Fakultät für Medizin

Subtyping psychiatric disorders using unsupervised learning methods

Helena Pelin

Vollständiger Abdruck der von der Fakultät für Medizin der Technischen Universität München zur Erlangung des akademischen Grades einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Josef Priller
Prüfende der Dissertation:	1. apl. Prof. Dr. Bertram Müller-Myhsok
	2. Prof. Dr. Julien Gagneur

Die Dissertation wurde am 26.08.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Medizin am 17.05.2022 angenommen.

Contents

Acknowledgments

> "Cultivate the habit of being grateful for every good thing that comes to you, and to give thanks continuously. And because all things have contributed to your advancement, you should include all things in your gratitude."
>
> *Ralph Waldo Emerson*

I would like to thank my supervisor, Prof. Bertram Müller-Myhsok, for giving me the opportunity to work in his group, for providing a healthy environment and enough intellectual freedom to conduct this work. Most of all, I would like to thank him for being very supportive and encouraging along the way. Moreover, I would like to thank all the former and current members of the Statistical genetics group, who were always open to give advice, help, discuss and who simply enriched my PhD journey and made me happy. A special thanks goes to the past group member, Dr. Till Andlauer, who helped me a lot to publish my first scientific paper, the key pillar of this thesis. Furthermore, I am also thankful to the members of my Thesis committee, Prof. Dr. Karsten Borgwardt and Prof. Dr. Julien Gagneur, for the support and useful discussions throughout the years.

I would like to acknowledge all my colleagues from the Max Planck Institute of Psychiatry who in any way made this journey fulfilling. Thank you for the enjoyable occasional coffees, lunches, discussions, activities, and most importantly for your helpfulness. Here, I would like to especially thank Dr. Marcus Ising, for his support and the knowledge he shared during the project. I would also like to thank all the external colleagues I collaborated with and who helped me achieve my goal.

My deepest gratitude goes to my parents who supported me throughout my whole life and have always been there for me. I also wish to thank the other members of my closest family, my brother and my grandparents, for the incredible support and care. Finally, I would like to thank the family I got along the way - my partner, who has supported me in this endeavor from the very first moment and has stayed on the journey with me.

I am very happy to have had the opportunity like this and I feel very grateful for everyone who was/is part of my life even for just a short bit. Thank you.

Abstract

Psychiatric disorders are highly heterogeneous regarding their symptoms and disease course. Research evidence has shown that the existing psychiatric nosology might not accurately reflect the biological processes, as well as different clinical manifestations and outcomes of the mental disorders.

The aim of this thesis was to apply computational methods to generate new knowledge on the classification of psychiatric diseases. To this end, we crossed the existing diagnostic boundaries and analyzed a transdiagnostic sample including healthy controls and patients diagnosed with depression, schizoaffective disorder, schizophrenia, bipolar disorder, and other psychiatric disorders such as anxiety and social phobia. High-dimensional data clustering was applied on a discovery sample of N=1250 individuals using a wide range of clinical variables. Supervised learning was further used to describe clusters based on genetic variables, polygenic risk scores, and family history. Alongside identifying new subtypes, we explored the performance of a feature selection algorithm in order to deal with extremely high-dimensional genetic data (single nucleotide polymorphisms (SNPs)) to better describe the differences between the identified subtypes or formal diagnostic categories.

Five diagnostically mixed clusters were identified and ranked based on a continuous severity scale: from 0, indicating a well-being or the lowest severity, to 4, indicating the highest severity. Cluster 0 contained most of the healthy controls and showed general well-being. Cluster 4, on the other end of the scale, contained most of the patients diagnosed with psychotic disorders and showed the highest severity in many examined measures. The Clusters 1–3 varied predominantly regarding depression levels, quality of life, parental bonding, and childhood maltreatment levels. Differences in polygenic risk scores and family

history were the strongest between the two extreme clusters 0 and 4. Moreover, we showed that the inclusion of polygenic risk scores into the model with family history might result in better individual predictions. In the replication analysis in a sample of N=622 individuals, all except for the smallest Cluster 1 replicated, which shows the stability of the cluster solution. The identified clusters and their characteristics show the importance of transdiagnostic approaches, emphasizing the need for symptom-specific rather than diagnosis-specific treatment.
The results of feature selection analysis showed poor generalizability of the prediction model and the unstable sets of SNPs chosen across different runs, both when predicting the identified Cluster 0, as well as the formal diagnostic label. This led us to conclude that more work is needed to develop methods that could capture the high degree of polygenic dependencies and a continuum of small effects present in psychiatric disorders and traits.

This thesis suggests that the data-driven approaches in psychiatry offer great advantages to the field, by uncovering the patterns and relations hidden in the data available nowadays. We demonstrated that performing the transdiagnostic clustering and assessing the level of symptom severity may identify groups of patients that share clinical symptoms and, hence, could benefit from similar treatments. Such approaches might contribute to a deeper understanding of the heterogeneity between and within psychiatric disorders and support the development of individualized treatment regimes.

Zusammenfassung

Psychiatrische Erkrankungen sind in Bezug auf Symptome und Krankheitsverlauf sehr heterogen. Forschungsergebnisse haben gezeigt, dass die bestehende psychiatrische Nosologie die biologischen Prozesse sowie die verschiedenen klinischen Manifestationen und Ergebnisse der psychischen Störungen möglicherweise nicht genau widerspiegelt.

Das Ziel dieser Arbeit war es, computergestützte Methoden anzuwenden, um neue Erkenntnisse auf dem Gebiet der Klassifikation psychiatrischer Erkrankungen zu generieren. Zu diesem Zweck haben wir uns darauf konzentriert, die bestehenden diagnostischen Grenzen zu überschreiten und eine transdiagnostische Stichprobe zu analysieren, die gesunde Kontrollpersonen und Patienten mit diagnostizierten Depressionen, schizoaffektiven Störungen, Schizophrenie, bipolaren Störungen und anderen psychiatrischen Störungen wie Angst und soziale Phobie umfasst. Hochdimensionales Datenclustering wurde verwendet, um die Cluster unter Verwendung einer breiten Palette klinischer Variablen zu identifizieren. Überwachtes Lernen wurde verwendet, um Cluster mit genetischen Variablen, polygenen Risikoscores und Familienanamnese zu beschreiben. Neben der Identifizierung neuer Subtypen untersuchten wir die Durchführung eines Feature selection Algorithmus mit dem Ziel, einen Umgang mit extrem hochdimensionalen genetischen Daten (single nucleotide polymorphisms (SNPs)) zu finden, um die Unterschiede zwischen den identifizierten Subtypen oder formalen diagnostischen Kategorien besser zu beschreiben.

Die Clusteranalyse einer Entdeckungsstichprobe von N=1250 Individuen identifizierte fünf diagnostisch gemischte Cluster, die entlang einer kontinuierlichen Schweregradskala eingestuft wurden. Cluster 0 enthielt die meisten gesunden Kontrollen und zeigte das allgemeine Wohlbefinden. Cluster 4 am anderen

Ende der Skala enthielt die meisten Patienten mit diagnostizierten psychotischen Störungen und wies bei vielen beobachteten Messgrößen den höchsten Schweregrad auf. Die Cluster 1–3 lagen zwischen diesen beiden Extremen und unterschieden sich hauptsächlich in Bezug auf das Ausmaß der Depression, die Lebensqualität, die elterliche Bindung und das Ausmaß der Misshandlung in der Kindheit. Die Unterschiede in den polygenen Risikoscores und der Familienanamnese waren zwischen den beiden Extremclustern 0 und 4 am stärksten. Darüber hinaus zeigten wir, dass die Einbeziehung polygener Risikoscores in das Modell mit Familienanamnese zu besseren individuellen Vorhersagen führen könnte. In der Replikationsanalyse in einer Stichprobe von N=622 Individuen, alle außer dem kleinsten Cluster 1 repliziert. Die identifizierten Cluster und ihre Charakteristika zeigen die Bedeutung transdiagnostischer Ansätze und betonen die Notwendigkeit einer symptomspezifischen statt einer diagnosespezifischen Behandlung.
Die Ergebnisse der Merkmalsauswahlanalyse zeigten eine schlechte Generalisierbarkeit des Vorhersagemodells und die instabilen Sätze von SNPs, die über verschiedene Durchläufe hinweg ausgewählt wurden, sowohl bei der Vorhersage des identifizierten Clusters 0 als auch der formalen diagnostischen Markierung. Dies legt nahe, dass mehr Arbeit erforderlich ist, um Methoden zu entwickeln, die den hohen Grad an polygenen Abhängigkeiten und ein Kontinuum kleiner Effekte bei psychiatrischen Störungen und Merkmalen erfassen können.

Die Arbeit in dieser Dissertation legt nahe, dass die datengetriebenen Ansätze in der Psychiatrie große Vorteile für das Feld bieten, indem sie Muster und Zusammenhänge aufdecken, die in den heute verfügbaren Daten verborgen sind. Wir haben gezeigt, dass die Durchführung des transdiagnostischen Clusterings und die Bewertung des Schweregrads der Symptome, Patientengruppen identifizieren können, die klinische Symptome teilen und daher von ähnlichen Behandlungen profitieren könnten. Solche Ansätze könnten zu einem tieferen Verständnis der Heterogenität zwischen und innerhalb psychiatrischer Erkrankungen beitragen personalisierter Behandlungen unterstützen.

