Genetic Analysis of Patients with Bipolar Disorder

Alberto Gomes Pinheira
Mestrado Integrado em Engenharia de Redes e Sistemas Informáticos
Departamento de Ciência de Computadores
2019

Orientador
Inês de Castro Dutra, Professora Auxiliar, Faculdade de Ciências da Universidade do Porto

Co-Orientador
Rodrigo Dias, Investigador, Faculdade de Medicina da Universidade de São Paulo
Camila Nascimento, Investigadora, Faculdade de Medicina da Universidade de São Paulo

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ______/______/_________

Abstract

Bipolar Disorder (BD) is a disease that causes unusual shifts in mood, energy, activity levels and the ability to carry out daily tasks. BD is a complex disorder in which several factors may play a role. It is clinically heterogeneous and we still do not have bio-markers to help in the diagnostic. It is heritable, although not monogenic. Single-Nucleotide Polymorphism (SNP)s were found, but were not able to explain all cases.

In fact, a recent study using questionnaires to assess mood disorders has shown that the positive predictive value ranges from 50% to 55%, with the sensitivity value being higher in bipolar type I (85.7%), than in bipolar type II (72.4%), or cyclothymia/not otherwise specified bipolar disorder (59.3%). Therefore there is quite a lot of room for improvements.

Few works have applied machine learning techniques to study bipolar disorder. They apply those techniques on small scaled clinical data, genetic data, Magnetic Resonance Imaging (MRI), etc. Moreover, the diagnosis of bipolar disorder is still based on behavioural symptoms, which can lead to a misdiagnosis and imply a wrong or an ineffective treatment.

We studied a Genome-Wide Association (GWA) data from the Wellcome Trust Case Control Consortium (WTCCC) and found a higher prevalence of patients that are homozygous compared with heterozygous when analyzing different SNPs in genes previously associated with bipolar disorder. Our results indicate that there is a group of patients that present pairs or triples of genotypes while others have the presence of only one.

We applied association rules to find combination of genotypes that can possibly lead to uncovering subgroups of patients with different genetic conditions which may be candidates to better diagnosis and treatment.

Our findings may shed the light on better classification of bipolar disorder patients allowing for representing their clinical heterogeneity, ultimately opening a new possibility for distinct forms of treatment. Additional clinical data and larger datasets are encouraged to reinforce our findings.

Resumo

O transtorno bipolar é uma desordem que causa mudanças incomuns no humor, energia, níveis de atividade e capacidade de realizar tarefas diárias. É um distúrbio mental complexo em que vários factores podem desempenhar um papel. É clinicamente heterogêneo e ainda não temos bio-marcadores para ajudar no diagnóstico. É hereditário, embora não seja monogênico. SNPs foram encontrados, mas não foram capazes de explicar todos os casos.

De facto, um estudo recente que usa questionários para avaliar transtornos do humor revela que o valor preditivo positivo varia entre os 50% e os 55%, sendo que o valor de sensibilidade é maior no transtorno bipolar do tipo I (85.7%), do que no transtorno bipolar do tipo II (72.4%), ou ciclotimia / transtorno bipolar não especificado (59.3%). Portanto há espaço para melhorias.

Poucos estudos aplicaram técnicas de *Machine Learning* para diagnosticar o transtorno bipolar. Alguns deles aplicaram estas técnicas em dados clínicos de pequena dimensão, em dados genéticos, ressonâncias magnéticas, entre outros. No entanto, o diagnóstico do transtorno bipolar ainda é baseado em sintomas comportamentais, o que, muitas vezes, pode levar a um mal diagnóstico e implicar a aplicação de um tratamento errado ou ineficaz.

Estudamos dados genéticos, mais concretamente, *Genome-Wide Association (GWA)*, pertencentes ao *Wellcome Trust Case Control Consortium (WTCCC)*, e observamos uma maior concentração de pacientes homozigóticos do que heterozigóticos, quando analisamos diferentes *Single-Nucleotide Polymorphism (SNP)s* em genes previamente associados com o transtorno bipolar. Resultados mostram que existe um grupo de pacientes que manifesta pares ou triplos de genótipos enquanto outros só têm a presença de um.

Aplicamos regras de associação para encontrar combinações de genótipos que possivelmente pudessem evidenciar subgrupos de pacientes com diferentes condições genéticas. Estes subgrupos poderiam ser candidatos a um melhor diagnóstico e tratamento.

Os nossos resultados, desta forma, indicam que uma melhor classificação dos pacientes com transtorno bipolar poderia ser conseguida, permitindo representar uma heterogeneidade clínica, abrindo uma nova possibilidade para formas distintas de tratamento. Dados clínicos e de maior dimensão são necessários para reforçar estes resultados.

Acknowledgments

First, I want to thank Professor Dr. Inês de Castro Dutra for advising me during the process of making of this thesis and the execution of this project. She was always there to help even through the hard and tough moments and those moments that I wanted to give up. A big thank you for picking me for executing this project and it was a great pleasure working with her.

To my follow co-advisors Rodrigo Dias and Camila Nascimento, thank you for helping me in this area, especially expanding my knowledge in this area of mental disorders and genetics. And also a big thank you for the help on the paper as well.

To my parents Manuel Pinheira and Maria Gomes, and my sister Jennifer Pinheira, I want to give a huge thank you for helping me during the journey of the academic world, which during the beginning it wasn't easy, but you never gave up on me and gave me motivation to keep moving on, and that helped me finishing this course. With all pleasure I dedicate this project to all of you as in another chapter of my life is completed.

To my academic godchildren Ana Coutinho and José Pedro, I never thought that I was going to have godchildren at all, but life chose otherwise, but I'm very grateful to meet both of you, because you helped me a lot and we did the curricular units together and we all succeeded together and without both of you I probably wouldn't be writing this at the moment.

To my friend Inês Martins, thank you so much for helping me planning my ideas for the thesis, and also for the journey that we faced together, and helping me to pass by my problems.

To my academic godmother Ana Martins, grandfather Artur Peniche, and my academic aunt Ana Germano, I want to thank all of you for supporting me during my academic journey which all of you know that it's wasn't an easy beginning. But thanks to all of you I grew up and lose my fear of the pressure I was going through. And now I'm a better man thanks to that. A big thank you to all.

And finally to my friends Pedro Cunha and Ruben Ferreira, Bruno Cabral and Ricardo Santos, a big thank you for all the support that you gave me and also without you I wouldn't be writing this. And through the bad moments, you were always there for me. Thank you for the help that you gave and the motivation to keep me going in this journey.

"This study makes use of data generated by the Wellcome Trust Case-Control Consortium.

A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project that generated this data was provided by the Wellcome Trust under award 076113 and 085475".

This work resulted in a publication in the 16th International Conference on Computational Intelligence methods for Bioinformatics and Biostatistics [23].

The code developed for this work is available in a GitHub repository, whose link is: https://github.com/WarriorofNothing/Tese-code, which can be accessed under request.

I dedicate to my parents, sister and my friends.

Contents

List of Tables

List of Figures

Acronyms

BD	Bipolar Disorder
DNA	Deoxyribonucleic acid
DALY	disability-adjusted life year
GRS	Genetic Risk Score
GWA	Genome-Wide Association
GBD	Global Burden of Diseases
IBS	Irritable Bowel Syndrome
KNN	K-nearest neighbours
MRI	Magnetic Resonance Imaging
SNP	Single-Nucleotide Polymorphism
SVM	Support Vector Machine
WTCCC	Wellcome Trust Case Control Consortium
YLD	years lost due to disability

Chapter 1

Introduction

1.1 Medical context

Bipolar Disorder (BD) is a brain disorder that causes unusual shifts in mood, energy, activity levels, and the ability to carry out daily tasks [3]. There are three types of this disorder, all of them have changes in mood, energy and activity levels. These range from periods of extremely high moments,when the level of energy is high, known as manic episode (elated), to very sad or hopeless periods, known as depressive episodes. Less severe manic periods are classified as hypomanic episodes. They are classified as Bipolar Type I, Type II and Cyclothymic Disorder [3].

BD affects approximately 2% of the world's population [40]. According to the World Health Organization, bipolar disorder is among the 10 leading causes of disability-adjusted life year (DALY) in people at a young age. Studies show that patients with bipolar disorder may commit suicide, especially men, (7.8% men vs 4.9% women) and the life time of a patient with bipolar disorder is reduced by 9 years [40].

The diagnostic system, according to Nunes et al. [30], is still based on the description of behavioral symptoms, resulting into delayed or inaccurate diagnosis leading to a delay or an ineffective treatment [30]. This ineffective treatment can lead to an average delay of 5.8 years between the initial symptoms and the formal diagnosis [40].

Questionnaires are usually used by psychiatrists to assess mood disorders. Some examples are the Structured Clinical Interview for DSM-IV-TR (SCID). According to a recent study [14], the Mood Disorder Questionnaire (MDQ) had satisfactory sensitivity (75%) and specificity (74%), but a low positive predictive value (55%).

According to the Global Burden of Diseases 2013 [16], bipolar disorder is a mental illness that affects approximately 48.8 million people (between 43.5 and 54.4 million) with peak of incidence between the ages of 20 and 34 (early adulthood). BD is the 54th leading cause of global days lived with disabilities (explaining 0.4% of DALYs) and also the 16th leading cause of years lived with disabilities (this explains the 1.3% total years lost due to disability (YLD)s). Among the mental and substance use disorders assessed in Global Burden of Diseases (GBD) 2013, bipolar disorder was the fifth leading cause of DALYs (after major depressive disorder, anxiety disorder, schizophrenia and alcohol use disorders, respectively). It accounted for 5.7% of the burden due to mental and substance use disorders [31] [8].

The main goal of this work is classifying the groups of patients within the bipolar group. We use Single-Nucleotide Polymorphism (SNP)s as the patient's characteristics to try to classify the bipolar patients in subgroups. This is important because it can help devising better treatment and diagnose of Bipolar Disorder.

Various works study the brain behaviour by processing brain images such as Magnetic Resonance Imaging (MRI). Others apply questionnaires to access mood disorders, especially psychiatrists. Others researched SNPs that are related to BD. Other works applied machine learning methods, that involve studying families that are affected or not affected with BD, or systematic reviews of the impact of machine learning techniques in the study of BD or finding pattern extractions by finding clinical-genetic similarities between subgroups of BD samples. It has been shown that machine learning algorithms may help researchers and doctors to get a better diagnosis, prognosis and also a personalized treatment.

1.2 Goals

Genetic information can play an important role on distinguishing different mental disorders. Various works have studied genetic patterns in order to identify the interaction among diseases. The main goal of this work is to characterize BD patients according to their genetic patterns. For example, by finding subgroups of patients with particular genetic patterns. Our main tool to approach this problem is the use of machine learning algorithms. Our first challenge is to handle more than 30 GBytes of genetic data. Our dataset consists of 1998 patients, but their genetic information consists of genetic millions of mutations that can happen in each one of the 23 chromosomes.

1.3 Motivation

Few works have applied machine learning techniques to diagnose bipolar disorder. They apply those techniques on small scaled clinical data, genetic data, MRIs, etc. Moreover, the diagnosis of bipolar disorder is still based on behavioural symptoms, which can lead to a misdiagnosis and imply a wrong or an ineffective treatment.

1.4 Organization

In Chapter 1, we will introduce basic concepts on bipolar and other mental disorders as well as basic concepts of human genetics and their role on metabolism and mutations. Far from being complete, this chapter summarizes topics that are needed to the understanding of the remaining chapters. Next, Chapter 2 classifies the most relevant works in the area of BD and also machine learning applied to the analysis and mining of BD patient data. Chapter 3 presents the dataset we use, discusses the methodology designed to handle this data and describes the experimental methods. Chapter 4 presents our results and highlights our main contribution. Finally, in the last chapter, we draw conclusions and present perspectives of future work.

Chapter 2

Basic Concepts

2.1 Human Genetics

2.1.1 Genes, alleles, genotypes

Genes are made up of Deoxyribonucleic acid (DNA) sequences. Every human has two copies of each gene, each one received by each parent [19].

Alleles are forms of the same gene but have small differences in the DNA sequence. These differences have a contribution on the physical features of each human [19].

A genotype is usually referred to a genetic make up of an organism, meaning that it describes an organism's complete set of genes [20]. For example, consider a person's blood type that was inherited from their parents. We know that the A, B and AB alleles are dominant (manifests absolutely in the person) in relation to the O alleles that are recessive (only manifests if both alleles are the same). Let's assume that person's father has the O type (the i^i allele) and his or her mother has the A type, but she has two different alleles (the i^a allele and the i^i allele). When we combine both father's and mother's alleles, we create a mendelian Table (2.1) (table that combines all alleles) with all possibilities. The columns represent the mother's alleles and the rows represent the father's alleles.

	i^a	i^i
i^i	i^i i^a	i^i i^i
i^i	i^i i^a	i^i i^i

Table 2.1: Allele combination of the type of blood.

From Table 2.1 our case study has a 50% chance of having the A type blood and 50% chance of having the O type blood. We also observe that the person has different alleles for the type A blood. This is a heterozygous case, which consists of having two different versions of the allele, one obtained from one parent, and the other from the other parent [21]. There is also a case of having a copy of the same allele, in this case we are dealing with a homozygous patient [21].

2.1.1.1 Single-Nucleotide Polymorphism

The modern unit of genetic variation is the Single-Nucleotide Polymorphism (SNP). SNPs are single base-pair changes in the DNA sequence that occur with high frequency in the human genome [26]. There are four nucleotides possible in a DNA sequence: Adenine (A), Thymine (T), Guanine (G), and Cytosine (C). SNPs normally occur throughout a person's DNA. Figure 2.1 shows that the Guanine and Cytosine were replaced by Adenine and Thymine in that stretch of the DNA sequence [18].

Although SNPs may not cause a disorder, they may be associated with certain disease risks. These associations allow scientists and doctors to look for SNPs in order to evaluate an individual's genetic predisposition to develop a disease [18].

Figure 2.1 shows an example of a mutation in a SNP.

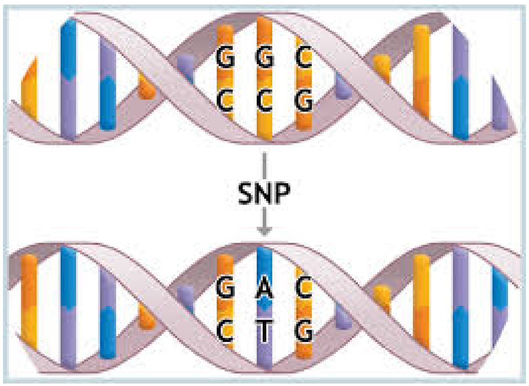


Figure 2.1: SNP example in a DNA sequence. [2]

SNPs typically have two alleles,meaning within a population there are two commonly occurring base-pair possibilities for a SNP location. The frequency of a SNP is given in terms of the minor allele frequency or the frequency of the less common allele. For example, a SNP with a minor allele (G) frequency of 0.40 implies that 40% of a population has the G allele versus the more common allele (the major allele), which is found in 60% of the population. These frequencies are important because risk scores for genes associated with some disease are derived from those.

2.2 Bipolar Disorder

As previously mentioned in the introduction, bipolar disorder is a disorder in the brain that causes unusual shifts in mood, energy, activity levels, and ability to carry out daily tasks. These range from periods of extremely high moments, when the level of energy is high, elated, known as manic episodes, to very sad or hopeless periods, known as depressive episodes. Less severe manic periods are classified as hypomanic episodes [3]. There are three types of this disorder, which will be described next.

2.2.1 Types

The first type is Bipolar Type I, which consists of manic episodes that last at least for 7 days of manic symptoms that are so severe that the patient needs immediate treatment at the hospital. Sometimes, depressive episodes happen, normally lasting for 2 weeks. It is also possible that episodes of depression with mixed features (depression and manic episodes occurring at the same time) may occur as well [3].

Figure 2.2 shows the comparison between the moods of a normal person and a person with bipolar type I.

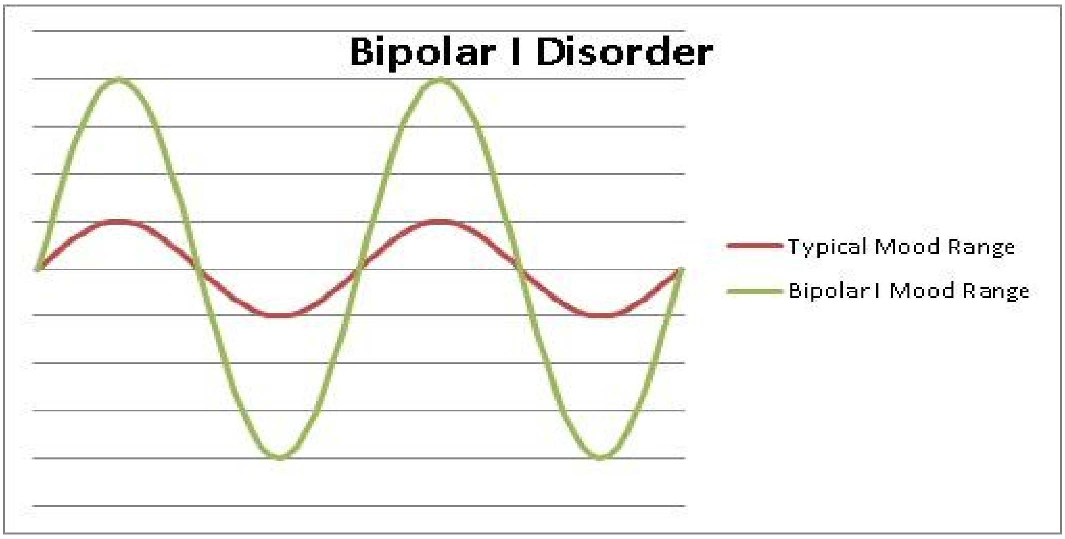


Figure 2.2: The comparison between moods of a normal person and Bipolar Disorder I [1].

