$.

In order to use a more recent reference panel, all target samples had their genome positions converted from NCBI build 36 to NCBI build 37 using the liftOver software (Hinrichs et al., 2006). To help facilitate this process, liftOverPlink (Ritchie, 2014), a liftOver wrapper, was used to assist in altering the genomic positions of the data files as liftOver cannot be used directly on plink data formats. Once the liftOver process was completed, genotypes were imputed using the software IMPUTE2 and the 1000 Genomes Phase 3, build 37 reference panel. IMPUTE2 best practices procedures were followed as a general workflow. Target samples were first separated by chromosome, excluding sex chromosomes X and Y, creating 22 subsets per sample to alleviate computation bottlenecks. To further decrease overall computation time and allow rapid future analyses, SHAPEIT2 (Delaneau, Zagury, & Marchini, 2013) was used to pre-phase target sample genotypes. During pre-phasing, estimated haplotypes were generated based on the sample's genotypes, these estimated haplotypes were then imputed using the reference panel (Howie, Fuchsberger, Stephens, Marchini, & Abecasis, 2012). Each chromosome was then imputed in five million base pair segments. After imputation, each chromosome was combined into a single file, per sample, and genotypes with an info score less than 0.3 were removed. The software GTOOL

(Freeman & Marchini, 2007) was then used to convert the genotype files into a PED format for downstream analyses. A threshold of 0.9 was set for GTOOL, meaning the probability of an allele pair must exceed 0.9 to be called as a genotype, otherwise the pair is set to unknown (0 0).

Once all pedigree files were produced, Affy 6, Affy 5, and i550 data sets were combined into a single data set to serve for training model purposes, as described in the Target Samples section. The pre-imputation quality control procedures stated above were then conducted once more to ensure high quality calls and reduce missingness. If an imputed SNP in the pedigree file was reported without an RSID, or if it was found to be triallelic, it was removed with PLINKs exclude function. The removal of non-identified SNPs was necessary as RSID is a needed component for matching between builds, especially during the process of risk score calculation.

RISK SCORE CALCULATION

Polygenic risk scores were calculated with the software PRSice (Euesden, Lewis, & O'Reilly, 2015), a program designed to automate the PRS calculation process. PRSice takes in GWAS summary statistics and uses the provided effect sizes to calculate risk scores for every individual of a given target sample. To reduce overfitting and selection bias, risk scores were calculated at ten different PTs: 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.3, 0.5, and 1. This approach was first implemented by So and Sham (2016) and was shown to be effective at estimating predictive power. PRSice's best fit function was not used as it was prone to gravitating towards the inclusion of all SNPs (i.e. PT = 1), a method that would be likely to overfit the training data.

Risk scores were calculated for each individual of every target sample for each of the 29 predictor data sets. A PRS for individual j is calculated at threshold P_T as:

$$PRS_{P_T,j} = \sum_{i=1}^{m} \beta_i G_{i,j}$$

where $PRS_{P_T j}$ represents individual j's PRS for every SNP i that has a p-value less than P_T. A SNPs genotype is represented as G and GWAS effect size estimate is represented as β.

To reduce bias in PRS calculation for SNPs in linkage disequilibrium (LD), clumping was performed using the default settings of PRSice, an r^2 threshold of 0.1, a p-value threshold of 1, and a clumping distance of 250 kilobases. When two SNPs are found to be in LD, the SNP with the lowest p-value is held while the leftover SNP is excluded from future risk score calculations. Any given SNP can only be included in a single clump, if it is included at all.

SINGLE POLYGENIC PREDICTOR MODELS

To assess the efficacy of a MPM in the prediction of BP, each of the 29 predictor variables were examined separately. For the MPM to be viable, it must first outperform each of the predictors singularly in the prediction of BP. Risk scores calculated for each of the predictor variables were entered into individual logistic regressions using the glm function of the rms R package. Area under the receiver operating characteristic (AUC) curve was used to assess the predictive validity of each of the 29 predictor data sets individually and was calculated using the prediction.obj function of the ROCR R package. The results of this process are shown in Table 3. As it became apparent that a large number of variables were gravitating towards large PTs, an additional approach was implemented whereby AUC values were rounded to the nearest hundredth decimal place in hopes of creating a more sparse SNP inclusion procedure to reduce problems of overfitting. For example, Verbal Numerical Reasoning's (VNR) highest AUC (0.51822) was at a PT of one, which led to the subsequent inclusion of 146,596 SNPs in the PRS calculation, however, when rounded, the highest AUC was 0.52 at the lowest PT of 0.0001, which lead to the inclusion

of 149 SNPs in the PRS calculation. Failure to have rounded the VNR AUC would have led to the inclusion of 146,447 more SNPs and an increase in AUC of 0.00217. The results of this process are shown in Table 4.

MULTI-POLYGENIC PREDICTOR MODEL

Risk scores calculated via different predictor data sets were incorporated as separate variables in elastic net regression. The threshold chosen for SNP inclusion of any given variable was that which provided the highest AUC, if the same AUC was given at different thresholds, the lowest threshold was always chosen. Elastic net regularized regression (Zou & Hastie, 2005), a method that penalizes the regression coefficients, was implemented in order to perform variable selection and model creation. Elastic net regression applies two regularization techniques, L1 regularization for variable selection and model shrinkage, and L2 regularization to combat issues of having more variables than sample size and perform grouping of correlated variables. Elastic net regression was primarily chosen as it selects groups of variables based on their collinearity, either including or excluding grouped variables together. The final model produced by the elastic net is both concise (elimination of unnecessary variables) and penalized, which serves to reduce the complications of selecting from large sets of variables, as well as to circumvent overfitting inherent in model construction.

MODEL TRAINING

To create a model with minimal bias and high accuracy, repeated 10-fold cross-validation with resampling was used (Kohavi, 1995). In this process, the training data set was first divided into ten equally sized segments, then nine of the ten segments were used to predict the training

sample while the remaining segment is withheld for validation. This process is then repeated ten times with each segment used only once for validation. After all ten folds have been used as validation, this process begins once more, until the 10-fold cross-validation has repeated a total of *n* times, with *n* representing the number of repeats chosen by the user. To reduce the total variance of the model, these cross-validation procedures were repeated 100 times. Cross-validation was performed using the trainControl function of the caret R package.

The hyperparameters of the elastic net regression, alpha and lambda, were derived during training. A grid of 10 alpha values was created by setting an ascending sequence from zero to one (i.e. 0.00, 0.11, 0.22, ..., 1.00) and a grid of 100 lambda values was created by setting a descending sequence from 100 to 0.01 (i.e. 100, 91.11, 83.02, …, 0.01). The alpha and lambda values were then given to the expand.grid R command to create a data frame with dimensions 1000 × 2 that listed all possible combinations of both alpha and lambda values. This grid was then provided in the train function of the caret R package under the tuneGrid argument. The best alpha and lambda values were then provided and used in model training. At this point, a cutoff value was chosen based on the highest AUC provided by the model and this cutoff value was used as a means to assess the models performance on the test sample.

The process stated above was then compared to the standard cv.glmnet function of the glmnet package in R. In all circumstances, cv.glmnet provided a model that produced a lower AUC than manually creating a grid and providing it to caret's train function. Therefore, cv.glmnet was not used for final model creation.

Results

Logistic regression was performed on each of the 29 variables at the 10 different PTs specified. An AUC was then calculated for every PT (Table 3 and Table 4). The highest AUC for a single variable (0.97081) was found when using the BP PGC GWAS with a PT of one, followed by SCZ (AUC = 0.77011, PT = 1), BP NIMH (AUC = 0.75175, PT = 1), anorexia (AUC = 0.60401, PT = 0.5), and MDD (AUC = 0.60164, PT = 0.3). Coefficients for the simple logistic regressions using variables BP PGC, BP NIMH, and SCZ are shown in Table 5. A high AUC for BP PGC was expected as the target and test samples used in this study were included in the analyses of the BP PGC GWAS and, as such, would be much more prone to overfitting due to data relatedness. The relatively high AUC for SCZ was also expected as this study, also conducted by the PGC, shared an overlap of controls with the BP PGC GWAS.

For a MPM to be viable it must outperform the single best predictor, in terms of classification accuracy, as such, multiple different models and approaches were examined. All regression coefficients and classification results of the model building process are shown in Tables 6 and 7, respectively. BP PGC was implemented in a model by itself on the training sample and provided an AUC of 0.97081, a training accuracy of 0.90240, and a prime cutoff value of 0.48863. This single predictor model, as well as the cutoff value, was then administered to the test sample, which provided a test accuracy of 0.97261. The above approach was also used on BP NIMH and SCZ to assess their efficacy as singular predictors, with their final test accuracy found to be 0.55141 and 0.54863, respectively.

All variables, with the chosen PT being that which provided the highest AUC in simple logistic regression, were included into an elastic net regression to create a classification model on PRSs derived from multiple traits. The AUC for classification of the training sample when using

all traits (using highest non-rounded AUC for inclusion criteria) was 0.98493, with a training accuracy of 0.93942, and a prime cutoff value of 0.49072. This multi-predictor model, as well as the derived cutoff value, was then applied to the test sample, which provided a test accuracy of 0.56689.

It was apparent that the highest AUC for 15 of the 28 variables was found when a PT of one was specified for risk score calculation, as such, to alleviate issues of overfitting inherent in this process, as well as to make an effort to include fewer SNPs in the risk score calculation, the AUCs were rounded to the nearest hundredth decimal place in an attempt to increase the generalizability of the model. The same steps stated above were then implemented to test the efficacy of the MPM approach when using the highest rounded AUC for inclusion criteria. For the training sample, an AUC of 0.98189, a training accuracy of 0.93058, and a prime cutoff value of 0.48008 was found. This rounded multi-predictor model and its cutoff value was then applied to the test sample which provided a test accuracy of 0.85113, showing a substantial improvement in terms of model generalizability.

Since the BP PGC variable was not independent of the training and test sample, it was decided that the singular BP PGC and rounded MPM classification results were most likely due to overfitting. As such, it was hypothesized that it would be beneficial to remove this biased variable and instead more heavily rely on the BP NIMH variable, which should, in theory, provide higher performance than the BP PGC, given its substantially larger sample size in the GWAS analysis. Non-rounded and rounded MPMs were implemented with the BP PGC variable excluded and resulted in test accuracies of 0.76419 and 0.55617, respectively. Surprisingly, the non-rounded model greatly outperformed the rounded model under these circumstances. It was possible that this was due, at least in part, to overfitting of the SCZ variable, which also shared a relationship

with the training and test samples. As such, non-rounded and rounded MPMs were once again created, this time excluding both the BP PGC and SCZ variables. These models provided test accuracies of 0.57840 (non-rounded) and 0.81064 (rounded).

Discussion

An objective measure for the diagnosis of psychiatric disorders is needed. With estimated heritability rates as high as 85% for BP (Smoller & Finn, 2003), it is a prime candidate for genetic based prediction. This project is one of the first to demonstrate that the prediction of BP, utilizing an individual's genotype information, is feasible. This research provides some of the first steps in forming an objective biological measure that could potentially be used alongside clinicians in a real world setting to help boost the accuracy of diagnosis. If genetic prediction theories were to be successfully implemented in practice, misdiagnosis would inevitably decline and lead to a decrease in patient treatment costs, as well as an increase in a patient's overall well-being.

The data suggests that genetic prediction for bipolar disorder is entirely feasible and further examination is warranted. It would also seem reasonable to assume that other highly heritable psychiatric disorders could be subject to similar protocol, however, the generalizability of this approach to other disorders was not examined in this study. The results reported herein also provide reasonable evidence for the incorporation of an MPM approach in psychiatric disorder genetic prediction. The test accuracy produced by the MPM, when excluding biased variables (BP PGC and SCZ), provides strong support for the MPM as it outperformed BP NIMH singularly. With that said, the MPM did fail to outperform BP PGC outright and, as such, should be interpreted with caution, however, this particular finding is most likely due to biased variables and data relatedness. While an accuracy of 81% is promising, there is still substantial room for improvement in future studies. Incorporating environmental factors, such as early life stressors and socioeconomic status, into the model building process could prove extremely beneficial as it is widely held that an individual's genetic composition is not the sole determinant in the development of BP or other psychiatric disorders.

While generating a single numerical score that estimates an individual's genetic risk for a particular disorder is an attractive concept, PRS calculation is most likely oversimplifying a complex biological process. With costs of genetic sequencing decreasing rapidly over time, it will become more and more feasible for scientists to move away from microarray analyses, and thus the implementation of PRSs, and focus more heavily on the examination of the multitudes of different sequencing applications. Whole-genome sequencing in particular will enable scientists to examine biological phenomenon in more detail and will also circumvent some of the current limitations encountered in this study, such as the inability of genotyping technologies to examine extremely rare (minor allele frequency < 0.01) variants. Multi-omic approaches, studies that make use of multiple different "omic" technologies, have become more widespread as the availability of this data continues to rise (Hasin, Seldin, & Lusis, 2017). It is possible that the integration of multi-omic information can be beneficial in the derivation of a psychiatric classification model. In a similar network, this proposed methodology has already been implemented in a number of studies examining classification of cancer types, such as that conducted by Rappoport & Shamir (2018). While the amount of available psychiatric multi-omics data is not as extensive as is that for cancer, this remains a plausible avenue of research for future application.