List of Figures

List of Tables

Acronyms

AAO	Age at onset
AC	ancestry component
ACE	adverse childhood experience
ADHD	Attention deficit/hyperactivity disorder
AIC	Akaike information criterion
ASD	Autism spectrum disorder
AUC	Area under the (ROC) curve
BALD	Blockwise Approach using Linkage Disequilibrium
BD	Bipolar disorder
BDI	Beck Depression Inventory
BMI	Body mass index
CD	cross-disorder
CS	continuous shrinkage
CTQ	Childhood trauma questionnaire
DNA	deoxyribonucleic acid.
DSM	Diagnostic and Statistical Manual of Mental Disorders
EA	Educational attainment
F-SozU	Fragebogen zur sozialen Unterstützung
FEB	Fragebogen zur elterlichen Bindung
FN	False negative
FOR2107	Forschergruppe 2107

FP	False positive
FWER	family-wise error rate
GAF	Global assesment of functioning
GMM	Gaussian mixture model
GWAS	Genome-wide association study
HAM-D	Hamilton Depression Rating Scale
HAMA	Hamilton Anxiety Rating Scale
HC	Healthy control
HDDA	High Dimensional Discriminant Analysis
HDDC	High dimensional data clustering
ICD	International Classification of Diseases
ICL	Integrated Completed Likelihood
IQ	Intelligence Quotient
LD	Linkage disequilibrium
LDA	Linear Discriminant Analysis
LEQ	Life Events Questionnaire
MDD	Major depressive disorder
NEO-FFI	NEO Five-Factor Inventory
n.s.	not significant
OCD	Obsessive-compulsive disorder
PCA	Principal component analysis
PGS	Polygenic risk score
PSS	Perceived Stress Scale
QDA	Quadratic Discriminant Analysis
ROC	receiver operator characterics

RS	Resilience scale
RSQ	Relationship scales questionnaire
RSS	Residual Sum of Squares
SANS	Scale for Assessment of Negative Symptoms
SAPS	Scale for Assessment of Positive Symptoms
SCID	Structured clinical interview for DSM
SCL	Symptom Checklist
SCZ	Schizophrenia
SD	standard deviation
SF36	Short Form Survey (36 items)
SHAPS	Snaith–Hamilton Pleasure Scale
SNP	Single nucleotide polymorphisms
SPQ-B	Schizotypal Personality Questionnaire (brief)
STAI	State-Trait Anxiety Inventory
STAI-S	State-Trait Anxiety Inventory - State
STAI-T	State-Trait Anxiety Inventory - Trait
SZA	Schizoaffective disorder
TN	True negative
TP	True positive
VLMT	Verbal Learning and memory test
W-Y	Westfall and Young (method)
WHO	World Health Organization
WS	window size
YMRS	Young mania rating scale

1 Introduction

1.1 Background and research motivation

Mental health is defined by the World Health Organization (WHO) as *"a state of well-being in which every individual realizes his or her own potential, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to her or his community"*. However, mental disorders pose a tremendous global burden, since 30% or more people experience a psychiatric illness during their lifetime (Steel et al., 2014). Their etiology is multifactorial, arising not only from the individual attributes, such as genetic predisposition, habits, behavior but also from the psychosocial, environmental, and cultural processes. Yet, it is difficult to determine any factor which is either necessary or sufficient for the development of any formal psychiatric disorder (Fried and Robinaugh, 2020). Over the years, the expert panels have put enormous efforts into defining the essential criteria to develop a scientifically based classification of psychiatric disorders (Kendler, 2009; Shorter, 2015).

The first attempts began in Europe in the late 19th and early 20th century. In that time, many experienced diagnosticians made a wide range of assumptions about which important features would constitute psychiatric disorders (Kendler, 2009). One of the proposing authors at the time, German psychiatrist Emil Kraepelin, had a significant impact on the development of psychiatry and, thus, set the foundation for the development of the Diagnostic and Statistical Manual of Mental Disorders (DSM) published in 1952 (DSM-I). The current, fifth edition of the manual (DSM-V) was published in 2013 (American Psychiatric Association, 2013). After extensive revision, it has grown to 947 pages and 541

diagnostic categories, compared to the 132 pages and 128 diagnostic categories in DSM-I (Blashfield et al., 2014).

The DSM, together with the International Classification of Diseases (ICD) (Khoury et al., 2017), maintained by the WHO, have evolved into the standard classification systems that define how mental health problems are diagnosed worldwide (Dalgleish et al., 2020). Due to the "criteria checklist" approach, DSM contributed toward one common international language for defining and conceptualizing psychiatric disorders. Theoretically, with such an approach, it is necessary to check the clinical features against the list of criteria to make a diagnosis. As a result of such diagnostic standardization, diagnostic reliability was increased. Moreover, publicly accessible diagnostic definitions fostered not only the communication between clinicians but also facilitated the communication with patients (Owen, 2014; Helzer et al., 2006). While DSM and categorical diagnostic criteria have certainly resulted in many improvements, there has been a widening recognition among psychopathology researchers that there is a need to revise the current nosology. As new knowledge and insights into the etiology and biology of the disorders emerge, it became more evident that the discrete categories do not cleanly map to the complexity of mental health (Barch, 2020). As a result, the current classification system has been continuously questioned. In the following, we will discuss some of the main challenges which are driving the crisis of confidence the current psychiatric nosology dived into. They are elaborated in more detail in Dalgleish et al., 2020; Zachar and Kendler, 2017, and Owen, 2014, which are all references for the next part of this section, if not stated otherwise.

The first challenge is an assumption that the domain of psychopathology could be better described with *dimensional models*. Within the current nosology, symptoms are grouped, marked as *present* or *absent* and some are put on the severity scale as *mild*, *moderate* or *severe*. Therefore, it is assumed that there is a qualitative difference between normal mood and abnormal mood. However, it is argued that mental health exists on a continuum of symptom experience, ranging from health on one end to illness on the other, as opposed to these distinct categorical entities (Seow et al., 2017). Indeed, it is known that mental health

is influenced by the various interactions between many different processes such as biology, environmental and social factors which are all regulated by an individual's lifelong experiences. Some individuals might not satisfy the criteria for a diagnosis of a specific psychiatric disorder, but they would exhibit some symptoms associated with it throughout their lifetime and potentially be at risk for developing a specific disorder (Seow et al., 2017; Johns and Os, 2001).

The second challenge is the *heterogeneity* of the disorders – the same disorder may be caused by various underlying mechanisms, and result in many different outcomes. The two individuals with the same diagnosis may have very different clinical presentations, treatment responses, and their experience may result in many different outcomes. It is argued that this kind of heterogeneity is incorporated in the diagnostic criteria, which consist of a group of conceptually similar symptoms. The criteria are met when one or more of these symptoms is present, and the final diagnosis is then dependent on a certain number of criteria that are satisfied. For example, to be diagnosed with major depressive disorder (MDD), a patient should satisfy at least five of eleven symptoms, including one of the two essential ones. Therefore, not all symptoms have to be satisfied and consequently, individuals with MDD diagnosis could have only a few symptoms in common, which introduces the heterogeneity within the same category.

The third challenge is high *comorbidity*, the occurrence of symptoms that meet the criteria for more than one formal diagnosis at the time. For example, 60% of the people with an anxiety disorder, investigated by Goldstein-Piekarski et al., 2016, also had depression or another additional anxiety diagnosis. The frequent co-occurrence of diagnoses emphasizes how uncommon it is to have a single, clearly differentiable clinical presentation (Kessler et al., 2005), and has been regarded as evidence against discrete disease entities (Maj, 2005). In line with this, there is gathering research support of symptomatic ambiguity, heterogeneity, and shared neurobiological and genetic underpinnings between different disorders (Lee et al., 2019a). For example, bipolar disorder (BD) and schizophrenia (SCZ) partly share clinical symptoms like psychosis and have overlapping genetic and neurobiological underpinnings (American Psychiatric

Association, 2013; Lichtenstein et al., 2009; American Psychiatric Association, 2013). This issue of setting the discrete, solid boundaries between the disorders has been addressed by the DSM itself, saying:

> *"There is no assumption that each category of mental disorder is a completely discrete entity with absolute boundaries dividing it from other mental disorders or from no mental disorder. There is also no assumption that all individuals described as having the same mental disorder are alike in all important ways"*. (American Psychiatric Association, 2013)

All the challenges the current classification system faces have motivated many researchers to aim for a revision and reformulation of the current psychiatric nosology. The next section will give an overview of the research done in the field and how this thesis builds on it.

1.2 Research overview

To better understand the complexity of psychiatric disorders and disease etiology, data-driven approaches have been used for some time now. In specific, the clustering methods have emerged as the dominant approach to partition the heterogeneous diagnostic categories and divide them into more homogeneous and clinically relevant subgroups. The development of technologies for assessing many aspects of chemical and biological diversities and the advances in analytical methods such as machine learning and statistics have been the reasons for renewed interest in employing clustering approaches in psychiatry nowadays (Marquand et al., 2016).

Studies that attempted to identify psychiatric disorder subtypes could be roughly divided into two groups – *single disorder subtyping studies*, trying to refine the definitions of one specific diagnostic category, and *transdiagnostic studies* focusing on finding cross-disorder subtypes, hence going beyond the existing diagnostic boundaries. So far, researchers have been dominantly focusing on the former. However, with mounting evidence of shared etiology between different disorders, the transdiagnostic studies have started to gain importance

(Barch, 2020; Dalgleish et al., 2020).

Many single disorder subtyping studies focus their attention on major depressive disorder, as it contributes significantly to the overall global disease burden and the increasing levels of mortality and morbidity. The review by Beijers et. al (Beijers et al., 2019) provides an interesting overview of many attempts to find more homogeneous subgroups of MDD by analyzing neuroimaging, psychopathology, genetics, or a combination of multiple domains. Nonetheless, efforts to refine other illnesses are manyfold, including schizophrenia (Geisler et al., 2015; Dwyer et al., 2018; Dickinson et al., 2017; Helmes and Landmark, 2003; Jablensky, 2006; Bell et al., 2011), autism spectrum disorder (ASD) (Giambattista et al., 2019; Ring et al., 2008; Veatch et al., 2013), attention-deficit/hyperactivity disorder (ADHD) (Mostert et al., 2018; Gates et al., 2014), eating disorders (Forbush et al., 2017; Grilo et al., 2002)...

While finding subtypes of a single disorder is important for a better understanding of it, these approaches cannot capture the overlap of symptoms and shared genetic and neurobiological underpinnings between different disorders, mentioned previously in this chapter. A promising avenue to tackle this issue has been put forward by the transdiagnostic approaches. Existing cross-disorder clustering studies support the existence of diagnostically diverse subtypes, either across two (Chan et al., 2017) or more disorders (Crouse et al., 2020; Dwyer et al., 2020; Grisanzio et al., 2018; Lewandowski et al., 2014). However, as pointed out in the review by Fusar-Poli et. al (Fusar-Poli et al., 2019), transdiagnostic studies are still limited in the number of observed disorders and often characterized by methodological weaknesses, such as small samples, biased models, and lack of validation.

Despite intensifying efforts to improve the current classification system of psychiatric disorders, there is still a long way to go. As nicely brought up by M. Maj (Maj, 2018), if we want to find a better categorization of psychiatric disorders, we need to be aware that diagnosing an individual with the specific disorder or the disorder subtype should only be the first step on the path to outcome prediction and personalized treatment regime. The other equally important

step is a detailed pathophysiological and molecular characterization of these individuals, including an assessment of clinical severity. Moreover, the detailed understanding of the single components that may drive the development of psychiatric disorders is not sufficient for the complete picture. Researches need to go further and also study the complex interactions among them (Fried and Robinaugh, 2020). So far, the vast majority of subtyping studies, irrespective of the approach they took, focused their clustering analysis on single data domains, such as neuroimaging (Drysdale et al., 2017; Cheng et al., 2014; Gould et al., 2014; Kaczkurkin et al., 2019; Dias et al., 2015; Sun et al., 2015), psychometric (Chan et al., 2017; Maglanoc et al., 2018; Fountain et al., 2012; Bell et al., 2011), biochemical markers (Haroon et al., 2018), and genetics (Yu et al., 2017; Howard et al., 2020). The field should focus on getting the holistic picture of an individual and try to find clinically relevant subgroups based on multiple domains (Maj, 2018). Additionally, it would be beneficial to include healthy controls into the analyses to assess the severity or detect the individuals at risk of developing a specific disorder, as they might experience some symptoms associated with mental illness in their life, but not formally meet the criteria for the diagnosis (Seow et al., 2017).

