PRS Evaluation

- As in any model, the importance of evaluation defines how well a model performs on an independent sample to determine the degree of overfitting with respect to the training process. In the present case, the training data known as base data is extracted from a genome wide association study (GWAS). The consequences of a mutation correlated with a specific condition are aggregated and quantified from multiple individuals' genomes to serve as input to calculate the Polygenic Risk Score (PRS).

Pre-Processing

- Before moving forward with PRS computation, a QC step (pre-processing) would be required to evaluate mutations (SNP) and samples. In general, the following steps are performed:

Sample Size
- The association testing step must a minimum of 100 samples involved in a case vs control manner.

Ref Genome
- All samples should have the genomic positions mapped to the same Genome [B38 -last version].

Ambiguous SNPs
- Different genotyping chips lead to mismatches between complementary alleles (C/G or A/T) may lead to ambiguity between base and target data.

Sex discrepancy
- Determine the difference between reported sex and found sex chromosomes to not reflect mislabeling of samples or overdraw conclusions from unreliable data.

Relatedness
- To reduce bias/inflation, a population structure procedure would identify a degree of relativeness between samples and those with high relatedness should be excluded from both base and target data to eliminate the risk.

Linkage Disequilibrium
- (LD), association of alleles at two or more sites on the same chromosome that are inherited - together more often than expected by chance.
- The LD coefficient represents the proportion of observations in which two specific pairs of alleles occur together.
- Measures the correlation between genetic variants that are more likely to be inherited together due to their physical proximity, leading to association within a population. Any SNPs that are not in approximate linkage disequilibrium have to be identified and removed by method such as “pruning”.

(PRS) Polygenic Risk Score

Minor Allele Frequency (MAF)

- Determine the frequency of the least often allele at a specific location
- SNPs with MAF are more susceptible to genotyping errors, and also, they have low power when performing association for given effect size.

The Hardy-Weinberg (dis)equilibrium [HWE] law

- indicates that allele and genotype frequencies in a stable population without evolutionary influences will stay constant between generations. Deviation from HWE indicates that genotype frequencies differ significantly from their expected values which could indicate genotyping errors, such variants are therefore often excluded from analyses.

- For the present data set, the LD could not be used since the founders are not present, only **HWE equilibrium and MAF** were calculated for presentation purpose, not statistical significance.

PRS Analysis

- Present formula is used by default in **PLINK**
- where the effect size of SNP i is S_i; the number of effect alleles observed in sample j is G_{ij}; the ploidy of the sample is P (is generally 2 for humans); the total number of SNPs included in the PRS is N; and the number of non-missing SNPs observed in sample j is M_j

