# Cross-trait polygenic risk score in schizophrenia

Bazaga Martínez, Gloria

Final degree project

Biotechnology degree

Universitat Rovira i Virgili

June 7, 2023

Academic advisor: Pujadas Anguiano, Gerard (PhD, Grup de Recerca en Quimioinformàtica i Nutrició, Departament de Bioquímica i Biotecnología, Univeristat Rovira i Virgili)

In cooperation with: Research department, Hospital Universitari Institut Pere Mata (HUIPM), Institut d'investigació sanitària Pere Virgili (IISPV).

Supervisor: Muntané Medina, Gerard. (PhD, Hospital Universitari Institut Pere Mata (HUIPM), gerard.muntane@urv.cat)

Me, Gloria Bazaga Martínez, with ID 39939140-P, I confirm that I am familiar with the URV guide to the prevention of plagiarism Prevention, detection, and treatment of plagiarism in teaching: Guide for students (approved in July 2017).

I affirm that this TFG does not constitute any of the behaviors considered as plagiarism by the URV.

Tarragona, June 7, 2023

## *Contents*

## *Acknowledgments*

Firstly, I would like to show my gratitude to my supervisor Dr. Gerard Muntané Medina for helping me with this journey in bioinformatic research, without his help I would not learned all the knowledge and the practice I have obtained thanks to this research.

His paper as my mentor have been impeccable, he always helped me with the doubts, explained me and help me to understand each process and encouraged me to think out of the box, to be resolutive and transmitted me his knowledge about bioinformatics and statistics.

I would like to regard the crew of the research group area of the Hospital Universitari Pere Mata (IPM) for all the advice and help that they showed me like a part of the team, teaching me the importance of the knowledge from diverse areas and how they are related in the same field, learning from each one of this areas.

Thanks to my academic supervisor, Dr Gerard Pujadas Anguiano, for helping me with the correction and the follow up of my work that permitted me obtaining the better of my work.

Finally, I would like to thank to the URV teachers, friends, and family members because without them I could have never been able to reach this point.

## *Information of the centre*

This project has been carried out in the research group Genetics and Environment in Psychiatry, GAP, at the Hospital Universitari Pere Mata and the Rovira i Virgili University (URV).

The GAP research group is consolidated by the Catalan Agency of Universities and Research of the Generalitat de Catalunya (AGAUR 2014 SGR 995), by the URV and is part of the Strategic Area of Neurosciences and Mental Health of the Pere Virgili Health Research Institute (IISPV). The responsible of the group is Elisabet Vilella.

They also are one of the research groups on the "Centro de Investigación Biomédica en Red", CIBERSAM.

## *Abstract*

**Background:** Schizophrenia is a neurodevelopmental disorder which presents genetic and environmental factors related to its aetiology with a high incidence among families. One of the most significant current topics of study is the contribution of the genetic liability to the development of schizophrenia, recent investigation in the field focused on genetic variants, such as copy number variations (CNVs), and single nucleotide polymorphisms (SNPs). These genetic variations permit to compute polygenic risk scores (PRS) for many psychiatric disorders and phenotyping traits, which has become a routine across biomedical research.

**Methods**: To date various methods have been developed and introduced to measure polygenic risk score, the PRS analysis is one of the practical ways to predict the course of an illness based on the genetical variations presented by the individual. Prior to commencing the study, it is important to proceed the Quality control analysis (QA). This method ensures that the results are consistent, comparable, accurate and valid. PRS method is based on Genome-wide association studies (GWAS) which are the base data of the proceed.

**Results:** It was found no significant association between schizophrenia PRS and the different phenotypic traits such as Positive and Negative Syndrome Scale, and Functional assessment short test among patients with schizophrenia spectrum disorder. As opposed to clinical characteristics, polygenic risk score does not contribute as an early detector of symptoms in the subclinical to clinical neurodevelopmental disorder.

**Conclusions:** Despite PRS scores are based on common genetic variants, environmental factors are also important for the development of the phenotypic traits and not only the influence from genetic factors implicated in schizophrenia.



**Kay words: Schizophrenia, SNPs, CNVs, PRS, GWAS, environmental factors, heritability**

## *Introduction*

The Global burden of disease are composed by depressive disorders, anxiety disorders, bipolar disorder, schizophrenia, autism spectrum disorders, conduct disorder, attention-deficit hyperactivity disorder, eating disorders, idiopathic developmental intellectual disability, and a residual category of other mental disorders (Ferrari, 2022) .

Individuals with mental disorders present a decrease between 10-15 years expectancy in life respect individuals without mental disorders. In this situation, early involvement at the first symptom of a mental disorder can be an improvement in the outcomes, with an alteration in the regular course of the disorder (Solmi et al., 2022).

Young people with attenuated symptoms for psychosis presents a 25% of probability to develop a mental disorder in a period of 3 years. However, clinical care for these individuals has an impact in the developing of psychosis. Therefore, preventive treatment and early interventions at first sign of mental disorders improve outcomes (Solmi et al., 2022). Despite this, the peak age at the onset of mental disorders it is not totally established yet (Zhang et al., 2022).

The prevalent cases in thousands by mental disorders in central Europe are 408.1, for eastern Europe are 710.8, for western Europe are 1447.3, and for Spain are 153.5, from 2019 ((Ferrari, 2022) e Table 7: Prevalent cases in thousands, with 95% uncertainty intervals, by mental disorder and location in 2019).

### *Schizophrenia. General concepts, and clinical information*

Schizophrenia is a neuropsychiatric disorder which throughout its historical background manifested diverse definitions of its clinical symptomatology. In its origins the mental disorder was named as “premature dementia”, afterwards in 1893 the psychiatrist Emil Kraepelin distinct the two psychoses: premature dementia and maniac depression. Is not until the 20th century that this disorder acquires the name Schizophrenia –split mind- which indicates “a split mind in two pieces”, subsequently, in the middle of the 20th century, a more specific definition is released by excluding personal hysteria from the onset definition and categorized this symptomatology as either a conversation reaction or a dissociative identity disorder (Blaylock & Faria, 2021).

At the moment, schizophrenia is fundamentally considered a disorder of neuronal activity related to diverse genetic factors involved in the synapsis process combined with several environmental factors, thus affect multiple brain regions and functions (Bachmann et al., 2022).

Moreover, this disorder is considered heterogeneous in its presentation, ethology, pathophysiology, and trajectory, with a shortage clear definition in its anatomic and biochemical implicated structures (Blaylock et al., 2021; Zick et al., 2022).

This neuropsychiatric disorder is also characterised by several alterations in the way reality is perceived, commonly manifest in late adolescence or early adulthood. These aspects can be related to different changes in behaviour classified as positive, negative, and cognitive symptoms. The positive symptoms are composed by any change in behaviour or thoughts, for example, hallucination or delusions. The negative symptoms may be aspects of behaviour that limits normal lifestyle aspects, such as social withdraw, decrease in productive or pleasurable activities, and even neglect personal hygiene or show itself as emotionless. Thus, are also associated with deficiencies in communication, social functioning, and affect (Bachmann et al., 2022; Correll & Schooler, 2020).

Although positive symptoms can be easily witnessed, negative symptoms may be unperceived signs that leads the necessity of new and effective treatments. In addition, many studies make a point in several phenotypic traits which ones could be related with the pharmacological treatment and not only due to genetical conditions in the individuals, one example is the overweight presented by the patients once their start the treatment (Bachmann et al., 2022; Correll & Schooler, 2020).

The evaluation of the symptoms is conducted by the Positive and negative syndrome scale and the Functional assessment short test.

### *Cognitive deficit*

The study of cognitive symptoms, more specifically, cognitive deficit, has been an increase area of study on the last decade since these neurocognitive impairments are the first witnessed symptoms in schizophrenia patients. These affects several cognitive domains like attention, working memory, visual and verbal learning, processing speed, problem-solving and social cognition (Martínez et al., 2021). Nevertheless, these symptoms cannot only be treated with pharmacological treatments such antipsychotic drugs, and the combination of antipsychotics and non-pharmacological therapies like cognitive training represents an improvement (Martínez et al., 2021).

## *Genetic studies and its implication in schizophrenia*

Most of the genetic studies in psychiatry focus on investigating schizophrenia and bipolar disorder, the focal point is based on the role of ultra-rare disruptive variants studies and studies of common genetic variants for multifactorial disease. However, the studies based on rare genetic variants associated with major depression are lower in number (Uranova et al., 2022).

Even though, these associations studies present a lower value heritability from common variants respect the estimate heritability from twin and family-based studies (Uranova et al., 2022).

### Genome-wide association studies

Genome-Wide Association Studies (GWAS) are genetic studies that aim to identify the relationship between single nucleotide polymorphisms (SNPs) and phenotypic traits (Marees et al., 2018). In addition, these studies allow investigating genetic risk factors related to human behaviour. Most of the concerns in these studies are related to the previous quality control process, which one involves a prior knowledge in genetics, statistics, and bioinformatics (Marees et al., 2018).

The quality control procedure is a set of previous and rigorous steps in GWAS that involves different filters to obtain a final and cleaned base data that implement further studies such as Polygenic Risk Score. The aim of this process is to remove any interference data that could compromise the integrity and veracity of the results obtained in the procedure of the study.

The principal objective of GWAS is to identify single nucleotide polymorphism (SNP's) of which presents a systematic variation in the allele frequency as a function of phenotypic trait values, as it is perceived between schizophrenia cases and healthy controls (Marees et al., 2018).

## *Polygenic Risk Score*

The Polygenic risk score (PRS) is a single value estimate of an individual's genetic liability to a phenotype, a predictive measure of the genetic susceptibility of an individual to a disease based on its genetic information (Tang et al., 2022). Polygenic risk score is composed by two main input data, a target data of genotypes and phenotypes from samples, and a base data with the summary statistics of single-nucleotide variants (GWAS) (Choi, S.W. et al.). PRS is composed by a set of independent risk variants associated with a disorder, constructed by present evidence from the genome-wide association studies related to the topic of study.

The basic equation for the PRS of an individual j is:

- Eq. (1): Standard equation to calculate a weighted polygenic risk score.

$$PRS_j = \sum N i \beta i * dosage_{ij}$$

The effect size of SNP i is Si; the number of effect alleles seen in the sample j is Gij; the ploidy of the sample is P (usually 2 for humans); the total number of SNPs presented in the PRS is N; and the number of non-missing SNPs observed in sample j is Mj.

This score is widely used on GWAS research due to the polygenic genetic architecture that common complex disorders present, allowing researchers to identify genetic variants associated with diseases. This unique score, related to the genetics of each individual, is obtained by weighting the effect size (for binary traits (OR) or beta coefficient for continuous traits), and this is based on the number of risk alleles transported in each variant of the individual (0, 1 or 2) (Cathryn M. Lewis et al.).

Although polygenic risk score may estimate the genetic risk of an individual to present an illness during its lifetime, in general populations terms, the PRS shows a lower ability in this research topic since traditional PRS cannot precisely model any gene-gene (epitasis) or gene-environment interactions (Kasap & Dwyer, 2022).

### *Positive and Negative Syndrome Scale*

Positive and Negative Syndrome Scale (PANSS) is a symptom-based rating scale that allows measure symptoms, syndromes, and general severity for mental disorders such as schizophrenia. That scale is based on 30 items, which ones are a combination of 18 items from the Brief Psychiatric Rating scale (BPRS), and another 12 items from the Psychopathology Rating Scale (PRS), the definitions easily match each graduation with its subsequently item, and the total score can vary from 30 to 120 (Buizza et al., 2022).

PANSS total score is used to evaluate schizophrenia psychopathology in clinical trials, with thresholds set up by cut-off scores for patients with acute schizophrenia. Despite this, the evaluation of negative symptoms with this rating scales may become a concern due to the difficulties to evaluate which are clinically significant changes (Leucht et al., 2019).

The total score of the positive, the negative, and the general symptoms is evaluated with the PANSS scale on the data we obtained from patients with Schizophrenia.

The PANSS scales are composed by the punctuation of each item regarding to different symptoms related to positive, negative, and general/cognitive symptoms presented by schizophrenia patients described below (Gopalakrishnan et al., 2020):

- Positive subscale items are composed by delusions, conceptual disorganization, hallucinations, excitement, grandiosity, suspiciousness, and hostility.
- Negative PANSS subscale items are composed by blunted effect, emotional withdrawal, poor rapport, passive social withdrawal, difficulty in abstract thinking, lack of spontaneity, stereotyped thinking.
- General PANSS subscale items are composed by Somatic concern, anxiety, guilt feelings, tension, mannerisms & posturing, depression, motor retardation, uncooperative, unusual thought content, disorientation, poor attention, lack of judgment, disturbance volition, poor impulse control, pre-occupation.

### *Functional assessment short test*

The Functional assessment short test (FAS or FAST) is an instrument to measure psychosocial functioning, this clinical-administrated assessment scale of psychosocial disfunction is widely used in individuals with bipolar disorder (Amoretti et al., 2021; Siegel-Ramsay et al., 2023).