Finally, we can conclude that in order to achieve the paradigm shift and move toward clinical care and individual treatment regimes, further research in the field is needed.

1.3 Thesis objective and approach

The aim of this thesis was to use data-driven methodologies to create new insights in the field of classification of psychiatric disorders and new subtypes discovery. To this end, we followed a transdiagnostic approach, analyzing a sample with healthy controls and patients diagnosed with major depressive disorder, schizoaffective disorder (SZA), schizophrenia, bipolar disorder, and other psychiatric disorders such as anxiety and social phobia. Hence, the analysis went beyond the existing boundaries not only between different disorders but also between health and disease. This transdiagnostic sample was subject

to cluster analysis and described with different data modalities, ranging from psychopathology, socio-demographic, cognition, environmental factors, genetics, etc. This part of the thesis is primarily based on the results published in Pelin et al., 2021.

Additionally, we wanted to explore the way we could work with extremely high-dimensional genetics data to better describe the differences between and within psychiatric disorders or identified subtypes. To this end, a feature selection algorithm was applied to the same transdiagnostic sample described above.

1.4 Overview of the thesis structure

Chapter 1 (Introduction) provides a brief introduction into the background of psychiatric nosology. The historical viewpoint and the challenges of the current classification system are given, as well as the overview of research efforts to improve it. Chapter 2 (Background) provides the theoretical background of the methodologies used in this thesis. By discussing the challenges that they face, it gives a rationale for the choice of the algorithms and approaches used in this study. Chapter 3 (Methods) covers the methodological framework, including samples, workflow, and tools used to perform the computational analyses. Chapter 4 (Results) reports the findings of the study, from the results of the clustering and identification of the potential subtypes, across characterization of the clusters to the feature selection process. In the final Chapter 5 (Discussion and Outlook) we discuss the findings, contributions, and limitations, and give perspectives for future work. Supplementary information of some methods and results is given in the Appendix.

2 Background

This chapter covers the background of the methodologies used in this thesis. By discussing the challenges they face, the chapter provides a rationale for the choice of the algorithms and approaches used to answer the research question. The first section provides an overview of the machine learning field and its subfields which enable us to discover intricate structures and inner relations in the datasets. The second section covers the background of the statistical testing, while the last section covers the background of genetics.

2.1 Machine learning background

The term *machine learning* refers to the science of programming machines in a way that they can automatically detect the meaningful patterns in the data, i.e. *learn from the data* (Gron, 2017; Mitchell, 1997). Machine learning is a subfield of computer science, however, it is closely associated with the mathematical disciplines of statistics, optimization, and information theory. The hope for machine learning is that these automated algorithms would be able to complement human intelligence and find meaningful patterns that might have been missed by the human observer (Shalev-Shwartz and Ben-David, 2014).

There are several subfields of machine learning, which vary in their approach to learning, the form of problem they are solving, or the type of data they use. In this thesis, both of the two most common methods were used - *supervised learning* and *unsupervised learning*. In this chapter, we cover the background of each type separately, with a focus on the algorithms and approaches used to

answer the research question.

2.1.1 Terminology

The dataset consists of observations, in our case individuals, for which the measures were collected. Throughout this and the following chapters, the dataset with N observations and p measures is represented as a matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$ with observations stored as rows and the measures as columns:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,p} \end{bmatrix}.$$

In supervised learning, which will be covered later, the above p column vectors are called the *predictors*, while the variable we want to predict is called the *response* or *outcome variable*, noted as a vector $\mathbf{Y} \in \mathbb{R}^N$:

$$\mathbf{Y} = (y_1, y_2, \dots, y_N).$$

In this thesis, the response variable was categorical, representing the cluster labels.

2.1.2 Unsupervised learning

Unsupervised learning is a technique in machine learning used to uncover the underlying structure or distribution in the input data, without the labeled responses. In other words, the goal is to find the structure without instructions as no classification of our observations is given. Some examples of unsupervised learning are clustering, anomaly and novelty detection, dimensionality reduction, association rule mining, etc.

As previously mentioned in the Introduction chapter, clustering has been the dominant approach in research efforts to discover more homogeneous psychi-

atric disorder subtypes and was used with the same purpose in this thesis.

2.1.2.1 Clustering

2.1.2.1.1 Definition and challenges

Clustering is a process of *grouping a set of objects* such that the one group (cluster) contains the objects that are more similar to one another than to the objects in other groups. The higher the similarity of objects within the formed clusters and the lower the similarity of objects between different clusters, the better the clustering. Objects subject to clustering are typically represented as points in a multi-dimensional space, with each dimension representing a distinct attribute (feature, variable) of the object.

Intuitively, this definition of clustering as a task of grouping sets of observations into more homogeneous groups is quite clear. However, there are challenges posed to the clustering process which need to be considered. The ones important for this thesis and the choice of the algorithm are discussed below.

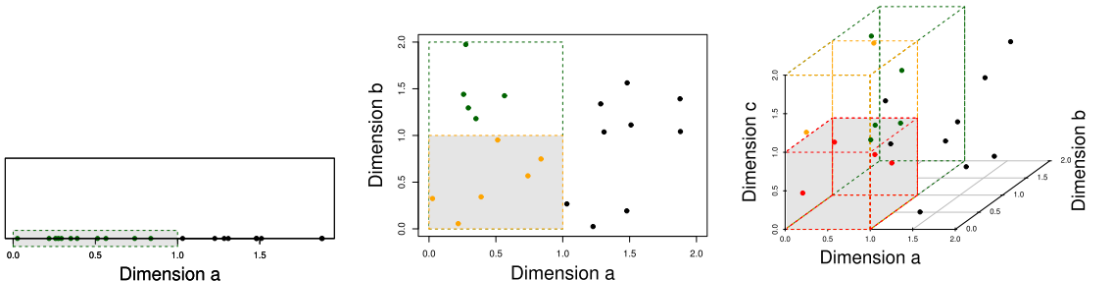
The challenge of the unknown ground truth and similarity metric
Clustering is an exploratory analysis, without the known *ground truth*. It is not *a priori* known how many clusters there are in the data and how are they supposed to look. For a given dataset, there may be a variety of possible clustering solutions. Consequently, there is a large number of clustering algorithms that can produce very different clustering results on the given input data (Shalev-Shwartz and Ben-David, 2014). There are, however, performance or evaluation metrics that can be used to infer a satisfying grouping and will be discussed later in this section. Furthermore, clustering, as per definition, aims to group the objects that are more *similar* to each other. But, what does a similarity between the two objects mean? For example, is the hair color the measure that determines the grouping of individuals or their height? Very probably, the clustering output in these two cases would result in very different clustering solutions, depending on the definition of similarity. Hence, the choice of similarity metric is very important for clustering and can result in many different groupings of the

objects. For a nice illustration of the similarity and the ground truth issues, please refer to Chapter 22 in Shalev-Shwartz and Ben-David, 2014.

The challenge of the high-dimensionality

Today's applications of clustering are faced with another very important issue, the so-called *curse of dimensionality*, a phrase coined by Richard Bellman (Bellman, 1954). This phenomenon is a result of an increasing amount of data generated nowadays and, simultaneously, an increasing number of attributes that describe our observations. Bellman himself used it to describe the rapid growth in the difficulty of problems arising as the number of variables increases. In high-dimensional datasets it is often the case that all the objects are practically equidistant from one another, hiding the real patterns in the data (Figure 2.1). What that means for the clustering procedure is that many irrelevant dimensions caused by high dimensionality can confuse the algorithms by masking the clusters in the noisy data. Traditional clustering algorithms, designed to work with spaces of lower dimensions, rely on assessing the similarity between pairs of groups of objects and fail to find important patterns and clusters in high-dimensional settings (Assent, 2012). Therefore, with the increasing amount of data generated, the algorithms have to be capable of handling the large number of dimensions that describe it. The standard approach for overcoming the *curse of dimensionality* is to apply the global dimensionality reduction techniques in the form of feature transformation or feature selection (Guyon and Elisseeff, 2003). The former creates the new set of variables that conveys a big part of the global information, while the latter finds the appropriate subset of variables, thus removing irrelevant and redundant dimensions. Feature selection will be covered later, in Section 2.1.4.1 of this chapter. Another approach in overcoming the curse of dimensionality in the clustering process is the application of *subspace clustering* algorithms. Subspace clustering algorithms localize their search for clusters, assuming they exist in the subspaces of lower dimension than the original one, unlike the global dimensionality reduction methods which observe the dataset as a whole. Hence, they could discover the clusters that live in several, potentially overlapping subspaces. For a detailed review on subspace clustering methods, please refer to Parsons et al., 2004.

Figure 2.1: *Curse of dimensionality* visualized
In only one dimension the data points are relatively close to each other (a). Adding a new dimension extends the points across it, moving them further apart (b). The third dimension moves the data points even further making the high-dimensional data very sparse (c). Figure taken, with permission, from Parsons et al., 2004.



(a) 11 Objects in One Unit Bin (b) 6 Objects in One Unit Bin (c) 4 Objects in One Unit Bin

The challenge of the stability of the solution

The basic idea of *cluster stability* is that the clustering algorithm should obtain similar results under resampling of the data or when applied to several other datasets. Hence, to evaluate the stability of a clustering algorithm, it needs to be run several times on the perturbed dataset. As explained in Von Luxburg, 2010, this can be done in a few different ways. In this thesis, the stability was evaluated by randomly subsampling the dataset without replacement, as it will be described later.