$$PRS_j = \frac{\sum_i^N S_i * G_{ij}}{P * M_j}$$

(PRS) Polygenic Risk Score

PRS Evaluation

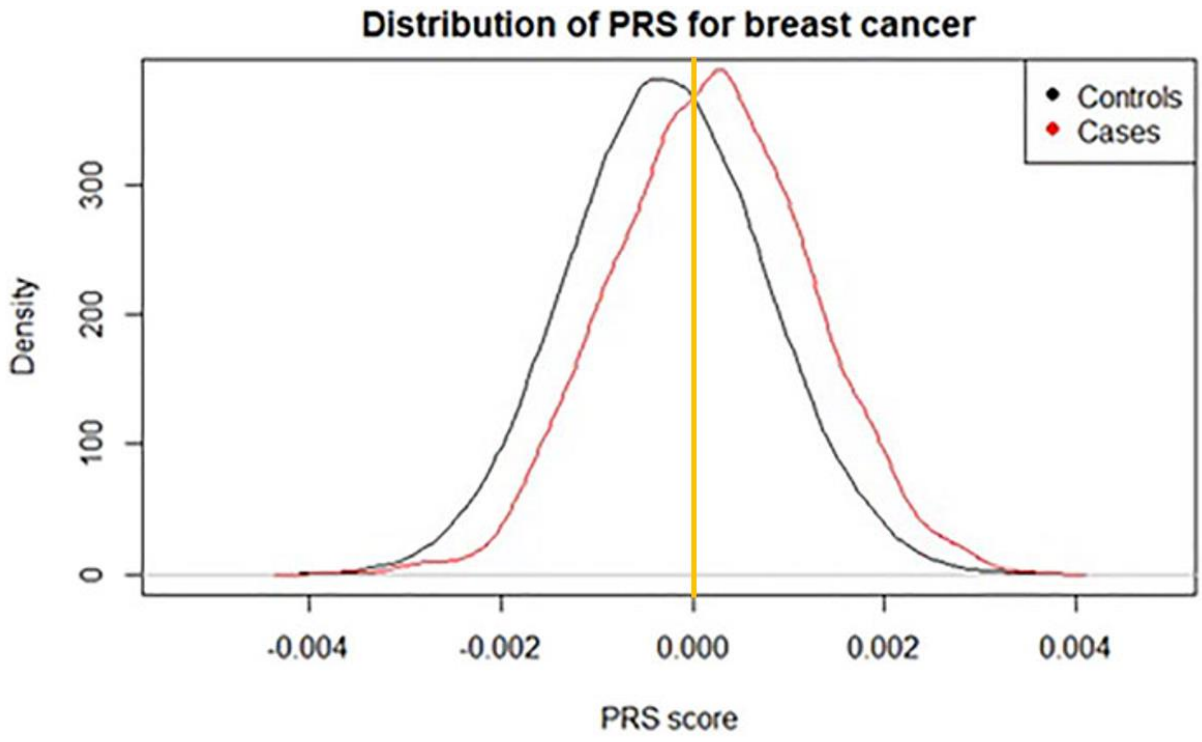


Fig. 1

The intersection of the 2 distributions (cases and controls) represents the threshold where higher scores indicate a higher risk (or higher association with a trait) based on the genotype information [Fig 1].

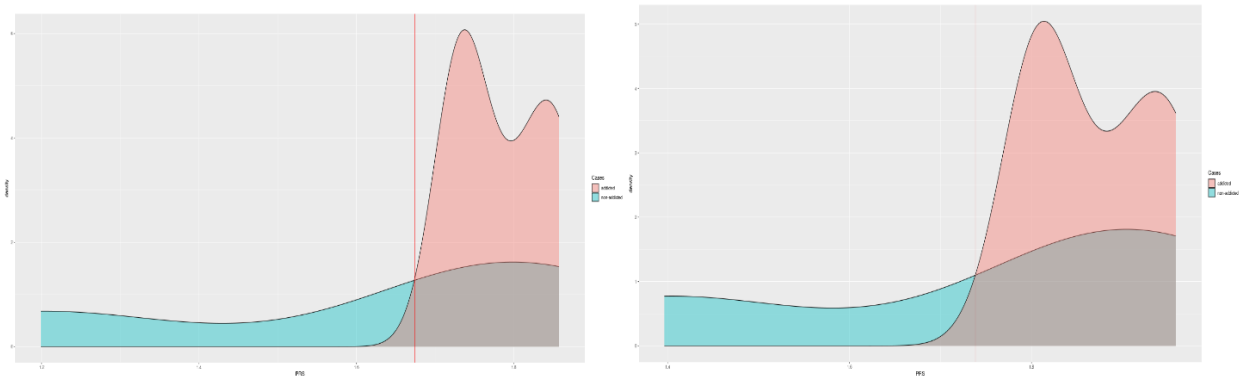


Fig. 2 Model A vs Model B

Differences in models are only between absolute values [Fig. 3], while the power of classification remains the same. Both models are expressing high bias on the **caffeine addiction trait,** where only 2 samples out of 7 from control group were classified correctly. This part could be tackled

with including covariates and weight variants in gene regions that influence this specific trait [BOADICEA]. Another step to reduce bias, is to perform most of the QC steps of a GWAS and have a well define boundary between case vs control.