This scale measures the different psychometric traits such as: internal consistency, concurrent validity, discriminant validity, factorial analyses, and sensitivity to change.

### *Genetic of Schizophrenia*

Despite the causes and biological consequences of schizophrenia are still unknown it presents a high heritability attributable to common risk alleles, rare copy number variants (CNVs), and rare coding variants (RCVs) (Trubetskoy et al., 2022). The most important risk genes associated to schizophrenia are disrupted-in-schizophrenia 1 (DISC1), catechol-O-methyl transferase (COMT), monoamine oxidases-A/B (MAO-A/B), glutamic acid decarboxylase 67 (GAD67), dysbindin-1, and neuregulin 1 (NRG1). (Năstase et al., 2022)

One of the latest GWAS in schizophrenia, *"Mapping genomic loci implicates genes and synaptic biology in schizophrenia"* (Trubetskoy et al., 2022), delights the diverse implications in schizophrenia patients related to mutations in ATP2A2, which encodes a sarcoplasmic/endoplasmic calcium pump on the reticulum. This mutation is the cause of Darier Disease, and it is associated with bipolar disorder and schizophrenia, more specifically, to the regulation of neuronal cytoplasmatic calcium levels conducted by ATP2A2 (Trubetskoy, Vassily et al.).

This GWAS was performed with EUR samples and identified moreover 287 genetic loci and 313 independent SNPs associated with schizophrenia in the European ancestry, despite this, there are still hundreds of loci undiscovered and the circumscribe region of studies implies a setback in the discovery of new loci related to schizophrenia (Lam et al., 2019; Trubetskoy et al., 2022). Many of these common risk alleles presents a minor effect, but most of them cause a third of genetic liability. These genetic factors play a role in the susceptibility of schizophrenia, and the heritability oscillates between 60-80 %, its majority is related to these common risk alleles (Lam et al., 2019).

Moreover, there are identified 8 rare copy number variants (CNVs) which vest an important individual risk (Trubetskoy et al., 2022). In addition, a recent study also identified 10 genes associated with rare disruptive mutations of large effect (Owen, 2023).

## *Hypothesis*

Given that patients with schizophrenia present an important habitability, the predisposition to schizophrenia could be predicted and score via Polygenic risk score based on their genetic variability and the relation with the phenotypic traits such as Positive and Negative syndrome scale and Functional assessment short test.

## *Objectives*

Polygenic risk scores (PRS) are mostly used to identify genetic associations to clinical symptoms in schizophrenia. In fact, the genetic architecture that characterized many common complex disorders permits to identify genetic variants associated with the disease.

This study aims to identify the association between schizophrenia PRS and many clinical symptoms (endophenotypes) present in initial stages of the disease, not only study the association between PRS and clinical traits but Positive and Negative Syndrome Scale related to schizophrenia spectrum disorder as well.

To calculate the PRS scores, it is essential to conduct a prior quality control procedure. This study focusses on comprehend the application of these analytical procedures to obtain an association between the calculated PRS score and the different variables of study.

## *Methodology and materials*

### *Participants*

The Target data were obtained from leucocyte blood-samples of schizophrenia patients from the repository of PSICMET, PSICOBANC, and PSICOBANC 2 genetic biobank. This target data contains the PANSS value of positive, negative, general, and total symptoms conducted by the different genetic studies form Pere Mata research group, and it is also the data used in the quality control procedure. It was also included on the study the FAS scale from the Target data participants.

The Base data were obtained from Psychiatric Genomics Consortium Wave 3 schizophrenia genome-wide association study meta-analysis summary statistics (PGC3SCZ wave 3) (Trubetskoy et al., 2022).

CIBERSAM dataset were collected by the authors and recruited from psychiatric patients units at seven different hospitals in Spain. Participation was approved by the ethical committees at the hospitals from the CIBERSAM samples. Additionally, samples were genotyped using the Illumina Infinium PsychArray at the Broad Institute as part of the PGC-SCZ wave 3.

The information of the study contains phenotypic, neuropsychologic, and omics data from patients with schizophrenia and other neuropsychologic disorders such as major depressive disorder and bipolar disorder.

*Tabla 1. Sample data information*

| DATASET | SOURCE | N | %MALES | %FEMALES | MEAN AGE |
|---|---|---|---|---|---|
| **GEMME** | Patient group from the Biomedical Research Network in mental Health (CIBERSAM) | 230 | 60.87 | 39.13 | 24 |
| **PSICOBANC** | | 150 | 67.33 | 32.67 | 24 |
| **PSICOBANC 2** | | 58 | 75.86 | 24.14 | 20 |
| **PSICMET** | | 45 | 68.89 | 31.11 | 18 |
| **HOMFOL** | Control group from the Biomedical Research Network in mental Health (CIBERSAM) | 381 | 52.23 | 47.77 | - |
| **GENUP** | | 394 | 43.90 | 56.10 | - |
| **TOTAL (In number)** | | 1258 | 678 | 561 | 21.5 |

## *Statistical and computational methods*

Initially there is a quality control analysis, and a posterior polygenic risk score analysis from the data obtained of the earlier quality control. After this, the base data with the positive, the negative, and the general scale values, included the polygenic risk score data obtained were used for the posterior cross-trait analysis (Marees et al., 2018).

Data management and analysis was performed using PLINK 1.09 version according to standard procedures of GWAS (Chang et al., 2015), R-studio (Posit | The Open-Source Data Science Company, n.d.), and with Python command base line tool PRS-CS (GitHub - Getian107/PRScs: Polygenic Prediction via Continuous Shrinkage Priors, n.d.).

## *Quality control procedure*

This statistic analysis and quality control procedure consists of a rigorous methodology based on several steps/filters which ones allows to obtain a better-quality genotype data with the desired characteristics. In addition, genetic data can be refined by deleting the individuals or the genetic variations that presents any inconsistencies related to the statistical analyse performed.

It has been used the free, open-source whole-genome association analysis toolset PLINK beta version 1.09 (the latest one) supplemented with other conventional statistical packages such as R, specifically to study the analytical results through their graphic representation. PLINK can either read binary files or text-format files, widely used to perform statistical analysis on huge genomic datasets.

The overview of the quality control procedure is commonly based on 7 steps that should be conducted prior genetic association analysis (Marees et al., 2018).

The two scripts used for the quality control procedure are presented on the annexes.

### Overview of quality control procedure:

1. Missingness of SNPs and individuals: Excludes the SNPs with a prominent level of missingness in a substantial proportion of the individuals. Individuals who have a high rate of genotype missingness are also excluded, and individuals with low genotype calls are removed. This step should be performed before individual filtering. It is usually a relaxed threshold >20% and a stringent threshold >2%.

2. Sex discrepancy: Examine the discrepancies between sex of the subjects noted in the dataset and their sex based on X chromosome heterozygosity/homozygosity rates. Males present a X chromosome homozygosity estimate >0.8 and females present a value <0.2. This filter may indicate sample mix-ups.
3. Minor allele frequency (MAF): The MAF parameter is a frequency of the least often occurring allele at a certain position. In this step there are only included SNPs above the set MAF threshold, which depend on the sample size. According to this, larger sample should use lower MAF threshold, for samples moreover N= 100.000, a threshold value of 0.01, and for moderate samples N= 1000, a threshold value of 0.05.
4. Hardy– Weinberg equilibrium (HWE): The HWE concerns the relation between the allele and genotype frequencies, assuming an indefinitely large population, with no mutation, migration, or no selection. Excludes genetic markers which deviate from Hardy–Weinberg equilibrium value. For binary traits it is suggest excluding HWE p value <1e−10 in cases and <1e−6 in controls. Common indicator of genotyping error.
5. Heterozygosity: This refers to carrying two different alleles of a particular SNP, the proportion of heterozygous genotypes, this may indicate a low sample quality. This step excludes subjects which presents low or prominent levels of heterozygosity rates. Deviations may indicate sample contamination, inbreeding.
6. Relatedness: This indicates how strongly two individuals are genetically related. Initially, it proceeds to calculate the identity by descendent (IBD) of all sample pairs, this step uses independent SNPs (pruning) for the analysis and limit to autosomal chromosomes only.

   Subsequently, the threshold value is established and a list of individuals with a relationship above a chosen threshold is generated. Suggesting that subjects related to each other, for example, present a pi-hat value >0.2 (second degree relatives).
7. Population stratification: Indicates the presence of multiple subpopulations, such as individuals with different ethnic background present in the data set. At first, calculates the identity by descendent (IBD) of all sample pairs (pruning). Secondly produces a k-dimensional (k is typically 10) representation of the structure in the data set, based on IBS (European population stratification plot). This step consists of multiple proceedings, described as a “control for population stratification”. Population stratification may lead to false positive associations regarding allele frequencies can differ between subpopulations.

The QC performed is outlined is summarize in the following steps: (1) SNPs with a genotype missing rate > 2% were removed from the data; (2) samples with a genotype missing rate > 2%, samples with absolute value of heterozygosity > 2%, or samples that failed sex checks were excluded; (3) SNPs with significant group associations with a $P< 1 x10^{-6}$ were removed, (4) markers with MAF < 5%, missing rate > 2%, significant deviation from HWE with a $P < 1 \times 10^{-10}$ , and relatedness > 20% (PI_HAT > 0.2) were removed before the population stratification step.

In this study the data target included participants from diverse populations, the population reference panel used was the 1KG European samples from the 1000 genomes project data set (1000 Genomes | A Deep Catalog of Human Genetic Variation, n.d.).

To infer the ancestry of the target data Participants. Following this it was conducted principal components (PCs) of the genotype data in the 1KG samples and trained a random forest model with 10 PCs from the super population labels (African [AFR], American [AMR], East Asian [EAS], European [EUR], and South Asian [SAS]). Following this, the random forest model was then applied to the Target data participants.

### *Population stratification results*

Once the first six steps of data filtering are conducted in the quality control, we proceed to the stratification of the population, selecting within our target data the individuals with European ethnicity. For this, a PCA were compounded for the genotyping data (1KG samples) with the different ethnicities that can be presented among the participants.

The principal component analysis (PCA) for the population stratification (Figure 1) presents the stratification based on the ancestry and the individuals from our target data (own) in a bidimensional plot. In the study the participants excluded were the individuals with an ethnicity different from European, and the participants (controls and patients) from the CIBERSAM target data are classified as (own). Related to the PCA results, the individuals which remained on the data were the ones above the threshold, which present European ethnicity and are classified as “own”.

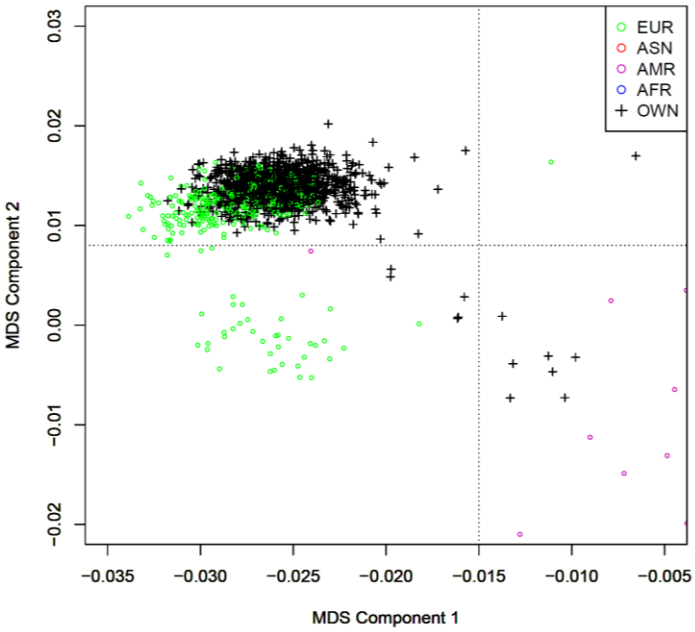



*Figure 1. Population stratification PCA plot. Green subjects represent European ethnicity, red subjects represent Asian ethnicity, pink subjects represent American Indian or Alaska Native ethnicity, blue subject. The black cross are the individuals from target data (own).*

### *Polygenic Risk Score*

After the quality control test, it proceeds to compute the Polygenic prediction via Bayesian Regression and continuous shrinkage (CS) method using GWAS summary statistics. The Polygenic Risk Score is performed using the data from the FAS and PANSS value (positive, negative, general, and total PANSS).

This PRS procedure is necessary for determining how each SNP is expected to contribute to the polygenic score for a specific trait ("weight"). In this case, it is a target sample more modest in size compared to the rule thumb (2,000 subjects).

For the analysis, the different value of trait-specific weights (betas for continuous traits and log of the odds ratios for binary traits) is obtained from GWAS related to the desired traits of study. In the case of the target sample, the PRS is calculated for each individual construct on the weighted sum of the amount of risk alleles that the individual carries multiplied by the trait-specific weights. On the case of complex traits, SNP effect sizes are from public access.