2.1.2.1.2 Clustering techniques and algorithms in this thesis

High dimensional data clustering

The main clustering algorithm used in this thesis is the *High dimensional data clustering* (HDDC), developed by Bouveyron et al., 2007. It is the subspace clustering algorithm that incorporates the Gaussian mixture model (GMM) framework for high-dimensional data. GMM is a model-based type of clustering, where observations are assumed to be a sample from a finite mixture of Gaussian probability distributions. The main advantage of the model-based approaches, compared to the widely used heuristic clustering algorithms such as K-means, is the variety of model choices, i.e., regularizations, that allow for modeling of clusters with different shapes, orientations, and sizes. As the authors of HDDC point out, the major drawback of the classical GMM is the poor performance when the dimensions of the dataset increase, i.e., when the curse of dimensionality arises. Their proposed HDDC relies on the *empty space* phenomenon, assuming that high-dimensional data exist in subspaces of lower dimension than the original one and it models the clusters accordingly. Moreover, since HDDC is designed to work with high-dimensional data, a prior application of global dimensionality reduction methods is not necessary. The theoretical background of the algorithm will be summarized in brief below, while all the details can be found in the previously referenced original paper.

As in the classical GMM, the HDDC also assumes that the class conditional densities are p-variate Gaussian $\mathcal{N}_p(\mu_i, \Sigma_i)$ for $i = 1, \ldots, K$, where μ_i defines the mean of the class$_i$ and the Σ_i is the covariance matrix, defining its width. The class conditional covariance matrix Δ_i is defined with:

$$\Delta_i = Q_i^T \Sigma_i Q_i, \tag{2.1}$$

where Q_i is orthogonal matrix with the eigenvectors of Σ_i. HDDC further assumes that the Δ_i is a two-block matrix:

$$\Delta_i = \left(\begin{array}{ccc|ccc} a_{i1} & & 0 & & & \\ & \ddots & & & \mathbf{0} & \\ 0 & & a_{id_i} & & & \\ \hline & & & b_i & & 0 \\ & \mathbf{0} & & & \ddots & \\ & & & & & \ddots \\ & & & 0 & & b_i \end{array}\right), \tag{2.2}$$

where $a_{ij} > b_i, j = 1, \ldots d_i$ and $d_i \in \{1, \ldots, p-1\}$ is unknown.
The parameters $a_{i1}, \ldots, a_{id_i}$ and b_i model the variance of the class$_i$ and the variance of the noise, respectively. In the context of subspaces, the parameter d_i is the intrinsic dimension of the subspace of the class$_i$ which is spanned by the first d_i column vectors of Q_i. When some of the above parameters are fixed to be common between or within classes, models corresponding to different regularizations are obtained. In this thesis, we empirically decided which of the possible regularizations is optimal for the data.

Consensus clustering

Consensus clustering is an approach that relies on multiple runs of a chosen clustering algorithm on subsamples of the dataset. By inducing variability with subsampling, it provides a consensus on parameter decisions (such as the number of clusters), on a cluster assignment for the observations (based on the assignments in all the runs of the algorithm), and on the assessment of the stability of the discovered clusters (Monti et al., 2003).

In this thesis, HDDC was wrapped in a consensus clustering framework to obtain the cluster solutions. The resampling scheme was the Leave-one-out Jackknife, a method introduced in 1949. by M. H. Quenouille (Quenouille, 1949). In general, it is applied to reduce bias and evaluate the variance of an estimator. The Leave-one-out Jackknife estimator of the parameter is found by sequentially removing a single observation in the dataset, then recomputing the desired statistic.

Clustering evaluation metrics

Due to the absence of the ground truth which would guide the unsupervised learning process, one of the main problems in the clustering process is evaluating the quality of the solution discovered by the clustering algorithm. For the heuristic clustering algorithms, there are many methods to evaluate the performance of the algorithm, all nicely summarized in Palacio-Niño and Berzal, 2019. These algorithms tend to create spherical clusters of equal volume and of the same within-cluster variance, the condition which is often not met in practice (Greve et al., 2016). Model-based clustering algorithms are overcoming this limitation by allowing clusters to vary in size, shape, and orientation. Because of that, the evaluation metrics for the heuristic algorithms are not reliable enough in the model-based settings. Popular criteria used for selecting the optimal model in a model-based clustering are the *Bayesian information criterion* (BIC) and *Integrated Completed Likelihood* (ICL). ICL is proposed by Biernacki et al., 2000 and claimed to be more reliable in estimating the optimal number of clusters, which was the reason for choosing it as the main criterion in this work.

Once the cluster labels are obtained by the clustering algorithm, the external validation metrics can be used to compare this result to a potentially different data partition. The popular measures to compare the two partitions are *Rand Index* and *Jaccard Index*. Rand Index calculates the proportion of correctly classified elements of all elements, while Jaccard Index ignores the pairs of elements that are in separate clusters for both partitions. Both indices range from 0 to 1, and the closer the value is to 1, the more similar the two clustering solutions are. For more details on comparing the two clustering solutions, please refer to Wagner and Wagner, 2007.

2.1.3 Supervised learning

Supervised learning is a technique in machine learning where the input variables ($\mathbf{X}$) are used to predict the values of the output variable ($\mathbf{Y}$). In other words, the technique tries to learn the mapping function f from $\mathbf{X}$ to $\mathbf{Y}$ ($f : \mathbf{X} \mapsto \mathbf{Y}$). Hence, unlike in the unsupervised learning settings, the algorithm learns on a labeled dataset, enabling the evaluation of the performance of the algorithm. Depending on the type of output variable, supervised learning algorithms can be further classified into a) *classification*, predicting classes, i.e., the qualitative (categorical) output and b) *regression*, predicting the quantitative output (James et al., 2014).

In this thesis, supervised learning in form of classification, i.e., prediction of the cluster labels, was used for the characterization of subtypes based on different sets of variables, to test the generalizability of the solution, and for the feature selection.

2.1.3.1 Classification

2.1.3.1.1 Definition and challenges

Classification is the most common type of supervised machine learning, intending to predict the qualitative class labels of new instances based on past examples. Depending on the number of different class labels a classification task has, it can be further categorized into the *binary classification*, with two class labels, and the *multiclass classification*, with three or more different class labels. The learning procedure usually entails randomly splitting the available set of observations (the input data $\mathbf{X}$), to the so-called *train* and *test* (*validation* or *hold-out*) sets. The algorithm is fitted on the train set, where it tries to learn the patterns which are distinguishing between different classes. To assess whether the machine successfully learned from the train data and how accurately it will be able to predict the labels of future observations, the error metrics need to be calculated. This is accomplished by using the trained model to predict the responses for the test set observations and calculating how far the predicted

responses are from the known true ones.

The high-dimensionality of the datasets in numerous practical and real-life applications today is posing some challenges to the supervised learning as well. Although it is easy to think that having more attributes describing our data is only beneficial for learning, the higher number of them does not necessarily mean a better prediction, especially if the training data consist of many irrelevant ones. These data points that represent the random chance and not the true properties of the data are called the *noise*. When present, the noise in the data could mask the real important features and relationships which become hard to unravel even with the complex supervised models. As a result, a model is likely to detect the patterns in the noise itself and won't generalize well to the new instances (Gron, 2017). This phenomenon of poor generalization, where a model learns the noise and the detail in the training data, but it performs poorly on the unseen data is called the model *overfitting*. There are few ways to try to avoid this, some of which are used in this thesis - the *resampling techniques* (such as cross-validation) and *regularization*.
Resampling methods are important for achieving better generalizability of the model and, thus, prevent overfitting. They involve drawing samples from a dataset multiple times and fitting a model on each sample to get more information about its average performance and generalizability, which would not be available if the model was fit only once. However, since these methods require fitting the same method several times using different subsets of the training data, they can be very computationally demanding (James et al., 2014). Regularization avoids overfitting by constraining a model to make it simpler, allowing for fewer degrees of freedom where the model can adapt to the training data. In this thesis, a regularized form of regression, called *Lasso regression*, was used and will be covered in more detail later.

Another challenge in many applications today is multiclass classification, classifying observations into $K > 2$ classes. Generally, binary problems ($K = 2$) are much easier to solve and many classification algorithms can be applied. A multiclass problem, however, is more complicated and requires particular strategies which can transform this type of problem into binary. The multiclass

problem strategy used in this thesis will be discussed in the following section.

2.1.3.1.2 Classification techniques and algorithms in this thesis

Multiclass classification strategies

As already mentioned, when the classification task consists of predicting the response variable with more than 2 classes, it is a good practice to transform this type of problem into a binary one. In this thesis, the two following types of techniques were used:

one-vs.-all

One-vs.-all strategy involves splitting the multiclass dataset into multiple binary classification problems. This means that instead of training the multiclass classifier to learn to distinguish between K classes, K binary classifiers are trained. Each of the K binary classifiers is trained to differentiate the single-class examples from the examples in all other classes (Rifkin and Klautau, 2004). The multiclass response variable $\mathbf{Y} = (y_1, y_2, \ldots, y_N)$ for N individuals $(\text{ind}_1, \text{ind}_2, \ldots, \text{ind}_N)$ grouped into K classes $(C_1, \ldots, C_K)$ is transformed to the K class-specific binary variables $\mathbf{Y}_{C_i}$, with elements y_j defined as:

$$y_j = \begin{cases} 1, & \text{ind}_j \in C_i \\ 0, & \text{ind}_j \notin C_i \end{cases}, \qquad i = 1, \ldots, K, \quad j = 1, \ldots, N.$$

Example: Let $N = 6$ individuals be grouped into the $K = 3$ different clusters $(C_1, C_2, \text{or } C_3)$. Let individuals 1 and 2 belong to the cluster C_1, individuals 3 and 6 to cluster C_2 and individuals 4 and 5 to cluster C_3. Therefore, the respective outcome variable is $\mathbf{Y} = [C_1, C_1, C_2, C_3, C_3, C_2]$. Instead of training the multiclass classifier with the outcome $\mathbf{Y}$, three separate binary classifiers are

trained with the following outcome variables:

$$\begin{aligned}\mathbf{Y}_{C_1} &= [1,1,0,0,0,0]\,,\\ \mathbf{Y}_{C_2} &= [0,0,1,0,0,1]\,,\\ \mathbf{Y}_{C_3} &= [0,0,0,1,1,0]\end{aligned}$$

In this thesis, the *one-vs.-all* strategy was the main strategy in the processes of cluster description and prediction.

one-vs.-one

In the *one-vs.-one* strategy, the $\binom{K}{2} = \frac{K(K-1)}{2}$ binary classifiers are trained instead of one multiclass classifier. Each binary classifier is trained on the samples from a pair of clusters from the full dataset ($\mathbf{X}$) in order to learn how to distinguish them. Let us follow the above example for the illustration of this approach:

Example: Assume the same settings as in the example above. Instead of training one multiclass classifier with response variable $\mathbf{Y} = [C_1, C_1, C_2, C_3, C_3, C_2]$ and the input dataset $\mathbf{X}$, consisting of 6 individuals ($[ind_1, ind_2, ind_3, ind_4, ind_5, ind_6]$), we will train $\binom{3}{2} = 3$ binary classifiers on the following pairs of individuals and response variables:

$$\begin{aligned}classifier_1 \quad &\rightarrow \quad \mathbf{X}_{C_1,C_2} = [ind_1, ind_2, ind_3, ind_6]\,,\ \mathbf{Y}_{C_1,C_2} = [1,1,0,0]\,,\\ classifier_2 \quad &\rightarrow \quad \mathbf{X}_{C_1,C_3} = [ind_1, ind_2, ind_4, ind_5]\,,\ \mathbf{Y}_{C_1,C_3} = [1,1,0,0]\,,\\ classifier_3 \quad &\rightarrow \quad \mathbf{X}_{C_2,C_3} = [ind_3, ind_4, ind_5, ind_6,]\,,\mathbf{Y}_{C_2,C_3} = [1,0,0,1]\,.\end{aligned}$$

Hence, $\mathbf{X}_{C_i,C_j}$ is a subset of the full dataset $\mathbf{X}$, containing individuals grouped in the clusters C_i and C_j, and $\mathbf{Y}_{C_i,C_j}$ is the corresponding subset of the response variable $\mathbf{Y}$.