Next one is Bipolar type II, which is defined by a pattern of depressive episodes and hypomanic episodes, but these episodes are not as impactful as in the type I disorder [3].

Figure 2.3 shows the comparison between the moods of a normal person and a person with bipolar type II.

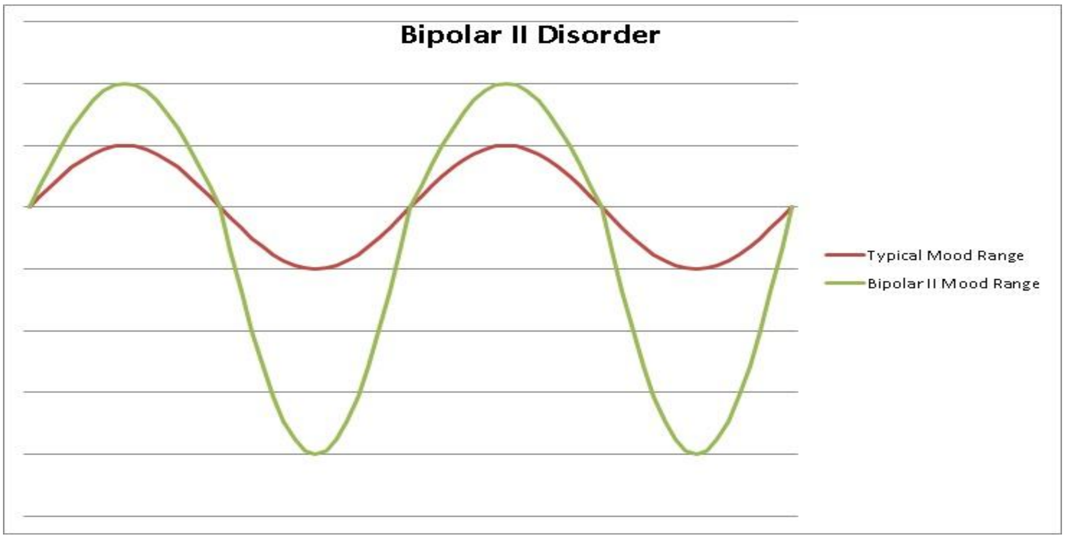


Figure 2.3: The comparison between moods of a normal person and Bipolar Disorder II [1].

The last one, known as the Cyclothymic Disorder, also denominated by cyclothymia, consists of numerous periods of hypomanic symptoms as well as depressive episodes, which could last for at least 2 years in adults, 1 year in children and adolescents. However, the symptoms there do not meet up the diagnosis requirements for a hypomanic or a depressive episode [3].

Figure 2.4 shows the comparison between the moods of a normal person and a person with cyclothymia.

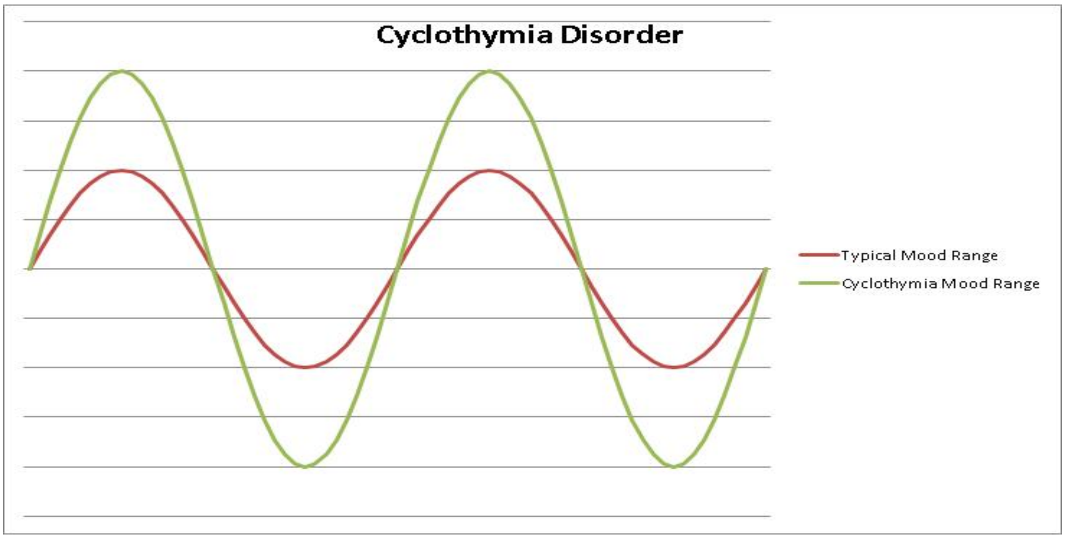


Figure 2.4: The comparison between moods of a normal person and cyclothymia [1].

The differences we observe in these figures are mostly related with the amplitude of the depressive or manic symptoms when compared with a typical mood range. While cyclothymia has equally distributed mood ranges along the time, bipolar disorder II has higher amplitudes for depression. Bipolar disorder I, as cyclothymia, has also equally distributed mood ranges along the time, but amplitudes are higher.

2.2.2 Risk factors

As in other psychiatric disorders, there are risks associated with Bipolar Disorder (BD). Some of them involve multiple risk factors, genetics, prenatal and perinatal factors, childhood, psychological stressors, substance misuse and medical comorbidities. Following up are some examples in each category [45].

1. Genetic:
   - Familial genetic risk;
   - Multiple SNPs.
2. Prenatal and perinatal:
   - Perinatal infections;
   - Obstetric complications;
   - Pregnancy, and post partum period increased the risk for BD [49].
3. Childhood
   - Childhood trauma;
   - Childhood trauma and outcomes in bipolar;
   - Childhood trauma and psychosis in bipolar.
4. Psychological stresses;
   - Life events prior to relapse;
   - Life events and first admission for mania;
   - Life events and mood episodes.
5. Substance misuse
   - Cannabis;
   - Opioids, tranquilizers, stimulants and sedatives;
   - Substance use disorders.
6. Medical comorbidities
   - Clinical risk factors;
   - Irritable bowel syndrome (IBS);
   - Asthma,
   - Obesity.

2.2.3 Genetic Risk Score

Complex diseases such as BD have numerous, well-established risk loci, and likely harbor many genetic determinants with effects too small to be detected at genome-wide levels of statistical significance. A simple and intuitive approach for converting genetic data to a predictive measure of disease susceptibility is to aggregate the risk effects of these loci into a single Genetic Risk Score (GRS). GRS is an estimate of the cumulative contribution of genetic factors to a specific outcome of interest in a person considering the reported risk alleles [27].

There are various softwares to calculate GRS, such as PredictABEL, PRSice, GPRS, PLINK [27]. The dataset we use has the GRS already calculated by the Chiamo software (https://mathgen.stats.ox.ac.uk/genetics_software/chiamo/chiamo.html). Chiamo uses allele frequencies, risk allele, minor allele, genotypic p-values and Bayesian factors, per SNP, to assess risk scores. Unfortunately, not many details are found about the algorithm used by Chiamo. A closer look at the source code could provide a better insight, however this was not the focus of this work.

2.2.4 Data Mining

According to Hand *et al.*[38], Data Mining is the analysis of observational data sets of large size, to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the owner of the data. These relationships and summaries derived through a data mining exercise are often referred to models or patterns. According to some authors, the task of generating models or patterns is also called Machine Learning.

2.2.5 Machine Learning

According to the literature, machine learning can be defined in several ways. According to Mitchell [41], Machine Learning is defined as follows: “A computer program is said to learn from experience E with respect to some class of tasks T and performance P, if its performance at tasks T, measured as P, improves with experience E.”

2.2.5.1 Algorithms of Machine Learning

In this section we focus on the most relevant algorithms mentioned and used in the area of bipolar disorder. We will separate the cases of the algorithms that used supervised and unsupervised learning.

Supervised Learning According to Russell and Norvig [46], supervised learning "is the agent observes some example, input–output pairs and learns a function that maps from input to output". The following methods are examples of supervised learning.

K-Nearest Neighbours This algorithm is used for predicting test samples according to a training model, which finds nearest neighbors to test the samples [47].

This method of classification is one of the simplest methods in machine learning. It's essentially classification by finding the most similar data points in the training data, and making informed guesses based on their classifications. This algorithm has seen wide applications in many domains, for example: recommendation systems or anomaly detection [13].

Unlike most other methods of classification, K-nearest neighbours (KNN) has no explicit training phase before classification. Substituting this, any attempts to generalize the data is made alongside classification. This means that we can start classifying immediately once we have our data. Although this is a good thing, there are some problems regarding this type of algorithm such as we have to keep the entire training set in memory unless we apply some type of reduction to the dataset. Performing classifications can be computationally expensive as the algorithm parses through all data points for each classification. Although some works in the literature use this algorithm in the area of bipolar disorder, we are not going to use it in our study because it works best in a small dataset, and we also don't have a class variable (labeled data) [13].

Decision Tree Decision tree learning is a method for approximating discrete-valued functions that is robust to noisy data and is competent to learn disjunctive expressions [41].

This algorithm is commonly used in Data Mining studies. It's been used as a tree-shaped model, which has a cleaner representation of results compared with other classification methods. The goal here is to create a model that classifies the target attribute based on input variables of the training set [47].

In terms of representation, decision trees classify instances by sorting them down the tree from the root to some node, which produces the classification of the instance. Each node in the tree indicates a test of some attribute of the instance, and each branch descending from the node in question corresponds to one of the possible values for this attribute [41]. Figure 2.5 shows an example of a decision tree, where the rectangles (nodes) are variables and the branches are the variables' values. Leaves are the final classification.

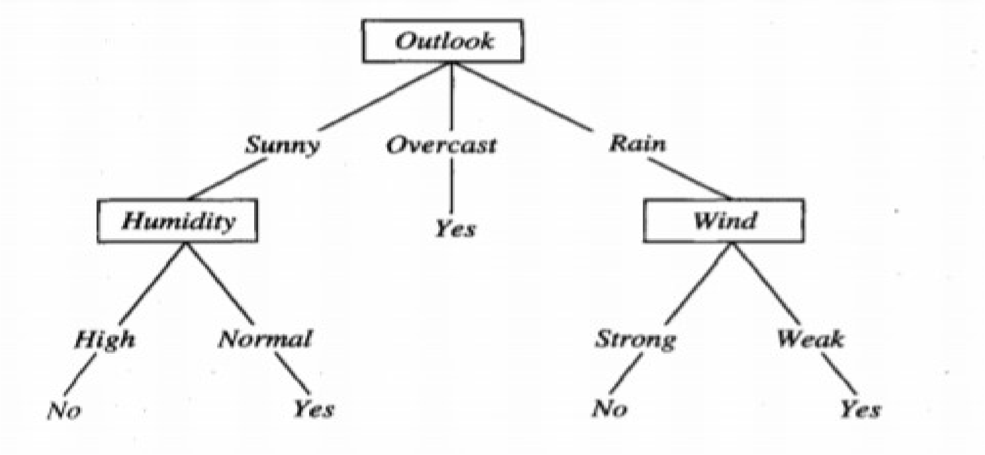
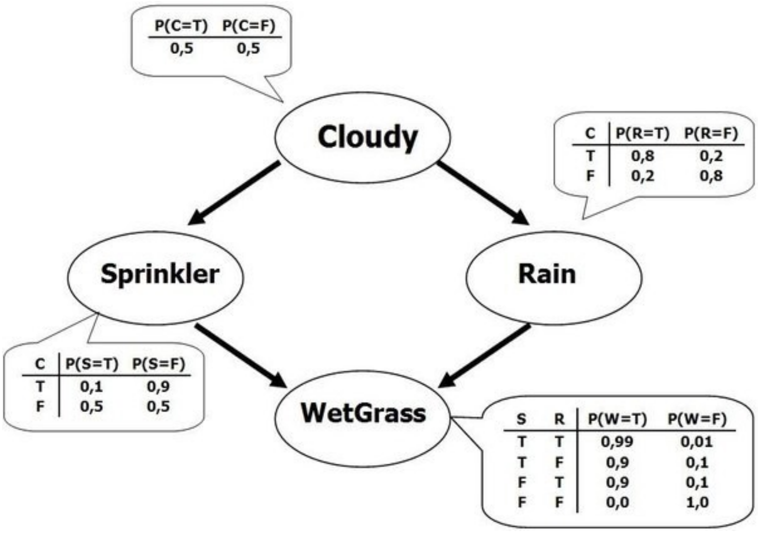


Figure 2.5: Example of a decision tree [41].

Bayes Network A Bayesian network is a graphical model that consists of a set of variables and a set of directed edges between variables. Each variable has a finite set of mutually exclusive states. Joining the variables and the directed edges form an acyclic directed group [39].

Figure 2.6 represents an example of a Bayesian Network, where Cloudy, Sprinkler, WetGrass and Rain are variables. The bubbles represent the Conditional Probability tables.



P(C=T)	P(C=F)
0,5	0,5

C	P(R=T)	P(R=F)
T	0,8	0,2
F	0,2	0,8

C	P(S=T)	P(S=F)
T	0,1	0,9
F	0,5	0,5

S	R	P(W=T)	P(W=F)
T	T	0,99	0,01
T	F	0,9	0,1
F	T	0,9	0,1
F	F	0,0	1,0

Figure 2.6: Bayesian Network [11].

Naive Bayes Classifier The naive Bayes classifier is a special case of a Bayesian Network. It considers as variables both class probabilities and conditional probabilities. This classifier assumes that each feature only depends on the class as shown in Figure 2.7. Meaning that each feature has only the class as a parent. Naive Bayes is useful for high dimensional data as the probability of each feature is estimated independently [51].

This algorithm classifies samples based on the Bayesian rule:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \tag{2.1}$$

In which C is a class of an observation X [51].

In the Naive Bayes Classifier, assuming that the features $X_1, X_2, \ldots, X_n$ are conditionally independent of each other given the class, we obtain [51]:

$$P(C|X) = \frac{\prod_{i=1}^{n} P(Xi|C)P(C)}{P(X)} \tag{2.2}$$

Figure 2.7 shows an example of a Naive Bayes.

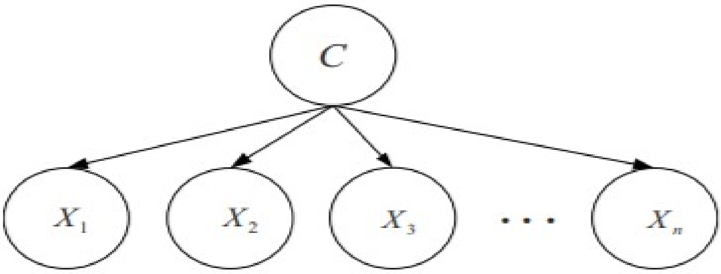


Figure 2.7: Naive bayes [51].

Support Vector Machines Support Vector Machine (SVM) is a supervised learning model with associated learning algorithms that recognizes patterns used for both classification and regression analysis [42].

The basic idea of SVMs is mapping the original data in a new high-dimensional space and where it's possible to apply linear models to obtain a separating hyper plane, which for classification tasks, separates the classes of the problem [52].

The hyper plane separation in a new dual representation is done by maximizing a separation margin between cases belonging to different classes. This problem of optimization is often solved with quadratic programming methods [52].

Gaussian process classifiers The Gaussian process classifiers focus on modeling the posterior probabilities by defining latent variables: let's assume f_i is the latent variable for pattern i [48].

Let's consider now a case of two classes: f_i is a measure of the degree of membership of class C_1, this means that if f_i is positive and big, this implies that pattern i belongs to class C_1 with a high probability. If f_i is negative and large in magnitude this means that pattern i belongs to class C_1 with a high probability. If f_i is close to zero, this means that the membership of the class is less certain [48].

Unsupervised Learning According to Russell and Norvig, [46], unsupervised learning is "the agent learns patterns in the input even though no explicit feedback is supplied."

K-means The K-means algorithm is one of the simplest and most popular unsupervised machine learning algorithms. Its main goal is to group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset.

This algorithm starts by deciding how many clusters that we would want to create from our data, calling this k. Normally k is a small number. After deciding k, we select k objects randomly, and these are treated as the centroids of k potential clusters, because at the moment those clusters have no members. We can select these points in anyway that we want, but a better idea that may work is if we pick k initial points that are far apart from each other. After that, one by one, we assign each point to the cluster which has the nearest centroid. After all objects are assigned to each cluster, we will have k clusters based on the original k centroids but those 'centroids' won't be anymore the true centroids of the clusters. Then we recalculate the centroids of the clusters, and then repeat the previous, assigning each object to the cluster with the nearest centroid [25].

Association Rules Association Rules is a ruled-based Machine Learning method for discovering interesting relations among variables in large databases. In general, association rules are interesting if they satisfy a maximum support and a maximum confidence [35].

An example of an association rule is:

$$ab \rightarrow c \tag{2.3}$$

Where the arrow reads as 'implies' and a and b are items that were brought together and c an item that was often brought as well [25].

The number of rules that are generated by a small database is potentially huge. Therefore, we need to find a way of deciding which rules to keep and which ones to discard. There are many ways to measure the interest of a rule, but the most common ways are support and confidence. In our study we are going to use support, confidence and lift [25].