With this said, a genetic-based prediction model can only be used in real-world settings if it is first validated and is able to show a high level of consistency in its ability to accurately separate affected and unaffected individuals. It must be noted that the goal of this study was not necessarily to develop a model that can be used directly in a clinical setting, but rather to provide theory and suggested application. A significant constraint of this study is the use of a binary model, as should we attempt to try and predict the diagnosis of a schizophrenic individual using the models generated here, the individual would most likely align with the bipolar disorder group, assuming

current theory is correct. For more realistic application, scientists will need to move away from binary classification and attempt to implement multi-classification models with the end goal being the separation of target samples with various different psychiatric disorders or possibly even heterogeneous subtyping.

As with all genetic prediction, there are many ethical and moral concerns to be thoroughly discussed. Prediction in early, or even prenatal, development is becoming more and more realistic as the field continues to advance at a rapid pace. While concerns such as these are warranted, the beneficial impact this research may have should not be ignored. If early life diagnosis becomes achievable, affected individuals would be able to undergo assessment and treatment at a much earlier age, potentially alleviating many of the difficulties faced in later stages of life. The concerns associated with early life prediction should be examined extensively and methodically before any implementation is considered.

Table 1. Individual sample information.

Sample	Ancestry	Case	Control	Platform
Systematic Treatment Enhancement Program for Bipolar Disorder	European-American	922	645	Affymetrix GeneChip Human Mapping 500K Array
University College London	British	457	495	Affymetrix GeneChip Human Mapping 500K Array
Systematic Treatment Enhancement Program for Bipolar Disorder	European-American	659	192	Affymetrix Genome-Wide Human SNP Array 5.0
Thematically Organized Psychosis Study	Norwegian	203	349	Affymetrix Genome-Wide Human SNP Array 6.0
Trinity College Dublin	Irish	150	797	Affymetrix Genome-Wide Human SNP Array 6.0
University of Edinburgh	Scottish	282	275	Affymetrix Genome-Wide Human SNP Array 6.0
Pritzker Neuropsychiatric Disorders Research Consortium	European-American	1130	718	Illumina HumanHap 550

Table 2. GWAS summary statistics information.

File name	Trait	Consortium	Ethnicity	Sample size	Number of SNPs	PMID	Publish Year
tag.logonset.tbl	Age of Smoking Initiation	TAG	European	47961	2457545	20418890	2010
pgc_alcdep.eur_unrel_genotyped.aug2018_release.txt	Alcohol Dependence	PGC	European	28757	9142831	30482948	2018
anxiety.meta.full.cc.tbl	Anxiety	ANGST	European	17310	6330995	26754954	2016
gcan_meta.out	Anorexia	GCAN	European	17767	1147629	24514567	2014
PGC.ASD.euro.all.25Mar2015.txt	Autism	PGC	European	10263	9499589	unpublished	2015
pgc.bip.full.2012-04.txt	Bipolar Disorder	PGC	European	16731	2427220	21926972	2011
BP_GWAS_Hou_et_al._2016_results.txt	Bipolar Disorder	NIMH	European	40255	9877008	27329760	2016
BMI.ACTIVE.ALL.European.txt	Body Mass Index	GIANT	European	123865	2692049	20935630	2010
tag.cpd.tbl	Cigarettes per Day	TAG	European	68028	2459118	20418890	2010
cad.add.160614.website.txt	Coronary Artery Disease	CARDIoGRAM	Mixed	184035	9455778	26343387	2015
EUR.CD.gwas_info03_filtered.assoc	Crohn's Disease	IBD Genetics	Mixed	51109	11002658	21102463	2010
DS_Full.txt	Depressive Symptoms	SSGAC	European	161460	6524474	27089181	2016
Davies2016_UKB_college_summary_results_22072016.txt	Educational Attainment	UK Biobank	European	111114	17344347	27046643	2016
tag.evrsmk.tbl	Ever Smoker	TAG	European	74035	2455846	20418890	2010
tag.former.tbl	Former Smoker	TAG	European	70675	2456554	20418890	2010
Hill2016_UKB_Income_summary_results_21112016.txt	Household Income	UK Biobank	European	96900	17344716	27818178	2016
EUR.IBD.gwas_info03_filtered.assoc	Inflammatory Bowel Disease	IBD Genetics	Mixed	96486	11555662	26192919	2015
MDD2018_ex23andMe	Major Depressive Disorder	PGC	European	307354	13554489	29700475	2018
Davies2016_UKB_Memory_summary_results_22072016.txt	Memory	UK Biobank	European	112067	17344579	27046643	2016
GPC-1.NEO-OPENNESS.full.txt	Openness to Experience	GPC	European	17375	2305640	21173776	2012
Neuroticism_Full.txt	Neuroticism	SSGAC	European	170911	6524432	27089181	2016
Hill2016_UKB_Income_summary_results_One_person_per_household_summary_results_21112016.txt	One Person Per Household Income	UK Biobank	European	88183	17345296	27818178	2016
meta3.INTsnplist_F.txt	Preschool Internalizing Problems	EAGLE	European	4596	2821734	24839885	2014
ckqny.scz2snpres	Schizophrenia	PGC	Mixed	150064	9444230	25056061	2014
SWB_Full.txt	Subjective Well-Being	SSGAC	European	298420	2268674	27089181	2016
jointGwasMc_TG.txt	Triglycerides	GLGC	European	96598	2438639	20686565	2010
EUR.UC.gwas_info03_filtered.assoc	Ulcerative Colitis	IBD Genetics	European	48950	11113952	21297633	2011
Davies2016_UKB_VNR_summary_results_22072016.txt	Verbal-Numerical Reasoning	UK Biobank	European	36035	17361492	27046643	2016
EduYears_Main.txt	Years of Schooling	SSGAC	European	293723	8146840	27225129	2016

ANGST: Anxiety NeuroGenetics STudy; CARDIoGRAM: Coronary ARtery DIsease Genome wide Replication And Meta-analysis; EAGLE: Early Genetics and Lifecourse Epidemiology; GCAN: Genetics Consortium for Anorexia Nervosa; GIANT: Genetic Investigation of ANthropometrix Traits; GLGC: Global Lipids Genetics Consortium; GPC: Genetics of Personality Consortium; IBD Genetics: Inflammatory Bowel Disease Consortium; NIMH: National Institute of Mental Health; PGC: Psychiatric Genomics Consortium; SSGAC: Social Science Genetic Association Consortium; TAG: Tobacco And Genetics Consortium; UK Biobank: United Kingdom Biobank.

Table 3. AUC derived on PRSs at multiple different p-value thresholds for multiple traits examined by GWASs.

	P-value Threshold									
	0.0001	0.0005	0.001	0.005	0.01	0.05	0.1	0.3	0.5	1
Trait										
Alcdep	0.49860	0.51179	0.51951	0.54691	0.55257	0.56935	0.57858	0.58319	0.58469	**0.58537**
Anorexia	0.49832	0.51331	0.51841	0.53666	0.54643	0.57884	0.59448	0.59573	**0.60401**	0.60341
Anxiety	0.51516	0.49986	0.50061	0.51864	0.52399	0.53693	**0.54026**	0.53501	0.53445	0.53383
Autism	0.50449	0.50409	0.50396	0.50678	0.50102	0.50530	0.50598	0.50422	0.50448	**0.50703**
BP PGC	0.62542	0.72724	0.76723	0.85916	0.89721	0.95038	0.96066	0.96816	0.97015	**0.97081**
BP NIMH	0.57688	0.60278	0.61319	0.65991	0.68614	0.73176	0.74150	0.74472	0.74934	**0.75175**
BMI	**0.53277**	0.51962	0.51691	0.50851	0.50078	0.50091	0.50422	0.50063	0.50163	0.50214
CAD	**0.51246**	0.50290	0.49905	0.50510	0.51039	0.50508	0.50187	0.50371	0.49947	0.49943
CD	0.52216	**0.52849**	0.52814	0.52499	0.52464	0.52519	0.52021	0.52534	0.52593	0.52519
College	0.51432	0.53895	0.54053	0.53185	0.53174	0.53531	0.53853	0.53913	**0.54439**	0.54283
CPD	0.50331	0.50472	0.51043	0.50670	0.51627	0.51823	0.51818	0.52444	0.52394	**0.52472**
DS	0.52789	0.52735	0.53849	0.55650	0.55987	0.56400	0.55649	**0.55976**	0.55929	0.55827
Evrsmk	0.50510	0.50810	0.51740	0.51365	0.52471	0.52907	0.53623	**0.54544**	0.54447	0.54238
Former	0.50565	0.51677	0.53066	0.53195	0.53265	0.54051	0.54226	0.55128	0.55191	**0.55204**
IBD	0.50611	0.51226	0.51330	0.51641	0.51848	0.52486	0.53737	0.54641	0.54908	**0.55047**
Income	0.51440	0.51224	0.50192	0.50601	0.50851	0.50424	0.51665	0.52228	**0.52234**	0.51958
INT	0.51446	0.50801	0.49994	0.51812	0.51750	0.51038	0.52260	**0.54036**	0.47049	0.46656
Logonset	0.50296	0.50169	0.50810	0.51026	0.52129	0.53404	0.54028	0.54094	0.54133	**0.54150**
MDD	0.56735	0.57012	0.56960	0.59586	0.59440	0.58037	0.59478	**0.60164**	0.58942	0.58372
Memory	0.51113	0.52321	0.52279	**0.52626**	0.51813	0.50201	0.49762	0.50508	0.50936	0.50686
Neuroticism	0.51226	0.51428	0.51747	0.52676	0.53073	**0.53309**	0.53035	0.52512	0.52604	0.52637
Openness	0.50508	0.50628	0.51825	0.53191	0.54008	0.54493	0.54837	0.54734	0.54899	**0.54952**
OPPH	0.51993	0.50662	0.50486	0.50377	0.50976	0.50777	0.51020	0.52247	0.52455	**0.52534**
SCZ	0.61725	0.63972	0.65740	0.69215	0.70626	0.73877	0.74958	0.76530	0.76941	**0.77011**
SWB	0.52151	0.52842	0.52770	0.53242	0.53021	0.54515	0.55060	**0.55434**	0.55336	0.55364
TG	0.54653	0.54375	0.54480	0.55323	0.54956	0.54975	0.55441	0.55514	0.55462	**0.55538**
UC	0.50676	0.50789	0.51060	0.51007	0.51578	0.52679	0.52691	0.52750	0.53117	**0.53140**
VNR	0.51605	0.50145	0.50507	0.52077	0.52226	0.50841	0.50317	0.51234	0.51732	**0.51822**
YoS	0.52523	0.53400	0.53758	0.54340	0.55199	0.57346	0.57549	0.57479	0.57426	**0.57584**

Highest AUC for a given trait is shown in bold. Alcdep: alcohol dependence; Anorexia: anorexia nervosa; Anxiety: anxiety disorders; Autism: autism spectrum disorders; BP PGC: bipolar disorder study conducted by the Psychiatric Genomics Consortium; BP NIMH: bipolar disorder study conducted by the National Institute of Mental Health Genetics Consortium; BMI: body mass index; CAD: coronary artery disease; CD: Crohn's disease; College: educational attainment; CPD: cigarettes smoked per day; DS: depressive symptoms; Evrsmk: ever smoked; Former: former smoker; IBD: inflammatory bowel disease: Income: household income; INT: preschool internalizing problems; Logonset: age of smoking initiation; MDD: major depressive disorder; Memory: declarative memory; Neuroticism: Big 5 neuroticism; Openness: Big 5 neo-openness; OPPH: one person per household income; SCZ: schizophrenia; SWB: subjective well-being; TG: Triglycerides; UC: ulcerative colitis; VNR: verbal-numerical reasoning; YoS: total years of schooling.

Table 4. Rounded AUCs derived on PRSs at multiple different p-value thresholds for multiple traits examined by GWASs.