## *Statistical analyses*

Initially, the study consisted of 1,258 individuals (n controls = 775, n patients = 483) with a total of 603,132 variables (loaded from binary files). Once the quality control was conducted, a final number of 1,048 individuals (controls and patients) and 239,470 variables was obtained.

The number of patients obtained at the end of the quality control procedure data set were 195 individuals and 239470 variants, with the PANSS scale information and FAS scale information, these are the final data set used to compute the statistical analysis.

The polygenic risk score was assessed with R-studio and python software, and the statistical analysis was assessed with R-studio and Jamovi analyse software, studying the correlation between the FAS and PANSS scales value (positive PANSS, negative PANSS, general PANSS, and total PANSS) and the Schizophrenia PRS score. Spearman's correlation coefficient (for not normal distributions) used in this analysis with a min-max normalization method (feature scaling), for an accordance between the ten Principal components and schizophrenia PRS score, and the rest of variables such as age, sex and the 5 psychological scales studied.

It was used a generalized linear model to evaluate PRS associations on case samples using Spearman's rank-order correlation (Spearman's rho), which is more suitable for nonparametric variables, using clinical traits as the dependent variable and the schizophrenia PRS with the 10 principal components as the independent variables and covariables. The clinical outcomes variables considered were the FAS and the PANSS scales. Spearman's rho measures the strength and direction of the monotonic relationship (less restrictive) between two variables and not the strength and direction of the linear relationship between the variables.

### *Performing polygenic risk prediction analysis*

The calculated scores from target sample can be used in a regression analysis to predict the trait which is expected to manifest genetic overlap with the trait of study. It can be used $R^2$ measure of the regression analysis for expressing the prediction accuracy, in addition certain parameters must be considered, such as the population stratification (controlled by including few MDS components as covariates for the regression analysis).

Obtaining a good accuracy in PRS prediction depends to a great extent on the (co-)heritability of the analysed traits, the number of SNPs, and the size of the sample. Despite this, to obtain a significant R2 it is enough the presence of a few thousand subjects in the target sample if the (co-)heritability of the trait(s) of the study and the sample size of the discovery sample used are sufficiently large.

The number of samples used on the PRS analysis were a total of 1048 individuals and 239,470 variants (patients and controls). From that target data it has been selected the 195 patients and variants for the linear model because this are the samples with PANSS (negative, positive, general, and total) and FAS additional information.

The software used was the PRS-cs Python based command line tool for studying the SNP effect sizes under continuous shrinkage priors using GWAS statistics and an external Linkage Disequilibrium reference panels constructed using the 1000 Genomes Project phase 3 samples (European reference used) (Ge et al., 2019).

## *Results*

### *Polygenic risk score evaluation*

The evaluation of the polygenic risk was conducted comparing the score value between the patients and control groups, it was used a density histogram for a better understanding of the distribution of the schizophrenia PRS score on each group. It can be appreciated how the patient group obtained a higher PRS score than the control group.

This fact shows us that the PRS procedure has been well conducted, since it is logical that already diagnosed patients obtain a higher score compared to controls (healthy individuals without symptoms associated with schizophrenia). In summary, this density histogram shows us that the PRS score for schizophrenia it has been correctly calculated and it is useful for value verification.

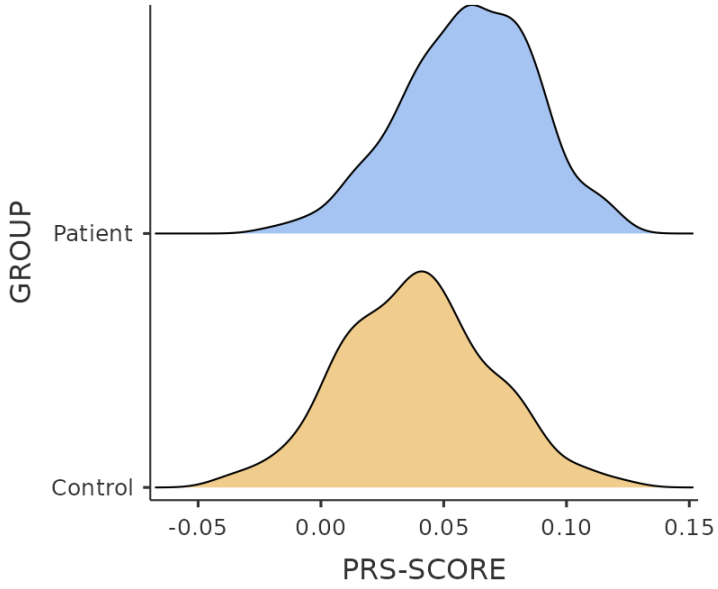



*Figure 2.Polygenic risk score in schizophrenia evaluated among the patients and control group. It can be seen how the PRS punctuation presented by the controls is lower compared to the PRS punctuation presented by the patients with schizophrenia in general term.*

### *Association between PRS and symptoms*

The association between the schizophrenia PRS and the symptoms scales are presented in Table 2. The Spearman correlation coefficient, rho, indicates the association of ranks. The association of the Positive PANSS and the PRS (rho = -0.097) indicates a negative association, the association between Negative PANSS and the PRS (rho = 0.136) indicates a positive association between the ranks, General PANSS and the PRS (rho = -0.075) indicates a negative association between the ranks, the Total PANSS and the PRS (rho = -0.089) indicates a negative association between the ranks, the FAS and the PRS (rho = 0.011) indicates no association between the ranks.

Positive associations indicate a positive relationship between the ranks individuals obtained in the Negative PANSS, the observed correlation demonstrate a higher rank in schizophrenia PRS is related to a higher Negative PANSS scale.

Negative associations indicate a lower relationship between the ranks, and the association of the FAS scale (rho = 0.011) indicates no relationship between the ranks. In general terms these association are not as robust as expected, despite this there is a delicate correlation between the Negative PANSS scale and the schizophrenia PRS.

Related to the Spearman's correlation method applied it can be appreciated the correlation scatter plots from the associations between the schizophrenia PRS and the symptoms scales separated by gender (Table 3.). In general terms the trend presented demonstrate no correlation, or a delicate correlation between the variables in the case of the Negative PANSS scale on the female group.

The different endophenotypes studied were the age, the gender, the 10 Principal components (Genotype data in 1KG samples) and the symptoms were measured using the schizophrenia polygenic risk score, based on the FAS scale and the different PANSS scores.

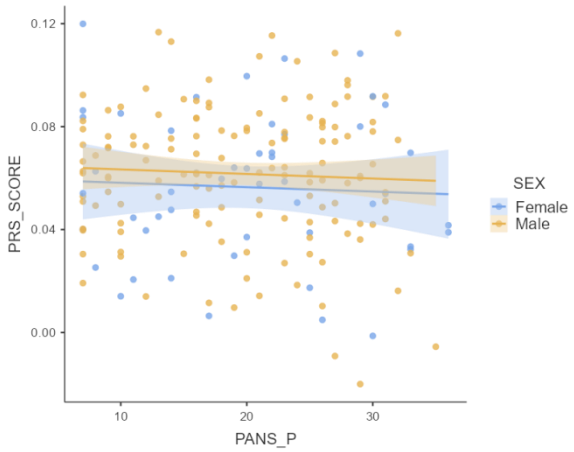
*Table 2. Correlation matrix between the different scales FAS and PANSS and the Schizophrenia PRS score with the Spearman correlation method.*

Spearman's correlation matrix

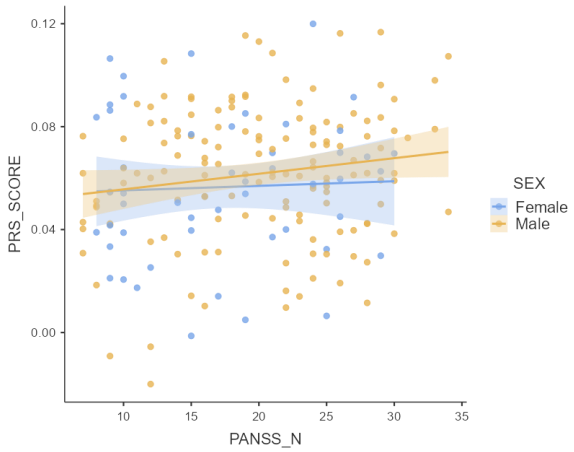
| | | Positive PANSS | Negative PANSS | General PANSS | Total PANSS | FAS | Schizophrenia PRS |
|---|---|---|---|---|---|---|---|
| Positive PANSS | Spearman's rho | — | | | | | |
| | df | — | | | | | |
| | p-value | — | | | | | |
| Negative PANSS | Spearman's rho | -0.4325 *** | — | | | | |
| | df | 107 | — | | | | |
| | p-value | < .001 | — | | | | |
| General PANSS | Spearman's rho | 0.4521 *** | 0.0689 | — | | | |
| | df | 107 | 107 | — | | | |
| | p-value | < .001 | 0.476 | — | | | |
| Total PANSS | Spearman's rho | 0.5643 *** | 0.3188 *** | 0.8592 *** | — | | |
| | df | 107 | 107 | 107 | — | | |
| | p-value | < .001 | < .001 | < .001 | — | | |
| FAS | Spearman's rho | 0.0609 | -0.2112 * | -0.0250 | -0.0772 | — | |
| | df | 107 | 107 | 107 | 107 | — | |
| | p-value | 0.530 | 0.027 | 0.796 | 0.425 | — | |
| Schizophrenia PRS | Spearman's rho | -0.0970 | 0.1362 | -0.0756 | -0.0896 | 0.0113 | — |
| | df | 107 | 107 | 107 | 107 | 107 | — |
| | p-value | 0.316 | 0.158 | 0.435 | 0.354 | 0.908 | — |

Note. * p < .05, ** p < .01, *** p < .001, N = 110

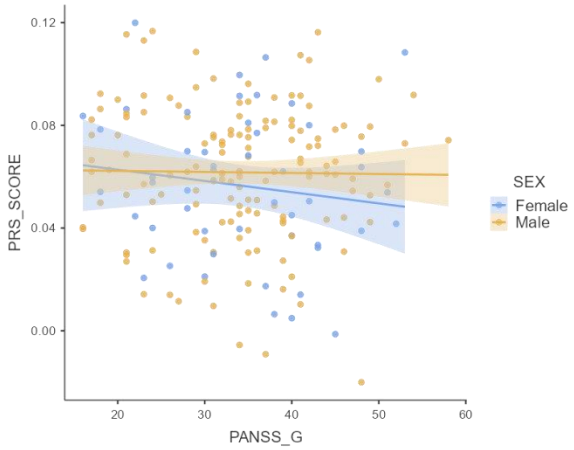
*Table 3. Correlation scatter plots between the Schizophrenia PRS score and each scale value (Positive, Negative, General and Total PANSS, and FAS scales). Spearman correlation method used.*



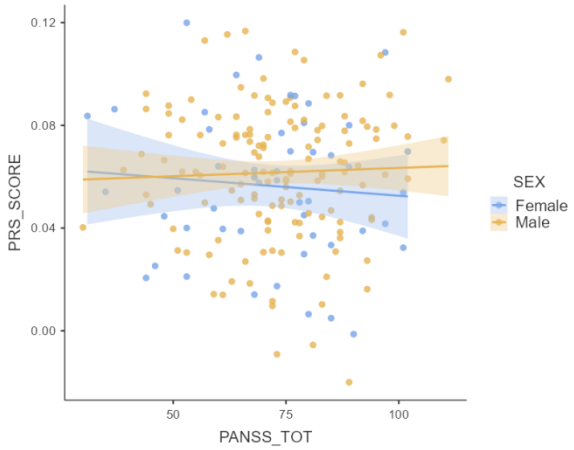

*(A) Scatterplot: Positive PANSS vs Scz-PRS*



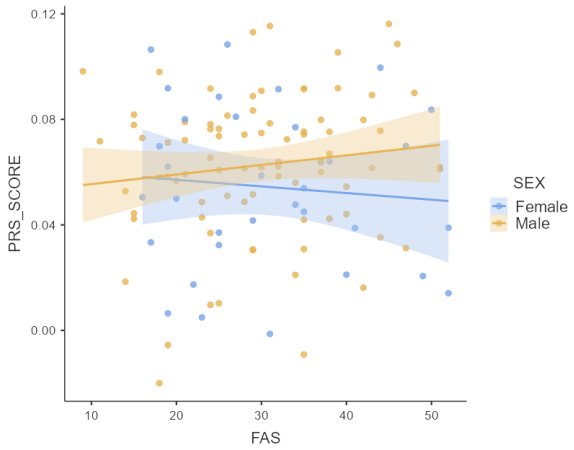

*(B) Scatterplot: Negative PANSS vs Scz-PRS*




*(C)Scatterplot: General PANSS vs Scz-PRS*




*(D) Scatterplot: Total PANSS vs Scz-PRS*




*(E) Scatterplot: FAS vs Scz-PRS*

*(A); Correlation coefficient (r) = -0.097, p=0.32*
*No relationship presented in general, in females there's a delicate negative linear relationship.*
*(B); Correlation coefficient (r) = 0.14, p=0.16*
*No relationship presented in general, female present a delicate positive correlation and males a delicate negative relationship.*
*(C); Correlation coefficient (r) = -0.076, p=0.43*
*There is no correlation presented.*
*(D); Correlation coefficient (r) = -0.09, p=0.35*
*There is no correlation presented.*
*(E); Correlation coefficient (r) = 0.011, p= 0.91*
*There is no correlation presented.*

## *Linear model*

The linear correlation and the simple linear regression are statistical methods which study the lineal relationship between two variable, and the linear regression consist of generate a model which permits explaining the linear relationship among diverse variables. The dependent variable is called y, however the predictor or independent variables are represented as x.