The support for a rule $A \rightarrow B$, is the relative frequency that item A and item B occur together in the dataset, meaning $A \cup B$ [25].

$$support(A \rightarrow B) = support(A \cup B) \tag{2.4}$$

The confidence of a rule indicates the relative frequency of items A and B in an itemset in relation to the frequency of A [50].

$$confidence(A \rightarrow B) = \frac{support(A \cup B)}{support(A)} \tag{2.5}$$

The higher the amount of transactions that contain A and also contains B, the higher the confidence of the rule.

The lift measures the (in)dependence of items in the antecedent and in the consequent of the rule [25].

$$lift(A \rightarrow B) = \frac{support(A \cup B)}{support(A) * support(B)} \tag{2.6}$$

If the value of the lift is greater than one, this indicates that A and B appear more often together than expected, meaning that the occurrence of B has a positive effect on the occurrence of A [17].

If the value of the lift is less than one, this indicates that A and B appear less often together than expected, meaning that the occurrence of B has a negative effect on the occurrence of A [17].

If the value of the lift is close to one, this indicates that A and B appear almost as often together as expected, meaning that the occurrence of B has almost no effect on the occurrence of A, meaning that A and B are statistically independent [17].

Among the association rules algorithms, for example fp-growth [37], one of the most popular is the apriori [22]. It uses a bottom-up strategy and it works on categorical attributes. First, it finds the frequent itemsets of length 1, according to the values of minimum support, for example. From those frequent itemsets, new rules of size 2 are generated, and so on and so forth. Algorithm 1 shows how apriori works.

Algorithm 1: Apriori Algorithm

```
Input:
D: transaction database;
Min_sup: the minimum support threshold;
Output: frequent itemsets
Description:
Procedure Apriori_gen($L_{k-1}$: frequent($k-1$)-itemsets)
$L_1$= find_frequent_1-itemsets(DB);
for (k=2; $L_{k-1} = \varnothing; k++$) {
    $C_k$= Apriori_gen($L_{k-1}$);    // generates new candidates
    for each transaction $t \in DB$ {    // scan DB for counts
        $C_t$ = subset($C_k, t$);   // get the subsets of $t$ that are candidates
        for each candidate $c \in C_t$
            $c.count++$;
    }
    $L_k = \{c \in C_k | c.count \geq min\_sup\}$
}
return $L = \bigcup_k L_k$;
```

Rules are ranked according to their support, confidence and lift, explained already in this section [50].

Although k-means and association rules are considered to be unsupervised, given the labels, clusters and rules can be found in a supervised way. The same is true for the Bayesian networks.

2.2.6 Anomaly Detection

Anomaly Detection identifies rare items or observations, which can raise suspicious information by having a different value from the majority of the data [10]. With this method, we are able to pick out which data points can be considered anomalies, and check if these data points are interesting.

Some cases of anomalies are easier to detect, if the dataset is relatively small and data visualization can give important information [10].

Figure 2.8 shows two-dimensional data (X and Y), and when we observe this figure, it becomes quite easy to visually identify anomalies through data points located outside the typical distribution. But when looking at the graphs more to the right, it is not possible to identify the outlier directly from investigating one variable at a time. But if we combine the X and Y variables together, it is easier to identify the anomaly [10].

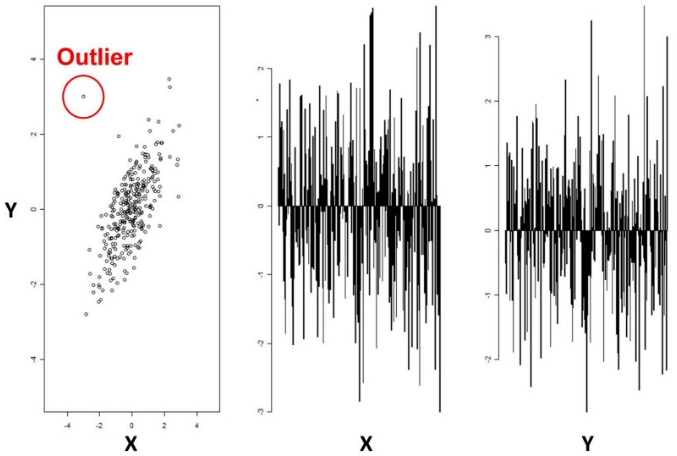


Figure 2.8: Anomaly detection for two variables [10].

Genetic data, specially in cohorts of patients with some disease, may present small subgroups which may be difficult to find using the conventional machine learning methodology. Most Machine Learning methods try to find common patterns while genetic data may present rare cases [32].

Chapter 3

Other studies on bipolar disorder

In this chapter, we will discuss some studies on Bipolar Disorder (BD) that are important to mention given the aim of our work. This chapter is divided in genetic studies and Machine Learning studies.

3.1 Genetic Studies

Orrù and Carta [44], had the objective of performing a synthetic description of the main Single-Nucleotide Polymorphism (SNP)s variants that were identified or confirmed by recent Whole Genome Sequencing analysis and also reconstruction of an *in vitro* mechanism or by amygdala activation protocol *in vivo* [44]. For the study, they used bibliographic data, genomic and protein Data Banks to carry out a cross genomic study for mutations, SNPs and chromosomal alterations described in these studies in BD patients. As the results, genes were discovered in this disorder, including ANK3, CACNA1c, NCAN, ODZ4, SYNE1 and TRANK1 [44].

3.2 Machine Learning Studies

Saylan *et al.* [47] tested several machine learning algorithms to classify Bipolar Disorder and Schizophrenia. They applied supervised methods such as K-nearest neighbors, Decision Tree and Naive Bayes Network [47]. The dataset used contains genetic information. In terms of size this dataset is very small (38.4 MB), containing only 7 subjects, 3 controls and 4 BD cases. Tables 3.1, 3.2 and Figure 3.1 show the results of this article, only for the bipolar group.

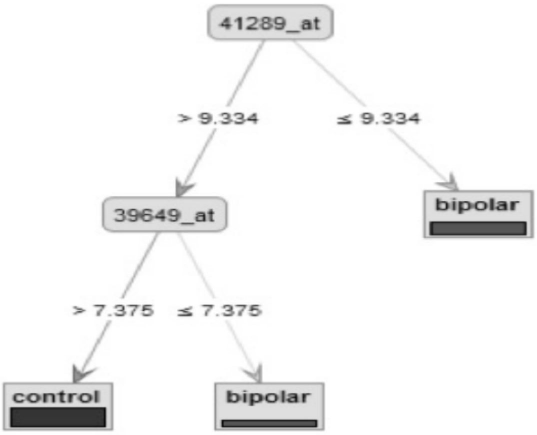


Figure 3.1: Decision Tree classification of bipolar disorder [47].

Table 3.1: Naive Bayes Classification of Bipolar Disorder [47].

	True Control	True Bipolar	Class Precision
Pred. Control	3	0	100.00%
Pred. Bipolar	1	3	75.00%
Class Recall	75.00%	100.00%	Accuracy: 87.71%

Table 3.2: k-NN Classification of Bipolar Disorder [47].

	True Control	True Bipolar	Class Precision
Pred. Control	4	1	80.00%
Pred. Bipolar	0	2	100.00%
Class Recall	100.00%	66.67%	Accuracy: 85.71%

Saylan *et al.* [47] concluded that different expressed genes used as classifying methods can be useful for revealing important genes and gene families that are related to bipolar disorder and schizophrenia.

Hajek *et al.* [36] applied supervised methods, Support Vector Machine (SVM) and Gaussian process classifiers, to structural Magnetic Resonance Imaging (MRI), which learning task was studying BD by using structural MRI to identify patients that have a genetic risk for BD. They also used cross-validation to validate their methods. They trained independently the Gaussian Process Classifiers in each site and the combined dataset.

Nunes *et al.* [30] with collaboration of the ENIGMA-BD Working Group, applied SVM to structural MRI data from 3020 participants (853 with BD and 2167 controls) recruited in 13 independent sites around the world. Results show that the study provided a realistic and fair estimate of classification performance in a large, multi-site sample of BD participants based on regional neurostructural measures.

Librenza-Garcia *et al.* [40] applied a systematic review of the impact of machine learning in the study of bipolar disorder. They searched the literature for articles that used machine learning techniques in BD patients. Those articles include MRI and SNP studies and whole-genome genotyping. Dataset sizes varied from 30 to 4488 patients. All surveyed articles applied both supervised and unsupervised methods. [40].

Moreira [43] used the K-means algorithm applied to the same genetic data we are going to use, (described later in chapter 4), but using a selection of 7 chromosomes. He found that there is a prevalence of BD in women and that the prevalence is in a particular region of the United Kingdom. He also discovered that SNPs rs1006737 and rs4765914 of CACNA1C gene have revealed a pattern in women [43]. The SNPs studied were the following:

- DISC1 (Chromosome 1) rs203368 and rs435136;
- ARPP21 (Chromosome 3) rs1523041;
- GABRB1 (Chromosome 4) rs7680321;
- ANKRD46 (Chromosome 8) rs80198067;
- ANK3 (Chromosome 10) rs10994336 and rs9804190;
- CACNA1c (Chromosome 12) rs1006737, rs4765913, rs4765914 and rs2239063;
- DUSP6 (Chromosome 12) rs769700, rs704076, rs770087, rs808820, and rs2279574;
- GRIND2B (Chromosome 12) rs1805502, rs1805247 and rs7301328;
- SYN3 (Chromosome 22) rs9621532;

Moreira [43] observed that the genes ARPP21 (rs1523041), GABRB1 (rs7680321), CACNA1c (rs10067379, rs4765914) and SYN3 (rs9621532) were present among our patients.

While there is a reasonable number of studies focusing on how each one of these SNPs affect individuals with BD, and how the same SNP can be involved in multiple mental disorders, for example the CACNA1c gene in SNP rs1006737 [24], to the best of our knowledge, investigation of the combination of these SNPs using machine learning techniques and their role on describing patients with BD has not yet been explored.

Table 3.3 summarizes the information of all the articles, including the type of data, size of the dataset, if possible, algorithms that were used and finally learning tasks and conclusions. N/A means that the data set is not available.

Table 3.3: Summary of state-of-the-art

Article	Size	Type of Data	Algorithm	Learning tasks and conclusions
Orrù and Carta (2018)	N/A	Genetic	N/A	A table of genes with their SNPs, including also the risk allele was created. They opened new possibilities in genomic research. However the results are still controversial due to analytical validity, clinical validity and utility and a reasonable cost for genetic analysis are not yet accessible [44].
Saylan and Yilancioglu (2016)	34.8MB	Genetic	KNN, Decision Tree and Naive Bayes	They classified BD and Schizophrenia using machine learning techniques, via a dataset GEO (Gene Expression Omnibus). According to the confusion matrix shown on the article, Naive Bayes algorithm has a better accuracy [47].
Hajek et al. (2015)	N/A	MRI	SVM and Gaussian process classifiers	They studied families that are unaffected and affected of BD probands. They distinguished unaffected participants at high and low genetic risk for BD [36].
Nunes et al. (2018)	N/A	MRI	SVM	Differentiate people with BD from control participants. The study provided a realistic and fair estimate of classification performance in a large, multi-site sample of BD participants based on regional neurostructural measures [30].
Librenza-Garcia et al. (2017)	N/A	Various	Various	They did a systematic review of the impact of machine learning techniques in the study of BD. Given the clinical heterogeneity of sample of patients with BD, machine learning algorithms may provide researchers and doctors with important information for diagnosis, prognosis and personalized treatment [40].
Moreira (2018)	30.0 GB	Genetic	K-means	Find pattern extractions by finding clinical-genetic similarities between subgroups of BD samples. He found that there is a prevalence in women and that the prevalence is in a particular region of the United Kingdom [43] .

Chapter 4

Identification of BD patients subgroups

4.1 WTCCC Bipolar Disorder (BD) Genetic Dataset

The database used is from the Wellcome Trust Case Control Consortium (WTCCC). Our institution has a signed agreement with WTCCC that allows the use of part of this data for research purposes. Genotype data was produced through blood samples using the Affymetrix 500K microarrays. This microarray comprises probes for more than 500,000 Single-Nucleotide Polymorphism (SNP)s. Specifically, Bipolar Disorder which contains genetic information about 1998 patients and is divided into two parts:

- Genotypes: This part contains the SNPs present in each patient (which is represented with a code), the genotype present in the SNP and the respective Genetic Risk Score (GRS).
- Meta-data: This data is the control data of the SNP.

The dataset contains 25 files of chromosomes where 24 files are genetic data and one is the demographic data.

Each file occupies approximately 4 GB, the genotype data has been partitioned according to each chromosome. Table 4.1 shows the information contained in each chromossome file, where column one is the SNP, the second column is the sample that was used, the third is the genotype of the sample and the last one is the Genetic Risk Score. The higher the score, the higher the risk of appearing the SNP [9]. Recalling from 2.2.3 that the GRS is in the continuous interval [0,1].

SNP	Sample	Genotype	Score
rs10488368	WTCCC65841	AG	1
rs10488368	WTCCC65569	AA	1
rs10488368	WTCCC65823	AG	1
rs10488368	WTCCC65845	AG	1

Table 4.1: WTCCC Genetic data information.

We also have demographic data that describes the information of the patient. Such information contains the gender (1- Male, 2 - Female), cohort (the disease/disorder that the patient has, which in our case is Bipolar disorder), the supplier, the plate, the region of the patient, their age of recruitment (1 → age between 10 and 19 and so on) and finally their age onset. Unfortunately, the age onset is missing, all values are "Unknown", due to not having this information. Table 4.2 illustrates the contents of the demographic file.

Table 4.2: WTCCC demographic data description.

Sample	Gender	Cohort	Supplier	Plate	Region	Age_Recruitment	Age_onset
WTCCC65841	2	BD	PMHWW	11142A1	Southwestern	3	Unknown
WTCCC65569	2	BD	PMHWW	11142A2	Southwestern	5	Unknown
WTCCC65777	2	BD	PMHWW	11142A3	Northwestern	6	Unknown
WTCCC65795	2	BD	PMHWW	11142A4	Southern	3	Unknown
WTCCC65810	2	BD	PMHWW	11142A5	Wales	7	Unknown

Table 4.3 shows the distribution, in percentage, of men and women given their age and region.

Table 4.3: Statistical information of the dataset

Age	**Quantity**	**Total Population %**	**Male**	**Female**
[10-19]	14	1.00%	28.57%	71.43%
[20-29]	195	9.70%	34.87%	65.13%
[30-39]	400	20.00%	35.50%	64.50%
[40-49]	575	28.70%	38.09%	61.91%
[50-59]	480	24.00%	36.87%	63.13%
[60-69]	277	13.80%	57.04%	42.96%
[70-79]	56	2.50%	44.00%	66.00%
[80-89]	7	0.30%	71.43%	28.57%
Region				
East + West Ridings	26	1.30%	34.62%	65.38%
Eastern	63	3.10%	33.33%	66.67%
London	133	6.60%	42.10%	57.90%
Midlands	475	24.00%	35.60%	64.40%
North Midlands	118	5.90%	33.05%	66.95%
Northern	176	8.80%	39.20%	60.80%
Northwestern	68	3.40%	32.35%	67.65%
Scotland	199	10.00%	43.72%	56.28%
Southeastern	95	4.70%	31.58%	68.42%
Southern	116	5.80%	37.07%	62.93%
Southwestern	115	5.70%	30.43%	69.57%
Wales	414	20.70%	41.30%	59.70%

All experiments were run in Python 3.7 using Jupyter Notebook. We used sklearn and pandas for implementing our project, apyori to generate association rules, time, os, sys and tqdm for operating system calls such as get timing and checkpointing our long run jobs. We also used matplotlib and matplotlib_venn for data visualization and graphical plotting.

4.2 Methodology and experiments

4.2.1 Methodology

Our challenges with this dataset were:

1. to understand the files and format. The WTCCC provides a good documentation [1], although not everything is explained. For example, there is no pointer or link to a description of how the GRS was calculated. We had to study the Chiamo software (https://mathgen.stats.ox.ac.uk/genetics_software/chiamo/chiamo.html) to better understand how these scores were calculated. This was relevant to understand distributions of patients with varying scores.

2. to select suitable variables. In order to select suitable variables, we had to go through the literature to identify the SNPs relevant to BD since this dataset has other SNPs that are not directly related with it. We also had to decide what to include in the study.

3. to select suitable algorithms. Besides using descriptive statistics, we used a machine learning method to identify subgroups of patients with BD. Among the various machine learning methods that we could use, we opted for association rules, since it is non supervised (we don't have a target variable), it can relate multiple variables and it can give an explanation (explicit pattern) for the subgroups.

4. to prepare the data. Since we were working with Gigabytes of data, we resorted to multiprocessing.

5. to generate the model. Since association rules tend to favor high support and high confidence rules, and high lift, we had to manually tune these parameters in order to identify small sets of patients with common patterns.

6. to evaluate the results. We had to select an experimental methodology and appropriate metrics and data visualization in order to evaluate the results. We opted for using Venn diagrams, matrix correlations, and used cross-validation to validate our results.

   The specific steps to handle these challenges are shown below and depicted in Figure 4.1.