	P-value Threshold									
	0.0001	0.0005	0.001	0.005	0.01	0.05	0.1	0.3	0.5	1
Trait										
Alcdep	0.50	0.51	0.52	0.55	0.55	0.57	0.58	0.58	0.58	**0.59**
Anorexia	0.50	0.51	0.52	0.54	0.55	0.58	0.59	**0.60**	0.60	0.60
Anxiety	0.52	0.50	0.50	0.52	0.52	**0.54**	0.54	0.54	0.53	0.53
Autism	0.50	0.50	0.50	**0.51**	0.50	0.51	0.51	0.50	0.50	0.51
BP PGC	0.63	0.73	0.77	0.86	0.90	0.95	0.96	**0.97**	0.97	0.97
BP NIMH	0.58	0.60	0.61	0.66	0.69	0.73	0.74	0.74	**0.75**	0.75
BMI	**0.53**	0.52	0.52	0.51	0.50	0.50	0.50	0.50	0.50	0.50
CAD	**0.51**	0.50	0.50	0.51	0.51	0.51	0.50	0.50	0.50	0.50
CD	0.52	**0.53**	0.53	0.52	0.52	0.53	0.52	0.53	0.53	0.53
College	0.51	**0.54**	0.54	0.53	0.53	0.54	0.54	0.54	0.54	0.54
CPD	0.50	0.50	0.51	0.51	**0.52**	0.52	0.52	0.52	0.52	0.52
DS	0.53	0.53	0.54	**0.56**	0.56	0.56	0.56	0.56	0.56	0.56
Evrsmk	0.51	0.51	0.52	0.51	0.52	0.53	0.54	**0.55**	0.54	0.54
Former	0.51	0.52	0.53	0.53	0.53	0.54	0.54	**0.55**	0.55	0.55
IBD	0.51	0.51	0.51	0.52	0.52	0.52	0.54	**0.55**	0.55	0.55
Income	0.51	0.51	0.50	0.51	0.51	0.50	**0.52**	0.52	0.52	0.52
INT	0.51	0.51	0.50	0.52	0.52	0.51	0.52	**0.54**	0.47	0.47
Logonset	0.50	0.50	0.51	0.51	0.52	0.53	**0.54**	0.54	0.54	0.54
MDD	0.57	0.57	0.57	**0.60**	0.59	0.58	0.59	0.60	0.59	0.58
Memory	0.51	0.52	0.52	**0.53**	0.52	0.50	0.50	0.51	0.51	0.51
Neuroticism	0.51	0.51	0.52	**0.53**	0.53	0.53	0.53	0.53	0.53	0.53
Openness	0.51	0.51	0.52	0.53	0.54	0.54	**0.55**	0.55	0.55	0.55
OPPH	0.52	0.51	0.50	0.50	0.51	0.51	0.51	0.52	0.52	**0.53**
SCZ	0.62	0.64	0.66	0.69	0.71	0.74	0.75	**0.77**	0.77	0.77
SWB	0.52	0.53	0.53	0.53	0.53	**0.55**	0.55	0.55	0.55	0.55
TG	0.55	0.54	0.54	0.55	0.55	0.55	0.55	**0.56**	0.55	0.56
UC	0.51	0.51	0.51	0.51	0.52	**0.53**	0.53	0.53	0.53	0.53
VNR	**0.52**	0.50	0.51	0.52	0.52	0.51	0.50	0.51	0.52	0.52
YoS	0.53	0.53	0.54	0.54	0.55	0.57	**0.58**	0.57	0.57	0.58

Highest AUC for a given trait is shown in bold. Alcdep: alcohol dependence; Anorexia: anorexia nervosa; Anxiety: anxiety disorders; Autism: autism spectrum disorders; BP PGC: bipolar disorder study conducted by the Psychiatric Genomics Consortium; BP NIMH: bipolar disorder study conducted by the National Institute of Mental Health Genetics Consortium; BMI: body mass index; CAD: coronary artery disease; CD: Crohn's disease; College: educational attainment; CPD: cigarettes smoked per day; DS: depressive symptoms; Evrsmk: ever smoked; Former: former smoker; IBD: inflammatory bowel disease: Income: household income; INT: preschool internalizing problems; Logonset: age of smoking initiation; MDD: major depressive disorder; Memory: declarative memory; Neuroticism: Big 5 neuroticism; Openness: Big 5 neo-openness; OPPH: one person per household income; SCZ: schizophrenia; SWB: subjective well-being; TG: Triglycerides; UC: ulcerative colitis; VNR: verbal-numerical reasoning; YoS: total years of schooling.

Table 5. Regression coefficients for simple logistic regression models.

Trait	Non-Rounded BP (PGC)	Rounded BP (PGC)	Non-Rounded BP (NIMH)	Rounded BP (NIMH)	Non-Rounded SCZ	Rounded SCZ
Intercept	-17.5607	-16.1519	23.86	14.68	39.859	39.07
BP PGC	38151.084	20735.5446	NA	NA	NA	NA
BP NIMH	NA	NA	21450	15340	NA	NA
SCZ	NA	NA	NA	NA	46348.409	26250.104

BP PGC: bipolar disorder study conducted by the Psychiatric Genomics Consortium; BP NIMH: bipolar disorder study conducted by the National Institute of Mental Health Genetics Consortium; SCZ: schizophrenia.

Table 6. Regression coefficients for multi-polygenic predictor models.

Trait	Non-Rounded MPM	Rounded MPM	Non-Rounded MPM - BP (PGC) excluded	Rounded MPM - BP (PGC) excluded	Non-Rounded MPM - BP (PGC) and SCZ excluded	Rounded MPM - BP (PGC) and SCZ excluded
Intercept	-96.94444	-74.979202	-10.3858666	-7.4328533	-34.13627	-28.32416
Alcdep	2459.68875	2800.942491	1875.953059	2253.591377	1706.71899	2346.15849
Anorexia	3989.38227	2996.63294	1066.568552	802.5648813	2009.351	1416.92933
Anxiety	.	12.606423	-0.9031498	43.5836368	.	23.98228
Autism	.	-63.224565	-155.746319	-36.4274775	.	-35.18637
BP (PGC)	26463.83938	14303.11812	NA	NA	NA	NA
BP (NIMH)	-56.58548	-99.028914	13109.57106	8376.498009	-121.15683	11616.57034
BMI	.	7.102537	-107.0921427	-99.3551526	32.35455	-98.23423
CAD	.	-8.113078	35.4877234	31.7325285	.	21.17856
CD	6750.08084	-21.658604	-3.5385627	0.2440534	3876.8225	.
College	-129.04479	-18.619458	5548.230201	206.3624924	138.64419	184.04608
CPD	3272.5812	650.321552	201.6442959	4.9343829	8290.32759	.
DS	-5044.31038	-4893.224933	5924.258851	698.4235454	-1020.71742	994.98831
Evrsmk	1979.71442	1304.471172	-1872.590716	-1648.607391	1319.72505	-988.33835
Former	6600.99984	3312.284547	712.9202237	355.6200735	3545.63115	661.28323
IBD	1257.49852	.	3815.824048	1280.361676	2850.34165	543.63331
Income	73.40921	103.680915	3269.849015	234.9909525	108.61418	.
INT	-5613.95989	-2511.205372	95.78244	90.1016555	-7648.7379	89.57707
Logonset	.	247.035386	-2825.350564	-1562.332739	5075.82959	-2328.79962
MDD	.	.	3015.529478	528.6195298	367.16917	886.29848
Memory	.	-387.262055	217.7386108	187.3721036	.	242.01629
Neuroticism	.	.	-71.8485767	-109.7288885	18575.59833	.
Openness	450.1478	174.601369	21.9920758	8376.498009	412.02422	130.41048
OPPH	6178.0403	10348.48981	5086.003194	53.23667	4585.90657	3871.84333
SCZ	14160.9248	7815.776374	27097.76169	5953.961511	NA	NA
SWB	-6299.58581	-1817.719839	-5277.865037	13634.66003	-7021.5495	-1647.02413
TG	15368.19751	8498.062691	8365.368878	-1304.539482	9952.87307	5025.86637
UC	-1723.24221	-204.309772	-3007.690578	4321.31047	-3429.35404	-108.38843
VNR	.	13.707697	-2323.844887	-346.2942198	-6183.58599	.
YoS	.	2541.127777	22255.3316	17.2981481	26709.3741	7982.79392

Variables excluded by the elastic net are denoted as a period. Alcdep: alcohol dependence; Anorexia: anorexia nervosa; Anxiety: anxiety disorders; Autism: autism spectrum disorders; BP PGC: bipolar disorder study conducted by the Psychiatric Genomics Consortium; BP NIMH: bipolar disorder study conducted by the National Institute of Mental Health Genetics Consortium; BMI: body mass index; CAD: coronary artery disease; CD: Crohn's disease; College: educational attainment; CPD: cigarettes smoked per day; DS: depressive symptoms; Evrsmk: ever smoked; Former: former smoker; IBD: inflammatory bowel disease: Income: household income; INT: preschool internalizing problems; Logonset: age of smoking initiation; MDD: major depressive disorder; Memory: declarative memory; Neuroticism: Big 5 neuroticism; Openness: Big 5 neo-openness; OPPH: one person per household income; SCZ: schizophrenia; SWB: subjective well-being; TG: Triglycerides; UC: ulcerative colitis; VNR: verbal-numerical reasoning; YoS: total years of schooling. MPM: classification model consisting of all variables at their best AUC as reported from simple logistic regression; MPM – BP PGC excluded: classification model consisting of all variables, except BP PGC, at their best AUC as reported from simple logistic regression; MPM – BP PGC and SCZ excluded: classification model consisting of all variables, except BP PGC and SCZ, at their best AUC as reported from simple logistic regression. Non-Rounded and Rounded refer to the AUC derived at each p-value threshold, Non-Rounded models included variables chosen at the best AUC, whereas Rounded models included variables chosen at the best AUC after rounding the AUC to the nearest hundredth decimal place.

Table 7. Classification results for each model examined.

Model	Train AUC	Train Accuracy	Prime Cutoff	Train MSE	Test Accuracy
Non-Rounded BP PGC	0.97081	0.90240	0.48863	0.06904	0.97261
Rounded BP PGC	0.96816	0.89945	0.47581	0.07181	0.96189
Non-Rounded BP NIMH	0.75175	0.67143	0.58080	0.19899	0.55141
Rounded BP NIMH	0.74934	0.66786	0.53182	0.20017	0.58992
Non-Rounded SCZ	0.77011	0.69983	0.51562	0.19503	0.54863
Rounded SCZ	0.76530	0.69499	0.50593	0.19693	0.54744
Non-Rounded MPM	0.98493	0.93942	0.49072	0.05082	0.56689
Rounded MPM	0.98189	0.93058	0.48008	0.05540	0.85113
Non-Rounded MPM - BP PGC excluded	0.84435	0.77072	0.50192	0.16266	0.76419
Rounded MPM - BP PGC excluded	0.84043	0.76525	0.49999	0.16659	0.55617
Non-Rounded MPM - BP PGC and SCZ excluded	0.80091	0.72592	0.51853	0.18113	0.57840
Rounded MPM - BP PGC and SCZ excluded	0.79833	0.72402	0.54511	0.18374	0.81064

BP PGC: bipolar disorder study conducted by the Psychiatric Genomics Consortium; BP NIMH: bipolar disorder study conducted by the National Institute of Mental Health Genetics Consortium; SCZ: schizophrenia study conducted by the Psychiatric Genomics Consortium; MPM: classification model consisting of all variables at their best AUC as reported from simple logistic regression; MPM – BP PGC excluded: classification model consisting of all variables, except BP PGC, at their best AUC as reported from simple logistic regression; MPM – BP PGC and SCZ excluded: classification model consisting of all variables, except BP PGC and SCZ, at their best AUC as reported from simple logistic regression. Non-Rounded and Rounded refer to the AUC derived at each p-value threshold, Non-Rounded models included variables chosen at the best AUC, whereas Rounded models included variables chosen at the best AUC after rounding the AUC to the nearest hundredth decimal place.