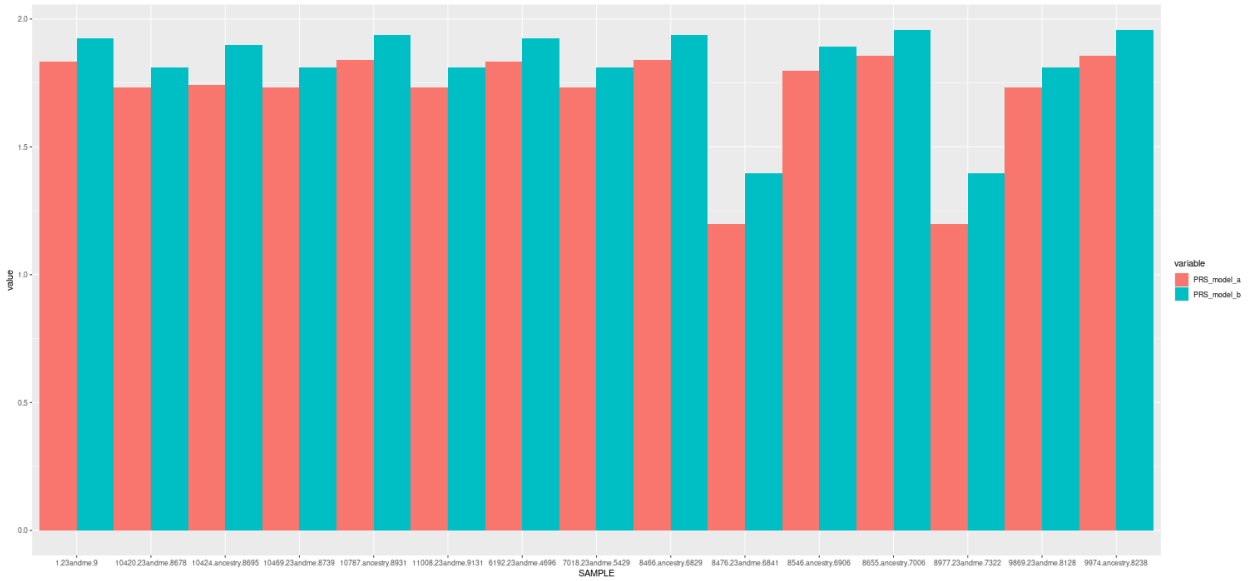


Fig. 3 Orange – model a / Blue – model b

Based on Label and classification, both models have the same values for evaluation metrics:

Confusion Matrix

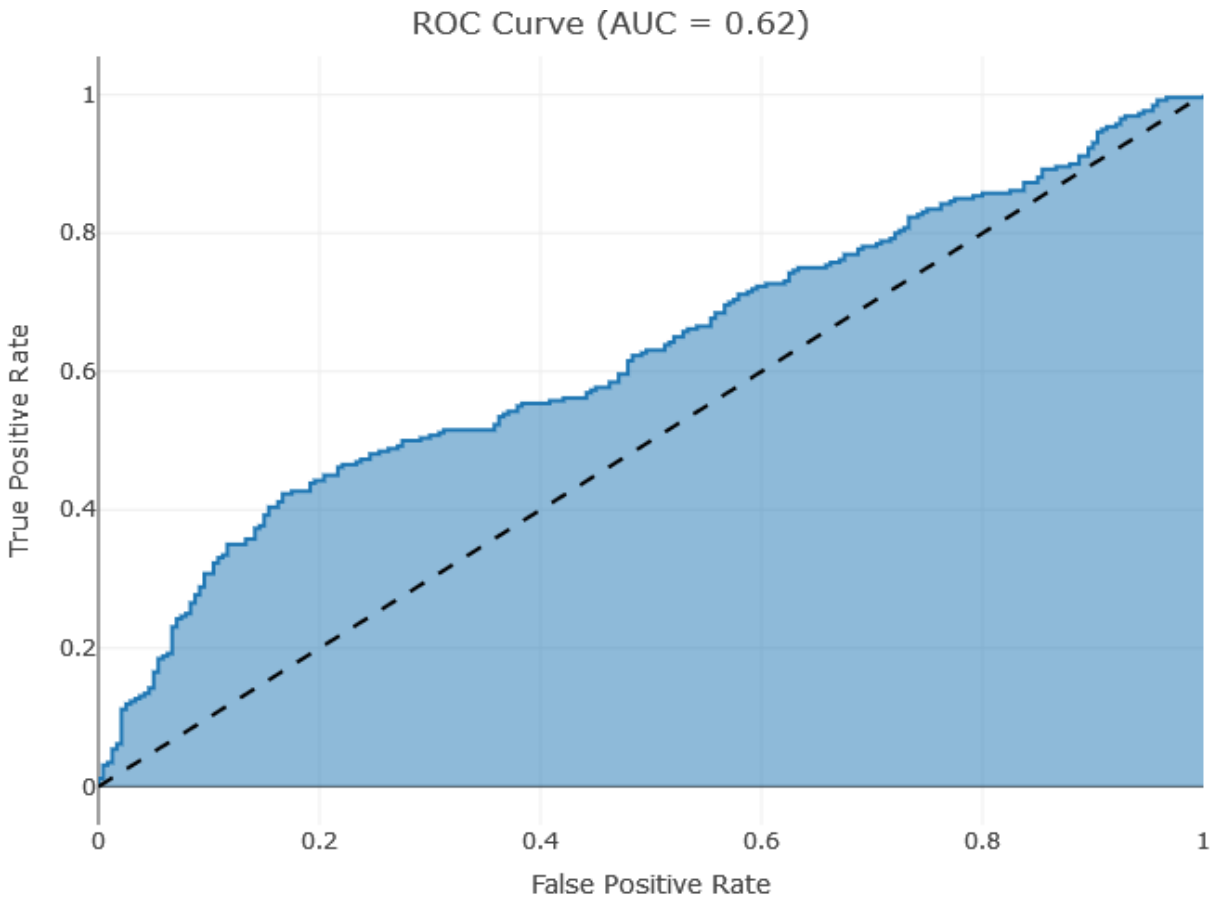
Prediction	**Reference**	
	0	**1**
0	*2*	*0*
1	*5*	*8*