In this thesis, the *one-vs.-one* strategy was used as an additional analysis to refine the cluster differences.

High Dimensional Discriminant Analysis

High Dimensional Discriminant Analysis (HDDA) was proposed by Bouveyron et al., 2005, as previously mentioned HDDC. Along the same lines, it assumes that the high-dimensional data exist in the subspaces with a lower number of dimensions than the original one.

In general, the goal of a discriminant analysis is to identify a group of prediction equations based on the independent variables that are used to classify observations into K classes ($K \geq 2$). Classes are known *a priori* and their densities are Gaussian $\mathcal{N}(\mu_i, \sigma_i), \forall i = 1, \ldots, K$. The classical methods of Discriminant analysis are Quadratic Discriminant Analysis (QDA) and Linear Discriminant Analysis (LDA), explained in detail elsewhere (James et al., 2014). However, they have disappointing behavior when the number of features p increases and especially when the number of observations N is much smaller than p ($p \gg N$). The HDDA adapts discriminant analysis to high dimensional data by working in the class-specific subspaces with lower dimensionality. It estimates the intrinsic dimension of each class, reducing the number of parameters that need to be estimated. The advantage of this is that the technique does not require prior dimensionality reduction and, thus, avoids the potential information loss. The formulation of Gaussian models for high-dimensional data classification is the same as already covered in Section 2.1.2 for the corresponding clustering method and the details can be found in the original paper (Bouveyron et al., 2005).

Lasso regularized regression

Lasso regularized regression, developed by R. Tibshirani (Tibshirani, 1996), is the regularized regression method that introduces a constraint on the coefficients, with an effect of shrinking them or even setting some to zero.
Mathematics behind Lasso involves the usual linear regression settings, where the aim is to approximate the outcome variable $\mathbf{Y}$ using a linear combination of the features in $\mathbf{X}$:

$$\mathbf{Y} \approx \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j. \tag{2.3}$$

Here, $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ is the vector of regression weights that are

parametrizing the model and the $\beta_0 \in \mathbb{R}$ is an intercept or "bias" term. The linear regression model is fitted with the *least square* method which picks the coefficients β such that they minimize the Residual Sum of Squares (RSS), defined as:

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 . \tag{2.4}$$

However, the prediction accuracy and interpretation of this approach are often not satisfying and the model often does not generalize well (Hastie et al., 2015), motivating the introduction of the regularized type of regression. Lasso incorporates the least-squares loss with the l_1-constraint ($||\cdot||_1$), i.e., a constrain on the sum of coefficients' absolute values:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda||\beta||_1 = \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.5}$$

where $\lambda \geq 0$ is a tuning parameter, controlling the shrinkage amount.

The characteristics of shrinking the coefficients of some features to zero played a major role in this thesis, as with this property we were able to depict the important genetic risk factors for groups of patients. For more details on Lasso, please refer to Hastie et al., 2015.

Cross-validation

Cross-validation is a class of methods for evaluating the machine learning models. While in the fitting process, it holds out part of the training observations and estimates the error rate by applying the trained model to those held out observations.

In k-fold cross-validation, the set of N observations is randomly split into k folds (groups) of approximately same size (N / k). Then, the first fold is used as a (hold-out) validation set, while the union of the remaining k-1 folds are treated as training set on which the method is fitted. The classification error is subsequently calculated on the observations in the held-out fold and the final error estimate is calculated as the average of all k error values computed

throughout the process. (James et al., 2014)

The k-fold cross-validation technique is very common approach for parameter tuning and model selection. When the optimal parameter for the model is selected, the algorithm is retrained on the whole training dataset by using the chosen parameter (Shalev-Shwartz and Ben-David, 2014).

In this thesis, k-fold cross-validation was used to tune the λ parameter in the Lasso prediction models.

Figure 2.2: A scheme for k-fold cross-validation
Set is randomly split into k non-overlapping groups. In the k-th iteration, the group k, representing the 1 / k-th of the data, acts as a test set, and the rest as a training set.
Image by Gufosowa - Own work, CC BY-SA 4.0, taken from
https://commons.wikimedia.org/w/index.php?curid=82298768

Classification evaluation metrics

After the model is trained, its performance needs to be assessed. In binary classification models, the performance can be summarized in the confusion matrix, counting the observations correctly and incorrectly predicted by the model:

		True	
		class 1	class 0
Predicted	class 1	True positive (TP)	False positive (FP)
	class 0	False negative (FN)	True negative (TN)

From the confusion matrix, many metrics for the performance evaluation can be calculated. The ones listed below were used in this thesis.

True positive rate or *Sensitivity* is the proportion of the instances which are known to be positive (in class 1) and are predicted as such:

$$\text{Sensitivity} = \frac{TP}{TP + FN}.$$

True negative rate or *Specificity* is the proportion of the instances which are known to be negative (in class 0) and are predicted as such:

$$\text{Specificity} = \frac{TN}{TN + FP}.$$

Accordingly, the false positive rate is defined as:

$$\text{False positive rate} = 1 - \text{Specificity} = \frac{FP}{TN + FP}.$$

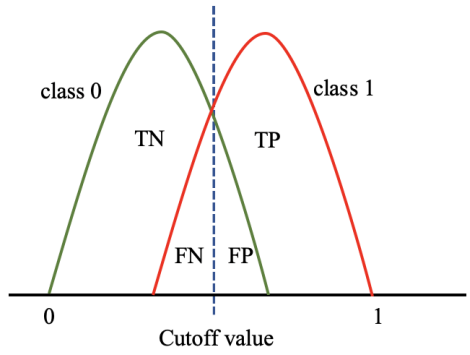


Figure 2.3: Sensitivity and Specificity trade-off

A predicted value for each individual, returned by the algorithm, is generally a numerical value (probability). To determine the predicted class, the decision threshold (the cut-off value) has to be applied. For a given cut-off value, class membership is determined for each observation - if the predicted value is less (greater) than the cut-off, the predicted class is la-

beled as negative, class 0 (positive, class 1). Most often the two distributions will overlap, as illustrated in Figure 2.3. Therefore, for every possible cut-off value selected to discriminate between the two groups, there will be some observations from both classes correctly and incorrectly classified, forming four possible outcomes represented in the confusion matrix above and illustrated in the Figure 2.3. Hence, both specificity and sensitivity of the model prediction depend on the chosen threshold and there exists a trade-off between the two. With the higher cut-off value, the proportion of false negatives (the observations in class 1 incorrectly classified to class 0) will increase, therefore increasing specificity and decreasing sensitivity. When the cut-off value is lower, the sensitivity increases, while specificity decreases. The optimal cut-off value can be chosen with the *receiver operator characteristics (ROC) curve*, where the sensitivity is plotted against the false positive rate for the different thresholds. The performance of the model is then assessed by *Area under the ROC curve* (AUC). The closer the AUC value is to 1, the better the model performs.

In this work, all of the above-mentioned performance metrics were used for the assessment of the model performance. The cut-off value used for the calculation of the sensitivity and specificity was chosen such that both metrics are maximized, as it will be explained in the Methods chapter.

2.1.4 Dimensionality reduction

The curse of dimensionality is posing a lot of challenges to the machine learning algorithms, both unsupervised and supervised. With high-dimensional data, as discussed previously, it becomes more difficult to detect meaningful patterns in the data and to detect the relationships among features. *Dimensionality reduction methods* are the unsupervised type of machine learning with the main objective to reduce the dimensionality of the dataset, while still keeping the most important information from the data. Generally, they can be in the form of *feature transformation* or *feature selection* methods. Feature transformation creates a new, smaller set of variables that capture most of the meaningful properties of the original data. The most common types of feature transformation methods

are Principal component analysis (PCA), LDA, Autoencoders, etc. On the other hand, the feature selection method selects subsets of original features from the dataset that are useful for further analysis. In the following, we will focus on the feature selection techniques, as they were the main method for the dimensionality reduction analysis in this thesis.

2.1.4.1 Feature selection

2.1.4.1.1 Definition and challenges

Feature selection methods, as mentioned above, are the type of dimensionality reduction methods that select subsets of predictor variables. Hence, they result in a set of original features and not the new, transformed ones. In this way, the meanings of the original feature sets are preserved, offering better readability and interpretability of the model by a domain expert, making it a big advantage over the feature transformation methods.

Feature selection methods can be categorized into unsupervised and supervised approaches with the main difference being the availability of the outcome variable, e.g., class label or some continuous response. Supervised feature selection methods choose features that are predictive or correlated with the outcome. Unsupervised feature selection is applied for a clustering task and it evaluates the feature subset importance by some clustering quality measure or intrinsic property of the data (Dy and Brodley, 2000).
Feature selection methods can be further divided into *wrapper*, *filter*, and *embedded* methods. While the details and challenges of each are explained elsewhere (Ullah et al., 2017), here we will briefly mention the differences between them, with focus on the supervised feature selection:

A *Filter methods* pick up the intrinsic properties of the data (distance, dependency, correlation...); hence, relying on the statistical criteria.

B *Wrapper methods* depend on the accuracy of classification algorithm while selecting the features. By doing so, they incorporate the effect of features into the learning process and select them based on the outcome

variable. However, they are very computationally expensive.

C *Embedded methods* are developed to overcome the gaps between the first two methods. The feature selection is integrated as part of the intrinsic model learning and is less computationally demanding compared to the wrapper type.