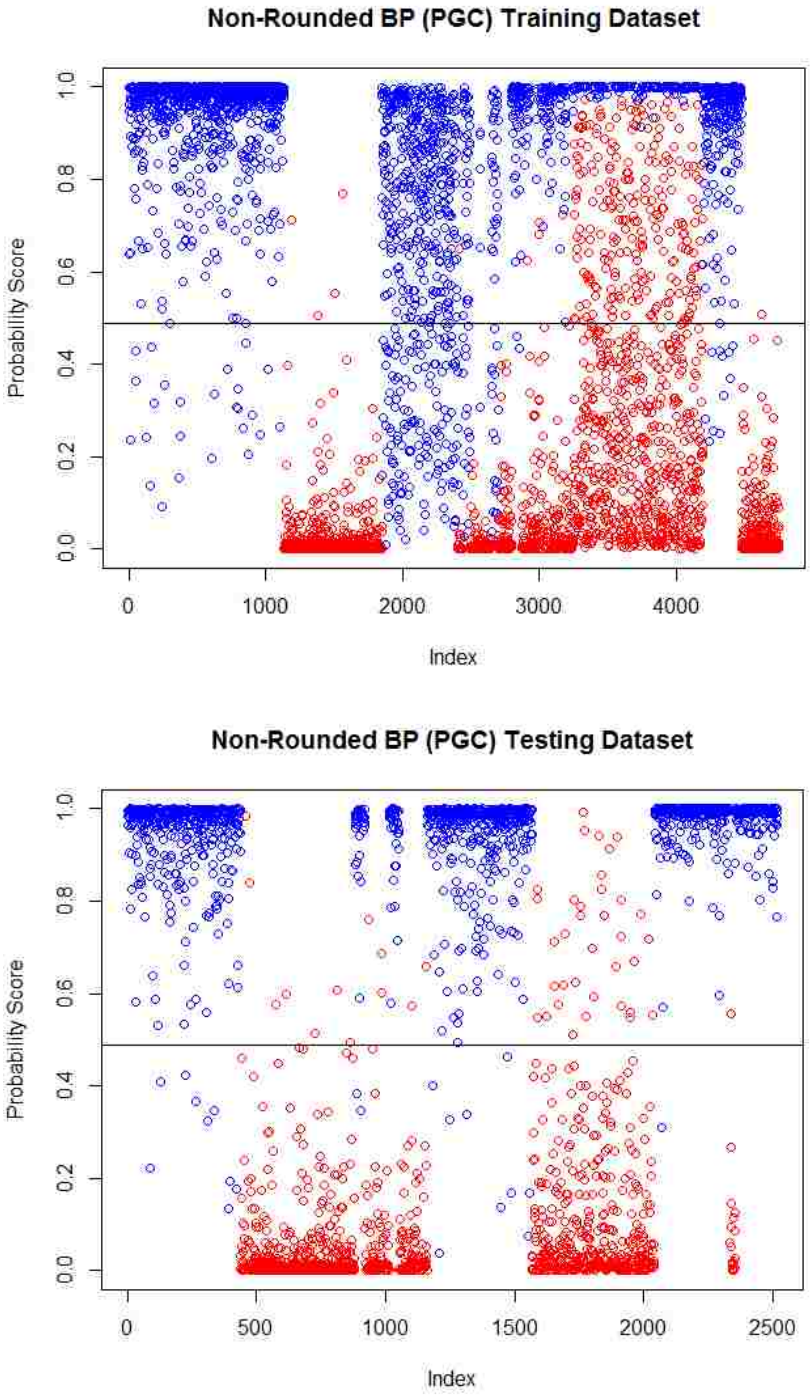


Figure 1. Probability scores for simple logistic regression using only the BP PGC variable at the non-rounded highest AUC. Blue denotes case and red denotes control.

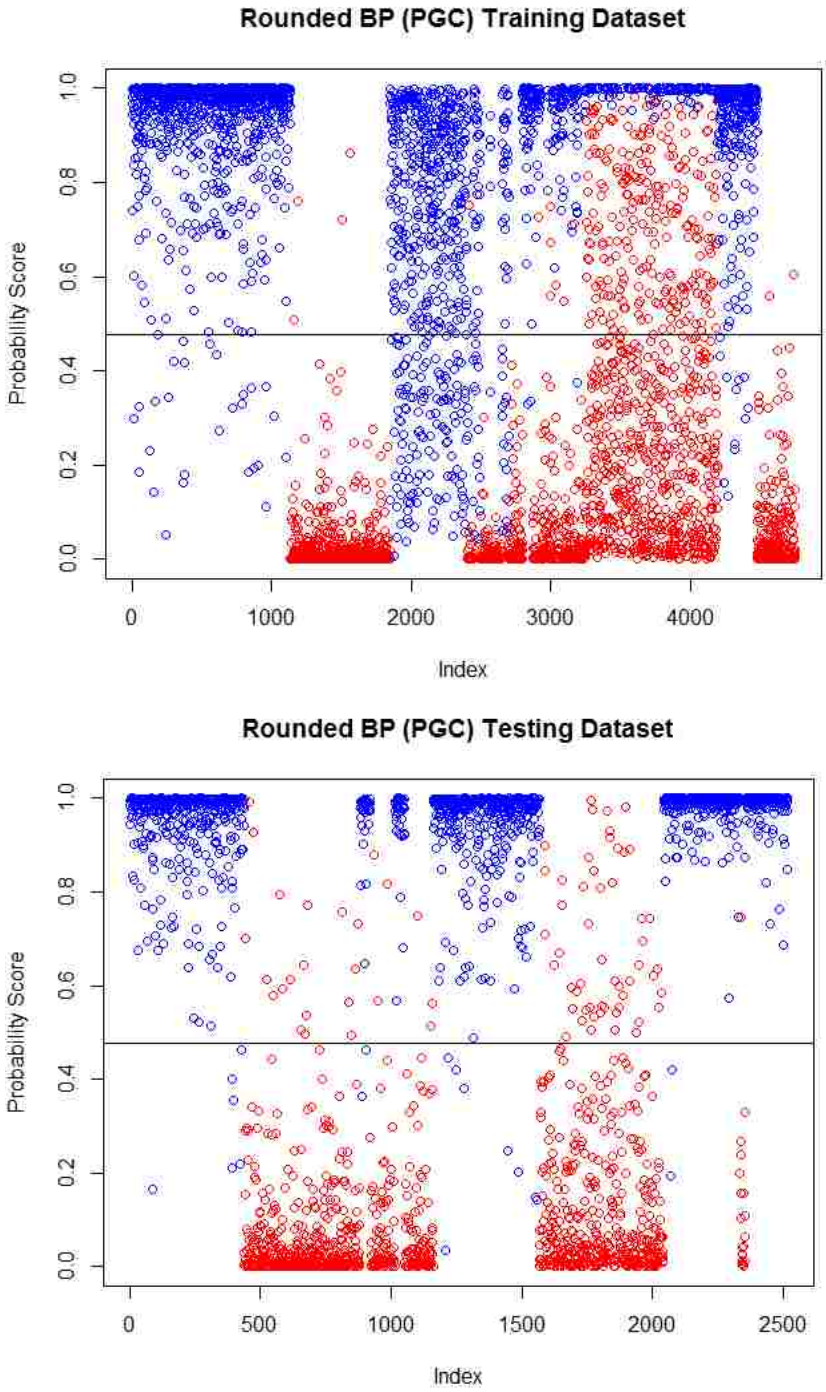


Figure 2. Probability scores for simple logistic regression using only the BP PGC variable at the rounded highest AUC. Blue denotes case and red denotes control.

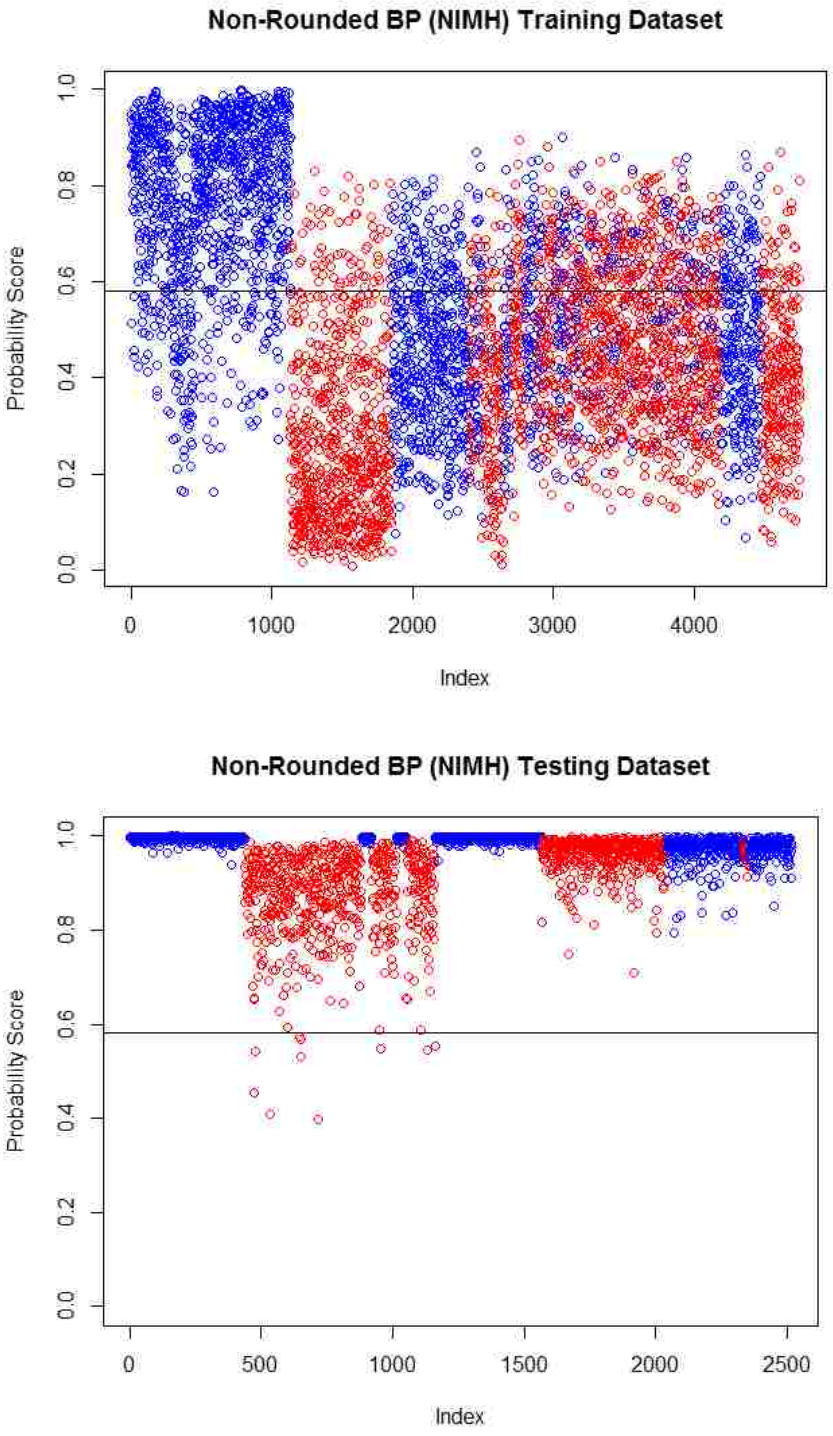


Figure 3. Probability scores for simple logistic regression using only the BP NIMH variable at the non-rounded highest AUC. Blue denotes case and red denotes control.

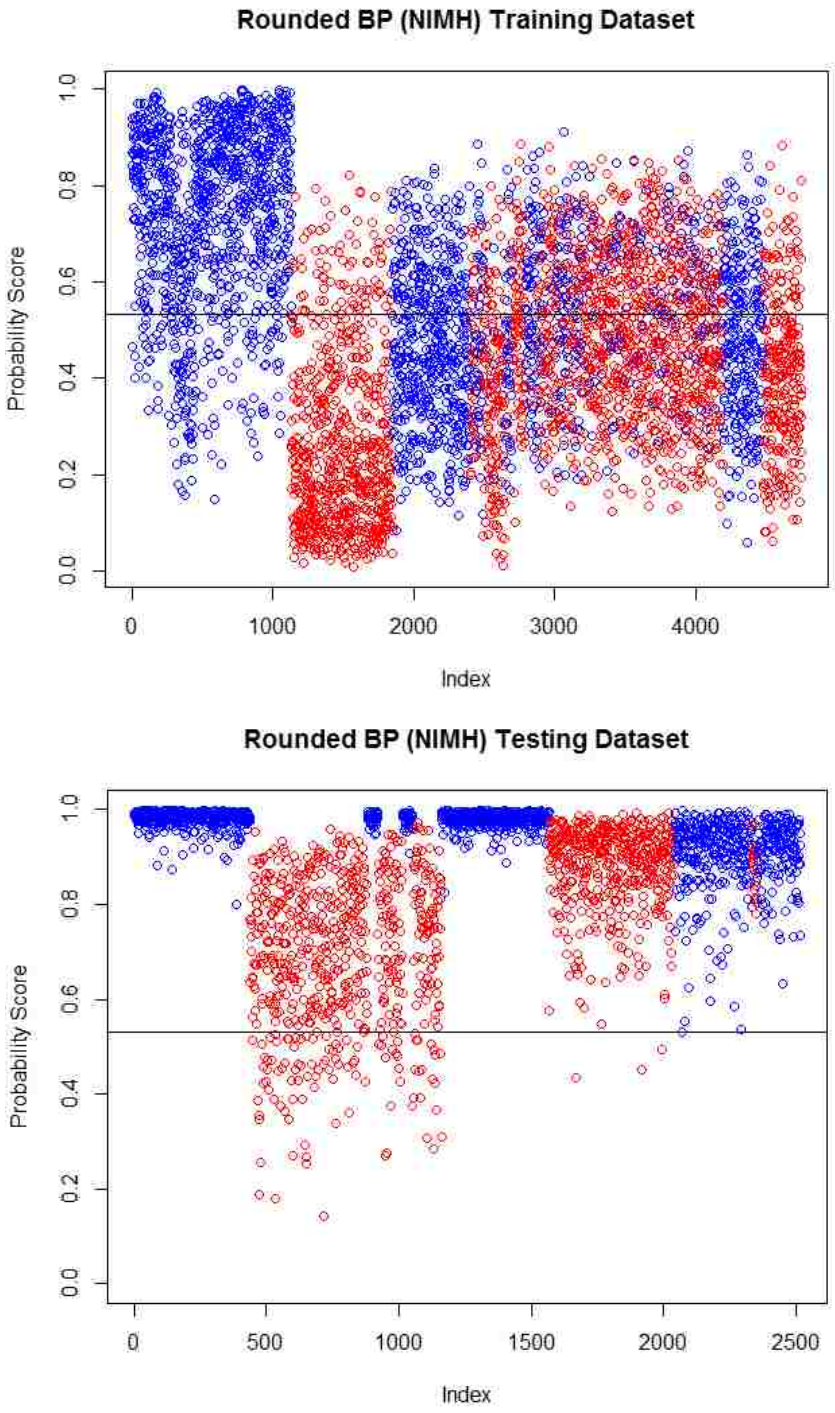


Figure 4. Probability scores for simple logistic regression using only the BP NIMH variable at the rounded highest AUC. Blue denotes case and red denotes control.