Evaluation Metrics

Accuracy	0.6667
95% CI	(0.3838, 0.8818)
Sensitivity	0.287
Specificity	1.00
Prevalence	0.4667

(PRS) Polygenic Risk Score

The receiver operating characteristic (ROC) curve [Fig. 4] shows how well the classifier performed on a given dataset. The performance of a classifier is given by how closer to the top left corner is the curve. If it's getting in the opposite direction, the less accurate the classifier is. The area under the curve (AUC) determines how well a model predicts classes. For the present models, the AUC is 0.36 and express a low performance indicating that the model does not distinguish classes 1 from 0 in most of the cases.



Fi. 4 ROC Curve performance example

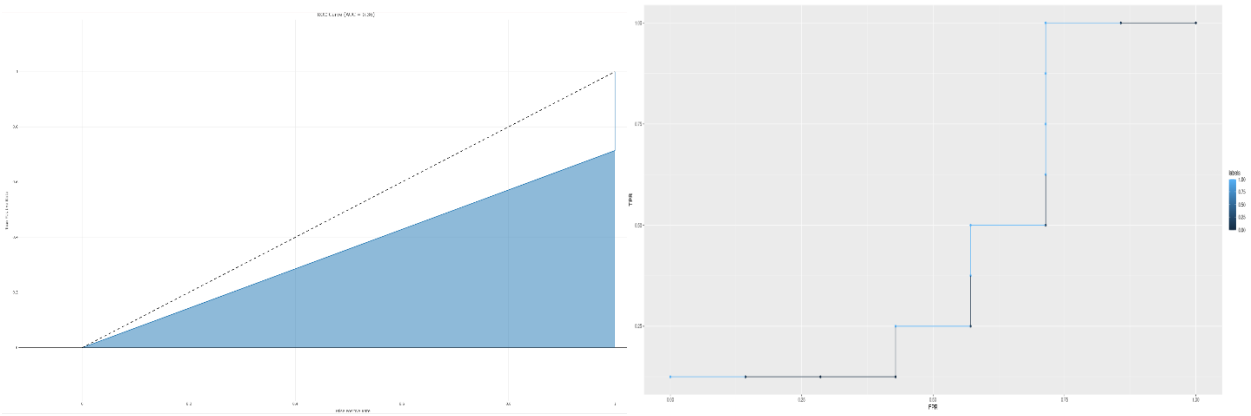


Fig. 5 ROC Cure with AUC = 0.36

Future Perspectives

Balance Bias/Variance:

- determine the impact of population
- keep same size date sets between case vs controls
- review any pruning based on LD or other adjustments made
- apply resampling or bootstrap

One of the key elements to increase the performance of a PRS model is to search in literature of similar studies (GWAS) and run other models on the target data, overlap between SNPs and find the most effective rs-IDs across studies with respect to a single population (e.g. EUROPEAN Population).

A well-established method to calculate PRS, demonstrated by **BOADICEA,** is to incorporate known **factors/covariates** from metadata (sex, age, alcohol, sleeping hours, environment etc.) in the model to increase the performance and also predict an approximate point in time when an individual would be the most susceptible to a given trait.

Nevertheless, the PRS model is influenced by the precedent GWAS analysis, the more samples with minimal/null relatedness and the maker are better provided to perform correlation, the more sense the PRS predicted value would have.

(PRS) Polygenic Risk Score

Quantitative traits/factors that could be included the present study:

- SEX is a very strong factor, so it has to serves as a covariate during GWAS.
- Two PRS models are built for male and female, respectively.
- Body mass index (BMI)
- Blood pressure: systolic blood pressure (SBP), diastolic blood pressure (DBP)
- Blood lipids: triacelglycerol (TG), low-density lipoprotein (LDL), high-density lipoprotein (HDL), low-density lipoprotein cholesterol (LDL-C)
- Glycated hemoglobin (HbA1c)
- Uric acid

Resources

- Jennifer A. Collister*, Xiaonan Liu: Calculating Polygenic Risk Scores (PRS) in UK Biobank: A Practical Guide for Epidemiologists
- William S. Bush, Jason H. Moore: Chapter 11: Genome-Wide Association Studies
- Andrew Lee, Nasim Mavaddat: BOADICEA-a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors
- Choi, S. W., Mak, T. S.-H., & O'Reilly, P. F. (2020). Tutorial: A guide to performing polygenic risk score analyses. Nature Protocols, 15(9), 2759–2772. https://doi.org/10.1038/s41596-020-0353-1
- https://github.com/apriha/snps