Independent of the type of filter selection methods used, there are some common challenges posed to the process to be considered (Li and Liu, 2017). First, more often than not, features are correlated and appear in various kinds of structures, for example in the group structure (e.g., genes acting together, correlated SNPs in Linkage disequilibrium (LD) blocks). Therefore, taking this into account should be important during the feature selection. The second important challenge to consider is the one posed to almost all machine learning applications, namely, the stability of the algorithms. In the case of feature selection, this means the selection of the same set of features even after the data perturbation.

In this thesis, the regularization regression model was used to select the features. Regularization models are embedded types of feature selection methods because they use an objective function to reduce overfitting errors and, at the same time, force the coefficients of the irrelevant features to be zero. Hence, the feature selection and the learning process interact.

2.1.4.1.2 Feature selection techniques and algorithms in this thesis

Group Lasso regression

Lasso regression, as discussed previously, applies the penalty to the coefficients of the features and shrinks the coefficients of unimportant variables to zero (Equation (2.5)). If features exhibit the group structure, i.e., are divided into m different groups, a solution that finds a sparse set of groups is needed. To solve this problem, Yuan and Lin, 2006 suggested the Group Lasso, with the

following formulation:

$$\min_{\beta} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{g=1}^{m}\sqrt{d_g}||\beta^{(g)}||_2. \tag{2.6}$$

Here, $\mathbf{X}$ is a data matrix of N individuals and p features, $\mathbf{Y}$ is an outcome variable and d_g is a size of the feature group $g = 1, \ldots, m$.

With this approach, when the group is included in the model, all the features in it are automatically included as well, i.e., have a non-zero β coefficient. Hence, the sparsity *between* the groups is imposed. However, sometimes it would be useful to additionally enforce the sparsity *within* each group, which was a motivation to develop the *Sparse group Lasso* algorithm, used in this work. It solves the following problem:

$$\min_{\beta} \frac{1}{2N}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda_1||\beta||_1 + \lambda_2\sum_{g=1}^{m}\sqrt{d_g}||\beta^{(g)}||_2, \tag{2.7}$$

where λ_1 is the parameter-wise regularisation penalty and λ_2 the group-wise regularisation penalty (Yuan and Lin, 2006; Simon et al., 2013).

2.2 Statistical testing background

2.2.1 Definition and challenges

Statistical analysis is a useful tool for exploring the data and discover the underlying patterns and dependencies between the variables. The main aim of statistical analysis is to answer the research question and to provide confidence in the answer. This research question or a research claim is called the *hypothesis* and the statistical method used to infer how likely it is to be true is called the *statistical hypothesis testing*.

The first step in hypothesis testing is formulating the null hypothesis (H_0) and the alternative hypothesis (H_1). For instance, H_0 can be a statement about a population parameter, or no difference between two measured variables, or that the two samples come from the same population. The alternative hypothesis H_1 is complementary to H_0 and it states what is thought to be wrong about the H_0. Then, the data is observed and test statistic specific to the hypothesis is computed. The test statistic provides the likelihood of obtaining sample outcomes if the null hypothesis was true. Finally, the decision of whether the H_0 is rejected or accepted is linked to the p-value. The p-value is the probability of obtaining the measured data or more extreme results, given that the H_0 is true. The null hypothesis is rejected if the p-value is below a predefined significance level α, normally chosen to be at 0.05.

While performing the hypothesis testing, two types of errors can occur: type I error, when the H_0 is rejected, that is actually true; and type II error, when the H_0 is accepted, that is actually false. The type I error produces the false-positive result where it is concluded that an effect exists when it actually does not. The probability for this occurring is the level of significance α set for the testing. When performing several hypothesis testing on the same dataset, the probability of producing false-positive results increases. This phenomenon is called *multiple testing problem* and it should be addressed. The probability of making one or more false-positive results when performing multiple testing is called *family-wise error rate* (FWER) and there are various controlling procedures

for it. They can be classified in three distinct groups - *step-down, step-up* and *single-step* procedures and the details of each can be found in Ge et al., 2003. The widely used correction for multiple testing is Bonferroni method, providing a very strong FWER correction by rejecting any hypothesis H_i with p-value $\leq \frac{\alpha}{M}$ ($i = 1, \ldots, M$, where M is the number of hypotheses tested).

In this thesis, the main method for statistical testing was the less conservative, step-down, procedure developed by Westfall and Young (Westfall and Young, 1993). The method controls for the FWER while taking the possible dependence structure of the variables into account.

2.3 Genetics background

2.3.1 Genome-wide association studies and polygenic risk scores

A *genome-wide association study* (GWAS) is a favored approach used in genetics for determining the genetic variants, named single nucleotide polymorphisms (SNPs), that are associated with particular diseases, or a certain treatment outcome. SNPs are variations at a certain position in the DNA that occurred during the evolution and were passed down to new generations, explaining the significant portion of the genetic diversity within the human population. Identification of SNPs associated with the specific disorder may advance our understanding of complex diseases or perform early diagnosis.
The GWA study involves scanning the genomes from many different people and comparing the allele frequencies of common genetic variants between cases (people affected by the disease) and controls. So far, GWAS have identified many SNPs that are associated with psychiatric disorders, such as major depressive disorder (Howard et al., 2019; Wray et al., 2018), schizophrenia (Psychiatric Genomics Consortium et al., 2014; Pardiñas et al., 2018), and bipolar disorder (Stahl et al., 2019; Sklar et al., 2011). However, it has become evident that those complex disorders have a genetic underpinning that is highly polygenic, meaning that hundreds or thousands of genetic variants influence disease risk

and that the single SNPs, in most cases, will not be helpful for diagnosis. Hence, to identify those at high risk of polygenic disorder, a method for calculating *Polygenic risk score* (PGS) was developed. As Fullerton and Nurnberger, 2019 explain, PGS captures the cumulative effects of many genetic variants into a single quantitative metric by adding up the effects of individually associated SNPs from independent GWA studies, counting how many risk alleles does that individual carry at each locus, and weighting each risk allele by its effect size. SNPs that enter the PGS calculation are typically selected based on their GWAS association strength (p-values). References for this whole section with respect to the more detailed explanation of PGS calculation and limitations are Fullerton and Nurnberger, 2019; Andlauer and Nöthen, 2020; Wray et al., 2020, if not stated otherwise.

PGS approach is a good and effective tool for medical research. However, it suffers from several issues which have limited their application in psychiatric disease prediction and clinical translation. Some of these limitations will be listed below, whereas a more detailed discussion can be found in the above-mentioned main references.

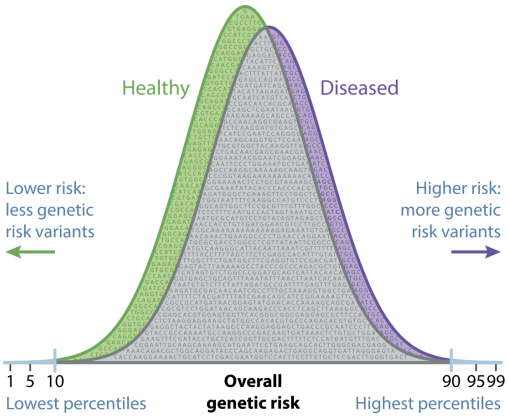


Figure 2.4: Distribution of PGS on a population level
Source:
Andlauer and Nöthen, 2020

First, PGS can be used for assessing the differences between cases and controls on the population level (So and Sham, 2017) but is not informative on the individual level. As Figure 2.4 shows, the distribution of PGS between cases and controls highly overlaps, thus PGS might have predictive value only for individuals in the lower and higher risk groups.
Second, PGS can explain only part of the genetic component of a given condition and, as a result, only a proportion of phenotypic variance (Stahl et al., 2019; Pardiñas et al., 2018; Howard et al., 2019). This is due to the fact that their construction depends on GWAS,

which are capturing only the contribution of common variants (mostly with the minor allele frequency in the population of at least 1%) to the disease risk.
Furthermore, PGS calculation is dependent on the selection of *p*-value threshold which is chosen empirically - the *p*-value at which the distinction between cases and controls is the best is selected as optimal. Several algorithms have been suggested to improve the selection and weighting of SNPs for the PGS calculation, among which is the PGS-CS method used in the analysis of this thesis. The method applies the continuous shrinkage (CS) priors on the effect sizes of the SNPs, where the amount of shrinkage applied to each SNP is modified by the strength of its GWAS association. For details, see Ge et al., 2019.
Finally, PGS is dependent on the GWAS discovery sample. With the larger sample sizes, the statistical power to detect association signals is higher, and therefore, the estimates of the effect sizes for SNPs are more accurate. Over time, the sample sizes are expected to increase, and consequently, also the variance explained by the PGS.

In summary, the prediction power of stand-alone PGSs in psychiatric analyses is still limited. To improve it, and at the same time, increase the chances of significant clinical implications, it has been suggested to combine the PGSs with other risk factors of an individual (Murray et al., 2020). This has already been shown to improve the accuracy outside of the field of psychiatry, in coronary artery disease prediction (Inouye et al., 2018). In psychiatry, for example, family history information can be used. It includes both genetic and non-genetic risk factors that the family members have in common, and it has been very good information guiding the clinical diagnosis and management so far (Wray et al., 2020). Observed together, PGSs and family history can provide additional information about the disease (Bigdeli et al., 2016; Hujoel et al., 2021). Additionally, PGSs may also be analyzed together with other non-genetic risk factors (trauma, stress, life events, brain imaging...) in order to answer many important questions including adverse outcomes and treatment response. However, to date, larger sample sizes and more research are needed to validate those factors.

Despite all the challenges and limitations of PGS now, they do show promising

clinical applications for the future. For example, the approach could be used to identify more severe disease subtypes (Power et al., 2017; Ruderfer et al., 2018; Dwyer et al., 2020) or to identify phenotypic correlations (Calafato et al., 2018; Consortium, 2009). Moreover, as suggested by Murray et al. in their review paper (Murray et al., 2020), the clinical decisions could already be slightly guided with PGS, just as they already are with the information on family history. This could be especially true in the initial stage of disease when patients experience general and non-specific symptoms that still do not satisfy the criteria for a formal diagnosis.

In conclusion, PGSs *per se* are still not ready to have significant usage in clinical practice. However, their combination with other risk information, the larger sample sizes, the methodological changes in their calculation, with the emphasis on machine learning, may offer great advances.

2.3.2 Feature selection and genetics

Genome-wide association studies require a lot of resources to be able to deal with the space and time complexity of the genetic data. A typical GWA study inspects millions of SNPs for association with the phenotype of interest, making it a very computationally expensive task. The standard approach in GWAS is univariate, where statistical association to the phenotype is inspected for each SNP separately. However, such an approach may miss the combinatorial factors between two or more SNPs, and therefore fail to take the polygenic effects into consideration. Polygenic risk scores, covered in the last section, represent the one way of addressing this issue, as they aggregate the information, i.e., the effect sizes, of many SNPs estimated in the univariate GWAS. However, due to the statistical testing-based approach in GWAS and many single tests performed, the chance of producing false positive results is high, and therefore possibly embedded into the PGS itself. As mentioned in the previous section, the quality of PGS should improve with larger sample sizes, but also with a more appropriate selection of SNPs, and adequate estimation of their effect sizes (Janssens, 2019).