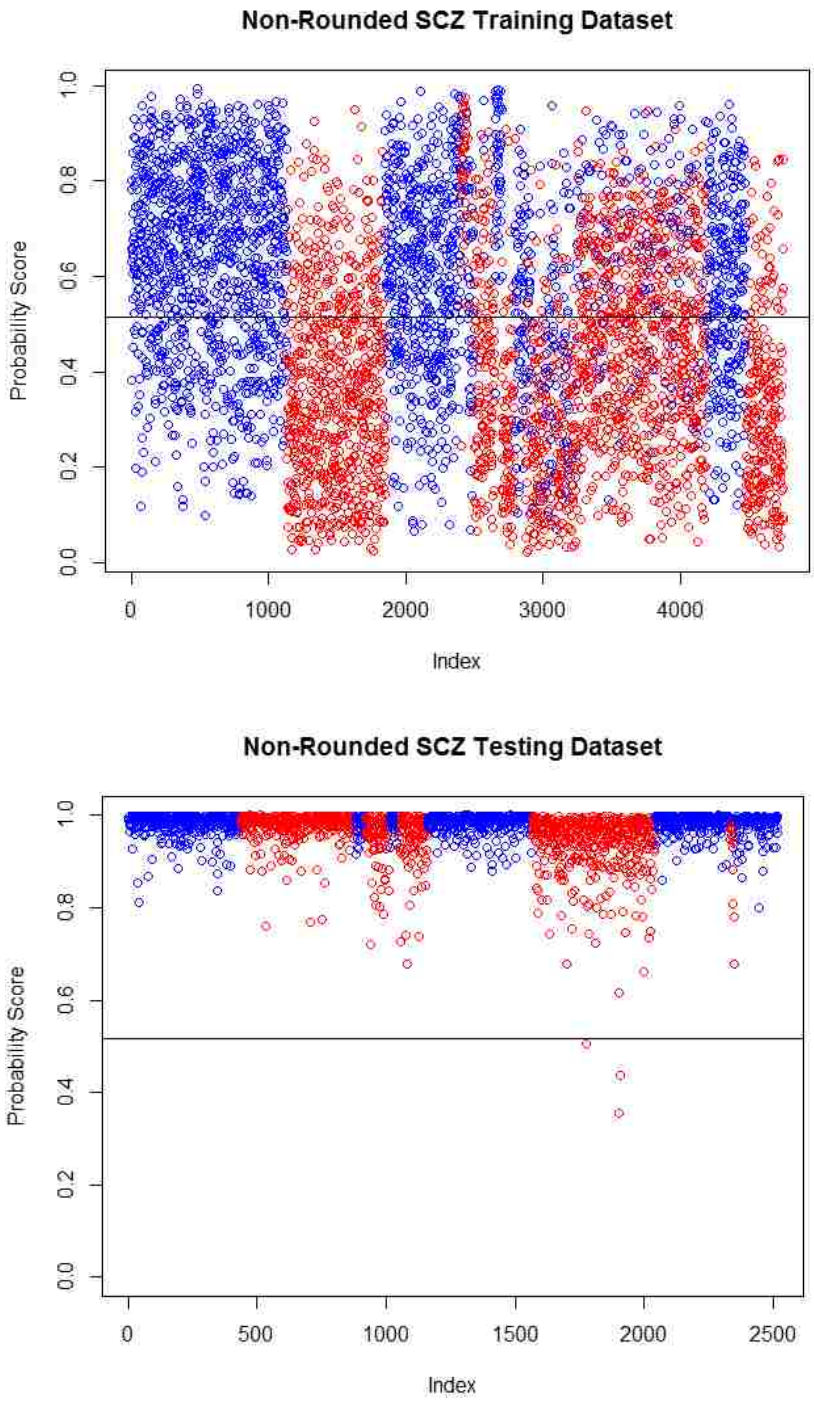


Figure 5. Probability scores for simple logistic regression using only the SCZ variable at the non-rounded highest AUC. Blue denotes case and red denotes control.

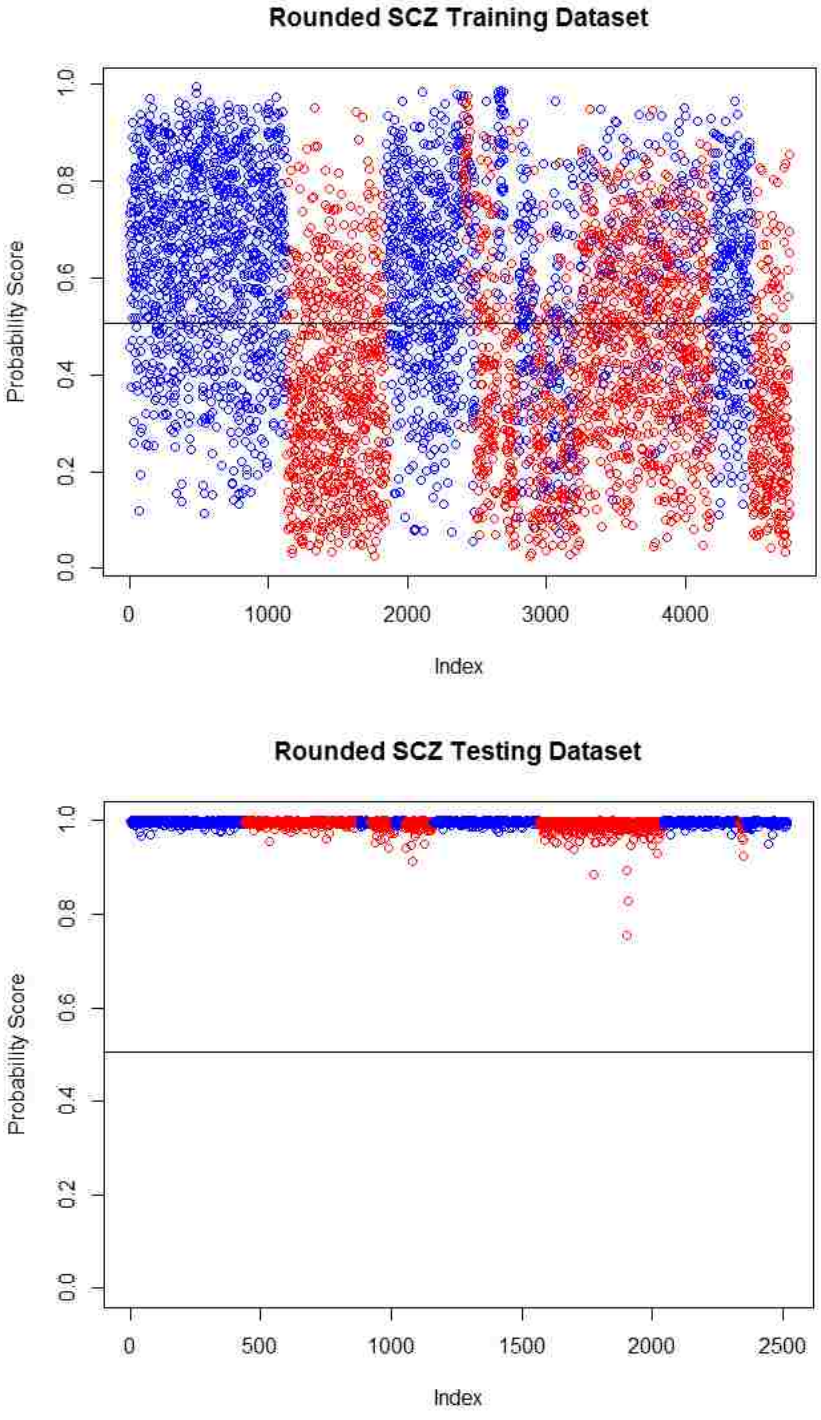


Figure 6. Probability scores for simple logistic regression using only the SCZ variable at the rounded highest AUC. Blue denotes case and red denotes control.

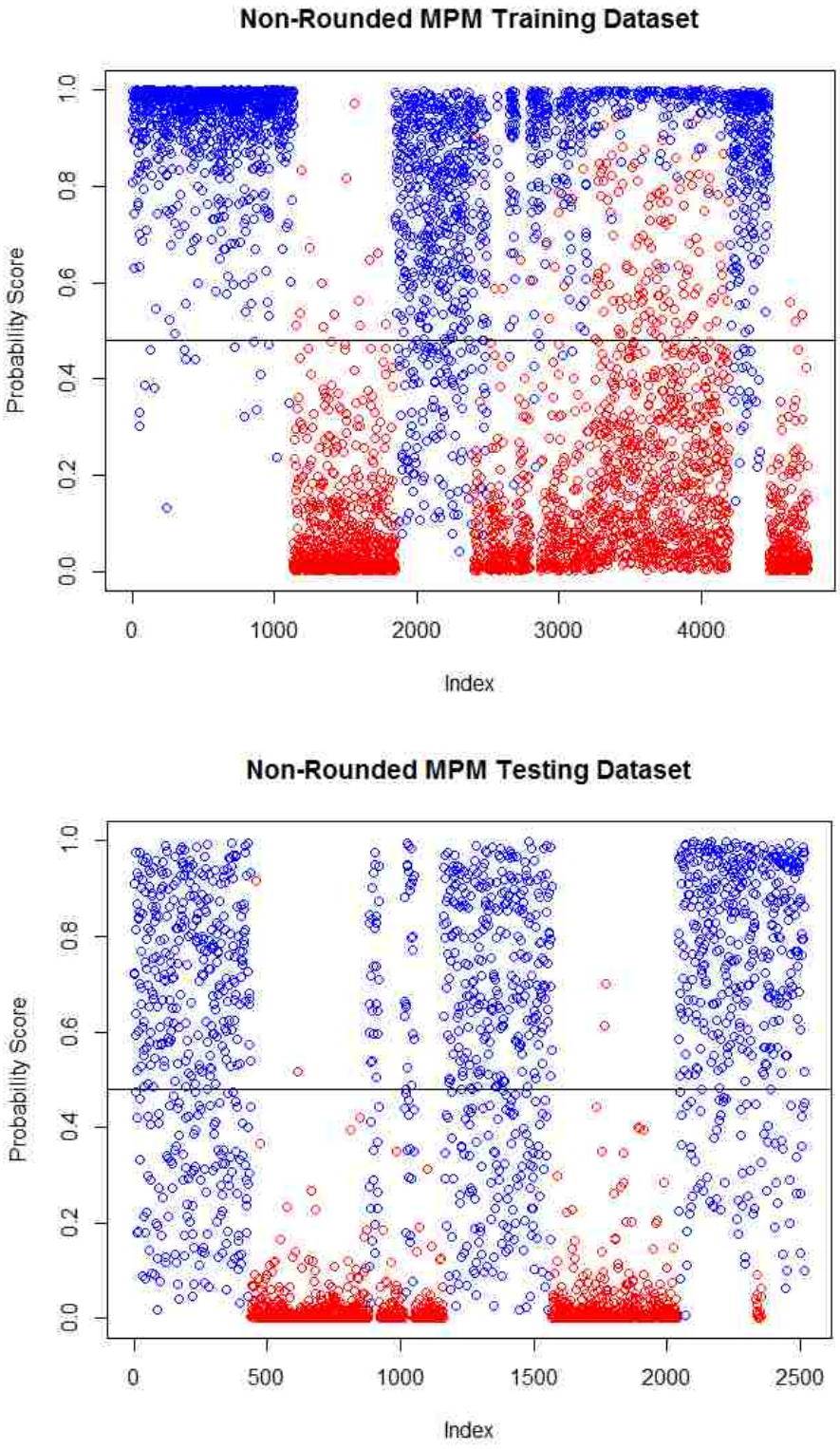


Figure 7. Probability scores for the multi-polygenic prediction model using all variables at the non-rounded highest AUC. Blue denotes case and red denotes control.