[1] https://www.wtccc.org.uk/info/data_formats.html

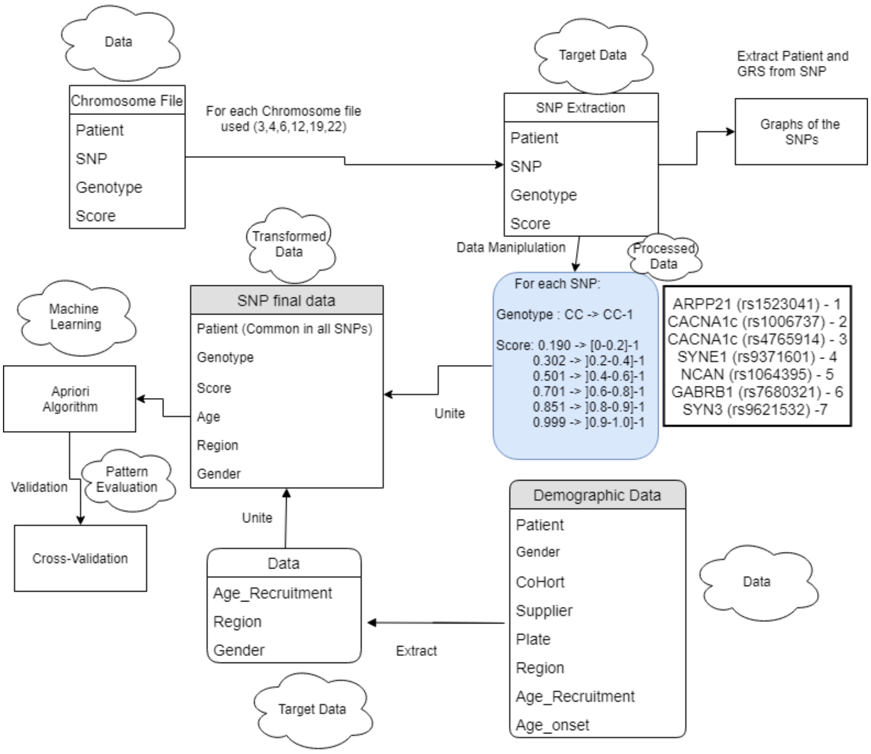


Figure 4.1: The methodology of our project.

Figure 4.1 shows the steps that we did in this project. First, from the original data set, we extracted the essential data and created our target data (SNP Extraction in the figure). Next, we did two things:

- We extracted from the SNP Extraction Table, the patients and the GRS and created plot graphs to study the behaviour of the GRS.
- We applied data manipulation to create a summarized data set and applied the following:
  1. For each SNP, in the genotype column, we assigned a number to distinguish what genotype comes from what SNP, for example for rs1523041, there is the CC genotype and we assigned this as CC-1.
  2. For each GRS column, we assigned the amount to a specific interval, defined in Figure 4.1. We also assigned a number to distinguish where the interval comes from, just like what we did with the genotypes.

After the data manipulation phase, we obtained a new data set with the selected SNPs, their genotype and GRS. We called this part the SNP final data.

For our data set to be complete we also have to deal with another file, the demographic data. From that, we extracted Age_Recruitment, Region and Gender.

Next, we aggregated data to produce the SNP final data.

To the SNP final data we applied the apriori algorithm.

There are various libraries in Python that implements the apriori algorithm, we will mention especially two of them: apyori (https://pypi.org/project/apyori/) and efficient-apriori (https://pypi.org/project/efficient-apriori/).

The advantages of the apyori library is the fact that it's not dependent from other libraries, its implementation is simple and supports the JSON format. However, when we apply this library on large datasets, it takes a lot of time to generate the rules due to the big amount of transactions.

The advantages of the efficient-apriori library is the fact that its implementation is much more efficient and well tested. However, when we apply this library on large datasets, it takes a lot of time to generate the rules due to the big amount of transactions, just like the apyori library.

Due to the implementation being more simple, and according to Python Software Foundation this library has a better rating, we used the apyori library to generate our rules.

After this, we moved to the validation phase, and applied 10 times 10-fold cross-validation to validate our results.

4.2.2 Experiments

4.2.2.1 General table of the SNPs studied

First we created a general table with all the possibilities of genotypes in a SNP, they are 'AA', 'TT', 'CC', 'GG', 'CT', 'AG', 'CG', 'AC', 'GT', 'AT'. We also included the number of patients that has a GRS in a certain interval.

4.2.2.2 Graph Visualization

For each SNP discovered in our dataset we created a graph of the GRS of each patient.

4.2.2.3 Creation of the summarized dataset

After analyzing the behaviour of the genotypes of our patients, we created a summarized dataset containing the genes that are related to BD. This dataset contains the genotype of each gene as well as the interval of the GRS. First of all, we changed the GRS values into intervals, [0-0.2],]0.2-0.4],]0.4-0.6],]0.6-0.8],]0.8-0.9] and]0.9-1.0]. We separated the interval]0.8-1.0] into two cases because we wanted to see cases of a *very high* GRS and *high* GRS. In order to distinguish among GRS of several SNPs, we associated a numerical code to each SNP. For example, a valid genotype for rs1006737 is AA. As AA can appear in other SNPs, we coded this as AA-2. We also modified the GRS accordingly appending a 2 to the interval: [0.9-1.0]-2. We repeated this to all SNPs using the numerical coding shown in Table 4.4.

Table 4.4: Connection coding for each SNP

SNP	Number Assigned
rs1523041	1
rs1006737	2
rs4765914	3
rs9371601	4
rs1064395	5
rs7680321	6
rs9621532	7

4.2.2.4 Finding Association Rules

After creating the new dataset, we apply the apriori algorithm. We tested two situations:

- We tried smaller values of support trying to capture not very frequent itemsets (rare subgroups of items). We chose minsupport = 3% in order to produce as many diverse itemsets as possible. We chose minconfidence = 60% to compensate the choice of using very low support. We chose minlift = 3 in order to capture itemsets where items have high probability of appearing together and having a positive dependency;
- We tried higher values of support to capture frequent itemsets, to find common subgroups of items. We chose minsupport = 50%, minconfidence = 50% and minlift = 1 to find independent items.

We generated rules for the entire dataset. For the first case we selected the top rules according to the algorithm generation order. All subsequent generated rules had all the same or less values of support, confidence and lift. For the second group we sorted the rules from the highest to the lowest support.

4.2.2.5 Cross-Validation

After the exploratory phase we moved to a validation phase where 10 times cross-validation was used to generate rules. We separated the data into the following: 80% training data and 20% test per fold. We chose minsupport = 2.5%, minconfidence = 50% and minlift = 3.

After applying the apriori algorithm in both cases (training and test), we moved on to a fold selection phase where we selected a fold from the training models and ranked the generated rules from best to worst confidence. We selected the first ten rules and for each fold, we checked if those rules are present or not.

Chapter 5

Results and analysis

5.1 Exploratory Analysis

5.1.1 Single-Nucleotide Polymorphism (SNP)s found in the literature

Works in the literature highlight the genes shown in table 5.1.

Table 5.1: Genes related to BD, SNPs and chromosome region previously associated with BD.

Gene	SNP	Chromossome
CACNA1c	rs1006737 / rs4765914	12
ANK3	rs1099413	10
NCAN	rs1064395	19
ODZ4	rs12576775 / rs17138171	11
TRANK1 / LBA1	rs9834970	3
SYNE1	rs9371601	6
DISC1	rs203368 / rs435136	1
ARPP21	rs1523041	3
GABRB1	rs7680321	4
ANKRD46	rs80198067	8
DUSP6	rs2279574 / rs769700 / rs704076 rs770087 / rs808820	12
GRIND2B	rs1805502 / rs1805247 / rs7301328	12
SYN3	rs9621532	22

From those, only seven of them are present in our dataset:

- ARPP21 (rs1523041);
- CACNA1c (rs1006737 and rs4765914);

- SYNE1 (rs9371601);
- NCAN (rs1064395);
- GABRB1 (rs7680321);
- SYN3 (rs9621532)

ARPP21 is a protein coding gene. Other diseases related to this gene include Ureteral Benign Neoplasm and Calcific Tendinitis [4].

The CACNA1c gene is responsible for making calcium channels, which transport positively charged calcium ions into cells. It plays a role in a cell's ability to generate and transmit electrical signals [12].

The SYNE1 gene is responsible for producing a protein called SYNE-1 which is critical due to playing a major role in the brain. This protein maintains the part of the brain that coordinates movement. Other health conditions include Emery-Dreifuss muscular dystrophy [15].

The NCAN gene is a protein coding gene. Other diseases include Diseases of glycosylation [6].

The GABRB1 gene is a protein coding gene. Other diseases include Schizoaffective disorder [5].

The SYN3 gene. Other diseases include Visual Epilepsy and Pseudoinflammatory Fundus Dystrophy [7].

Table 5.2 shows a few lines of the final dataset after all preprocessing. Each column corresponds to the genotype possible for the SNP in study and the interval of the Genetic Risk Score (GRS). Each row corresponds to a patient.

Table 5.2: Summarized table of quantity of patients with SNPs

patient	rs1523041	score1	rs1006737	score2	rs4765914	score3	rs9371601	score4	rs1064395	score5	rs7680321	score6	rs9621532	score7	Age	Region	Gender
WTCCC65841	GG-1	[0.9-1.0]-1	AG-2	[0.9-1.0]-2	CT-3	[0.9-1.0]-3	GG-4	[0.9-1.0]-4	GG-5	[0.9-1.0]-5	TT-6	[0.9-1.0]-6	AC-7	[0.9-1.0]-7	3	Southwestern	female
WTCCC65569	CG-1	[0.9-1.0]-1	AG-2	[0.9-1.0]-2	CT-3	[0.9-1.0]-3	TT-4	[0.9-1.0]-4	AG-5	[0.9-1.0]-5	CT-6	[0.9-1.0]-6	AA-7	[0.9-1.0]-7	5	Southwestern	female
WTCCC65777	CG-1	[0.9-1.0]-1	AG-2	[0.9-1.0]-2	CC-3	[0.9-1.0]-3	GG-4	[0.9-1.0]-4	GG-5	[0.9-1.0]-5	TT-6	[0.9-1.0]-6	AA-7	[0.9-1.0]-7	6	Northwestern	female
WTCCC65795	GG-1	[0.9-1.0]-1	AG-2	[0.9-1.0]-2	CC-3	[0.9-1.0]-3	GG-4	[0.9-1.0]-4	GG-5	[0.9-1.0]-5	TT-6	[0.9-1.0]-6	AA-7	[0.9-1.0]-7	3	Southern	male

5.1.2 GRS Distribution per SNP

Figure 5.1 shows an example of how the GRS is distributed among the patients for the CACNA1c SNP at rs1006737, where the X-axis represents the patients and the Y-axis the GRS values.

Patients exhibit quite distinct GRS distribution depending on the gene. For this population, most patients have very high risk for genes ARPP21, CACNA1c, GABRB1, SYNE1 while having lower risk for NCAN and SYN3. Figures for all genes are shown in Appendix A.

These differences do not distinguish homozygous or heterozygous genotypes. Next, we show the same information, but now with a separation per genotype.

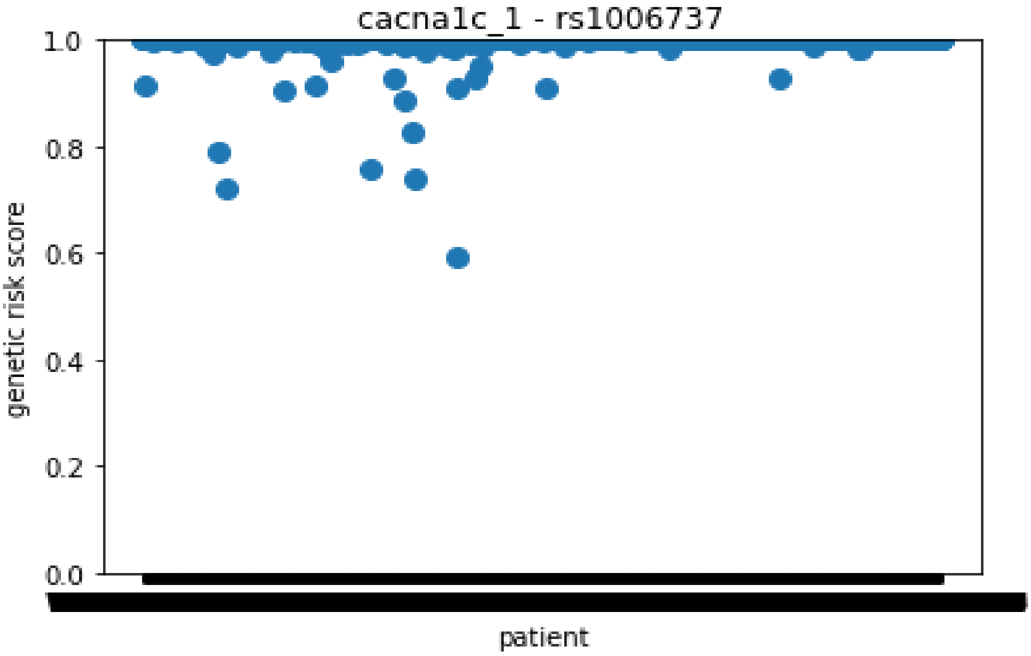
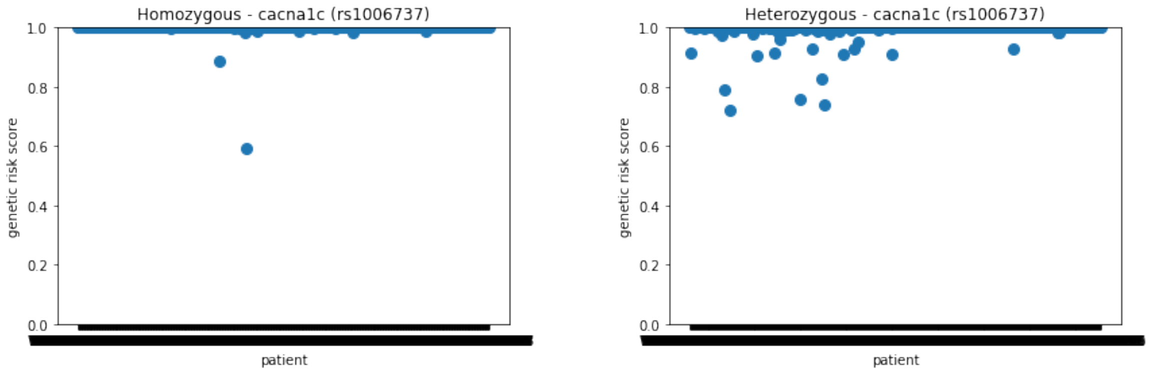


Figure 5.1: GRS of patients with CACNA1c - rs1006737

5.1.3 GRS Distribution per SNP, homozygous and heterozygous

Figure 5.2 shows an example of a comparison of CACNA1c (rs1006737) between homozygous and heterozygous patients.



(a) Homozygous patients

(b) Heterozygous patients

Figure 5.2: GRS of homozygous and heterozygous patients with CACNA1c (rs1006737)

There is a higher dispersion of GRS in heterozygous patients than homozygous. Most of the SNPs follows the same pattern, however GABRB1 is an exception: the GRS dispersion for this gene is higher for homozygous patients. All GRS plots are shown in Appendix A.

5.1.4 Summary table creation

Table 5.3 shows a summarized table of the SNPs in study. Each column corresponds to a SNP. For each SNP we show homozygous and heterozygous total counters and counters per genotype. We also show their genetic risk score represented in intervals.

Table 5.3: Homozygous and Heterozygous genotypes per SNP

Gene	ARPP21	CACNA1c	CACNA1c	SYNE1	NCAN	GABRB1	SYN3
SNP	rs1523041	rs1006737	rs4765914	rs9371601	rs1064395	rs7680321	rs9621532
Tot. Patients	1998	1998	1998	1998	1998	1998	1998
Tot. Homo	1091	1075	1276	1069	1444	1641	1796
Tot. Hetero	907	923	722	929	554	357	202
Gen. A	CC	AA	CC	TT	AA	CC	AA
# A	791	248	1180	259	65	26	1788
Gen. B	GG	GG	TT	GG	GG	TT	CC
#B	300	827	96	810	1377	1615	8
Gen. C	CG	AG	CT	GT	AG	CT	AC
#C	907	923	722	929	554	357	202
[0-0.2]	6	0	2	1	3	1	8
]0.2-0.4]	0	0	0	0	1	0	0
]0.4-0.6]	0	1	1	0	2	0	2
]0.6-0.8]	1	4	1	0	1	0	1
]0.8-0.9]	2	2	3	1	0	1	3
]0.9-1.0]	1989	1991	1991	1996	1991	1996	1984
Tot <0.9	9	7	7	2	7	2	14

As we observe in Table 5.3, most genotype patterns for bipolar disorder are homozygous. We also see that genes SYN3 and GABRB1 have a greater imbalance between homozygous and heterozygous genotypes, (90% homozygous x 10% heterozygous and 82% homozygous x 18% heterozygous respectively). At the position rs4765914 of the gene CACNA1c, we can see that this SNP has a high rate of genotype CC (92%) compared to TT (8%) among the homozygous patients, and SYN3 has a very high rate of AA (99%) compared to CC (1%), once again among the homozygous. We can see that patients with the gene SYN3 at the position rs9621532 are distinct from other patients because we have more patients (14) with GRS below 0.9. A single patient with the gene SYNE1 at the position rs9371601 has a very low GRS compared to the other 1997 patients with the genotype GT (heterozygous) and another single patient with the gene GABRB1 at the position rs7680321 has a very low GRS among the other 1997 patients with genotype TT (homozygous). In association with the GRS plots, Table 5.3 shows the number of patients for each GRS interval.