Another way to capture the possible polygenic effects is to use the multivariate approach, where more or all of the SNPs are observed together in the model (Zhang et al., 2009; Porter and O'Reilly, 2017). However, in a typical genomic dataset, a number of features p is much higher than the number of samples N ($p \gg N$). This makes the estimated parameters of the multivariate models unreliable and may cause overfitting (Dubitzky et al., 2007). As a result, the feature selection methods became important to identify a subset of SNPs that is informative enough, while remaining sufficiently small to minimize the complexity of the association study.

Feature selection approaches have been applied in big genomic data analyses, covering all three types of feature selection methods, explained in Section 2.1.4.1: wrapper approach (Shah and Kusiak, 2004; Li et al., 2001; He et al., 2015; Long et al., 2009), filter approach (Lee and Shatkay, 2006; Phuong et al., 2005; Halldorsson et al., 2004), and the embedded approach (Zhang et al., 2018b; Sasikala et al., 2015; He and Zelikovsky, 2006). As already mentioned in Section 2.1.4.1, the embedded type of feature selection is explored in this thesis due to its advantages over the wrapper and filter approaches. The focus was put on the penalized regression models which, according to some studies, may be adequate to identify the additive effect of several SNPs and allow for the reliable estimation of the parameters in the high-dimensional settings with $p \gg N$ (De Maturana et al., 2014; Abraham et al., 2013; Waldmann et al., 2013).

Another important concept in the process of identification of important SNPs is the dependence between them induced by Linkage disequilibrium (LD). Briefly, LD is the non-random association of alleles at different loci in the genome and is the base for the association mapping methods (Weir, 1979; Balding et al., 2008). It is important in GWAS because it allows for the identification of genetic markers that tag the real causal variants in complex human disorders (Joiret et al., 2019). Consequently, when analyzing genome-wide genetic variants, it has to be considered that some regions of the genome might be over-represented due to higher levels of LD, masking the patterns in the genome-wide data. This concept has not been addressed well enough in the feature selection algorithms for genomic prediction (Bermingham et al., 2015). As a result, we decided to

explore the feature selection algorithm that takes the LD structure into account, by inferring the LD-blocks of SNPs and applying the Group Lasso regression for the feature selection and prediction (Dehman et al., 2015). The theoretical details of the algorithm were covered in Section 2.1.4.1 of this chapter, whereas the workflow will be presented in the Methods chapter.

3 Methods

This chapter covers the methodological framework, including samples, workflow, and tools used to perform the computational analyses. The first section of this chapter introduces the cohort used for the analyses in this thesis. The second section covers the workflow for the identification and characterization of transdiagnostic clusters of psychiatric disorders. It is primarily based on the results published in Pelin et al., 2021. The third section covers the workflow for the process of feature selection with genetic data applied to find the subsets of SNPs that could be important for the identified subtypes or formal diagnostic categories.

3.1 Samples for the analyses

The data for this thesis was collected from participants that are part of FOR2107 cohort, an ongoing multi-center study recruiting from the areas of Marburg and Münster in Germany (Kircher et al., 2018). To date, the study provides a sample of up to 2500 participants which is including healthy controls and patients suffering from affective disorders – major depressive disorder and bipolar disorder, extended by subsamples of schizophrenia and schizoaffective disorder patients.
The study was approved by the ethics committees of the Universities of Marburg and Münster, following the Declaration of Helsinki (Kircher et al., 2018; Pelin et al., 2021). All participants of the study went through the structured clinical interview based on DSM-IV-TR (SCID-I) (Wittchen et al., 1997), administered by trained clinical raters. The SCID is a semi-structured interview used to deter-

mine the major DSM-IV Axis I diagnoses, including substance use disorders. The output of the SCID interview is the presence or absence of each of the disorders covered in the diagnostic manual (Spitzer et al., 1992).

3.2 Transdiagnostic subtypes discovery with unsupervised learning

3.2.1 Participants

For the clustering analysis, sample was divided into discovery and replication, based on the data availability at the time this analysis started (Pelin et al., 2021). The details on both samples are provided below and in the Methods A1 in the Appendix.

Discovery sample

The discovery sample included all individuals recruited during the study's first phase and whose data was available when the analysis began (N=1 619). Participants who had withdrawn from the study and individuals with missing diagnostic information were excluded from all analyses. Next, of each pair of relatives, one individual with the lower missing call rate was kept in the analyses. Finally, individuals with missing data in any of the variables of interest were omitted. The final discovery sample consisted of N=1 250 individuals, with n=590 healthy controls, n=477 MDD, n=75 BD, n=25 SZA, and n=53 SCZ cases, as well as n=30 patients with a different diagnosis, such as social phobia or anxiety disorder.

Replication sample

All N=852 individuals recruited subsequently were included in the replication sample. After the filtering steps described above, N=622 individuals remained for the analyses, of which n=240 were healthy controls, n=283 MDD, n=44 BD, n=13 SZA, and n=17 SCZ cases, and n=25 patients presented with a different diagnosis (Table 4.1).

3.2.2 Measures for cluster identification and description

3.2.2.1 Clinical data

Clinical variables were used to form the clusters of psychiatric patients and healthy controls. The data collected throughout the SCID interview underwent thorough quality control as explained in detail in Kircher et al., 2018. The main aim was to include the characterization of the individuals with respect to many relevant psychopathological dimensions. As suggested by Maj, 2018, we combined the evaluation of disease progression with variables not only capturing the current stage and symptoms profile, but also some antecedent events, such as early environmental factors and parental factors, as well as concomitant variables such as resilience, personality traits, and cognitive functioning. Only the main measures of psychometric questionnaires and clinical rating scales were used to prevent over-representation of any diagnostic aspects. A total of 57 variables were used to form the clusters (Table A5.1 in the Appendix). Variables that had high differentiation between specific diagnoses and with healthy controls were excluded from the clustering and used in a post-hoc analysis (e.g., medication, lifestyle, sociodemographic...) (Table A5.2 in the Appendix).

After the clusters were formed, they were ranked by the Global assessment of functioning scale (GAF), presented in the DSM (American Psychiatric Association, 2013). The scale measures how much an individual's symptoms affect their daily life, and how serious a mental illness may be in general. The maximum score on the scale is 100 (extremely high functioning) and the minimum is 0 (severely impaired). GAF measure was used in this analysis as a severity proxy, a continuum along which the clusters were ranked. The higher cluster number indicated the higher disease severity.

3.2.2.2 Genetic data

Genetic analyses in the cluster characterization step used a set of PGSs combined with the family history. Covariates in all genetic analyses were age, gender, and

eight ancestry components (AC), used to determine genetic outliers (Pelin et al., 2021).

PGSs were provided for the participants of the FOR2107 study and calculated with the PGS-CS method using training summary statistics from the published GWAS. PGSs for 10 different disorders were used in this analysis and are listed below together with the reference to the corresponding GWA study:

1. Cross psychiatric disorder (Lee et al., 2019b)
2. ADHD (Demontis et al., 2019)
3. ASD (Grove et al., 2019)
4. BD (Stahl et al., 2019)
5. MDD (Howard et al., 2019)
6. SCZ (Pardiñas et al., 2018)
7. Educational attainment (EA) (Okbay et al., 2016)
8. Extraversion (Berg et al., 2016)
9. Hedonic well-being (Baselmans and Bartels, 2018)
10. Neuroticism (Luciano et al., 2018)

Moreover, four family history variables were used. They were self-reported and capturing whether the individual had cases of any psychiatric disorder or specifically of MDD, BD, and SCZ/SZA in the family (up to the second-degree relatives).

1. Family history of any psychiatric disorder
2. Family history of MDD
3. Family history of BD
4. Family history of SCZ/SZA

Finally, the dataset subject to analyses consisted of another 10 covariates - age, gender, and 8 ACs, as mentioned before. In total, 24 variables were used. Since some of the individuals present in the clustering analysis were missing some of the genetic measures needed, the discovery sample size for genetic analyses

was n=1137, while replication n=542 (Pelin et al., 2021).

3.2.3 Clustering analysis

The clustering analysis was conducted in R v3.6.0 using the discovery sample. Clinical variables were scaled and submitted to the HDDC algorithm implemented in the R package *HDclassif* (Bergé et al., 2012). The clustering pipeline consisted of four steps (Pelin et al., 2021):

Step 1 - determining the right regularization. As mentioned before, some parameters of the HDDC can be fixed or allowed to vary within or between the classes. To find the best regularization for our data, the observations were randomly sampled 100 times, and each time 80% of the data was submitted to the HDDC. All possible regularizations were fit for number of clusters ranging from $K = 2$ to $K = 15$. In all runs, the ICL value was computed. At the end of the first step, the model with the best median ICL value was chosen.

Step 2 - determining the optimal number of clusters. The HDDC with the regularization chosen in Step 1 was fit according to the *Leave-one out Jackknife* method. Hence, the chosen model (regularization) type was fit N times (N being the number of individuals in the dataset), each time working with N-1 individuals. Throughout the runs, for each K in the range from $K = 2$ to $K = 15$, ICL was calculated and the cluster number with the best median ICL value was chosen as the optimal one.

Step 3 - determining the final cluster solution. The Jackknife runs from Step 2 resulted in the optimal number of clusters and generated N clusters assignments for each individual. The final consensus on the cluster assignment was reached via *majority voting*, the method that combines the clustering results generated using different data subsamples. The individual is assigned to the cluster where it was grouped most often across subsampling. Majority voting is implemented in the package *diceR* (Chiu and Talhouk, 2018) and ran with argument is.relabelled = False, i.e., the data was relabeled using the first clustering iteration as the reference.