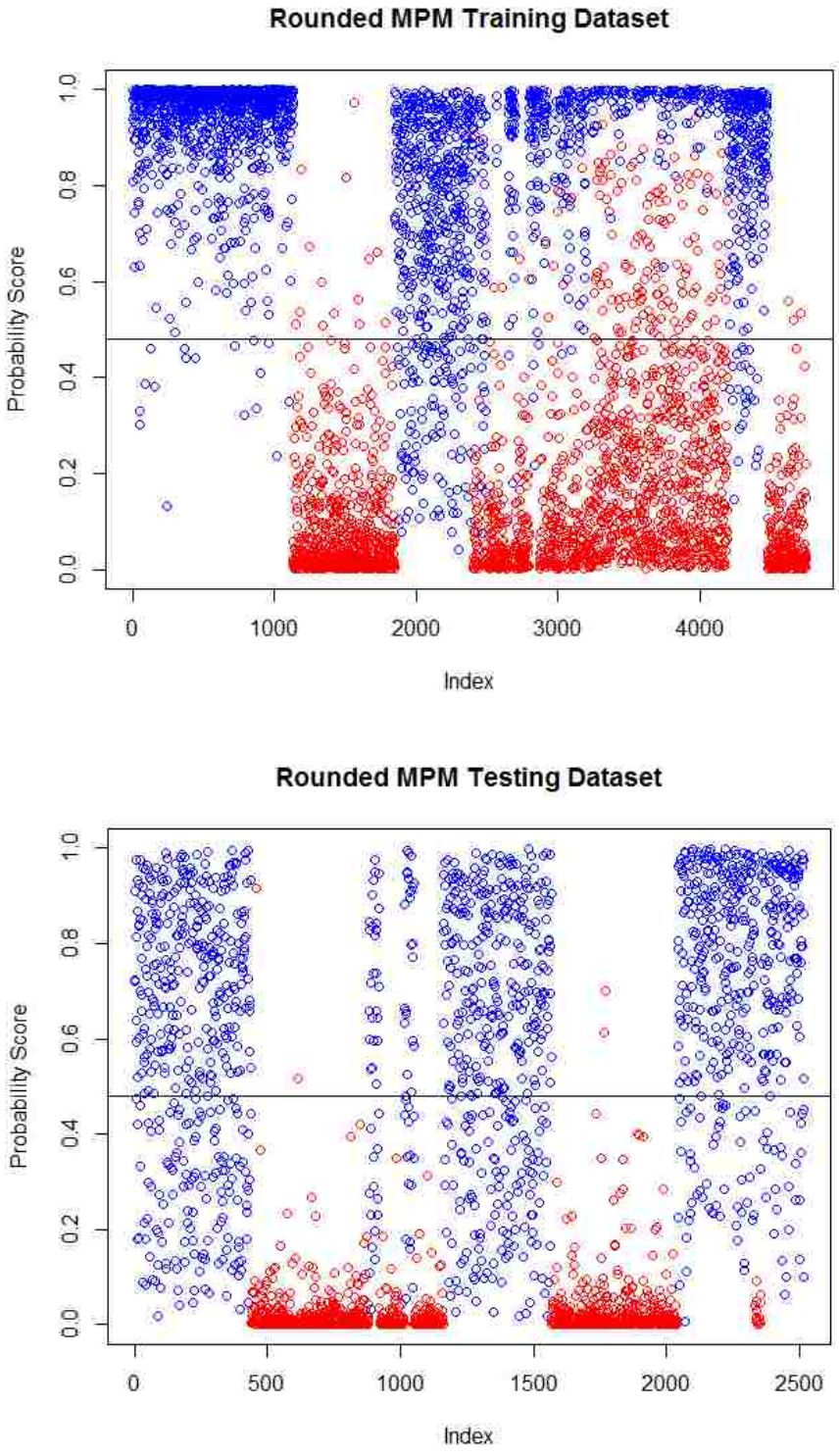


Figure 8. Probability scores for the multi-polygenic prediction model using all variables at the rounded highest AUC. Blue denotes case and red denotes control.

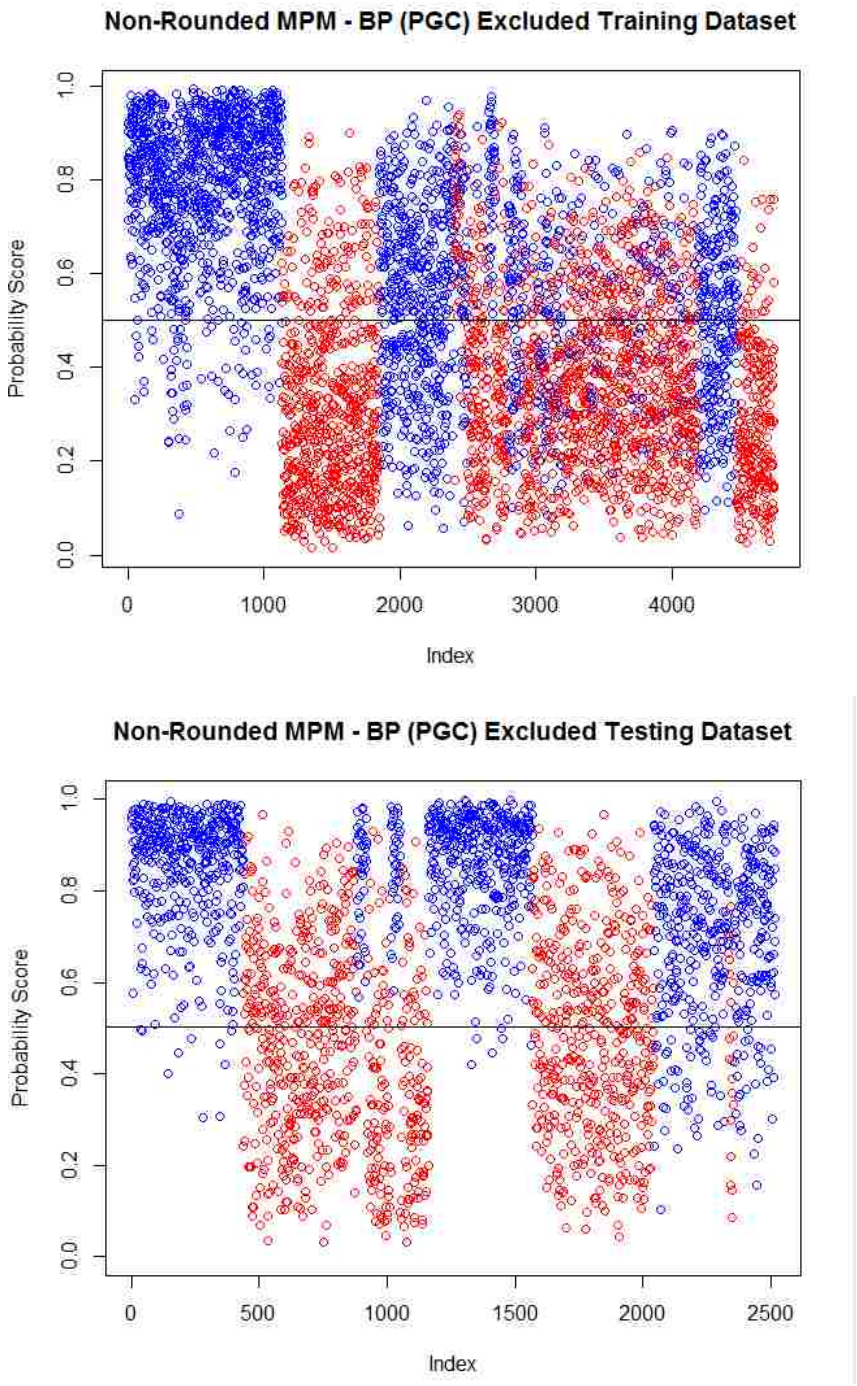
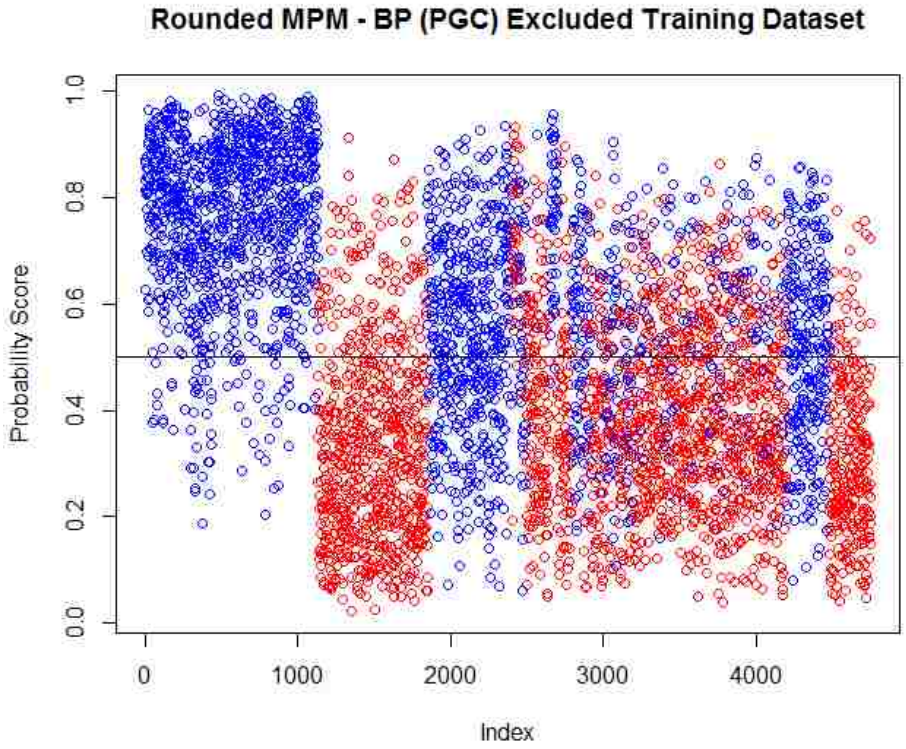


Figure 9. Probability scores for the multi-polygenic prediction model using all variables, except BP PGC, at the non-rounded highest AUC. Blue denotes case and red denotes control.



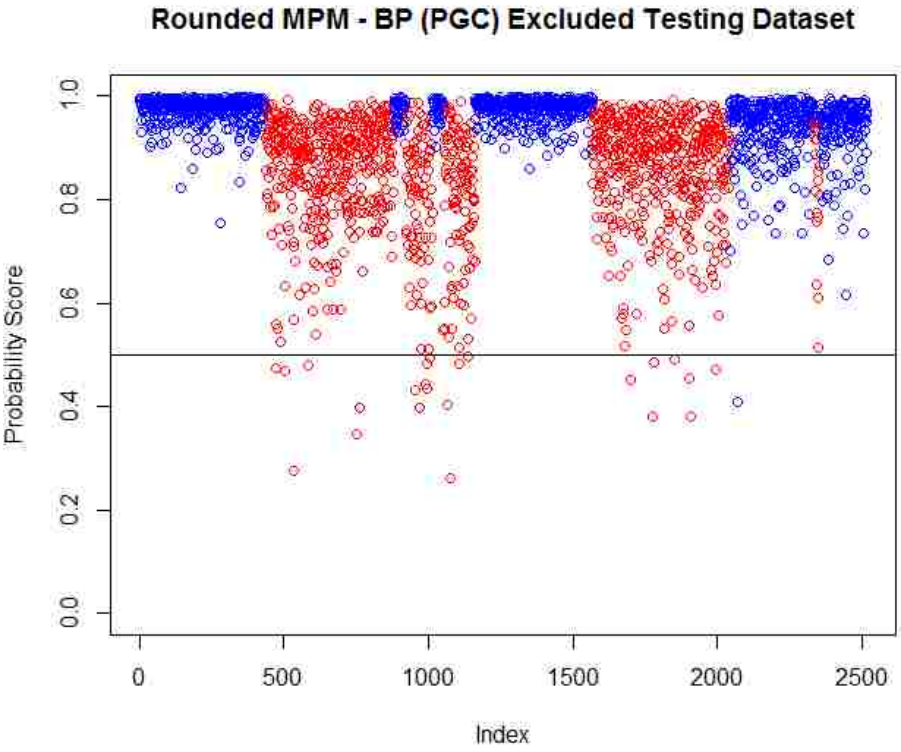


Figure 10. Probability scores for the multi-polygenic prediction model using all variables, except BP PGC, at the rounded highest AUC. Blue denotes case and red denotes control.