5.2 Intersection of two genotypes

Before we apply the apriori algorithm, we decided to study the intersection of two genotypes to see if relevant information is obtained. Table 5.4 shows those intersections.

Table 5.4: Intersection of Genotypes of Genes of Bipolar Disorder

	GG-1	CC-1	CG-1	AA-2	GG-2	AG-2	CC-3	TT-3	CT-3	GG-4	TT-4	GT-4	AA-5	GG-5	AG-5	TT-6	CC-6	CT-6
AA-2	32	111	105															
GG-2	127	329	371															
AG-2	141	351	431															
CC-3	172	467	541	48	770	362												
TT-3	13	48	35	74	0	22												
CT-3	115	276	331	126	57	539												
GG-4	125	327	358	96	352	362	483	40	287									
TT-4	36	104	119	24	97	138	146	6	107									
GT-4	139	360	430	128	378	423	551	50	328									
AA-5	14	28	25	7	29	31	43	4	20	25	14	28						
GG-5	204	520	653	170	567	640	802	67	508	571	175	631						
AG-5	82	243	229	71	231	252	335	25	194	214	70	270						
TT-6	249	642	724	209	655	751	945	81	589	653	210	752	55	1114	446			
CC-6	6	12	8	1	14	11	20	0	6	13	4	9	1	22	3			
CT-6	45	137	175	38	158	161	215	15	127	144	45	168	11	241	105			
AA-7	270	722	796	216	745	827	1055	86	647	726	232	830	61	1235	492	1436	24	328
CC-7	3	2	3	0	5	3	6	1	1	3	0	5	1	5	2	5	1	2
AC-7	21	67	108	32	77	93	119	9	74	81	27	94	5	137	60	174	1	27

Some of the pairs are very frequent. Examples are the pairs CC-3 (CACNA1c, rs4765914) and GG-2 (CACNA1c, rs1006737), AA-7 (SYN3) and CG-1 (ARPP21) and CC-3 (CACNA1c, rs4765914) and AA-7 (SYN3).

5.3 Finding Association Rules

Table 5.4 already has shown some subgroups of patients that, in the entire dataset, share pairs of genotypes. In this Section, we show the results of the apriori algorithm with rules that correlate at most 3 genotypes and other data.

5.3.1 Case 1: Low Support, High Confidence and High Lift

From the first test, low support, high confidence and high lift, we obtained 384 rules, however, we are going to just use 2 of them because the remaining rules are subsets of the rules used in figure 5.3, and the other rules have a less than equal than the rules used in 5.3.

```
=====================================
Rule: ['TT-3', 'AA-7'] -> ['AA-2']
Support: 0.03303303303303303
Confidence: 0.7674418604651162
Lift: 6.1828582145536375
=====================================
Rule: ['TT-6', 'TT-3'] -> ['AA-2']
Support: 0.032532532532532535
Confidence: 0.8024691358024691
Lift: 6.46505376344086
=====================================
```

Figure 5.3: Top two ranked association rules found in our dataset, from the first test.

The first rule says that given that the patient has both TT-3 (has TT as its genotype at rs4765914, which gene is CACNA1c) and AA-7 (has AA as its genotype at rs9621532, which gene is SYN3), the probability of also having AA-2 (has AA as its genotype at rs1006737, which gene is CACNA1c) is high due to its confidence (76%) and its lift (6.18)

The second rule says that the patient has both TT-3 (has TT as its genotype at rs4765914, which gene is CACNA1c) and TT-6 (has TT as its genotype at rs7680321, which gene is GABRB1), the probability of also having AA-2 (has AA as its genotype at rs1006737, which gene is CACNA1c) is high due to its confidence (80%), and its lift (6.46).

Figure 5.4 shows the intersection of both rules. The red circle represents the rule [TT-3,AA-7] → AA-2 and the green circle represents the rule [TT-6,TT-3] → AA-2.

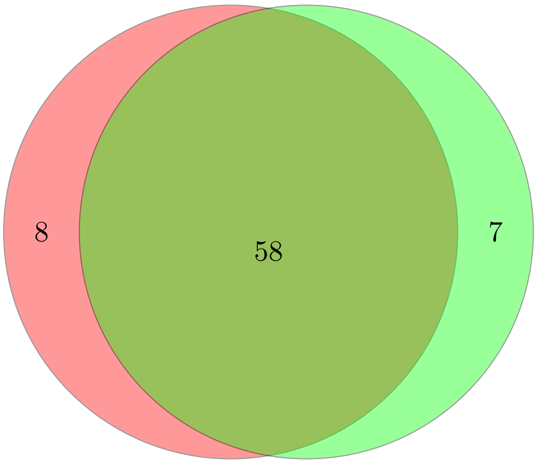


Figure 5.4: Venn-Diagram of the top ranked rules in the first case

Figure 5.4 shows the following:

Group 1 : 58 patients have all four homozygous genotypes (TT-3, TT-6, AA-2 and AA-7);

Group 2 : 8 patients have TT-3, AA-2, AA-7 but don't have TT-6 as their genotypes;

Group 3 : 7 patients have TT-3, AA-2, TT-6 but don't have AA-7 as their genotypes;

It's important to contrast the numbers shown in Table 5.4 with the produced rules. Some not so evident associations were revealed. We also see that there's is a few amount of patients when intersecting the CC-6 and also the CC-7 genotype.

Now we are going to fuse this information with the demographic information. From the first group, we see that from the 58 patients, 19 of them are male and 39 are female. Figures 5.5, 5.6 shows the distribution of the age and region of the first group.

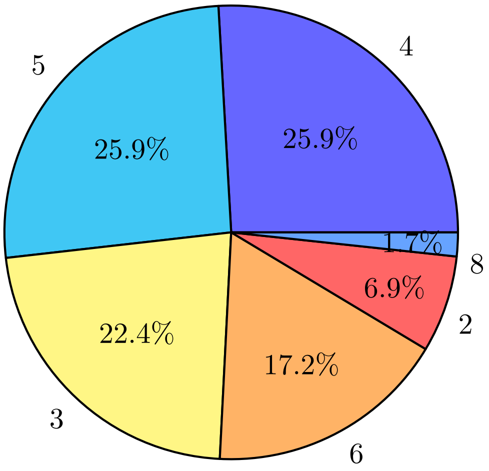


Figure 5.5: Age of the patients from the first group in case 1 (n = 58)

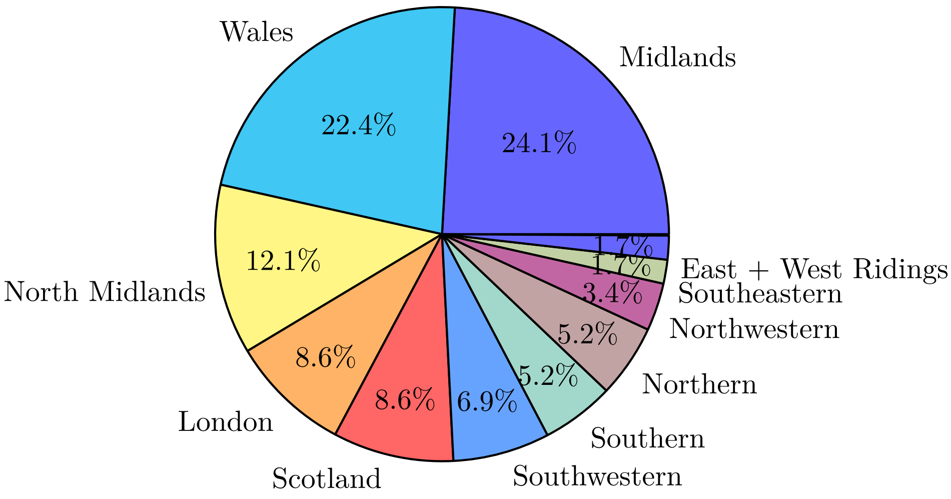


Figure 5.6: Region of the patients from the first group in case 1 (n = 58)

As we see from Figures 5.5 and 5.6, the most frequent age intervals are 40-49 and 50-59, with 15 patients in each subgroup, and the majority of the patients are from the Midlands.

For the second group, from the 8 patients, 5 patients are female and the other 3 are male.

Figures 5.7 and 5.8 show the distribution of age and region for the second group.

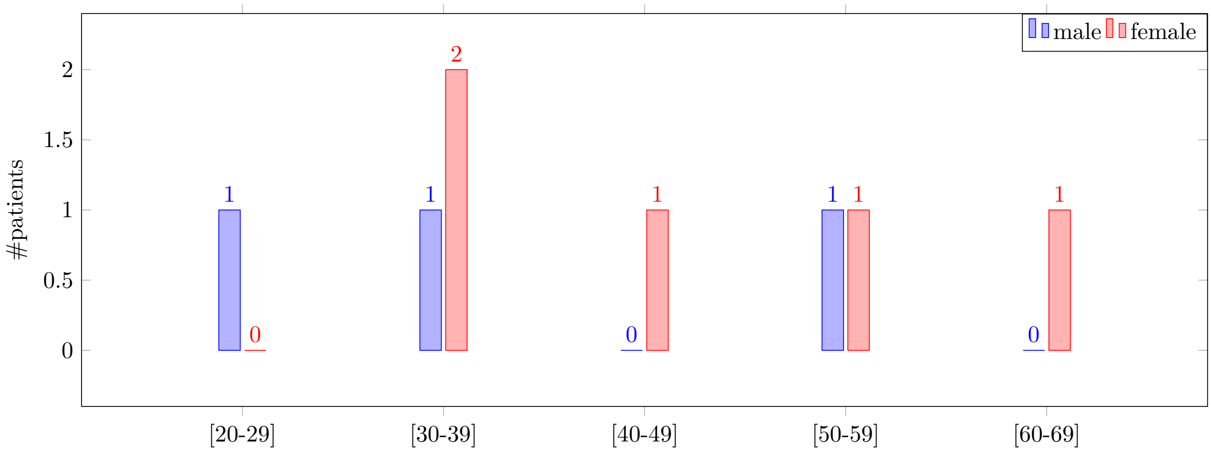


Figure 5.7: Age of the patients from the second group in case 1 (n = 8)

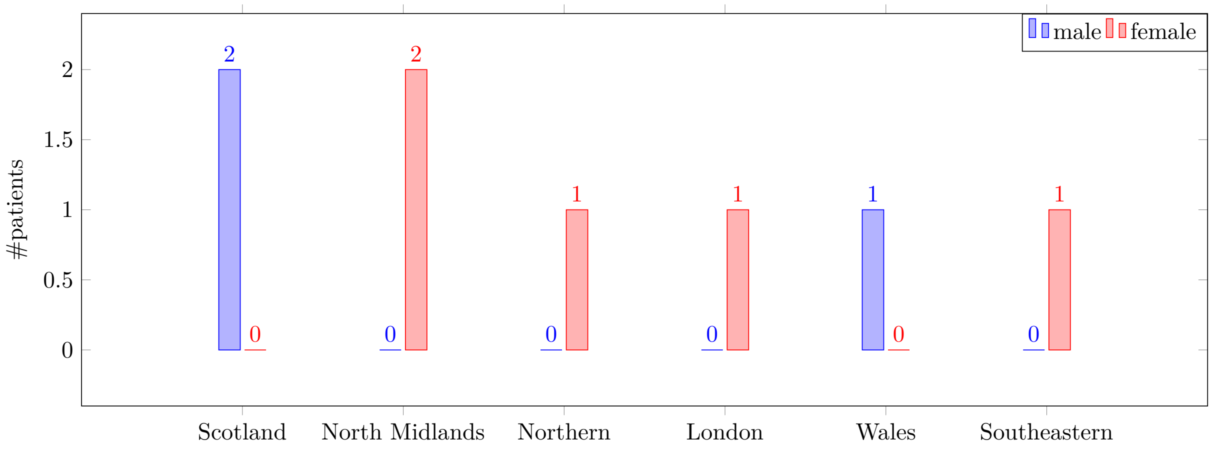


Figure 5.8: Region of the patients from the second group in case 1 (n = 8)

As we see from the previous two figures, the majority of the patients are in their 30's and came from North Midlands and Scotland.

For the third group we observe that from the 7 patients, 4 patients are female and the other 3 are male.

Figures 5.9 and 5.10 show the distribution of age and region for group 3.

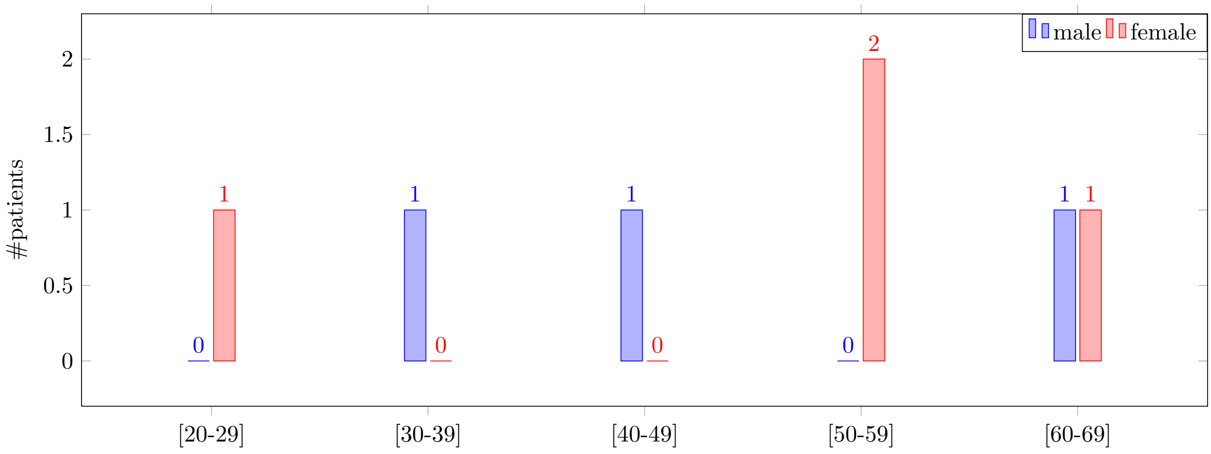


Figure 5.9: Age of the patients from the third group in case 1 (n = 7)

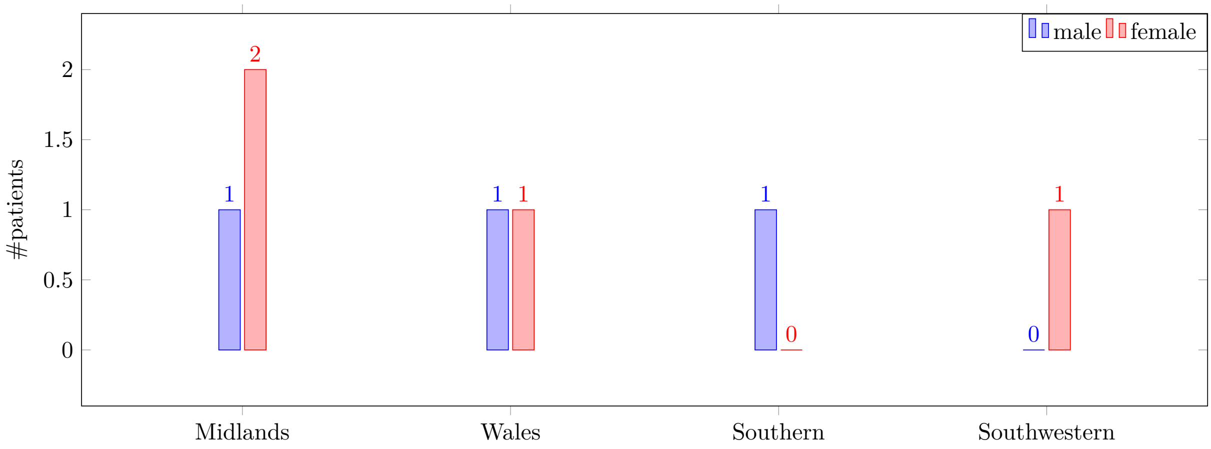


Figure 5.10: Region of the patients from the third group in case 1 (n = 7)

From the previous two figures, we see that the patients are in their 50s or in their 60s. And once again, the majority of the patients come from the Midlands.

5.3.2 Case 2: High Support, High Confidence and Lift= 1 (Independent items)

From the second case, we obtained 1481 rules. We sorted the results from highest to lowest support. Figure 5.11 shows the top 2 ranked association rules according to the highest support.

```
=====================================
Rule: ['[0.9-1.0]-3'] -> ['[0.9-1.0]-5']
Support: 0.993993993993994
Confidence: 0.9974886991461577
Lift: 1.000995691056767
=====================================
Rule: ['[0.9-1.0]-1'] -> ['[0.9-1.0]-3']
Support: 0.9934934934934935
Confidence: 0.9979889391654096
Lift: 1.0014976898304815
=====================================
```

Figure 5.11: Top two ranked association rules found in our dataset, from the second case.

The first rule says that given that the patient has a GRS between [0.9-1.0]-3 (at SNP rs4765914, which gene is CACNA1c), the frequency of also having a GRS between [0.9-1.0]-5 (at rs1064395, which gene is NCAN) is high due to its high support (99.3%). The confidence is also high (99.7%) due to the high number of transactions. Items are very weakly dependent on each other due to its lift (1.001).

The second rule says that given that the patient has a GRS between [0.9-1.0]-1 (at SNP rs1523041, which gene is ARPP21), the frequency of also having a GRS between [0.9-1.0]-3 (at

SNP rs4765914, which gene is CACNA1c) is high due to its support (99.3%). The confidence is high (99.8%), due to the high number of transactions. Once again, items are very weakly dependent on each other due to its lift (1.001).