A basic lineal model follows the equation presented below (a.).

(a.) $yi=\beta0+\beta1x1i+ \ldots +\beta nxni+\epsilon i$

Where $x_{1\ldots n}$ represents each predictor variable and β1 represent the coefficients/parameters to be estimated. The effect of each coefficient (for example, β2) should be interpreted as the change given by a unitary change in the predictor variable associated with that coefficient (x2) provided that the other variables (x) are kept constant. Furthermore, we have an ordinate to the origin (or β0) which represents the overall mean.

To evaluate the linear relationship between the phenotypic traits such as PANSS and FAS scale with the different variables of the study (the 10 principal components and the schizophrenia PRS) it is made a linear model with spearman's correlation. The linear model equations from each variable of study were obtained using R-studio software and are described below.

1. Linear model: Positive PANSS and the covariables (10 PC, PRS schizophrenia, Age and sex):

$$Positive\ PANSS = 19.60 - 76.31C1 - 60.06C2 + 21.90C3 - 147.69C4 + 30.87C5 - 61.39C6 - 110.24C7 + 137.72C8 + 112.30C9 - 49.80C10 + 0.064Age - 0.65Sex - 23.66PRS$$

2. Linear model: Negative PANSS and the covariables (10 PC, PRS schizophrenia, Age and sex):

$$Negative\ PANSS = 5.264 + 91.27C1 - 126.62C2 - 156.87C5 + 47.46C4 + 71.29C5 + 22.83C6 + 87.43C7 + 143.21C8 - 40.86C9 + 8.01 + 0.22Age + 2.89Sex + 29.75PRS$$

3. Linear model: General PANSS and the covariables (10 PC, PRS schizophrenia, Age and sex):

$$General\ PANSS = 39.007 - 108.11C1 - 71.94C2 + 36.72C3 - 35.03C4 + 126.05C5 - 91.02C6 - 216.98C7 + 192.93C8 + 95.54C9 - 116.00C10 - 0.011Age - 1.68Sex - 29.14\ PRS$$

4. Linear model: Total PANSS and the covariables (10 PC, PRS schizophrenia, Age and sex):

$$Total\ PANSS = 63.87 - 93.14C1 - 285.62C2 - 98.25C3 - 135.26C4 + 228.20C5 - 129.58C6 - 239.78C7 + 473.86C8 + 166.98C9 - 157.80C10 + 0.27Age - 0.56Sex - 23.05PRS$$

5. Linear model: FAS and the covariables (10 PC, PRS schizophrenia, Age and sex):

$$FAS = 38.28 + 253.96C1 - 367.44C2 + 78.67C3 - 28.54C4 - 326.61C5 + 9.94C6 + 111.90C7 - 31.28C8 - 153.69C9 - 30.42C10 - 0.21Age - 1.53Sex + 30.38PRS$$

The graphics and the correlation tests show there is no linear relation and with no significance, despite this can be appreciated that negative coefficients present a negative relationship (decrease its value) with the dependent variable, and positive coefficients present a positive relationship (increase its value).

Regarding this the covariables may increase or decrease the value from the clinical variables, despite this the PRS covariable it increases the value of the Negative PANSS scale and FAS scale, it is logical because patients with schizophrenia present a higher PRS score.

*Table 4. PRS correlation with clinical variables*

| Lineal model descriptives | | | | | |
|---|---|---|---|---|---|
| **Schizophrenia Polygenic risk score** | **Clinical variables** | **N** | **P-value** | **Multiple $R^2$** | **Adjusted $R^2$** |
| | Positive PANSS | 195 | 0.6736 | 0.05348 | 0.6736 |
| | Negative PANSS | 195 | **5.889e-05** | 0.203 | 0.1457 |
| | General PANSS | 195 | 0.4846 | 0.06501 | -0.002142 |
| | Total PANSS | 195 | 0.1713 | 0.09033 | 0.025 |
| | FAS | 110 | 0.4791 | 0.1088 | -0.002587 |

In the lineal model of Negative PANSS with the PRS, there is a significance related to the p-value (5.889e-05), in addition the model contains several significant covariables such as Age and Sex, with an important significance (***, *), and C1, C3 and C8 with a moderate significance (**.**) (table 4).

That results obtained in the lineal model delights that Negative PANSS presents a linear relationship with the Age and Sex from the patients but not so important with the PRS punctuation from each individual. This scenario is common among the different clinical variables studied (PANSS and FAS scale), so, the PRS do not show any significance as a co-variable in the six lineal models evaluated in this study.

On the results obtained by R-studio, on the estimate column returns the estimated value for the parameters of the linear model equation that are equal to the intercept and slope. In the linear model generated for the Negative PANSS is; Negative PANSS = 5.264 +91.27C1 -126.62C2 -156.87C5 +47.46C4 +71.29C5 +22.83C6 +87.43C7 +143.21C8 -40.86C9 +8.01 +0.22Age +2.89Sex +29.75PRS, which indicates that for each unit that the number of PRS increases, the number of the Negative PANSS increases by an average of 29.75 units.

Likewise, together with the regression results, the standard deviations of the parameters are shown, just with the value of the t statistic and the p-value of each of the parameters. For the Negative PANSS generated model, both the intercept and the slope are significant (p-values are less than 0.05).

For its part, the value of $R^2$ indicates that the calculated model explains 20.30% of the variability present in the response variables (phenotypic variables) through the independent variables (Negative PANSS). In this sense, it is evident that the goodness of fit of the model is not very high.

Finally, the p-value obtained in the test (5.889e-05) determines that the variance explained by the model is significantly higher than the total variance. It is the parameter that determines if the model is significant and therefore can be accepted.

*Figure 4. Lineal model results with R-studio. Spearman's correlation method, Negative PANSS and clinical covariables.*

```
Call:
lm(formula = G28_ALL$PANSS_N ~ G28_ALL$C1 + G28_ALL$C2 + G28_ALL$C3 +
    G28_ALL$C4 + G28_ALL$C5 + G28_ALL$C6 + G28_ALL$C7 + G28_ALL$C8 +
    G28_ALL$C9 + G28_ALL$C10 + G28_ALL$Age + G28_ALL$SEX + G28_ALL$PRS_SCORE)
Residuals:
    Min      1Q  Median      3Q     Max
-15.0364  -4.7235   0.7422   4.7892  22.3185
Coefficients:                    Estimate   Std. Error t value Pr(>|t|)
(Intercept)              5.26479    2.78891   1.888   0.0607 .
G28_ALL$C1              91.27785  115.89233   0.788   0.4320
G28_ALL$C2            -126.61709  125.05327  -1.013   0.3126
G28_ALL$C3            -156.86827   79.72479  -1.968   0.0506 .
G28_ALL$C4              47.45615   98.84628   0.480   0.6317
G28_ALL$C5              71.29089   92.24261   0.773   0.4406
G28_ALL$C6              22.82724   92.89635   0.246   0.8062
G28_ALL$C7              87.43415   79.79489   1.096   0.2747
G28_ALL$C8             143.20550   84.20601   1.701   0.0907 .
G28_ALL$C9             -40.86163   78.57284  -0.520   0.6037
G28_ALL$C10              8.00787   80.29784   0.100   0.9207
G28_ALL$Age              0.21620    0.04158   5.200 5.36e-07 ***
G28_ALL$SEX              2.88576    1.13059   2.552   0.0115 *
G28_ALL$PRS_SCORE       29.74612   18.89113   1.575   0.1171
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 6.815 on 181 degrees of freedom
Multiple R-squared:  0.203,     Adjusted R-squared:  0.1457
F-statistic: 3.546 on 13 and 181 DF,  p-value: 5.889e-05
```

## *Discussion*

The statistical analysis in this study demonstrate that it is found no statistically significant associations between schizophrenia PRS and the across traits in the patients with schizophrenia spectrum disorder (the samples with values for the across traits studied). This indicates that the schizophrenia PRS did not contribute to the prediction of the Positive and negative syndrome scales, and FAS scale in this study as it is explained in the results.

These findings demonstrate a weak detectable association between polygenic risk of schizophrenia and the Positive and negative syndrome scale and the Functional assessment short test. Despite this, there is an existent significant association between the age and sex of the patients, and the Negative PANSS scale as it is presented in the results (Table 4, figure 4).

However, this lacking significant associations with PRS could have some suitable explanations. Firstly, common variants estimated to explain at best 30-50% of schizophrenia's heritability(Trubetskoy et al., 2022). The current schizophrenia PRS do not represent an important amount of schizophrenia's genotypic variance such as copy number variants, rare or de novo variants, and small deletions and insertions (Lyngstad et al., 2020).

Most studies in the field of PRS associated to clinical covariables have been focused from research discovery to clinical research studies. And although PRS-CS provides a substantial improvement over existing methods for polygenic prediction, current prediction accuracy of PRS is still lower than what can be considered clinically useful, and much work is needed to further improve the predictive performance and translational value of PRS. In theory, the utility of PRS depends on multiple factors, including the GWAS sample size, and the heritability and genetic architecture of the disease (Ge et al., 2019).

In addition, recent studies PRS is not powerful enough to predict the relationship between PRS schizophrenia and apathy (Lyngstad et al., 2020; Ranlund et al., 2018). In our case, a difference can be observed in the PRS score obtained (Figure 2), although an overlap can be observed between the control group and the group of patients, this slight difference shows that the PRS has been calculated correctly. This overlap may be because the sample size used was small (n = 195), so a larger sample size would add more individuals, and therefore, would add more power to the study (Sample Size and Power, n.d.).

Polygenic risk scores (PRS) have been described in a large quantity of common complex disease since their first appearance in the context of schizophrenia and bipolar disorder in 2009. However, the clinical utility of PRS is quite limited because PRS usually only take in account the heritable part of a trait and ignore the ethological role of environment and lifestyle (Koch et al., 2023).

The most interesting finding was that the PRS verify a probabilistic susceptibility between the participants, the control group presented a lower punctuation compared to the group of patients (individuals with diagnosed schizophrenia). These findings explain the genetic association between PRS and the neuropsychiatric disorder.

## *Limitations*

The principal limitation of the study is the moderate sample size for used, compared to other studies, the sample in this instigation (n = 110) is quite low in number, this explained the no association between the PRS and the multiple phenotypic traits (FAS and PANSS scales).

Polygenic risk score requires using two independent data sets, the genotype level data, and the base data (training set usually a GWAS study). A limitation in this field is that genotype data is not usually publicly available due to patient confidentiality.

Despite this issue, we were able to analyse samples of unrelated individuals with schizophrenia and healthy controls obtaining a PRS indicating that the PRS demonstrate that genetic predisposition to schizophrenia can be predicted.

## *Conclusions*

Despite PRS are based on common genetic variants, environmental factors may be important for the development of schizophrenia and not only the influence from genetic factors.

In summary, considering our results and waiting for future studies that replicate and amplify the work realized in this study with larger and more heterogeneous samples, it is important to focus on environmental risk factors such as maternal behaviour and health which increase the risk, many studies have obtained important results with a good correlation and association between schizophrenia PRS, and the phenotypes related to the neuropsychiatric disorder.