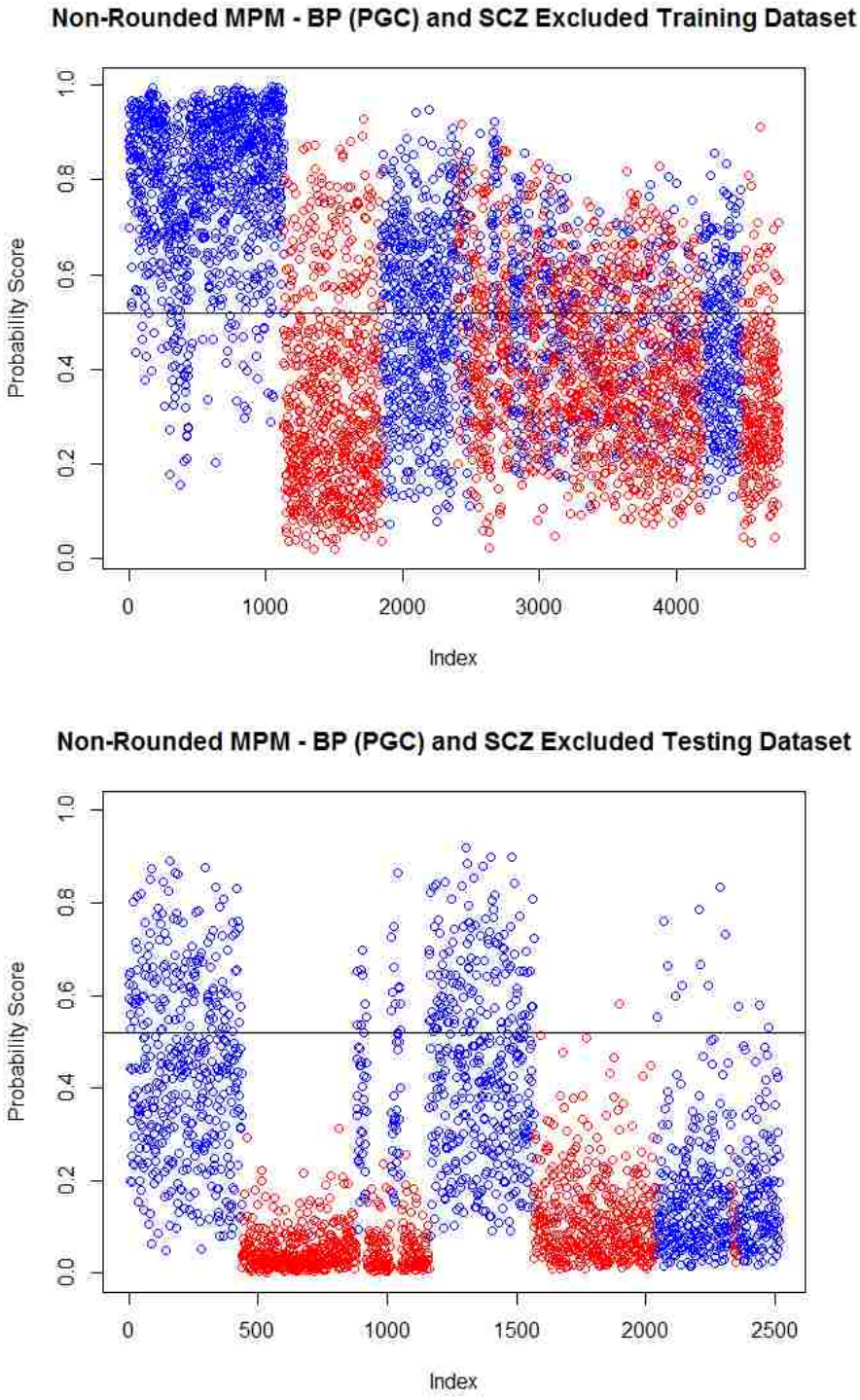


Figure 11. Probability scores for the multi-polygenic prediction model using all variables, except BP PGC and SCZ, at the non-rounded highest AUC. Blue denotes case and red denotes control.

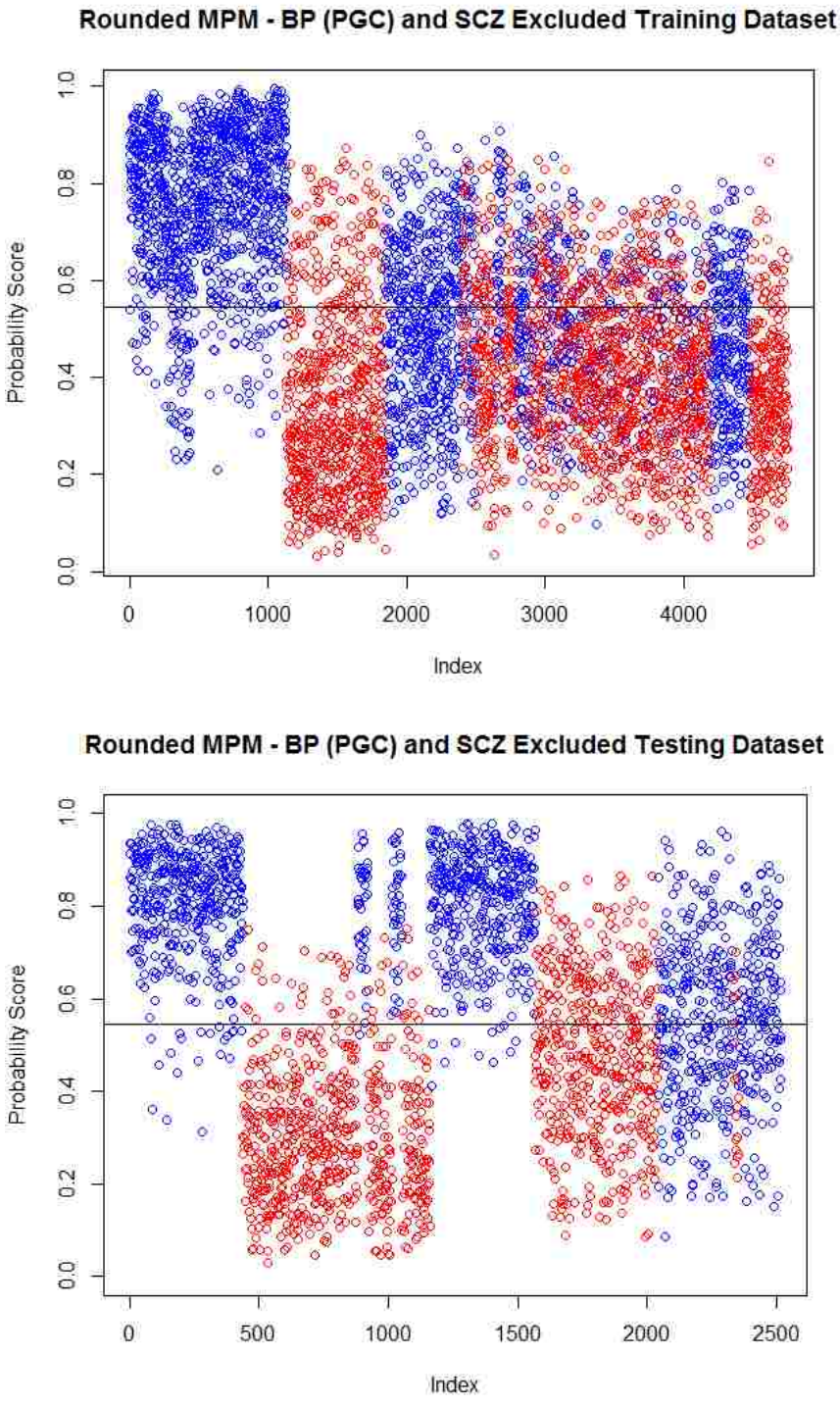


Figure 12. Probability scores for the multi-polygenic prediction model using all variables, except BP PGC and SCZ, at the rounded highest AUC. Blue denotes case and red denotes control.

References

Abdolmaleky, H. M., Thiagalingam, S. & Wilcox, M. (2005). Genetics and Epigenetics in Major Psychiatric Disorders. *Am J Pharmacogenomics, 5*, 149-160. https://doi.org/10.2165/00129785-200505030-00002

American Psychiatric Association. Diagnostic and statistical manual of mental disorders, 5th ed. (DSM-5). Washington, DC: American Psychiatric Publishing; 2013.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols, 5*, 1564-1573. https://doi.org/10.1038/nprot.2010.116

Anderson, C. A., Boucher, G., Lees, C. W., Franke, A., D'Amato, M., Taylor, K. D., … Rioux, J. D. (2011). Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature Genetics*, *43*(3), 246–252. https://doi.org/10.1038/ng.764

Arlot, S. & Celisse A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys, 4*, 40-79. https://doi.org/10.1214/09-SS054

Barret, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., … Daly, M.J. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature genetics, 40*(8), 955-62. https://doi.org/10.1038/ng.175

Boraska, V., Franklin, C. S., Floyd, J. a. B., Thornton, L. M., Huckins, L. M., Southam, L., … Bulik, C. M. (2014). A genome-wide association study of anorexia nervosa. *Molecular Psychiatry*, *19*(10), 1085–1094. https://doi.org/10.1038/mp.2013.187

Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Monterey, CA: Wadsworth & Brooks; 1984.

Bumgarner R. (2013). Overview of DNA microarrays: types, applications, and their future. *Current protocols in molecular biology*, *Chapter 22*, Unit–22.1.. doi:10.1002/0471142727.mb2201s101

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS computational biology*, *8*(12), e1002822. doi:10.1371/journal.pcbi.1002822

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., … Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nature Genetics, 22, 231-238. https://doi.org/10.1038/10290

Chan, M. K., Krebs, M. O., Cox, D., Guest, P. C., Yolken, R. H., … Bahn, S. (2015). Development of a blood-based molecular biomarker test for identification of schizophrenia before disease onset. *Translational Psychiatry, 5*. https://doi.org/10.1038/tp.2015.91

Charney, A. W., Ruderfer, D. M., Stahl, E. A., Moran, J. L., Chambert, K., … Sklar, P. (2017). Evidence for genetic heterogeneity between clinical subtypes of bipolar disorder. *Translational Psychiatry, 7*. https://doi.org/10.1038/tp.2016.242

Chatterjee, N., Shi, J., & García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature reviews. Genetics*, *17*(7), 392–406. doi:10.1038/nrg.2016.27

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

Craddock, N. & Jones, I. (1999). Genetics of bipolar disorder. *Journal of medical genetics*, *36*(8), 585–594. doi:10.1136/jmg.36.8.585

Davies, G., Marioni, R. E., Liewald, D. C., Hill, W. D., Hagenaars, S. P., Harris, S. E., … Deary, I. J. (2016). Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N=112 151). *Molecular Psychiatry*, *21*(6), 758–767. https://doi.org/10.1038/mp.2016.45

Delaneau, O., Zagury, J. F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods, 10*, 5-6. https://doi.org/10.1038/nmeth.2307

Elastic Net Regression in R. (2017). Retrieved from https://educationalresearchtechniques.com/2017/04/14/elastic-net-regression-in-r/

Evaluating Logistic Regression Models. (2015). Retrieved from https://www.r-bloggers.com/evaluating-logistic-regression-models/

Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2015). PRSice: Polygenic Risk Score software. *Bioinformatics*, *31*(9), 1466–1468. https://doi.org/10.1093/bioinformatics/btu848

Ferreira, M. A., O'Donovan, M. C., Meng, Y. A., Jones, I. R., Ruderfer, D. M., Jones, L., … Wellcome Trust Case Control Consortium (2008). Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nature genetics*, *40*(9), 1056–1058. doi:10.1038/ng.209

Franke, A., McGovern, D. P. B., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., … Parkes, M. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics*, *42*(12), 1118–1125. https://doi.org/10.1038/ng.717

Freeman, C. & Marchini, J. (2007) GTOOL: A program for transforming sets of genotype data for use with the programs SNPTEST and IMPUTE. Retrieved from http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html

Furberg, H., Kim, Y., Dackor, J., Boerwinkle, E., Franceschini, N., … Sullivan, P. F. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics*, *42*(5), 441–447. https://doi.org/10.1038/ng.571

Gejman, P., Sanders, A., & Duan, J. (2010). The Role of Genetics in the Etiology of Schizophrenia. *The Psychiatric Clinics of North America*, *33*(1), 35–66. https://doi.org/10.1016/j.psc.2009.12.003

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., … Tanaka, T. (2003). The International HapMap Project. *Nature, 426*, 789-796. https://doi.org/10.1038/nature02168

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology, 18*. https://doi.org/10.1186/s13059-017-1215-1

Hill, W. D., Hagenaars, S. P., Marioni, R. E., Harris, S. E., Liewald, D. C. M., Davies, G., … Deary, I. J. (2016). Molecular Genetic Contributions to Social Deprivation and Household Income in UK Biobank. *Current Biology: CB*, *26*(22), 3083–3089. https://doi.org/10.1016/j.cub.2016.09.035

Hinrichs A. S., Karolchik D., Baertsch R., Barber G. P., Bejerano G., … Kent W. J. (2006) The UCSC Genome Browser Database. *Nucleic Acids Res.* 1(34). (Database issue): D590-8

Hirschfeld, R. M. A., Lewis, L. & Vornik, L. A. (2003). Perceptions and impact of bipolar disorder: How far have we really come? Results of the national depressive and manic-depressive association 2000 survey of individuals with bipolar disorder. Journal of Clinical Psychiatry, 64(2), 161-174.