Figure 5.12 shows the intersection of both rules. The red circle represents the rule [0.9-1.0]-3 → [0.9-1.0]-5 and the green circle represents the rule [0.9-1.0]-1 → [0.9-1.0]-3.

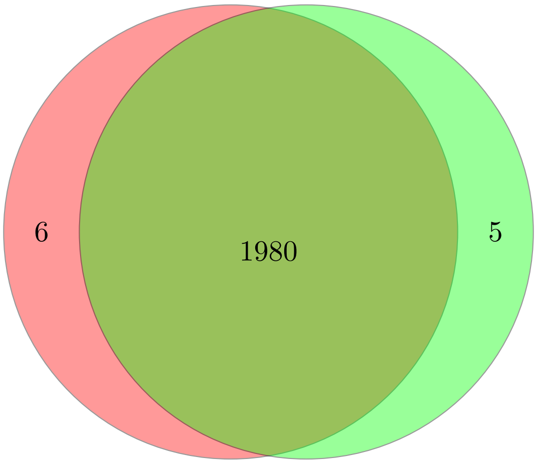


Figure 5.12: Venn-Diagram of the top ranked rules in the second case

Figure 5.12 shows the following:

Group 1 : 1980 patients have all three GRS ([0.9-1.0]-1,[0.9-1.0]-3,[0.9-1.0]-5);

Group 2 : 6 patients have [0.9-1.0]-3, [0.9-1.0]-5 but don't have GRS between [0.9-1.0]-1;

Group 3 : 5 patients have [0.9-1.0]-1, [0.9-1.0]-3 but don't have GRS between [0.9-1.0]-5;

Now we are going to fuse this information with the demographic information. From the first group, we see that from the 1980 patients, 744 of them are male and 1236 are female. Figures 5.13 and 5.14 show the distribution of the age and region of the first group.

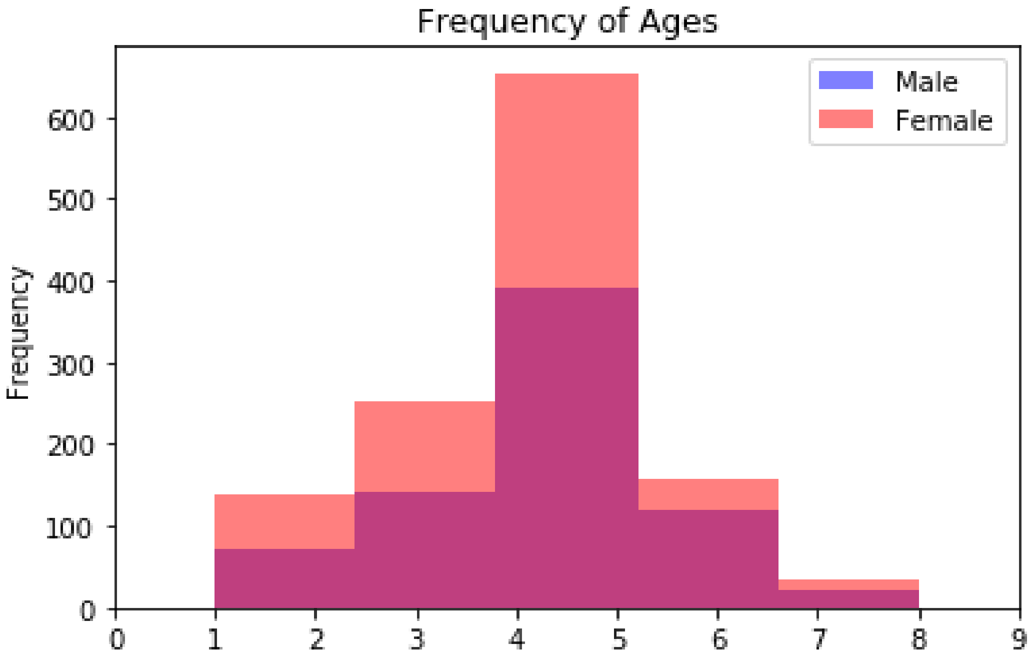


Figure 5.13: Age distribution of patients from the first group in the second case (n= 1980)

Figure 5.14 has the following regions: Midlands (Mid), London, North Midlands (N. Mid), Wales, Southern (South), Southwestern (SW), Scotland (Scot), Northern (North), Southeastern (SE), Northwestern(NW), Eastern (E) and East + West Ridings (E+W).

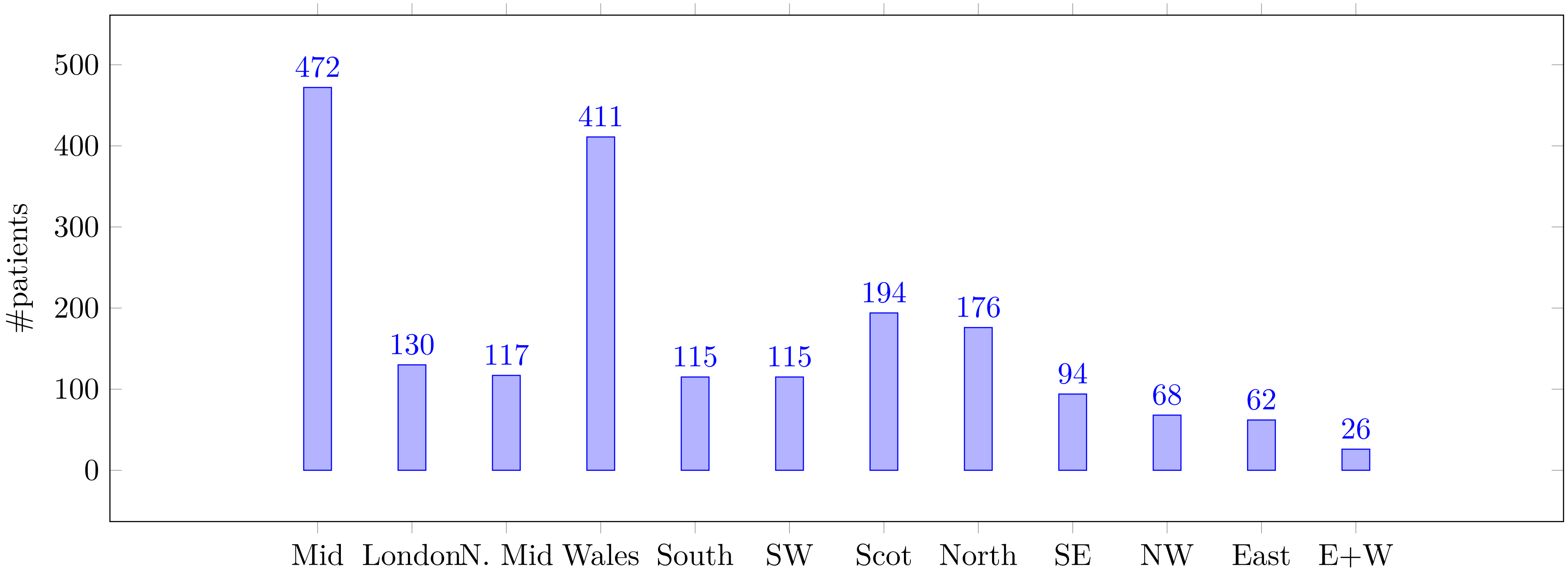


Figure 5.14: Region of the patients from the first group in case 2 (n= 1980)

As we see from Figures 5.13 and 5.14, the most frequent age interval is 40-49 in both genders, and the majority of the patients is from the Midlands.

For the second group, from the 6 patients, 3 patients are female and the other 3 are male. Figures 5.15 and 5.16 show the distribution of age and region for group 2.

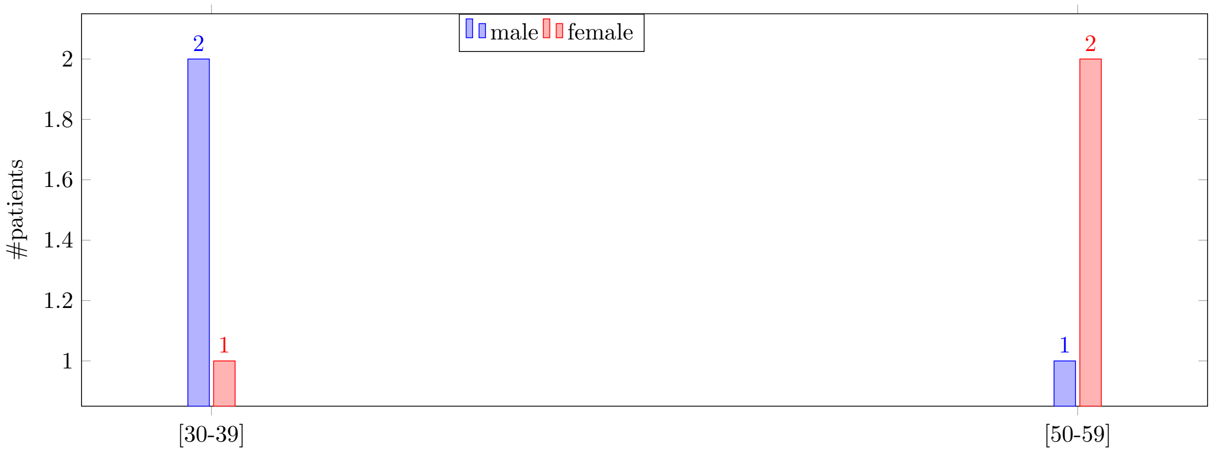


Figure 5.15: Age of the patients from the second group in case 2 (n=6)

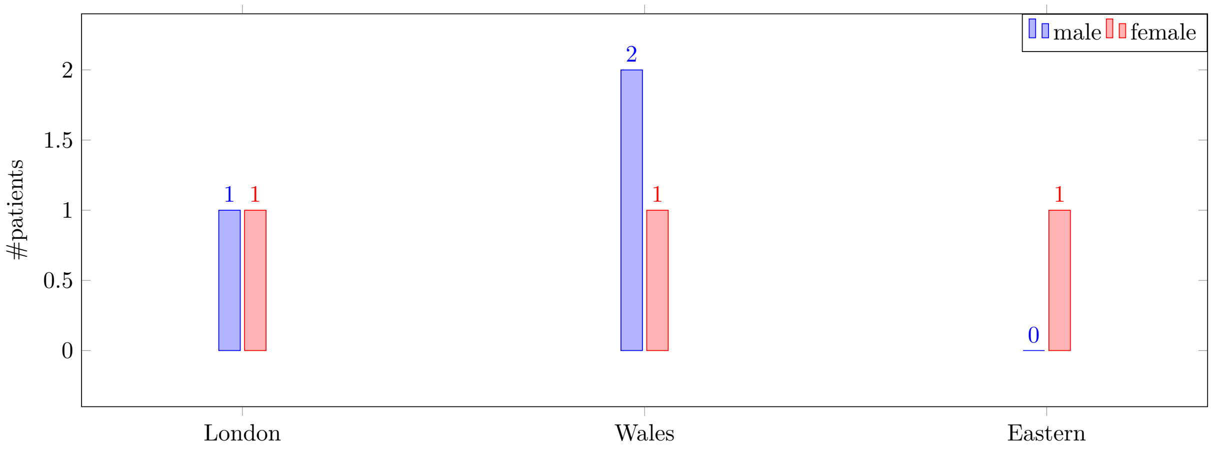


Figure 5.16: Region of the patients from the second group in case 2 (n=6)

Figures 5.15 and 5.16 have a 50/50 distribution in terms of age, however there are more men in their 30's and more women in their 50's.

For the third group we observe that from the 5 patients, 2 patients are female and the other 3 are male.

Figures 5.17 and 5.18 show the distribution of age and region for group 3.

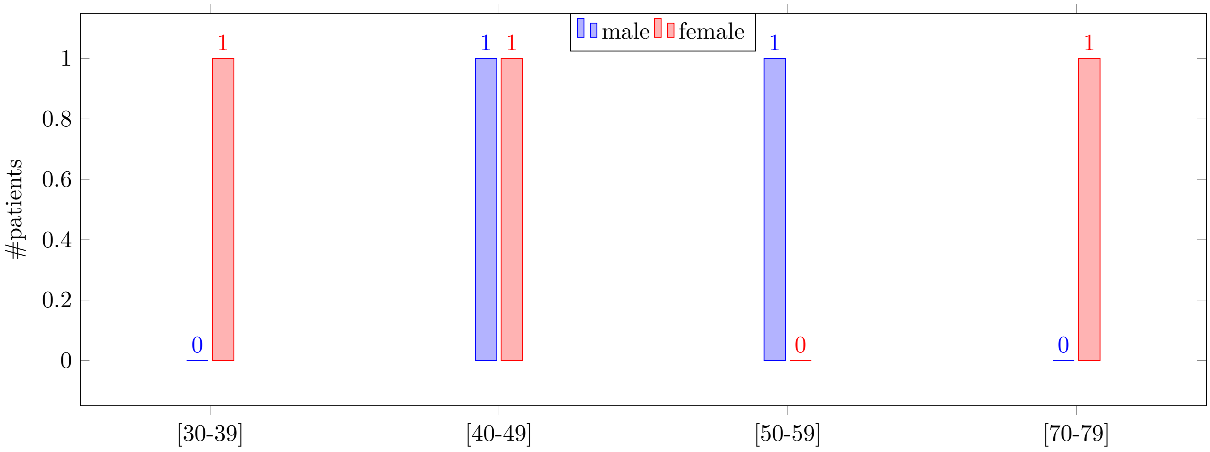


Figure 5.17: Age of the patients from the third group in case 2 (n=5)

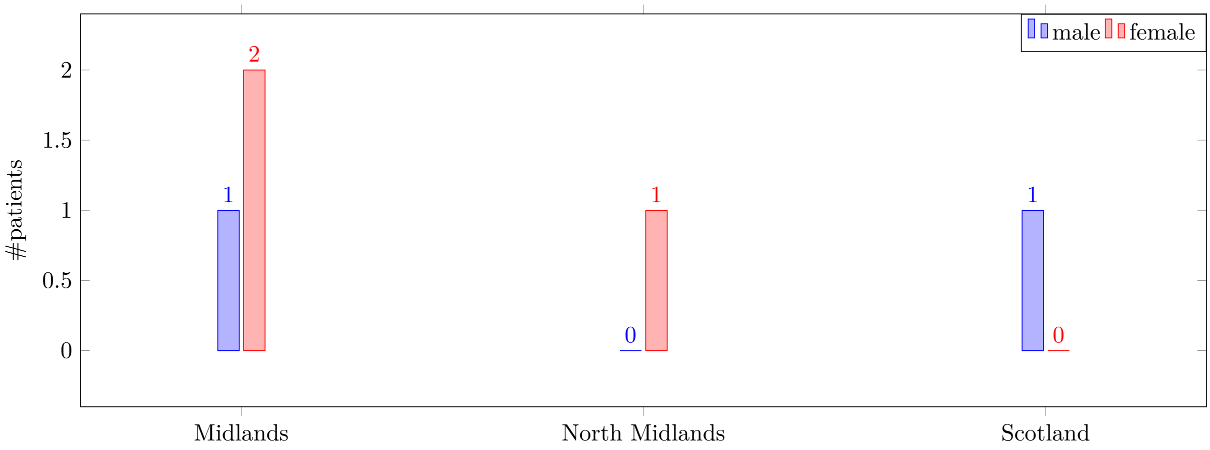


Figure 5.18: Region of the patients from the third group in case 2 (n=5)

From the previous two figures, there is a 50/50% distribution in gender when the patients are in their 40's, however there are cases where there are males in an age interval and no females and vice-versa. The majority of the patients come from the Midlands.

5.4 Validation

We selected the following ten rules from a random fold:

```
1- ['TT-3', 'TT-6'] -> [AA-2]
2- ['TT-3', '[0.9-1.0]-1', 'TT-6'] -> ['AA-2']
3- ['TT-3', '[0.9-1.0]-2', 'TT-6'] -> ['AA-2']
4- ['TT-3', 'TT-6', '[0.9-1.0]-3'] -> ['AA-2']
5- ['[0.9-1.0]-4', 'TT-3', 'TT-6'] -> ['AA-2']
6- ['TT-3', '[0.9-1.0]-5', 'TT-6'] -> ['AA-2']
```

```
7- [’TT-3’, ’TT-6’, ’[0.9-1.0]-6’] -> [’AA-2’]
8- [’TT-3’, ’[0.9-1.0]-2’, ’[0.9-1.0]-1’, ’TT-6’] -> [’AA-2’]
9- [’TT-3’, ’[0.9-1.0]-1’, ’TT-6’, ’[0.9-1.0]-3’] -> [’AA-2’]
10- [’[0.9-1.0]-4’, ’TT-3’, ’[0.9-1.0]-1’, ’TT-6’] -> [’AA-2’]
```

After selecting the rules we did an intersection with all folds and checked if these rules were present in the other folds.

Tables 5.5 and 5.6 show the results while intersecting the ten rules shown above with the training folds and the test folds. X means that the rule is present in that fold, RNP means that rule in that certain fold is not present and ES means that a certain fold is empty.