## *References*


*1000 Genomes | A Deep Catalog of Human Genetic Variation*. (n.d.). Retrieved 6 June 2023, from https://www.internationalgenome.org/

Amoretti, S., Mezquida, G., Rosa, A. R., Bioque, M., Cuesta, M. J., Pina-Camacho, L., Garcia-Rizo, C., Barcones, F., González-Pinto, A., Merchán-Naranjo, J., Corripio, I., Vieta, E., Baeza, I., Cortizo, R., Bonnín, C. M., Torrent, C., & Bernardo, M. (2021). The functioning assessment short test (FAST) applied to first-episode psychosis: Psychometric properties and severity thresholds. *European Neuropsychopharmacology*, *47*, 98–111. https://doi.org/10.1016/J.EURONEURO.2021.02.007

Bachmann, S., Resch, F., & Mundt, C. (2022). Psychological Treatments for Psychosis: History and Overview. *Https://Doi.Org/10.1521/Pdps.2022.50.1.24*, *50*(1), 24–42. https://doi.org/10.1521/PDPS.2022.50.1.24

Blaylock, R. L., & Faria, M. (2021). New concepts in the development of schizophrenia, autism spectrum disorders, and degenerative brain diseases based on chronic inflammation: A working hypothesis from continued advances in neuroscience research. *Surgical Neurology International*, *12*(556). https://doi.org/10.25259/SNI_1007_2021

Blaylock, R. L., Faria, M., & Blaylock, R. (2021). New concepts in the development of schizophrenia, autism spectrum disorders, and degenerative brain diseases based on chronic inflammation: A working hypothesis from continued advances in neuroscience research HISTORICAL BACKGROUND AND CLINICAL PRESENTATION. *Surgical Neurology International • 2021 •*, *12*(556). https://doi.org/10.25259/SNI_1007_2021

Buizza, C., Strozza, C., Sbravati, G., de Girolamo, G., Ferrari, C., Iozzino, L., Macis, A., Kennedy, H. G., & Candini, V. (2022). Positive and negative syndrome scale in forensic patients with schizophrenia spectrum disorders: a systematic review and meta-analysis. *Annals of General Psychiatry*, *21*(1). https://doi.org/10.1186/S12991-022-00413-2

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1). https://doi.org/10.1186/S13742-015-0047-8

Correll, C. U., & Schooler, N. R. (2020). Negative Symptoms in Schizophrenia: A Review and Clinical Guide for Recognition, Assessment, and Treatment. *Neuropsychiatric Disease and Treatment*, *16*, 519. https://doi.org/10.2147/NDT.S225643

Ferrari, A. (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry*, *9*(2), 137–150. https://doi.org/10.1016/S2215-0366(21)00395-3

Ge, T., Chen, C. Y., Ni, Y., Feng, Y. C. A., & Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, *10*(1). https://doi.org/10.1038/S41467-019-09718-5

*GitHub - getian107/PRScs: Polygenic prediction via continuous shrinkage priors*. (n.d.). Retrieved 6 June 2023, from https://github.com/getian107/PRScs

Gopalakrishnan, M., Farchione, T., Mathis, M., Zhu, H., Mehta, M., Uppoor, R., & Younis, I. (2020). Shortened Positive and Negative Symptom Scale as an Alternate Clinical Endpoint for Acute Schizophrenia Trials: Analysis from the US Food & Drug Administration. *Https://Doi.Org/10.1176/Appi.Prcp.20200003*, *3*(1), 38–45. https://doi.org/10.1176/APPI.PRCP.20200003

Kasap, M., & Dwyer, D. S. (2022). How Variation in Risk Allele Output and Gene Interactions Shape the Genetic Architecture of Schizophrenia. *Genes*, *13*(6). https://doi.org/10.3390/GENES13061040

Koch, S., Schmidtke, J., Krawczak, M., & Caliebe, A. (2023). Clinical utility of polygenic risk scores: a critical 2023 appraisal. *Journal of Community Genetics*. https://doi.org/10.1007/S12687-023-00645-Z

Lam, M., Chen, C. Y., Li, Z., Martin, A. R., Bryois, J., Ma, X., Gaspar, H., Ikeda, M., Benyamin, B., Brown, B. C., Liu, R., Zhou, W., Guan, L., Kamatani, Y., Kim, S. W., Kubo, M., Kusumawardhani, A. A. A. A., Liu, C. M., Ma, H., … Huang, H. (2019). Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nature Genetics*, *51*(12), 1670–1678. https://doi.org/10.1038/S41588-019-0512-X

Leucht, S., Barabássy, Á., Laszlovszky, I., Szatmári, B., Acsai, K., Szalai, E., Harsányi, J., Earley, W., & Németh, G. (2019). Linking PANSS negative symptom scores with the Clinical Global Impressions Scale: understanding negative symptom scores in schizophrenia. *Neuropsychopharmacology 2019 44:9*, *44*(9), 1589–1596. https://doi.org/10.1038/s41386-019-0363-2

Lyngstad, S. H., Bettella, F., Aminoff, S. R., Athanasiu, L., Andreassen, O. A., Færden, A., & Melle, I. (2020). Associations between schizophrenia polygenic risk and apathy in schizophrenia spectrum disorders and healthy controls. *Acta Psychiatrica Scandinavica*, *141*(5), 452–464. https://doi.org/10.1111/ACPS.13167

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, *27*(2). https://doi.org/10.1002/MPR.1608

Martínez, A. L., Brea, J., Rico, S., de los Frailes, M. T., & Loza, M. I. (2021). Cognitive Deficit in Schizophrenia: From Etiology to Novel Treatments. *International Journal of Molecular Sciences 2021, Vol. 22, Page 9905*, *22*(18), 9905. https://doi.org/10.3390/IJMS22189905

Năstase, M. G., Vlaicu, I., & Trifu, S. C. (2022). Genetic polymorphism and neuroanatomical changes in schizophrenia. *Romanian Journal of Morphology and Embryology*, *63*(2), 307. https://doi.org/10.47162/RJME.63.2.03

Owen, M. J. (2023). Genomic insights into schizophrenia. *Royal Society Open Science*, *10*(2). https://doi.org/10.1098/RSOS.230125

*Posit | The Open-Source Data Science Company*. (n.d.). Retrieved 6 June 2023, from https://posit.co/

*Sample size and power*. (n.d.). Retrieved 6 June 2023, from https://www.iwh.on.ca/what-researchers-mean-by/sample-size-and-power

Siegel-Ramsay, J. E., Wu, B., Kapczinski, F., Lanza di Scalea, T., David, S., Frey, B. N., Strakowski, S. M., & Almeida, J. R. C. (2023). Functional assessment short test (FAST): Self-administration in outpatient mental health settings. *Journal of Psychiatric Research*, *160*, 258–262. https://doi.org/10.1016/J.JPSYCHIRES.2023.02.029

Solmi, M., Radua, J., Olivola, M., Croce, E., Soardo, L., Salazar de Pablo, G., Il Shin, J., Kirkbride, J. B., Jones, P., Kim, J. H., Kim, J. Y., Carvalho, A. F., Seeman, M. V., Correll, C. U., & Fusar-Poli, P. (2022). Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. *Molecular Psychiatry*, *27*(1), 281. https://doi.org/10.1038/S41380-021-01161-7

Tang, Y., You, D., Yi, H., Yang, S., & Zhao, Y. (2022). IPRS: Leveraging Gene-Environment Interaction to Reconstruct Polygenic Risk Score. *Frontiers in Genetics*, *13*, 368. https://doi.org/10.3389/FGENE.2022.801397/BIBTEX

Trubetskoy, V., Pardiñas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T. B., Bryois, J., Chen, C. Y., Dennison, C. A., Hall, L. S., Lam, M., Watanabe, K., Frei, O., Ge, T., Harwood, J. C., Koopmans, F., Magnusson, S., Richards, A. L., Sidorenko, J., … van Os, J. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature*, *604*(7906), 502–508. https://doi.org/10.1038/S41586-022-04434-5

Uranova, N. A., Vikhreva, O. V., & Rakhmanova, V. I. (2022). [Specific interactions between microglia and oligodendrocytes in white matter in continuous schizophrenia]. *Zhurnal Nevrologii i Psikhiatrii Imeni S.S. Korsakova*, *122*(12), 128–137. https://doi.org/10.17116/JNEVRO2022122121128

Zhang, R., Sjölander, A., Ploner, A., Lu, D., Bulik, C. M., & Bergen, S. E. (2022). Novel disease associations with schizophrenia genetic risk revealed in ~400 000 UK Biobank participants. *Molecular Psychiatry*, *27*(3), 1448. https://doi.org/10.1038/S41380-021-01387-5

Zick, J. L., Staglin, B., & Vinogradov, S. (2022). Eliminate Schizophrenia. *Schizophrenia Research*, *242*, 147. https://doi.org/10.1016/J.SCHRES.2022.01.004

## *Self-assessment*

Initially, when I started in the GAP group, I would have never imagined that bioinformatics has such an important paper in the psychiatric disorders. My first three months I learned how to use PLINK and how to conduct a quality control procedure using the bioinformatics tools (Marees et al., 2018). This tutorial from Marres A. helped me to understand the importance of this previous procedure.

Once I ended the quality control, I learned how to calculate the Polygenic risk score associated to schizophrenia, using the tutorial explained in the Git-Hub platform by Tian Ge (GitHub - Getian107/PRScs: Polygenic Prediction via Continuous Shrinkage Priors, n.d.).

Followed by this step, I learned the application of statistics to explain the different relation between the PRS and the different phenotypic traits studied (PANSS and FAS scale).

Furthermore, I learned how to be resolutive on my own and work independently, I also learned to do research work based on bioinformatics, which was my intention and thanks to this work I have learned and valued the work behind each investigation and the role of bioinformatics.

## *Keywords*

**Schizophrenia:** Neuropsychological disorder characterized by the abnormal way people interpret reality. This complex disorder presents a combination of hallucinations, delusions, and intense disorder thinking and behavior that affects lifestyle individuals.

**SNPs:** Single nucleotide polymorphism, called SNPSs. Is a germline variation that occurs when a single nucleotide at a specific position in the genome.

**CNVs:** Copy number variants, called CNVs. Genetic trait involving the number of copies of a particular gene present in the genome of a person.

**PRS:** Polygenic risk score, called PRS. It estimates the genetic risk of an individual for a specific disease or trait.

**GWAS:** Genome-wide association study, called GWAS. Research approach used to detect genomic variants that statistically associated with the risk for a disease or particular trait.

**Environmental factors:** These are factors related to genetics are defined to exposures to substances from the ambient we lived or work, behaviors (such smoking) that generates an increase an individual's risk of disease.

**Heritability:** The proportion of variation among the population that is attributed to inherited genetic factors.

## *Annexes*

**Script 1 for the quality control procedure (filter 1-6)**

```
## this script needs 3 arguments
## arg 1 --> folder where the plink is executable
## arg 2 --> folder with the genotype data
## arg 3 --> folder where the files are saved
# Create arguments
#plinkfolder=$1
plinkfolder="/home/gloria/Downloads/plink_linux_x86_64_20230116/plink"
#genofolder=$2
genofolder="/home/gloria/Desktop/carpeta_Gloria/Genotips/G28"
#outfolder=$3
outfolder="/home/gloria/Desktop/carpeta_Gloria/QC"
### Step 1 ### Investigate missingness per individual and per SNP and make
histograms.
$plinkfolder --bfile $genofolder --geno 0.2 --make-bed --out $outfolder/geno_0.2
$plinkfolder --bfile $outfolder/geno_0.2 --mind 0.2 --make-bed --out
$outfolder/mind_0.2
$plinkfolder --bfile $outfolder/mind_0.2 --geno 0.02 --make-bed --out
$outfolder/geno_0.02
$plinkfolder --bfile $outfolder/geno_0.02 --mind 0.02 --make-bed --out
$outfolder/mind_0.02
### Step 2 ### Check for sex discrepancy.
$plinkfolder --bfile $outfolder/mind_0.02 --check-sex --make-bed --out
$outfolder/G28_sexcheck
# 1) Delete individuals with sex discrepancy.
grep "PROBLEM" $outfolder/G28_sexcheck.sexcheck| awk '{print$1,$2}'>
$outfolder/sex_discrepancy.txt
# The following two scripts can be used to deal with individuals with a sex
discrepancy.
```

```
# Note, please use one of the !TWO OPTIONS! below to generate the bfile, this
file we will use in the next step of this tutorial.
# 1_1)This command removes the list of individuals with the status PROBLEM.
$plinkfolder --bfile $outfolder/mind_0.02 --remove
$outfolder/sex_discrepancy.txt --make-bed --out $outfolder/G28_6_sex
# 1_2) impute-sex.
# This imputes the sex based on the genotype information into your data set.
$plinkfolder --bfile $outfolder/mind_0.02 --impute-sex --make-bed --out
$outfolder/G28_6_sex_2
#exclude chromosome not on the listed chromosomes (from 1-22)
$plinkfolder --bfile $outfolder/G28_6_sex --chr 1-22 --make-bed --out
$outfolder/G28_1-22
### Step 3 ### Generate a bfile with autosomal SNPs only and delete SNPs with a
low minor allele frequency (MAF).
awk '{ if ($1 >= 1 && $1 <= 22) print $2 }' $outfolder/G28_1-22.bim >
$outfolder/snp_1_22.txt
$plinkfolder --bfile $outfolder/G28_1-22 --extract snp_1_22.txt --make-bed --out
$outfolder/G28_autosomic
$plinkfolder --bfile $outfolder/G28_autosomic --freq --out $outfolder/MAF_check
Rscript --no-save MAF_check.R
# Remove SNPs with a low MAF frequency.
$plinkfolder --bfile $outfolder/G28_autosomic --maf 0.05 --make-bed --out
$outfolder/G28_maf0.05
### Step 4 ###
# Delete SNPs which are not in Hardy-Weinberg equilibrium (HWE).
# Check the distribution of HWE p-values of all SNPs.
$plinkfolder --bfile $outfolder/G28_maf0.05 --hardy --out $outfolder/hardy
awk '{ if ($9 <0.00001) print $0 }' $outfolder/hardy.hwe >
/$outfolder/hardy.hwezoomhwe.hwe
$plinkfolder --bfile $outfolder/G28_maf0.05 --hwe 1e-6 --make-bed --out
```

```
$outfolder/G28_hwe_filter_step1
$plinkfolder --bfile $outfolder/G28_hwe_filter_step1 --hwe 1e-10 --hwe-all
--make-bed --out $outfolder/G28_hwe_1e-10
### step 5 ###
# Generate a plot of the distribution of the heterozygosity rate of your
subjects.
# And remove individuals with a heterozygosity rate deviating more than 3 sd
from the mean.
$plinkfolder --bfile $outfolder/G28_hwe_1e-10 --exclude $outfolder/inversion.txt
--range --indep-pairwise 50 5 0.2 --out $outfolder/indepSNP
$plinkfolder --bfile $outfolder/G28_hwe_1e-10 --extract
$outfolder/indepSNP.prune.in --het --out $outfolder/R_check
# This file contains your pruned data set.($outfolder/R_check)
# The following code generates a list of individuals who deviate more than 3
standard deviations from the heterozygosity rate mean.
# For data manipulation we recommend using UNIX. However, when performing
statistical calculations R might be more convenient, hence the use of the
Rscript for this step:
# Plot of the heterozygosity rate distribution
Rscript --no-save heterozygosity_outliers_list.R
# Output of the command above: fail-het-qc.txt
sed 's/"// g' $outfolder/fail-het-qc.txt | awk '{print$1, $2}'>
$outfolder/het_fail_ind.txt
# Remove heterozygosity rate outliers.
$plinkfolder --bfile $outfolder/G28_hwe_1e-10 --remove
$outfolder/het_fail_ind.txt --make-bed --out $outfolder/hetfile_1
### step 6 ###
# It is essential to check datasets you analyse for cryptic relatedness.
# Assuming a random population sample we are going to exclude all individuals
above the pihat threshold of 0.2 in this tutorial.
```

```
# Check for relationships between individuals with a pihat > 0.2.
$plinkfolder --bfile $outfolder/hetfile_1 --extract $outfolder/indepSNP.prune.in
--genome --min 0.2 --out $outfolder/pihat_min0.2
# The HapMap dataset is known to contain parent-offspring relations.
# The following commands will visualize specifically these parent-offspring
relations, using the z values.
awk '{ if ($8 >0.9) print $0 }'
$outfolder/pihat_min0.2.genome>$outfolder/zoom_pihat.genome
Rscript --no-save Relatedness.R
# To demonstrate that the majority of the relatedness was due to
parent-offspring we only include founders (individuals without parents in the
dataset).
$plinkfolder --bfile $outfolder/hetfile_1 --filter-founders --make-bed --out
$outfolder/G28_11
#Now we will look again for individuals with a pihat >0.2
$plinkfolder --bfile $outfolder/G28_11 --extract $outfolder/indepSNP.prune.in
--genome --min 0.2 --out $outfolder/pihat_min0.2_in_founders
# For each pair of 'related' individuals with a pihat > 0.2, we recommend to
remove the individual with the lowest call rate.
$plinkfolder --bfile $outfolder/G28_11 --missing --out $outfolder/G28_miss
# Generate a list of FID and IID of the individual(s) with a Pihat above 0.2, to
check who had the lower call rate of the pair.
#vi 0.2_low_call_rate_pihat.txt
# Delete the individuals with the lowest call rate in 'related' pairs with a
pihat > 0.2
$plinkfolder --bfile $outfolder/G28_11 --remove $outfolder/pihat_mino0.2_clean_1
--make-bed --out $outfolder/G28_0.2_pihat
```