Hirschfeld, R. M. A. & Vornik, L. A. (2004). Recognition and Diagnosis of Bipolar Disorder. *Journal of Clinical Psychiatry, 65*. Retrieved from: https://www.psychiatrist.com/jcp/article/Pages/2004/v65s15/v65s1503.aspx

Hou, L., Bergen, S. E., Akula, N., Song, J., Hultman, C. M., Landén, M., … McMahon, F. J. (2016). Genome-wide association study of 40,000 individuals identifies two novel loci associated with bipolar disorder. *Human Molecular Genetics*, *25*(15), 3383–3394. https://doi.org/10.1093/hmg/ddw181

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, *44*(8), 955–959. https://doi.org/10.1038/ng.2354

Howie, B., Marchini, J., & Stephens, M. (2011). Genotype Imputation with Thousands of Genomes. *G3: Genes, Genomes, Genetics*, *1*(6), 457–470. https://doi.org/10.1534/g3.111.001198

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics 5*(6). https://doi.org/10.1371/journal.pgen.1000529

Krapohl, E., Patel, H., Newhouse, S., Curtis, C. J., von Stumm, S., Dale, P. S., … Plomin, R. (2017). Multi-polygenic score approach to trait prediction. *Molecular Psychiatry*, 1-7. https://doi.org/10.1038/mp.2017.163

Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V., Purcell, S. M., … Wray, N. R. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*, *45*(9), 984–994. https://doi.org/10.1038/ng.2711

Lee, S. H., Yang, J., Goddard, M. E., Vissher, P. M., & Wray, N. R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Biofinformatics, 28*(19), 2540-2542. doi: 10.1093/bioinformatics/bts474

Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., … Weersma, R. K. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics*, *47*(9), 979–986. https://doi.org/10.1038/ng.3359

Lydall, G. J., Bass, N. J., McQuillin, A., Lawrence, J., Anjorin, A., Kandaswamy, R., … Gurling, H. M. (2011). Confirmation of prior evidence of genetic susceptibility to alcoholism in a genome wide association study of comorbid alcoholism and bipolar disorder. *Psychiatric Genetics*, *21*(6), 294–306. https://doi.org/10.1097/YPG.0b013e32834915c2

Manji, H. K., Quiroz, J. A., Payne, J. L., Singh, J., Lopes, B. P., Viegas, J. S., & Zarate, C. A. (2003). The underlying neurobiology of bipolar disorder. *World psychiatry : official journal of the World Psychiatric Association (WPA)*, *2*(3), 136–146.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., … Marchini, J. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics, 48*, 1279-1283. https://doi.org/10.1038/ng.3643

Moor, M. H. M., Costa, P. T., Terracciano, A., Krueger, R. F., de Geus, E. J. C., Toshiko, T., … Boomsma, D. I. (2012). Meta-analysis of genome-wide association studies for personality. *Molecular Psychiatry*, *17*(3), 337–349. https://doi.org/10.1038/mp.2010.128

Okbay, A., Baselmans, B. M. L., De Neve, J.-E., Turley, P., Nivard, M. G., Fontana, M. A., … Cesarini, D. (2016). Genetic variants associated with subjective well-being, depressive

symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics*, *48*(6), 624–633. https://doi.org/10.1038/ng.3552

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., … Benjamin, D. J. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, *533*(7604), 539–542. https://doi.org/10.1038/nature17671

Park, S. T., & Kim, J. (2016). Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. *International neurourology journal*, *20*(2), S76–S83. doi:10.5213/inj.1632742.371

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., … Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), 748–752. https://doi.org/10.1038/nature08185

Rajkowska G., Halaris A., Selemon L. D. Reductions in neuronal and glial density characterize the dorsolateral prefrontal cortex in bipolar disorder. *Biol Psychiatry.* 2001;49:741–752

Rapport, N. & Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research, 46*(20), 10546-10562. https://doi.org/10.1093/nar/gky889

Ripke, S., Neale, B. N., Corvin, A., Walters, J. T. R., Farh, K., … O'Donovan M. C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*(7510), 421–427. https://doi.org/10.1038/nature13595

Ripke, S., Wray, N. R., Lewis, C. M., Hamilton, S. P., Weissman, M. M., … Sullivan, P. F. (2013). A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry*, *18*(4), 497–511. https://doi.org/10.1038/mp.2012.21

Ritchie, S. (2014). liftOverPlink. Retrieved from https://github.com/sritchie73/liftOverPlink on December 18th, 2018. *PNAS, 104*(28), 11694-11699.

Rzhetsky A., Wajngurt D., Park N., & Zheng, T. (2007). Probing genetic overlap among complex human phenotypes. https://doi.org/10.1073/pnas.0704820104

Shastry, B. S. (2009). SNPs: Impact on Gene Function and Phenotype. In: Komar A. (eds) Single Nucleotide Polymorphisms. Methods in Molecular Biology (Methods and Protocols), vol 578. Humana Press, Totowa, NJ. doi: 10.1007/978-1-60327-411-1_1

Sklar, P., Ripke, S., Scott, L. J., Andreassen, O. A., Cichon, S., … Purcell, S. M. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near *ODZ4*. *Nature Genetics*, *43*(10), 977–983. https://doi.org/10.1038/ng.943

Smoller J. W. & Finn C. T. (2003). Family, twin, and adoption studies of bipolar disorder. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, *123C*(1), 48–58. https://doi.org/10.1002/ajmg.c.20013

So, H. C. & Sham, P. C. (2017). Exploring the predictive power of polygenic scores derived from genome-wide association studies: a study of 10 complex traits. *Bioinformatics*, *33*(6), 886–892. https://doi.org/10.1093/bioinformatics/btw745

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., … Loos, R. J. F. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, *42*(11), 937–948. https://doi.org/10.1038/ng.686

Stahl, E. A., Breen, G., Forstner, A. J., McQuillin, A., Ripke, S., … Sklar, P. (2019). Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature Genetics, 51*, 793-803. https://doi.org/10.1038/s41588-019-0397-8

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature, 526*, 68-74. https://doi.org/10.1038/nature15393

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., … Kathiresan, S. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, *466*(7307), 707–713. https://doi.org/10.1038/nature09270

Zheng, J., Erzurumluoglu, A. M., Elsworth, B. L., Kemp, J. P., Howe, L., … Laurin, C. (2017). LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics, 33*(2), 272-279. https://doi.org/10.1093/bioinformatics/btw613

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Curriculum Vitae

Travis Mize

mizetrav@isu.edu

CURRENT POSITION

University of Nevada Las Vegas — Las Vegas, Nevada
Graduate Assistant, Department of Psychology — 2016 - Present

EDUCATION

B.S. Psychology — December 2014
Idaho State University, Pocatello, Idaho

HONORS AND AWARDS

Idaho Promise Scholarship ($400.00)	Fall 2013 – Spring 2014
Idaho Promise Scholarship ($400.00)	Fall 2012 – Spring 2013
Idaho Promise Scholarship ($495.00)	Fall 2011 – Spring 2012
Dean's List (ISU)	Spring 2012, Fall 2013, Spring 2014, Fall 2014

PEER-REVIEWED PUBLICATIONS

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

RESEARCH EXPERIENCE

Bioinformatics and Computational Biology

Graduate Research Assistant, University of Nevada Las Vegas — July 2018 – Current

- Supervisor: Mira Han, Ph.D.
- *Responsibilities*: Download/management of Genomic Data Commons open source cancer data, dataset manipulation, dimension-reduction, batch effect correction, and classification employing various machine learning techniques via Python and R.
- **Classification of Unknown Primary Cancer**: DNA methylation and RNA-seq data for various tissue types are being used in the classification of cancers of unknown primary.

Human Genetics and Big Data Analyses

Graduate Research Assistant, University of Nevada Las Vegas August 2016 – Current

- Supervisors: Xiangning "Sam" Chen, Ph.D. and Jingchun Chen, Ph.D.
- *Responsibilities*: DNA/RNA sequence analyses utilizing the GATK workflow, differential gene expression analyses, Linux server administration, statistical analyses in R, and data parsing with Python programming.
- **Whole-Genome Sequence Data Analysis**: Processing and analyzing whole-genome sequencing data of 187 individuals to discover genetic variants involved in the development of schizophrenia.
- **Polygenic Risk Score Prediction Model of Mental Health Disorders**: Open source genetic data was used to create a statistical model that allowed us to predict the diagnoses of mental health disorders in individuals.

Genetics of Major Depressive Disorder

Undergraduate Research Assistant, Idaho State University May 2014 – Dec. 2014

- Supervisors: Mark Austin, Ph.D. and Prabha Awale, Ph.D.
- *Responsibilities*: Genotyping, western blotting, and BCA protein assay.
- **Depression Development in SIRT6 Gene-Knockout Mice:** We assessed the development of depression in mice with a knockout of the SIRT6 gene.

Social Health and Neuroscience

Undergraduate Research Assistant, Idaho State University Jan. 2014 – Dec. 2014

- Supervisor: Xiaomeng "Mona" Xu, Ph.D.
- *Responsibilities*: Literature review, manuscript preparation, data entry, and data collection.
- **Self-Expansion in Art and Music Majors**: The role of self-expansion and its importance in Art and Music student's success in their field of study was scrutinized by examining self-report measures.

TEACHING EXPERIENCE

University of Nevada Las Vegas Las Vegas, Nevada
Instructor of Record, General Psychology August 2018 – Present

- Prepared lectures and classroom activities on a wide range of topics in psychology.
- Constructed course material and assessed student progress to increase likelihood of success.

CONFERENCE PRESENTATIONS

Mize, T. & Chen, X. (2018, May). *Multi-Polygenic Risk Score Prediction Model for Bipolar Disorder*. Oral presentation at the 2018 Nevada Institute of Personalized Medicine Symposium, Las Vegas, NV.

Mize, T., Thigpen, A., Moreno, M., Hamid, M., Bashy, B., Servin, F., Chen, J., & Chen, X. (2018, May). *Whole-Genome Sequencing of Schizophrenia Families with Multiple Affected Individuals*. Poster presentation at the 2018 Nevada Institute of Personalized Medicine Symposium, Las Vegas, NV.

Moreno, M., Hamid, M., Servin, F., Bashy, B., **Mize, T.**, Thigpen, A., Chen, X., & Chen, J. (2018, May). *Identification of CHST9 as a Candidate Gene for Schizophrenia from Whole-Genome Sequencing*. Poster presentation at the 2018 Nevada Institute of Personalized Medicine Symposium, Las Vegas, NV.

Chen, X., Wu, J., **Mize, T.**, & Chen, J. (2017, October). *Whole Genome Sequencing of Schizophrenia Families with Multiple Affected Individuals*. Poster presentation at the 25th World Congress of Psychiatric Genetics, Orlando, FL.

Chen, X., Wu, J., **Mize, T.**, & Chen, J. (2017, October). *Genetically Informed Subtyping of Schizophrenia*. Poster presentation at the 25th World Congress of Psychiatric Genetics, Orlando, FL.

Mize, T., Wu, J., Ingle, A., Chen, J., & Chen, X. (2017, May). *Polygenic Risk Score Prediction Model for Schizophrenia.* Poster presentation at the 26th Annual Nevada Psychological Association Conference, Las Vegas, NV.

Mize, T., Wu, J., Ingle, A., Chen, J., & Chen, X. (2017, February). *Polygenic Risk Score Prediction Model for Schizophrenia.* Poster presentation at the 2017 Nevada Institute of Personalized Medicine Symposium, Las Vegas, NV.

Mize, T., Tart-Zelvin, A., & Xu, X. (2014, April). *A Neuroimaging Approach to Working Memory Performance.* Oral presentation at the 3rd Annual Southeastern Idaho Psi Chi Psychology Conference, Pocatello, ID.

Savary, T. A., **Mize, T.**, Colman, D. E., & Letzring, T. D. (2014, April). *Correlates of the Satisfaction with Life Scale: A look at age, gender, ethnicity, religion, and social support.* Poster presentation at the 3rd Annual Southeastern Idaho Psi Chi Psychology Conference, Pocatello, ID.

MANUSCRIPTS IN PREPARATION OR UNDER REVIEW

Chen, J., **Mize, T.**, Hong, E., Nimgaonkar, V., Kendler, K., Allen, D., Oh, E., & Chen, X. (under review). *Genetically Informed Subtyping of Schizophrenia*.

Chen, J., Wu, J., **Mize, T.**, Moreno, M., Hamid, M., Servin, F., Bashy, B., Zhao, Z., Jia, P., Tsuang, M., Kendler, K., Xiong, M., & Chen, X. (under review). *Whole Genome Sequencing Identified CHST9 as a Schizophrenia Candidate Gene*.

PROFESSIONAL MEMBERSHIPS

Phi Kappa Phi Student Member
2018 – present

Psi Chi Student Member
2012 – present

REFERENCES

Xiangning "Sam" Chen, Ph.D.
Scientific Review Officer
NIH/CSR (Center for Scientific Review)
6701 Rockledge Dr., Bethesda, MD 20817
(302) 402-2664

Mira Han, Ph.D.
Assistant Professor
Department of Life Sciences, Nevada Institute of Personalized Medicine
University of Nevada Las Vegas, Las Vegas, NV 89118
(702) 774-1503

Rochelle Hines, Ph.D.
Assistant Professor
Department of Psychology
University of Nevada Las Vegas, Las Vegas, NV 89118
(702) 895-0187

Xiaomeng "Mona" Xu, Ph.D.
Associate Professor
Department of Psychology
Idaho State University, Pocatello, ID 83209
(208) 282-3541, xuxiao@isu.edu