Table 5.5: Intersection of the 10 rules with the training folds

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
Rule 1	X	X	X	X	X	X	RNP	X	X	X
Rule 2	X	X	X	X	X	X	RNP	X	X	X
Rule 3	X	X	X	X	X	X	RNP	X	X	X
Rule 4	X	X	X	X	X	X	RNP	X	X	X
Rule 5	X	X	X	X	X	X	RNP	X	X	X
Rule 6	X	X	X	X	X	X	RNP	X	X	X
Rule 7	X	X	X	X	X	X	RNP	X	X	X
Rule 8	X	X	X	X	X	X	RNP	X	X	X
Rule 9	X	X	X	X	X	X	RNP	X	X	X
Rule 10	X	X	X	X	X	X	RNP	X	X	X

Table 5.6: Intersection of the 10 rules with the test folds

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
Rule 1	X	X	RNP	RNP	X	ES	X	X	RNP	X
Rule 2	X	X	RNP	RNP	X	ES	X	X	RNP	X
Rule 3	X	X	RNP	RNP	X	ES	X	X	RNP	X
Rule 4	X	X	RNP	RNP	X	ES	X	X	RNP	X
Rule 5	X	X	RNP	RNP	X	ES	X	X	RNP	X
Rule 6	X	X	RNP	RNP	X	ES	X	X	RNP	X
Rule 7	X	X	RNP	RNP	X	ES	X	X	RNP	X
Rule 8	X	X	RNP	RNP	X	ES	X	X	RNP	X
Rule 9	X	X	RNP	RNP	X	ES	X	X	RNP	X
Rule 10	X	X	RNP	RNP	X	ES	X	X	RNP	X

Tables 5.5 and 5.6, show that the rules are present in most folds. Three test folds do not have intersection with our 10 rules, but have always TT-3 and AA-2 together, which is a subset of some of the ten rules. When the score is high ([0.9-1.0]) for folds 4 and 5, the rules indicate

a correlation of TT-3 and TT-6 and AA-2. Moreover the relations are always for homozygous patients.

Chapter 6

Conclusions and Future Work

From the literature, we still observe that Bipolar Disorder (BD) presents a highly heterogeneous clinical course, with not a single gene or a bio-marker as diagnostic tool for this disorder, even with the high heritable component observed within families [34]. Computational methods may help, but there are not many works in the area. Our work contributes towards explaining a little more about this disease by finding differences between homozygous and heterozygous groups and by finding subgroups of patients with combinations of Single-Nucleotide Polymorphism (SNP)s associated with BD. As far as we know this is the first work that applies association rules to BD in an unsupervised learning task.

Our main findings can be summarized as follows. As we studied the behaviour of the Genetic Risk Score (GRS) of each SNP per patient, we observed differences between homozygous and heterozygous patients. There are some cases where even though the majority of the patients have a very high GRS in most of the SNPs, some cases exhibit a very low GRS for a given SNP.

From the apriori algorithm results, in both cases, we found subgroups presenting specific patterns of genetic characteristics in BD, in particular among the TT genotype at SNP rs7680321, the AA genotype at SNP rs9621532, the TT genotype at SNP rs4565914 and the AA genotype at SNP rs1006737. Other associations involve the GRS in the interval [0.9-1.0] in SNPs rs476591, rs106439 and rs1523041, respectively. Because different age groups were clustered by specific SNPs, this would suggest that different age at onset for BD that is observed in the clinical settings would be associated to specific genetic variants. This information allied to clinical evaluation and other patient characteristics may help strength the effort in the precision medicine field devising better treatment.

While intersecting our results with the demographic data we observed that the female gender is more frequent than the male gender [29]. It could be because women usually seek more for medical care than men. Epidemiology data does not show higher prevalence of bipolar on women when compared to men. All of the patients are adults, and most of them older adults, it might be because most of the bipolar cases have the diagnostic of the disease around 30 to 40 years old [28][33]. These are all indications, but not strong conclusions, because we do not have clinical data to validate our conjectures. Future studies could benefit from additional demographic data to draw further conclusions.

As next steps we would like to work with a larger cohort of patients, ideally added by clinical information. Unfortunately, it is not easy to have access to this kind of data. We would also like to contrast our findings with the controls. Wellcome Trust Case Control Consortium (WTCCC) has another cohort with controls that we did not use in this study. It would be also interesting to confirm if the subgroups of patient genetic characteristics highlighted in this study also holds to other BD datasets. Last, we would like to contrast the genetic information of these subgroups with SNPs associated with other mental disorders.

Bibliography

[1] Experiencing mood swings is not bipolar, 2014.

[2] Single nucleotide polymorphism (snp) allele frequency dna pools, 2014.

[3] Bipolar disorder, 2016.

[4] Arpp21 gene, 2017.

[5] Gabrb1 gene, 2017.

[6] Ncan gene, 2017.

[7] Syn3 gene, 2017.

[8] Lift in an association rule, 2018.

[9] WTCCC - data formats and interpretation, 2018.

[10] How to use machine learning for anomaly detection and condition monitoring, 2018.

[11] Experiencing mood swings is not bipolar, 2018.

[12] Cacna1c gene, 2018.

[13] Introduction to k-nearest-neighbors, 2018.

[14] False positives in mood disorders questionnaire screening for bipolar disorder, 2018.

[15] Syne1 gene, 2018.

[16] Global burden of disease, 2019.

[17] Lift in an association rule, 2019.

[18] What are single nucleotide polymorphisms (SNPs)?, 2019.

[19] What is a gene?, 2019.

[20] genotype, 2019.

[21] Heterozygous vs. homozygous differences, 2019.

[22] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN: 1-55860-153-8.

[23] A.Pinheira, C. Nascimento, R. Dias, and I. Dutra. Characterizing the profile of bipolar disorder-associated single nucleotide polymorphisms in a large UK cohort. In *16th International Conference on Computational Intelligence methods for Bioinformatics and Biostatistic (CIBB)*, 2019.

[24] B. Arts, C. J.P. Simons, and J. v. Os. Evidence for the impact of the cacna1c risk allele rs1006737 on 2-year cognitive functioning in bipolar disorder. *Psychiatric Genetics 2013*, 23 (1):41–42, 2013. doi:10.1097/YPG.0b013e328358641c.

[25] M. Bramer. *Principles of Data Mining*. Springer, 2007.

[26] W. S. Bush and J. H. Moore. Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, 8(12), 2012. doi:10.1371/journal.pcbi.1002822.

[27] J. N. Cooke Bailey and R. P. Jr. Igo. Genetic risk scores. *Curr. Protoc. Hum. Genet.*, pages 1.29.1–1.29.9, October 2016. doi:10.1002/cphg.20.

[28] J Dagani, G Signorini, O Nielssen, and et al. 2017;62(4):247–258. doi:. Meta-analysis of the interval between the onset and management of bipolar disorder. *Can J Psychiatry.*, 62(4): 247–258, 2017. doi:10.1177/0706743716656607.

[29] A. Diflorio and I. Jones. Is sex important? gender differences in bipolar disorder. *International Review of Psychiatry*, 22(5):437–452, 2010. doi:10.3109/09540261.2010.514601.

[30] A. Nunes *et al.* for the ENIGMA Bipolar Disorders Working Group. Using structural MRI to identify bipolar disorders – 13 site machine learning study in 3020 individuals from the ENIGMA bipolar disorders working group. *Molecular Psychiatry*, 2018. doi:10.1038/s41380-018-0228-9.

[31] A. J. Ferrari, E. Stockings, J. P. Khoo, H. E. Erskine, L. Degenhardt, T. Vos, and H. A. Whiteford. The prevalence and burden of bipolar disorder: findings from the global burden of disease study 2013. *Bipolar Disorders An International Journal of Psychiatry and Neurosciences*, 18(5):440–450, 2016. doi:10.1111/bdi.12423.

[32] I. P. Gorlov, O. Y. Gorlova, S. R. Sunyaev, M. R. Spitz, and C. I. Amos. Shifting paradigm of association studies: Value of rare single-nucleotide polymorphisms. *The American Journal of Human Genetics*, 82(1):100 – 112, 2008. ISSN: 0002-9297. doi:10.1016/j.ajhg.2007.09.006.

[33] I. Grande, M. Berk, B. Birmaher, and Dr. E. Vieta. Is age of onset associated with severity, prognosis, and clinical features in bipolar disorder? a meta-analytic review. *Bipolar Disorders, an international journal of psychiatry and neurosciences*, 18(5):389–403, 2016. doi:10.1111/bdi.12419.

[34] I. Grande, M. Berk, B. Birmaher, and Dr. E. Vieta. Bipolar disorder. *The Lancet*, 387 (10027):1561–1572, 2016. doi:10.1016/S0140-6736(15)00241-X.

[35] C. Győrödi, R. Győrödi, and S. Holban. A comparative study of association rules mining algorithms. *ResearchGate*, 2004. doi:10.13140/2.1.1450.3365.

[36] T. Hajek, C.Cooke, M. Kopecek, T. Novak, C. Hoschl, and M. Alda. Using structural MRI to identify individuals at genetic risk for bipolar disorders: a 2-cohort, machine learning study. *J Psychiatry Neurosci*, 40(5):316–24, 2015. doi:10.1503/jpn.140142.

[37] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. *SIGMOD Rec.*, 29(2):1–12, May 2000. ISSN: 0163-5808. doi:10.1145/335191.335372.

[38] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining.* The MIT Press, 2001.

[39] F. V. Jensen and T. D. Nielson. *Bayesian Networks and Decision Graphs 2nd Edition.* Springer, 2007.

[40] D. Librenza-Garcia, B. J. Kotzian, J.Yang, B. Mwangi, B. Cao, L. N. P. Lima, M. B. Bermudez, M. V. Boeira, F.Kapczinski, and I. C. Passos. The impact of machine learning techniques in the study of bipolar disorder: A systematic review. *Neuroscience and Biobehavioral Reviews*, 80:538–554, 2017. doi:10.1016/j.neubiorev.2017.07.004.

[41] T. M. Mitchell. *Machine Learning.* McGraw-Hill Science/Engineering/Math, 1997.

[42] M. Mohammed, M. B. Khan, and E. B. M. Bashier. *Machine Learning Algorithms and Applications.* CRC Press, 2016.

[43] L. M. M. Moreira. Extraction of a bipolar disorder associated genetic pattern. Master's thesis, Faculdade de Ciências da Universidade do Porto, 2018.

[44] G. Orrù and M. G. Carta. Genetic variants involved in bipolar disorder, a rough road ahead. *Clinical Practice & Epidemiology in Mental Health*, 14:37–45, 2018. doi:10.2174/1745017901814010037.

[45] T.A. Rowland and S. Marwaha. Epidemiology and risk factors for bipolar disorder. *Therapeutic Advances in Psychopharmacology*, 8(9):251–269, September 2018.

[46] S. Russell and P. Norvig. *Artificial Intelligence, A Modern Approach.* Pearson Education, Inc., 2010.

[47] C. C. Saylan and K. Yilancioglu. Classification of schizophrenia and bipolar disorder by using machine learning algorithms. *The Journal of Neurobehavioral Sciences*, 3(3):92–95, 2016. doi:10.5455/JNBS.1471026038.

[48] F. Schwenker, H.M. Abbas, N. El Gayar, and E. Trentin. *Artificial Neural Networks in Pattern Recognition.* Springer, 2010.

[49] V. Sharma, V. Bergink, M. Berk, P.S. Chandra, T. Munk-Olsen, A.C. Viguera, and L.N. Yatham. Childbirth and prevention of bipolar disorder: an opportunity for change. *Lancet Psychiatry*, April 2019. doi:10.1016/S2215-0366(19)30092-6.

[50] Ms. Shweta and Dr. K. Garg. Mining efficient association rules through apriori algorithm using attributes and comparative analysis of various association rule algorithms. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3:306–312, 2013. ISSN: 2277 128X.

[51] S.Taheri and M. Mammadov. Learning the naive bayes classifier with operation models. *Int. J. Appl. Math. Comput. Sci.*, 23(4):787–795, 2013. doi:10.2478/amcs-2013-0059.

[52] L. Torgo. *Data Mining with R Learning with Case Studies.* CRC Press, 2011.

Appendix A

GRS for all patients per SNP

A.1 General GRS

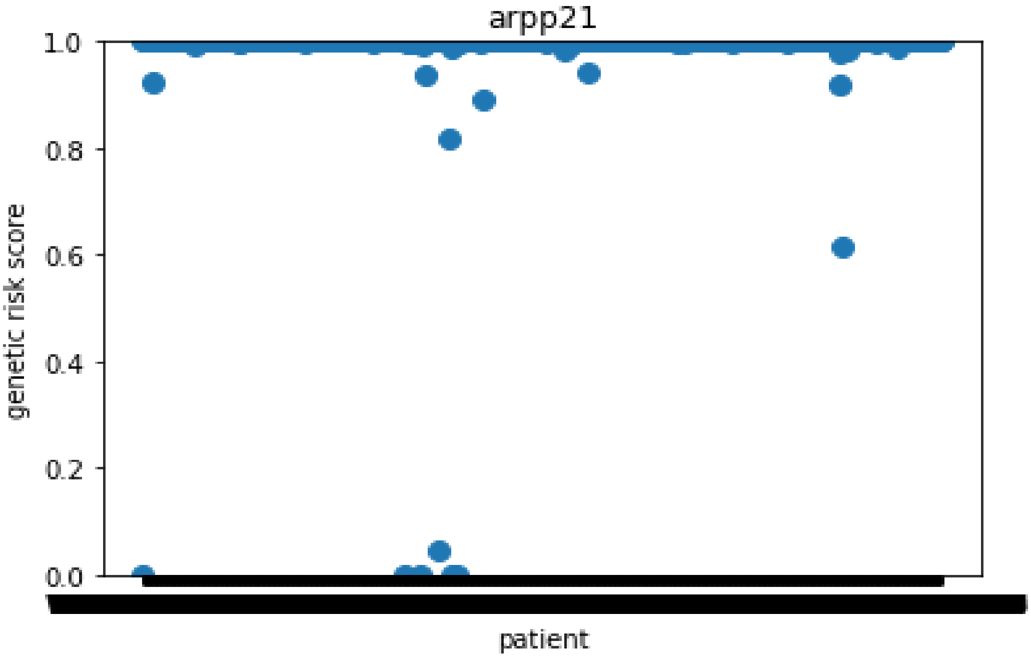


Figure A.1: Genetic Risk Score of patients with ARPP21

A.1.0.1 CACNA1c - rs4765914

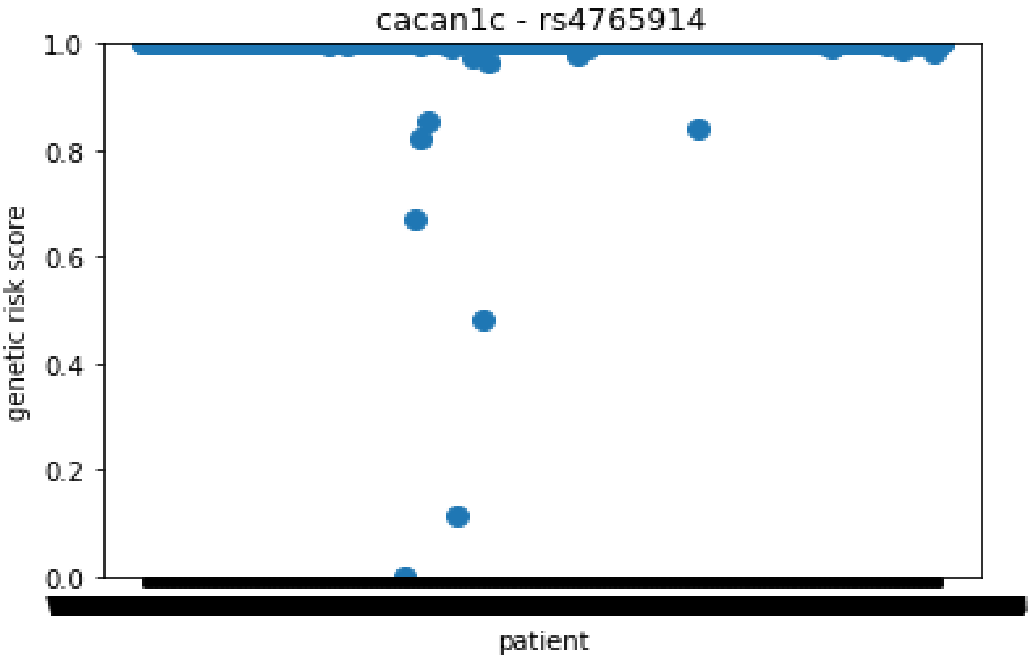


Figure A.2: Genetic Risk Score of patients with CACNA1c - rs4765914

A.1.0.2 SYNE1 (rs9371601)