**Script 2 for the quality control procedure (population stratification)**

```
## this script needs 4 arguments
## arg 1 --> folder where the plink executable is
## arg 2 --> folder with the genotype data
## arg 3 --> folder where the files are saved
## arg 4 --> folder where the Population stratification files are saved
# create arguments
#plinkfolder=$1
plinkfolder="/home/gloria/Downloads/plink_linux_x86_64_20230116/plink"
#genofolder=$2
genofolder="/home/gloria/Desktop/carpeta_Gloria/QC/Filter 1-6/G28_0.2_pihat"
#outfolder=$3
outfolder="/home/gloria/Desktop/carpeta_Gloria/proves"
#populationfolder=$4
populationfolder="//home/gloria/Desktop/carpeta_Gloria/QC/Population_stratificat
ion/proves_popes"
### Step 1 ###
## do it with the 1000 freely available genomic data to calculate the population stratification.
#wget
#ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20100804/ALL.2of4intersecti
on.20100804.genotypes.vcf.gz
# Convert vcf to Plink format.
/home/gloria/Downloads/plink_linux_x86_64_20230116/plink --vcf
/media/gloria/SeagateBasic/gloria/ALL.2of4intersection.20100804.genotypes.vcf.gz
--make-bed --out
/media/gloria/SeagateBasic/gloria/ALL.2of4intersection.20100804.genotypes
# However, for good practice, we will assign unique indentifiers to the SNPs
with a missing rs-identifier (i.e., the SNPs with ".").
/home/gloria/Downloads/plink_linux_x86_64_20230116/plink --bfile
/media/gloria/SeagateBasic/gloria/ALL.2of4intersection.20100804.genotypes
```

```
--set-missing-var-ids @:#[b37]\$1,\$2 --make-bed --out
/home/gloria/Desktop/carpeta_Gloria/QC/Population_stratification/ALL.2of4interse
ction.20100804.genotypes_no_missing_IDs
## QC on 1000 Genomes data.
# Remove variants based on missing genotype data.
$plinkfolder --bfile
/home/gloria/Desktop/carpeta_Gloria/QC/Population_stratification/ALL.2of4interse
ction.20100804.genotypes_no_missing_IDs --geno 0.2 --allow-no-sex --make-bed
--out $populationfolder/1kG_MDS
# Remove individuals based on missing genotype data.
$plinkfolder --bfile $populationfolder/1kG_MDS --mind 0.2 --allow-no-sex
--make-bed --out $populationfolder/1kG_MDS2
# Remove variants based on missing genotype data.
$plinkfolder --bfile $populationfolder/1kG_MDS2 --geno 0.02 --allow-no-sex
--make-bed --out $populationfolder/1kG_MDS3
# Remove individuals based on missing genotype data.
$plinkfolder --bfile $populationfolder/1kG_MDS3 --mind 0.02 --allow-no-sex
--make-bed --out $populationfolder/1kG_MDS4
# Remove variants based on MAF.
$plinkfolder --bfile $populationfolder/1kG_MDS4 --maf 0.05 --allow-no-sex
--make-bed --out $populationfolder/1kG_MDS5
# Extract the variants present in HapMap dataset from the 1000 genomes dataset.
awk '{print$2}' $populationfolder/G28_0.2_pihat.bim >
$populationfolder/HapMap_SNPs.txt
$plinkfolder --bfile $populationfolder/1kG_MDS5 --extract
$populationfolder/HapMap_SNPs.txt --make-bed --out $populationfolder/1kG_MDS6
# Extract the variants present in 1000 Genomes dataset from the HapMap dataset.
awk '{print$2}' $populationfolder/1kG_MDS6.bim >
$populationfolder/1kG_MDS6_SNPs.txt
$plinkfolder --bfile $populationfolder/G28_0.2_pihat --extract
```

```
$populationfolder/1kG_MDS6_SNPs.txt --recode --make-bed --out
$populationfolder/HapMap_MDS
#--recode pasa a formato .map
## The datasets must have the same build. Change the build 1000 Genomes data build.
awk '{print$2,$4}' $populationfolder/HapMap_MDS.map >
$populationfolder/buildhapmap.txt
# buildhapmap.txt contains one SNP-id and physical position per line.
$plinkfolder --bfile $populationfolder/1kG_MDS6 --update-map
$populationfolder/buildhapmap.txt --make-bed --out $populationfolder/1kG_MDS7
# 1kG_MDS7 and HapMap_MDS now have the same build.
## Merge the HapMap and 1000 Genomes data sets
# 1) Make sure the reference genome is similar in the HapMap and the 1000 Genomes Project datasets.
# 2) Resolve strand issues.
# 3) Remove the SNPs which after the previous two steps still differ between
datasets.
# 1) set reference genome
awk '{print$2,$5}'
/home/gloria/Desktop/carpeta_Gloria/proves/Population_stratification/proves_pope
s/1kG_MDS7.bim > $populationfolder/1kg_ref-list.txt
$plinkfolder --bfile $populationfolder/HapMap_MDS --reference-allele
$populationfolder/1kg_ref-list.txt --make-bed --out $populationfolder/HapMap-adj
# This command will generate some warnings for impossible A1 allele assignment.
# 2) Resolve strand issues.
# Check for potential strand issues.
awk '{print$2,$5,$6}' $populationfolder/1kG_MDS7.bim >
$populationfolder/1kG_MDS7_tmp
awk '{print$2,$5,$6}' $populationfolder/HapMap-adj.bim >
$populationfolder/HapMap-adj_tmp
sort $populationfolder/1kG_MDS7_tmp $populationfolder/HapMap-adj_tmp |uniq -u >
$populationfolder/all_differences.txt
```

```
# 80 differences between the files, some of these might be due to strand issues.
## Flip SNPs for resolving strand issues.
# Print SNP-identifier and remove duplicates.
awk '{print$1}' $populationfolder/all_differences.txt | sort -u >
$populationfolder/flip_list.txt
# Generates a file of 40 SNPs. These are the non-corresponding SNPs between the
two files.
# Flip the 40 non-corresponding SNPs.
$plinkfolder --bfile $populationfolder/HapMap-adj --flip
$populationfolder/flip_list.txt --reference-allele
$populationfolder/1kg_ref-list.txt --make-bed --out
$populationfolder/correct_hadmap
# Check for SNPs which are still problematic after they have been flipped.
awk '{print$2,$5,$6}' $populationfolder/correct_hadmap.bim >
$populationfolder/corrected_hadmap_tmp
sort $populationfolder/1kG_MDS7_tmp $populationfolder/correct_hadmap_tmp |uniq
-u > $populationfolder/uncorresponding_SNPs.txt
# This file demonstrates that there are 18 differences between the files.
# 3) Remove problematic SNPs from HapMap and 1000 Genomes.
awk '{print$1}' $populationfolder/uncorresponding_SNPs.txt | sort -u >
$populationfolder/SNPs_for_exlusion.txt
# The command above generates a list of the 9 SNPs which caused the 84
differences between the HapMap and the 1000 Genomes data sets after flipping and
setting of the reference genome.
# Remove the 9 problematic SNPs from both datasets.
$plinkfolder --bfile $populationfolder/corrected_hadmap --exclude
$populationfolder/SNPs_for_exlusion.txt --make-bed --out
$populationfolder/HapMap_MDS2
$plinkfolder --bfile $populationfolder/1kG_MDS7 --exclude
$populationfolder/SNPs_for_exlusion.txt --make-bed --out
```

```
$populationfolder/1KG_MDS8
# Merge HapMap with 1000 Genomes Data.
$plinkfolder --bfile $populationfolder/HapMap_MDS2 --bmerge
$populationfolder/1KG_MDS8.bed $populationfolder/1KG_MDS8.bim
$populationfolder/1KG_MDS8.fam --allow-no-sex --make-bed --out
$populationfolder/MDS_merge2
# Note, we are fully aware of the sample overlap between the HapMap and 1000
Genomes datasets. However, for the purpose of this tutorial this is not
important.
## Perform MDS on HapMap-CEU data anchored by 1000 Genomes data.
# Using a set of pruned SNPs
$plinkfolder --bfile $populationfolder/MDS_merge2 --extract
$outfolder/indepSNP.prune.in --genome --out $populationfolder/MDS_merge2
$plinkfolder --bfile $populationfolder/MDS_merge2 --read-genome
$populationfolder/MDS_merge2.genome --cluster --mds-plot 10 --out
$populationfolder/MDS_merge2
### MDS-plot
wget
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20100804/20100804.ALL.panel
# The file $populationfolder/20100804.ALL.panel contains population codes of the
individuals of 1000 genomes.
# Convert population codes into superpopulation codes (i.e., AFR,AMR,ASN, and
EUR).
awk '{print$1,$1,$2}' $populationfolder/20100804.ALL.panel >
$populationfolder/race_1KG.txt
sed 's/JPT/ASN/g' $populationfolder/race_1KG.txt>$populationfolder/race_2KG2.txt
sed 's/ASW/AFR/g'
$populationfolder/race_2KG2.txt>$populationfolder/race_1KG3.txt
sed 's/CEU/EUR/g'
$populationfolder/race_1KG3.txt>$populationfolder/race_1kG4.txt
```

```
sed 's/CHB/ASN/g'
$populationfolder/race_1kG4.txt>$populationfolder/race_1kG5.txt
sed 's/CHD/ASN/g'
$populationfolder/race_1kG5.txt>$populationfolder/race_1kG6.txt
sed 's/YRI/AFR/g'
$populationfolder/race_1kG6.txt>$populationfolder/race_1kG7.txt
sed 's/LWK/AFR/g'
$populationfolder/race_1kG7.txt>$populationfolder/race_1kG8.txt
sed 's/TSI/EUR/g'
$populationfolder/race_1kG8.txt>$populationfolder/race_1kG9.txt
sed 's/MXL/AMR/g'
$populationfolder/race_1kG9.txt>$populationfolder/race_1kG10.txt
sed 's/GBR/EUR/g'
$populationfolder/race_1kG10.txt>$populationfolder/race_1kG11.txt
sed 's/FIN/EUR/g'
$populationfolder/race_1kG11.txt>$populationfolder/race_1kG12.txt
sed 's/CHS/ASN/g'
$populationfolder/race_1kG12.txt>$populationfolder/race_1kG13.txt
sed 's/PUR/AMR/g'
$populationfolder/race_1kG13.txt>$populationfolder/race_1kG14.txt
# Create a racefile of your own data.
awk '{print$1,$2,"OWN"}'
$populationfolder/HapMap_MDS.fam>$populationfolder/racefile_own.txt
# Concatenate racefiles.
cat $populationfolder/race_1kG14.txt $populationfolder/racefile_own.txt | sed -e
'1i\FID IID race' > $populationfolder/racefile.txt
# Generate population stratification plot.
Rscript
/home/gloria/Desktop/carpeta_Gloria/proves/Population_stratification/r_scrip
## Exclude ethnic outliers.
```

```
# Select individuals in HapMap data below cut-off thresholds. The cut-off levels
are not fixed thresholds but have to be determined based on the visualization of
the first two dimensions. To exclude ethnic outliers, the thresholds need to be
set around the cluster of population of interest.
awk '{ if ($4 >-0.016) print $1,$2 }' $populationfolder/MDS_merge2.mds >
/home/gloria/Desktop/carpeta_Gloria/proves/Population_stratification/EUR_MDS_mer
ge2
# Extract these individuals in HapMap data.
$plinkfolder --bfile $populationfolder/G28_0.2_pihat --exclude
/home/gloria/Desktop/carpeta_Gloria/proves/Population_stratification/EUR_MDS_mer
ge2 --make-bed --out $populationfolder/HapMap_3_r3_13
# Note, since our HapMap data did include any ethnic outliers, no individuals
were removed at this step. However, if our data would have included individuals
outside of the thresholds we set, then these individuals would have been
removed.
## Create covariates based on MDS.
# Perform an MDS ONLY on HapMap data without ethnic outliers. The values of the
10 MDS dimensions are subsequently used as covariates in the association
analysis in the third tutorial.
$plinkfolder --bfile $populationfolder/HapMap_3_r3_13 --extract
$outfolder/indepSNP.prune.in --genome --out $populationfolder/HapMap_3_r3_13
$plinkfolder --bfile $populationfolder/HapMap_3_r3_13 --read-genome
$populationfolder/HapMap_3_r3_13.genome --cluster --mds-plot 10 --out
$populationfolder/HapMap_3_r3_13_mds
# Change the format of the .mds file into a plink covariate file.
awk '{print$1, $2, $4, $5, $6, $7, $8, $9, $10, $11, $12, $13}'
$populationfolder/HapMap_3_r3_13_mds.mds > $populationfolder/covar_mds.txt
```

**Quality control and filtering procedure. Results obtained in each filtering step.**

| INITIAL | N = 1251 individuals<br>(678 MALES, 561 FEMALES, 12 AMBIGUOUS) | 603132 variants loaded from .bim file.<br>TOTAL GENOTYPING RATE IS 0.987857 | |
|---|---|---|---|
| | **Filter 1** | **Filter 2** | **Filter 3** |
| | **Missignes of SNPS (0,2)** | **Sex discrepancy for Missignes of SNPS (0,02)** | **Minor allele frequency (0,05)** |
| | 6155 variants removed | 669 male | 669 male |
| | **Missignes of individuals (0,2)** | 546 females | 546 females |
| | 0 individuals removed | 547932 variants and 1215 people pass filters and QC. | 0 individuals removed |
| | **Missignes of SNPS (0,02)** | **Select autosomal SNPs only** | 287580 variants removed due to minor allele threshold(s) |
| | 49045 variants removed | 527343 out of 547932 variants loaded from .bim file. | |
| | **547932 variants and 1251 people pass filters and QC.** | 527343 variants remaining. | **239763** |
| | **Missignes of individuals (0,02)** | 25 individuals removed | |
| | 11 individuals removed | **527343 variants and 1215 people pass filters and QC.** | 239763 variants and 1215 people pass filters and QC. |
| **N** | **1240** | **1215** | **1215** |
| | **Filter 4** | **Filter 5** | **Filter 6** |
| | **Hardy-Weinberg equilibrium (1e-6)** | **Heterozigosity** | **Releatdness (pihat >0,2)** |
| | 239470 variants and 1215 people pass filters and QC. | 235894 variants remaining | 652 males |
| | 293 variants removed due to Hardy-Weinberg exact test | **Pruning complete: 168408 of 235894 variants removed**. | 536 females |
| | 239470 | 67486 variants and 1215 people pass filters and QC. | |
| | **Hardy-Weinberg equilibrium (1e-10)** | 67486 | |
| | 0 variants removed due to Hardy-Weinberg exact test | **Remove heterozygosity rate outliers.** | |
| | 239470 | 27 individuals removed (deviate +3SD) | 121 individuals removed with 0.2_low_call_rate_pihat |
| | 0 individuals removed | 239470 variants and 1188 people pass filters and QC. | 239470 variants and 1067 people pass filters and QC. |
| | 239470 variants and 1215 people pass filters and QC. | | |
| **N** | **1215** | **1188** | **1067** |
| | **Filter 7** | | |
| **N (FINAL)** | **239470 variants and 1048 people pass filters and QC.** | | |

**PRS-cs script:**

##python PRScs.py --ref_dir=PATH_TO_REFERENCE --bim_prefix=VALIDATION_BIM_PREFIX --sst_file=SUM_STATS_FILE --n_gwas=GWAS_SAMPLE_SIZE --out_dir=OUTPUT_DIR [--a=PARAM_A --b=PARAM_B --phi=PARAM_PHI --n_iter=MCMC_ITERATIONS --n_burnin=MCMC_BURNIN --thin=MCMC_THINNING_FACTOR --chrom=CHROM --beta_std=BETA_STD --seed=SEED]

##The test data contains GWAS summary statistics and a bim file for 1,000 SNPs on chromosome 22.An example to use the test data:

#python PRScs.py --ref_dir=path_to_ref/ldblk_1kg_eur --bim_prefix=path_to_bim/test --sst_file=path_to_sumstats/sumstats.txt --n_gwas=200000 --chrom=22 --phi=1e-2 --out_dir=path_to_output/eur

python PRScs.py --ref_dir=/home/gloria/Desktop/carpeta_Gloria/QC_G28/PRS_CS/ldblk_1kg_eur --bim_prefix=/home/gloria/Desktop/carpeta_Gloria/QC_G28/PRS_CS/PRScs-master/test_data/test --sst_file=/home/gloria/Desktop/carpeta_Gloria/QC_G28/PRS_CS/PRScs-master/test_data/sumstats.txt --n_gwas=200000 --chrom=22 --phi=1e-2 --out_dir=/home/gloria/Desktop/carpeta_Gloria/QC_G28/PRS_CS/data_test_EUR

sed -i 's/psy_//g' /home/gloria/Desktop/carpeta_Gloria/QC_G28/G28_CORRECT_PANSS.bim

#unify results in one document

cat '/home/gloria/Desktop/carpeta_Gloria/QC_G28/PRS_CS/G28_EUR_pst_eff_a1_b0.5_phiauto_chr22.txt' >> /home/gloria/Desktop/carpeta_Gloria/QC_G28/PRS_CS/all_files_prs

##Script to calculate the score bia plink, for patients

/home/gloria/Downloads/plink_linux_x86_64_20230116/plink --bfile '/home/gloria/Desktop/carpeta_Gloria/QC_G28/G28_CORRECT_PANSS' --score '/home/gloria/Desktop/carpeta_Gloria/QC_G28/PRS_CS/all_files_prs' 2 4 6

/home/gloria/Desktop/carpeta_Gloria/plink_linux_x86_64_20230116/plink --bfile '/home/gloria/Desktop/carpeta_Gloria/prs-cs_data/G28_CORRECT_PANSS' --score '/home/gloria/Desktop/carpeta_Gloria/prs-cs_data/concatenated_prs' 2 4 6

##for patients&controls

home/gloria/Desktop/carpeta_Gloria/plink_linux_x86_64_20230116/plink --bfile '/media/gloria/SeagateBasic/gloria/QC_G28/QC_G28/G28_PANSS_PATIENTS_CONTROLS' --score '/home/gloria/Desktop/carpeta_Gloria/prs-cs_data/concatenated_prs' 2 4 6

##plot for control&patients prs score, with R_STUDIO

data_all <- read.table("~/Desktop/carpeta_Gloria/prscs_data/results_prscs/data_all.profile", h=T)

> head(data_all)

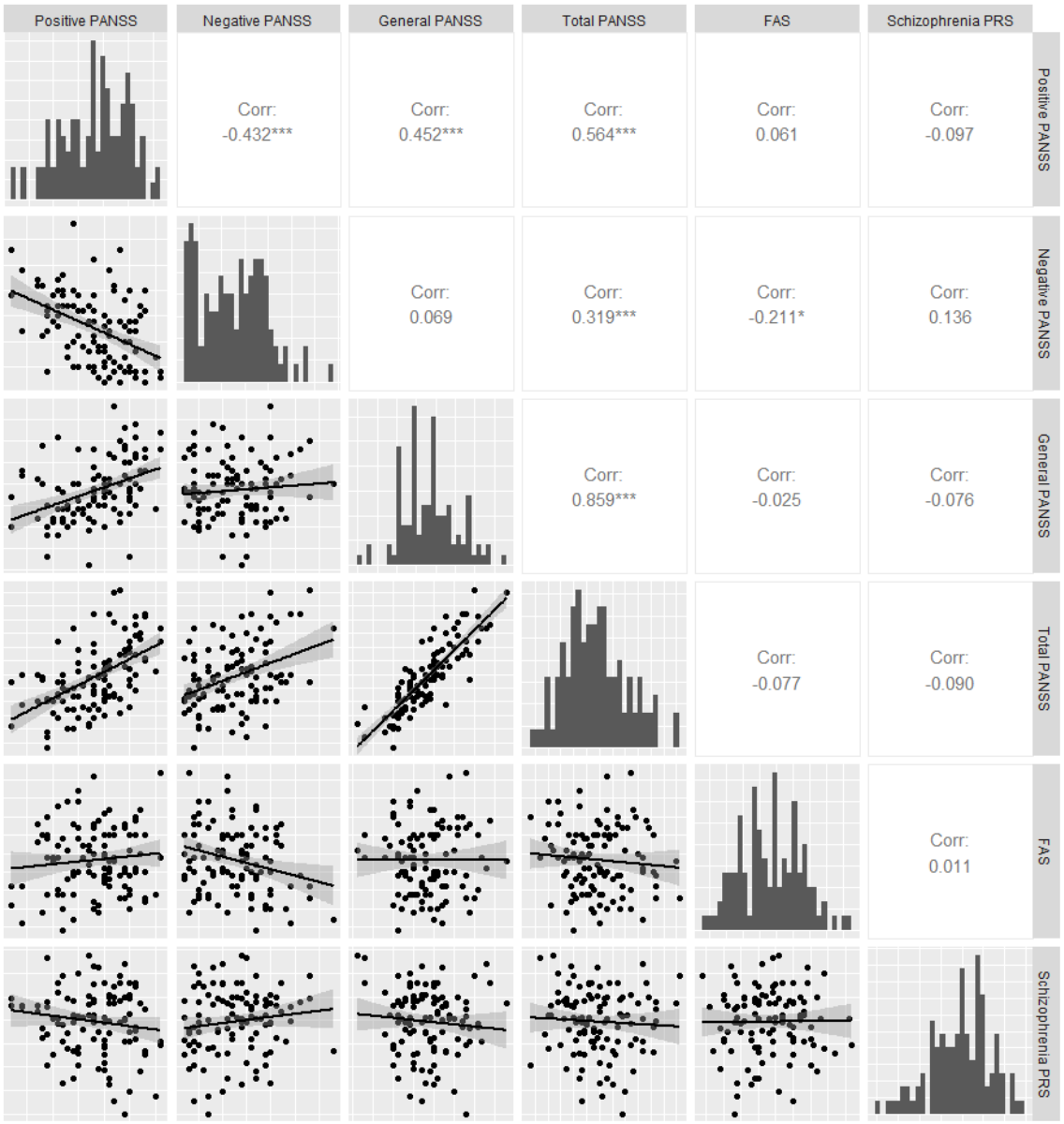
hist(data_all$SCORESUM)

> densityplot((data_all$SCORESUM))

plot(density(data_all$SCORESUM))

#*The density plot;  PRS results (patients and controls) were conducted using jamovi.