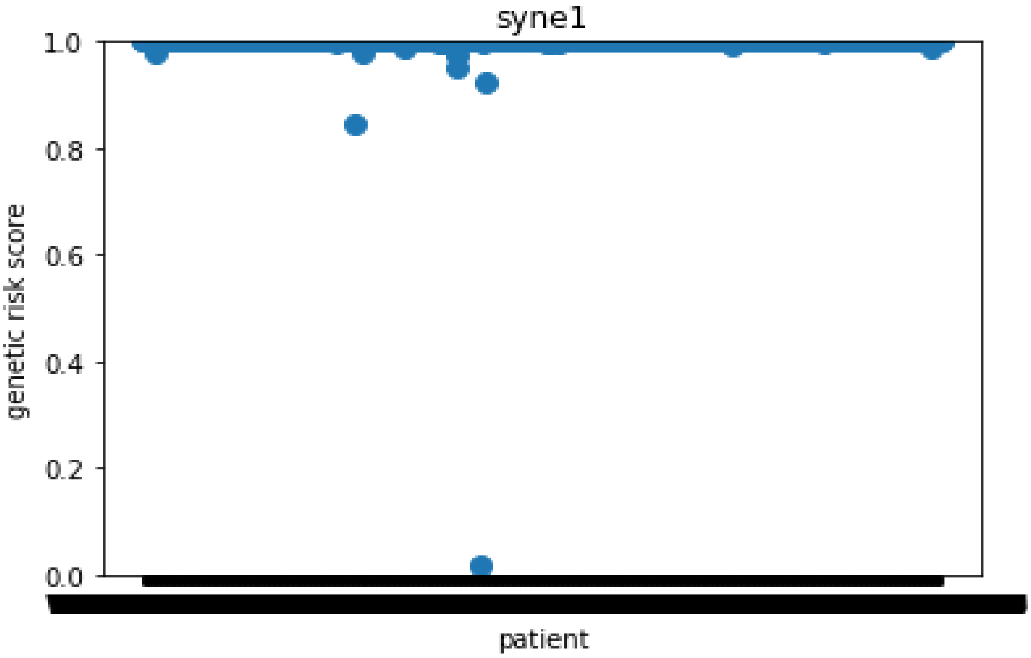


Figure A.3: Genetic Risk Score of patients with SYNE1

A.1.0.3 NCAN (rs1064395)

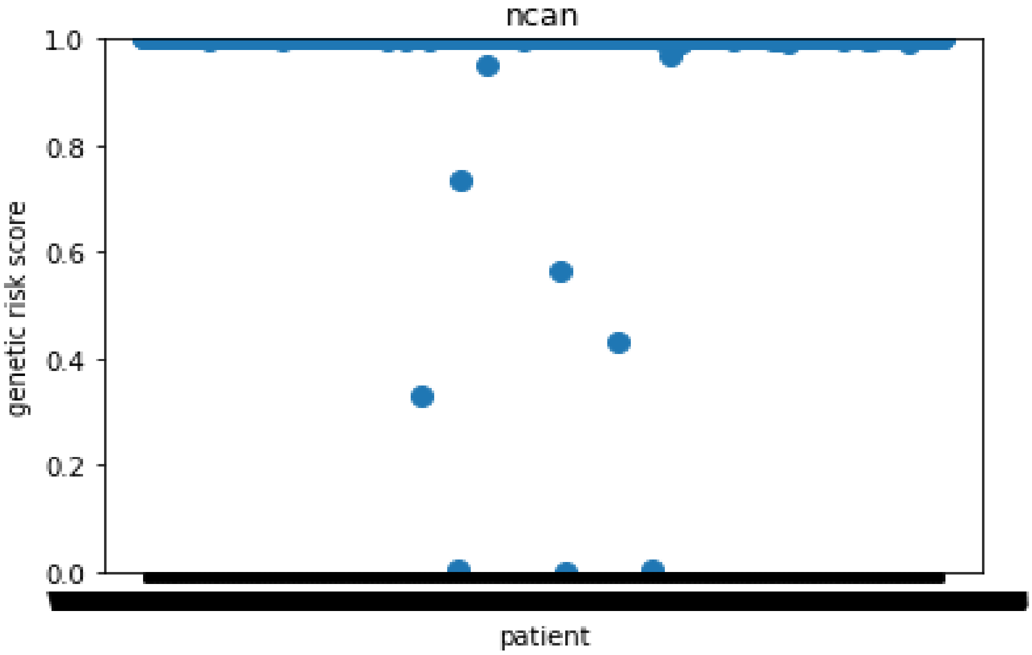


Figure A.4: Genetic Risk Score of patients with NCAN

A.1.0.4 GABRB1 (rs7680321)

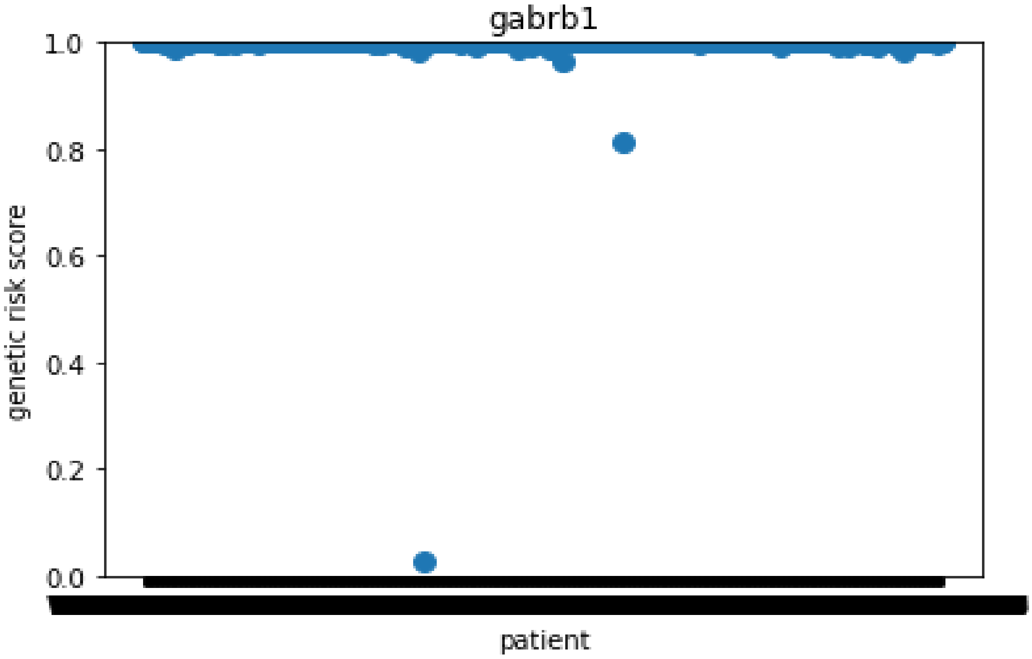


Figure A.5: Genetic Risk Score of patients with GABRB1

A.1.0.5 SYN3 (rs9621532)

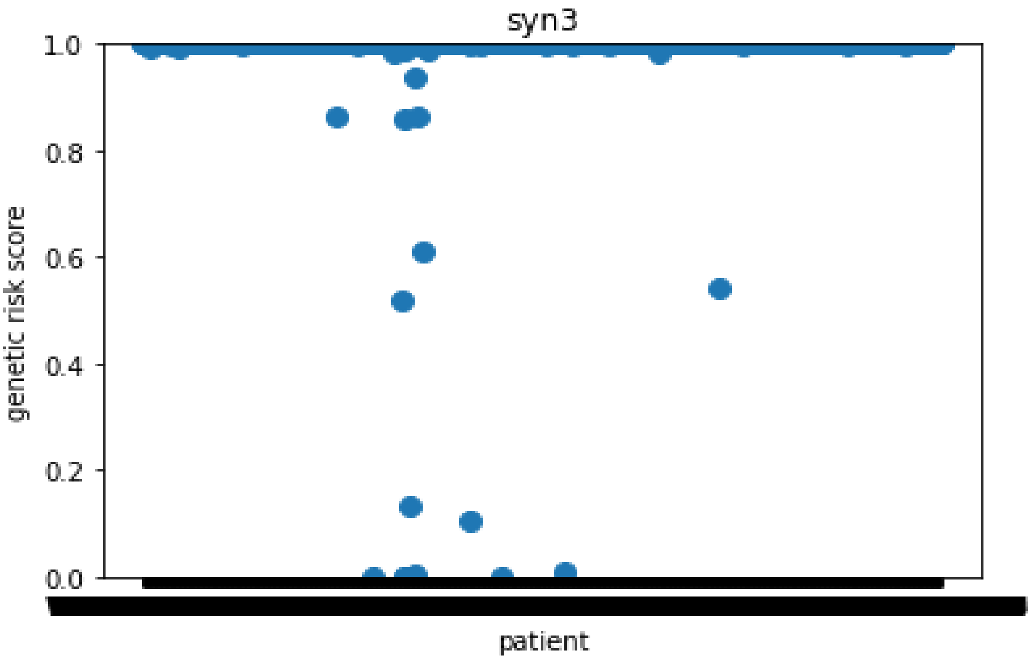
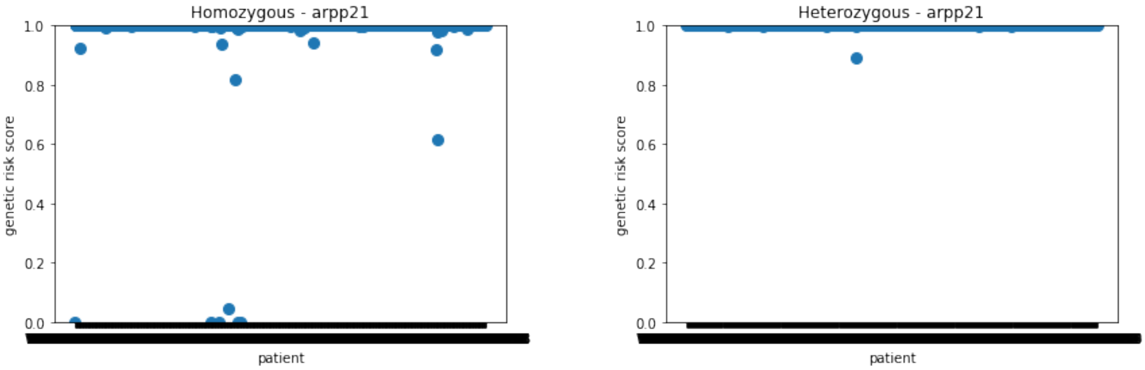


Figure A.6: Genetic Risk Score of patients with SYN3

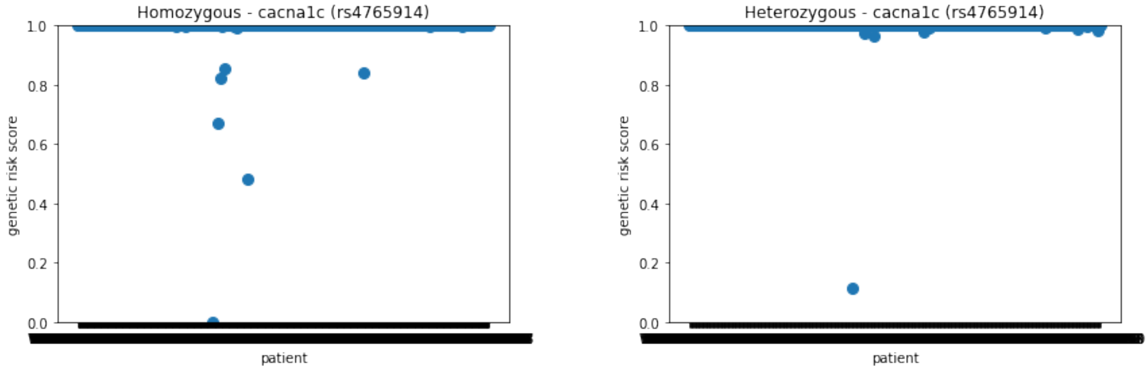
A.2 Homo vs. hetero GRS



(a) Homozygous patients

(b) Heterozygous patients

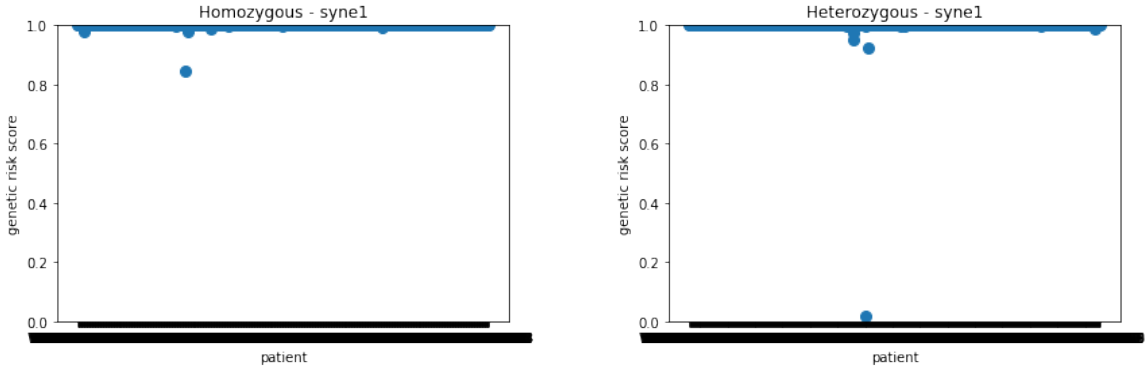
Figure A.7: GRS of Homozygous and Heterozygous patients with ARPP21



(a) Homozygous patients

(b) Hetrozygous patients

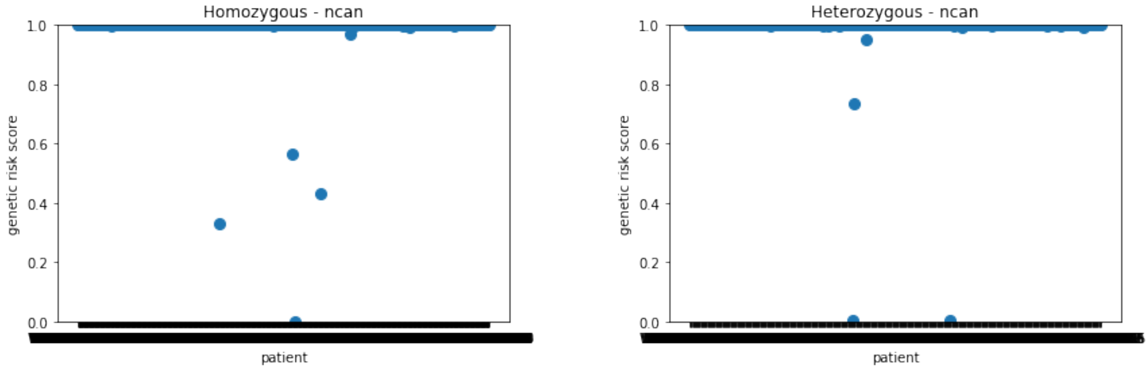
Figure A.8: GRS of Homozygous and Heterozygous patients with CACNA1c (rs4765914)



(a) Homozygous patients

(b) Heterozygous patients

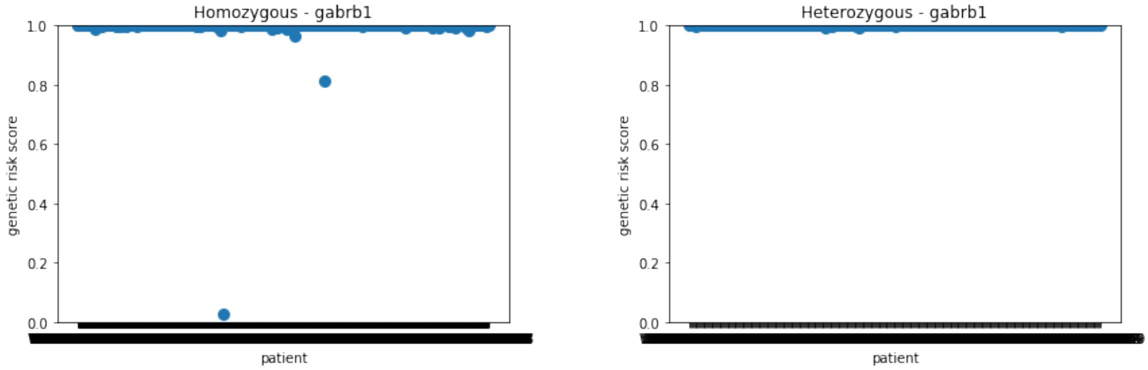
Figure A.9: GRS of Homozygous and Heterozygous patients with SYNE1



(a) Homozygous patients

(b) Heterozygous patients

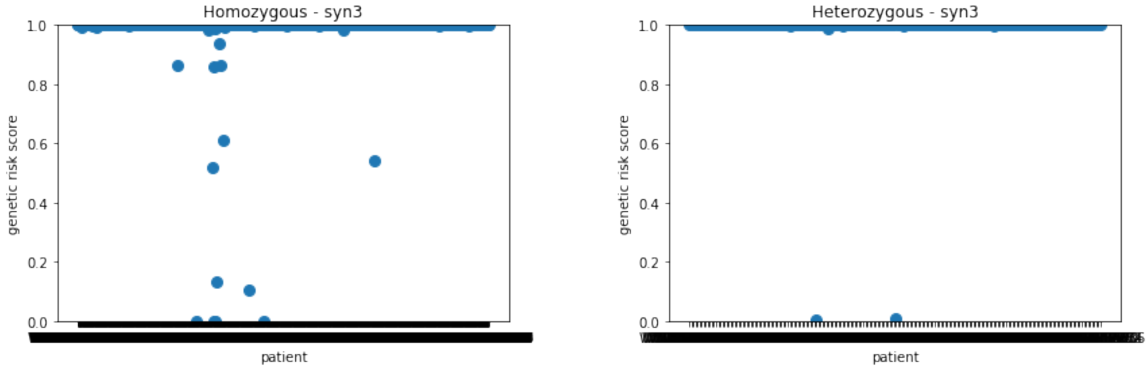
Figure A.10: GRS of Homozygous and Heterozygous patients with NCAN



(a) Homozygous patients

(b) Heterozygous patients

Figure A.11: GRS of Homozygous and Heterozygous patients with GABRB1



(a) Homozygous patients

(b) Heterozygous patients

Figure A.12: GRS of Homozygous and Heterozygous patients with SYN3