**Scatterplot matrix computed using R-studio, correlation between the PANSS and FAS scales, and the PRS calculated:**



## Script lineal model, PRS and PANSS and FAS scales

```
##Lienal model script

#Simplify names from each variable
PC1<- G28_ALL$C1
PC2<- G28_ALL$C2
PC3<- G28_ALL$C3
PC4<- G28_ALL$C4
PC5<- G28_ALL$C5
PC6<- G28_ALL$C6
PC7<- G28_ALL$C7
PC8<- G28_ALL$C8
PC9<- G28_ALL$C9
PC10<- G28_ALL$C10
Pos.PANSS<- G28_ALL$PANS_P
Neg.PANSS<- G28_ALL$PANSS_N
Gen.PANSS<- G28_ALL$PANSS_G
Tot.PANSS<- G28_ALL$PANSS_TOT
FAS<- G28_ALL$FAS
Scz.PRS<- G28_ALL$PRS_SCORE
Sex<- G28_ALL$SEX
Age<- G28_ALL$Age
ID_biobanc<- G28_ALL$ID_BIOBANC

#Generate a new data frame
df<- data.frame(PC1,PC2,PC3,PC4,PC5,PC6,PC7,PC8,PC9,PC10,Scz.PRS,Pos.PANSS,Neg.PANSS,Gen.PANSS,Tot.PANSS,FAS,Age,Sex,ID_bio
banc)
View(df)

#Calculate each lineal model, 5 in total

modelo_lineal_P_P <- lm(G28_ALL$PANS_P ~ G28_ALL$C1 + G28_ALL$C2 + G28_ALL$C3 + G28_ALL$C4 + G28_ALL$C5 + G28_ALL$C
6 + G28_ALL$C7 + G28_ALL$C8 + G28_ALL$C9 + G28_ALL$C10 + G28_ALL$Age + G28_ALL$SEX + G28_ALL$PRS_SCORE)
modelo_lineal_N_P <- lm(G28_ALL$PANSS_N ~ G28_ALL$C1 + G28_ALL$C2 + G28_ALL$C3 + G28_ALL$C4 + G28_ALL$C5 + G28_ALL
$C6 + G28_ALL$C7 + G28_ALL$C8 + G28_ALL$C9 + G28_ALL$C10 + G28_ALL$Age + G28_ALL$SEX + G28_ALL$PRS_SCORE)
modelo_lineal_G_P <- lm(G28_ALL$PANSS_G ~ G28_ALL$C1 + G28_ALL$C2 + G28_ALL$C3 + G28_ALL$C4 + G28_ALL$C5 + G28_ALL
$C6 + G28_ALL$C7 + G28_ALL$C8 + G28_ALL$C9 + G28_ALL$C10 + G28_ALL$Age + G28_ALL$SEX + G28_ALL$PRS_SCORE)
modelo_lineal_T_P <- lm(G28_ALL$PANSS_TOT ~ G28_ALL$C1 + G28_ALL$C2 + G28_ALL$C3 + G28_ALL$C4 + G28_ALL$C5 + G28_AL
L$C6 + G28_ALL$C7 + G28_ALL$C8 + G28_ALL$C9 + G28_ALL$C10 + G28_ALL$Age + G28_ALL$SEX + G28_ALL$PRS_SCORE)
modelo_lineal_FAS <- lm(G28_ALL$FAS ~ G28_ALL$C1 + G28_ALL$C2 + G28_ALL$C3 + G28_ALL$C4 + G28_ALL$C5 + G28_ALL$C6 +
G28_ALL$C7 + G28_ALL$C8 + G28_ALL$C9 + G28_ALL$C10 + G28_ALL$Age + G28_ALL$SEX + G28_ALL$PRS_SCORE)

summary(modelo_lineal_P_P)

Call:
lm(formula = G28_ALL$PANS_P ~ G28_ALL$C1 + G28_ALL$C2 + G28_ALL$C3 +
    G28_ALL$C4 + G28_ALL$C5 + G28_ALL$C6 + G28_ALL$C7 + G28_ALL$C8 +
    G28_ALL$C9 + G28_ALL$C10 + G28_ALL$Age + G28_ALL$SEX + G28_ALL$PRS_SCORE)

Residuals:
    Min      1Q  Median      3Q     Max
-13.9524  -6.2099  -0.3894   6.3072  17.8970

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      19.59577    3.28523   5.965 1.26e-08 ***
G28_ALL$C1      -76.30787  136.51671  -0.559    0.577
G28_ALL$C2      -60.06431  147.30795  -0.408    0.684
G28_ALL$C3       21.90152   93.91274   0.233    0.816
G28_ALL$C4     -147.68838  116.43712  -1.268    0.206
G28_ALL$C5       30.86092  108.65825   0.284    0.777
G28_ALL$C6      -61.39119  109.42833  -0.561    0.575
G28_ALL$C7     -110.23646   93.99532  -1.173    0.242
G28_ALL$C8      137.72338   99.19144   1.388    0.167
G28_ALL$C9      112.30481   92.55579   1.213    0.227
G28_ALL$C10     -49.80130   94.58777  -0.527    0.599
G28_ALL$Age       0.06470    0.04898   1.321    0.188
G28_ALL$SEX      -0.65240    1.33179  -0.490    0.625
G28_ALL$PRS_SCORE  -23.65798   22.25303  -1.063    0.289
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.028 on 181 degrees of freedom
Multiple R-squared:  0.05348, Adjusted R-squared:  -0.0145
F-statistic: 0.7867 on 13 and 181 DF,  p-value: 0.6736
```

```
summary(modelo_lineal_N_P)

Call:
lm(formula = G28_ALL$PANSS_N ~ G28_ALL$C1 + G28_ALL$C2 + G28_ALL$C3 +
    G28_ALL$C4 + G28_ALL$C5 + G28_ALL$C6 + G28_ALL$C7 + G28_ALL$C8 +
    G28_ALL$C9 + G28_ALL$C10 + G28_ALL$Age + G28_ALL$SEX + G28_ALL$PRS_SCORE)

Residuals:
    Min      1Q  Median      3Q     Max
-15.0364  -4.7235   0.7422   4.7892  22.3185

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.26479    2.78891   1.888   0.0607 .
G28_ALL$C1       91.27785  115.89233   0.788   0.4320
G28_ALL$C2     -126.61709  125.05327  -1.013   0.3126
G28_ALL$C3     -156.86827   79.72479  -1.968   0.0506 .
G28_ALL$C4       47.45615   98.84628   0.480   0.6317
G28_ALL$C5       71.29089   92.24261   0.773   0.4406
G28_ALL$C6       22.82724   92.89635   0.246   0.8062
G28_ALL$C7       87.43415   79.79489   1.096   0.2747
G28_ALL$C8      143.20550   84.20601   1.701   0.0907 .
G28_ALL$C9      -40.86163   78.57284  -0.520   0.6037
G28_ALL$C10       8.00787   80.29784   0.100   0.9207
G28_ALL$Age       0.21620    0.04158   5.200 5.36e-07 ***
G28_ALL$SEX       2.88576    1.13059   2.552   0.0115 *
G28_ALL$PRS_SCORE  29.74612   18.89113   1.575   0.1171
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.815 on 181 degrees of freedom
Multiple R-squared:  0.203,    Adjusted R-squared:  0.1457
F-statistic: 3.546 on 13 and 181 DF,  p-value: 5.889e-05


summary(modelo_lineal_G_P)

Call:
lm(formula = G28_ALL$PANSS_G ~ G28_ALL$C1 + G28_ALL$C2 + G28_ALL$C3 +
    G28_ALL$C4 + G28_ALL$C5 + G28_ALL$C6 + G28_ALL$C7 + G28_ALL$C8 +
    G28_ALL$C9 + G28_ALL$C10 + G28_ALL$Age + G28_ALL$SEX + G28_ALL$PRS_SCORE)

Residuals:
   Min     1Q Median     3Q    Max
-23.088  -6.352   0.140   6.165  45.805

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      39.00733    3.98569   9.787   <2e-16 ***
G28_ALL$C1     -108.10548  165.62418  -0.653   0.5148
G28_ALL$C2      -71.93699  178.71629  -0.403   0.6878
G28_ALL$C3       36.71926  113.93638   0.322   0.7476
G28_ALL$C4      -35.02751  141.26332  -0.248   0.8044
G28_ALL$C5      126.05171  131.82587   0.956   0.3402
G28_ALL$C6      -91.01808  132.76014  -0.686   0.4939
G28_ALL$C7     -216.97608  114.03657  -1.903   0.0587 .
G28_ALL$C8      192.92797  120.34060   1.603   0.1106
G28_ALL$C9       95.53652  112.29012   0.851   0.3960
G28_ALL$C10    -115.99831  114.75535  -1.011   0.3134
G28_ALL$Age      -0.01057    0.05942  -0.178   0.8590
G28_ALL$SEX      -1.67598    1.61575  -1.037   0.3010
G28_ALL$PRS_SCORE  -29.14127   26.99772  -1.079   0.2818
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.74 on 181 degrees of freedom
Multiple R-squared:  0.06501, Adjusted R-squared:  -0.002142
F-statistic: 0.9681 on 13 and 181 DF,  p-value: 0.4846
```

```
summary(modelo_lineal_T_P)

Call:
lm(formula = G28_ALL$PANSS_TOT ~ G28_ALL$C1 + G28_ALL$C2 + G28_ALL$C3 +
    G28_ALL$C4 + G28_ALL$C5 + G28_ALL$C6 + G28_ALL$C7 + G28_ALL$C8 +
    G28_ALL$C9 + G28_ALL$C10 + G28_ALL$Age + G28_ALL$SEX + G28_ALL$PRS_SCORE)

Residuals:
    Min      1Q  Median      3Q     Max
-45.454 -11.278  -0.987  10.984  73.876

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        63.8679     6.8376   9.341  < 2e-16 ***
G28_ALL$C1        -93.1355   284.1344  -0.328  0.74345
G28_ALL$C2       -258.6184   306.5944  -0.844  0.40005
G28_ALL$C3        -98.2475   195.4621  -0.503  0.61583
G28_ALL$C4       -135.2597   242.3425  -0.558  0.57744
G28_ALL$C5        228.2035   226.1522   1.009  0.31429
G28_ALL$C6       -129.5820   227.7549  -0.569  0.57009
G28_ALL$C7       -239.7784   195.6340  -1.226  0.22192
G28_ALL$C8        473.8569   206.4487   2.295  0.02286 *
G28_ALL$C9        166.9797   192.6379   0.867  0.38720
G28_ALL$C10      -157.7917   196.8671  -0.802  0.42388
G28_ALL$Age         0.2703     0.1019   2.652  0.00872 **
G28_ALL$SEX         0.5574     2.7719   0.201  0.84086
G28_ALL$PRS_SCORE  -23.0531    46.3156  -0.498  0.61927
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.71 on 181 degrees of freedom
Multiple R-squared:  0.09033, Adjusted R-squared:  0.025
F-statistic: 1.383 on 13 and 181 DF,  p-value: 0.1713

summary(modelo_lineal_FAS)

Call:
lm(formula = G28_ALL$FAS ~ G28_ALL$C1 + G28_ALL$C2 + G28_ALL$C3 +
    G28_ALL$C4 + G28_ALL$C5 + G28_ALL$C6 + G28_ALL$C7 + G28_ALL$C8 +
    G28_ALL$C9 + G28_ALL$C10 + G28_ALL$Age + G28_ALL$SEX + G28_ALL$PRS_SCORE)

Residuals:
     Min       1Q   Median       3Q      Max
-20.4374  -6.7845  -0.7653   6.7865  21.3113

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       38.28404    5.77697   6.627 1.55e-09 ***
G28_ALL$C1       253.95613  231.91132   1.095   0.2760
G28_ALL$C2      -367.44309  228.52481  -1.608   0.1109
G28_ALL$C3        78.67170  155.59444   0.506   0.6142
G28_ALL$C4       -28.53625  192.38110  -0.148   0.8824
G28_ALL$C5      -326.61379  183.69914  -1.778   0.0783 .
G28_ALL$C6         9.94069  182.38735   0.055   0.9566
G28_ALL$C7       111.90304  165.80819   0.675   0.5012
G28_ALL$C8       -31.28090  171.37998  -0.183   0.8555
G28_ALL$C9      -153.68696  156.42903  -0.982   0.3281
G28_ALL$C10      -30.41631  152.90312  -0.199   0.8427
G28_ALL$Age       -0.21371    0.09949  -2.148   0.0340 *
G28_ALL$SEX       -1.53070    2.19807  -0.696   0.4877
G28_ALL$PRS_SCORE  30.37923   34.89992   0.870   0.3861
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.29 on 104 degrees of freedom
  (77 observations deleted due to missingness)
Multiple R-squared:  0.1088,   Adjusted R-squared:  -0.002587
F-statistic: 0.9768 on 13 and 104 DF,  p-value: 0.4791
```