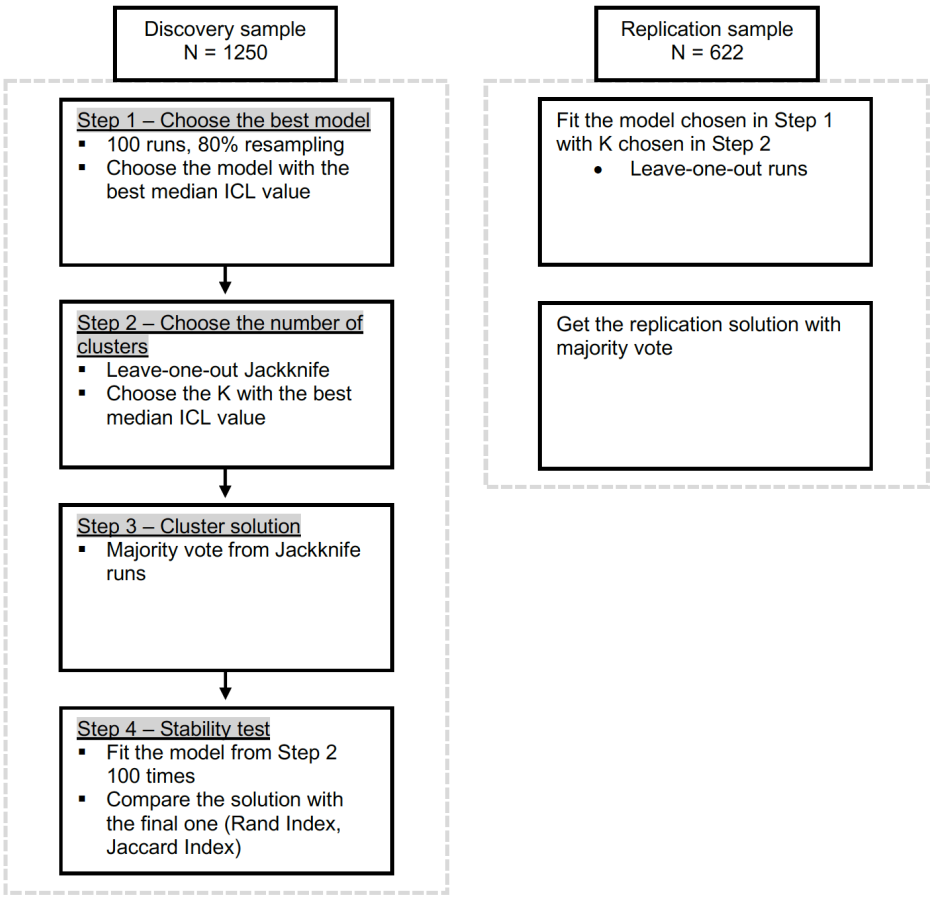
Step 4 - determining the stability of the cluster solution. The chosen regularization was fit on the 95% resampled dataset and the ICL values were recorded throughout 100 runs. The stability solution was compared with the final solution from Step 3 by using the *Rand* and *Jaccard indices*. Note that the 5% holdout for the stability analysis was chosen to generate an additional view of the solution, using the different holdout proportion of 1/20, in-between the 1/5 and 1/1250 employed for the model and K choice, respectively.

Three different resampling strategies were used in the different steps to reduce the risk of overfitting, which is higher when employing only one approach.

External validation of the clustering solution
To check if the clustering solution stays stable on the new, previously unseen, dataset, the algorithm was applied to the replication sample. The model type and cluster number determined in Step 1 and Step 2 of the pipeline were fit with the Leave-one-out Jackknife method to the replication sample and the replication solution was determined by majority voting.

The illustration of the complete four-step clustering workflow with the external validation is shown in Figure 3.1.

Figure 3.1: Clustering pipeline illustration



Source: Pelin et al., 2021

3.2.4 Cluster characterization methods

Clusters were characterized by using supervised learning in the form of classification and statistical testing. Both *one-vs.-all* and *one-vs.-one* strategies were used to characterize each cluster separately.

3.2.4.1 HDDA for identification of important features per cluster

HDDA was used in a post-hoc analysis of clusters obtained by the corresponding HDDC to detect the important features that describe each cluster (Pelin et al., 2021). The model regularization chosen in Step 1 of the clustering pipeline was fit to the same dataset, in the *one-vs.-all* fashion. The prediction model was run in 100 iterations, each time randomly splitting the discovery set to 70% vs. 30% train and test set, respectively. In each run, the AUC was calculated and the average across 100 runs was used as the final measure of classification success. In order to detect the important variables distinguishing each cluster from the others, the variable importance was assessed. It was measured by calculating the decrease in the model's AUC metric after the random permutation of the values of the respective variable. The greater the AUC decrease after the permutation, the more important the variable. The variables were ranked using the average AUC drop after 100 runs. For the calculation of AUC the R package *pROC* was used (Robin et al., 2011).

3.2.4.2 Lasso for cluster prediction using genetic variables

Lasso analyses were conducted using the R package *glmnet* (Friedman et al., 2010). The dataset used consisted of 24 predictor variables (four family history variables, ten PGS variables, 8 ACs, age, and gender).

The cross-validation was used for tuning the λ parameter (Equation 2.5). The sample was split 1000 times, using stratification based on cluster labels, into 70% training and 30% test sets. On each of the 1000 training sets, λ was tuned via 3-fold cross-validation and λ minimizing the cross-validation error, i.e., maximizing the AUC (λ_{min}), was used in each run to obtain the classification metric and Lasso coefficients on the test set. In each run, metrics AUC, sensitivity, and specificity were calculated and the average of all runs was reported. The cut-off value for the sensitivity and specificity was determined using the *MaxSpSe* method from the *OptimalCutpoints* package (López-Ratón et al., 2014).

Finally, Lasso was fit to the full dataset with λ equal to the median value of all

λ_{min} chosen during the tuning procedure. In this way, we obtained the final model, that is, the sets of variables and their corresponding coefficients for the respective cluster prediction (Pelin et al., 2021).

3.2.4.3 Significance testing with genetic variables

Statistical testing was used to infer the significant differences between the clusters with respect to genetic data (Pelin et al., 2021). Moreover, different models were compared for the significant information gain achieved by PGS. The later was done by observing the all cluster labels at the same time, hence with the multiclass outcome variable.

Westfall and Young significance testing

Westfall and Young (W-Y) method was used to detect significant differences among clusters with respect to genetic data. In order to take the possible population structure into account, age, gender, and 8 ancestry components were included in the analysis. The analysis was conducted using the mt.minP function from the Bioconductor's *multtest* package (Pollard et al., 2004). The adjusted p-values were directly estimated via 20 000 permutations with Welch's t statistic. These p-values were additionally corrected for multiple testing, i.e., the number of comparisons made, using Bonferroni's correction. The number of comparisons corresponded to the number of clusters K in the *one-vs.-all* analyses and to $\frac{K(K-1)}{2}$ for *one-vs.-one* analyses. To report the association, a significance threshold $\alpha = 0.05$ was used for the final p-values.

Assessment of PGS information gain

Multinomial logistic regression was used to asses if PGS provide the information gain when describing the clusters. Multinomial logistic regression is a generalization of logistic regression to multiclass classification problem. The outcome variable $\mathbf{Y}$ was representing cluster labels, hence having K different discrete outcomes. The four models with the following sets of variables were compared:

A: PGSs and ancestry components (ACs)

B: family history only
C: family history and ACs
D: the full model with PGS, family history, and ACs

The null-model for the comparison was the one containing only age and gender as predictors. Information gain (the "bonus effect") was measured in terms of Nagelkerke R^2 (Nagelkerke et al., 1991) and Akaike information criterion (AIC) (Akaike, 1998), along with the likelihood ratio test. The likelihood ratio test is a hypothesis test that compares the two *nested models*. Nested models are the models where a more complicated model with more variables can be transformed into the simpler one with less variables. In our example, models A, B, and C are all nested with the full model D, because we can transform the model D into any other by removing the additional predictors. The null hypothesis H_0 of the likelihood ratio test is that the smaller model is "better". Hence, if the H_0 is rejected, then the larger model is a significant improvement over the smaller one. The null model for R^2 calculation was the model including only the two baseline covariates - age and gender. This part of the analysis was done using the R package *VGAM* (Yee, 2010).

3.2.5 Replication analysis

To delineate the pattern of replication, a correspondence between the discovery-stage and the replication clusters, i.e., pairing of each, is required. To identify this, each of the binary *one-vs.-all* HDDA classification models, trained on the discovery dataset, were used for all possible *one-vs.-all* predictions in the replication dataset, thereby conducting K^2 comparisons (Pelin et al., 2021). The discovery-stage model producing the best prediction metric above 70% was chosen to assign the cluster identity in the replication sample. After the discovery and replication clusters were matched, further statistical analysis (Westfall and Young significance testing), was used as a secondary analysis to validate the matching and homogeneity of paired discovery-replication clusters. As for the analysis with genetic data in the discovery sample, p-values adjusted by the Westfall and Young procedure were further corrected using Bonferroni's

method for the five comparisons made.

We analyzed the performance of the discovery-stage Lasso regression models of genetic variables in the replication sample. Here, the models that were trained on the full discovery set with the optimized λ were used. (Pelin et al., 2021)

3.3 Feature selection with genetic data

3.3.1 Participants

The purpose of the feature selection analysis was to explore if the algorithm was able to successfully reduce the set of SNPs in a supervised way (Section 2.3.2). Due to the exploratory nature of the analysis and the computational complexity, we decided to first test the framework on the two outcome variables.

First, we chose the cluster label from the previous analysis providing the biggest balance between classes, which is important for the classification task. Hence, we used the discovery sample from the clustering analysis and the binary cluster label *Cluster 0 vs. all* as a phenotype of interest. Due to missing data in some SNPs, a total of N=1120 individuals were subject to the analysis.
Additionally, we applied the algorithm on the healthy controls and patients diagnosed with MDD (N=1776) from the full FOR2107 sample, which was available at the time this analysis started. The rationale for this step was to have a fair comparison to the already established MDD GWA studies. Apart from that, healthy controls and MDD patients were the two biggest diagnostic categories.

3.3.2 Feature selection analysis

As opposed to GWA studies, which are based on univariate analysis to detect important SNPs, our approach was multivariate, i.e., dealing with more than one SNP in the model (Section 2.3.2). The workflow of the analysis follows the one developed by Dehman et al., 2015 and implemented in the corresponding

BALD (Blockwise Approach using Linkage Disequilibrium) package. First, groups of SNPs were identified by using a clustering algorithm based on the LD metric. Second, Group Lasso was used to select SNPs carrying the information about the phenotype, given their group structure from the clustering step. Even though this step follows the idea from the previously mentioned paper and package, we used the modified version of Group Lasso called Sparse group Lasso, developed by Simon et al., 2013 and covered in Section 2.1.4.1.2 of the Background chapter. The reason for choosing this modified version was that it works with the binary phenotype and enables a selection of SNPs within the groups as well. The Sparse group Lasso package with its documentation can be found on `https://group-lasso.readthedocs.io/en/latest/maths.html`.

3.3.2.1 Clustering of SNPs based on LD

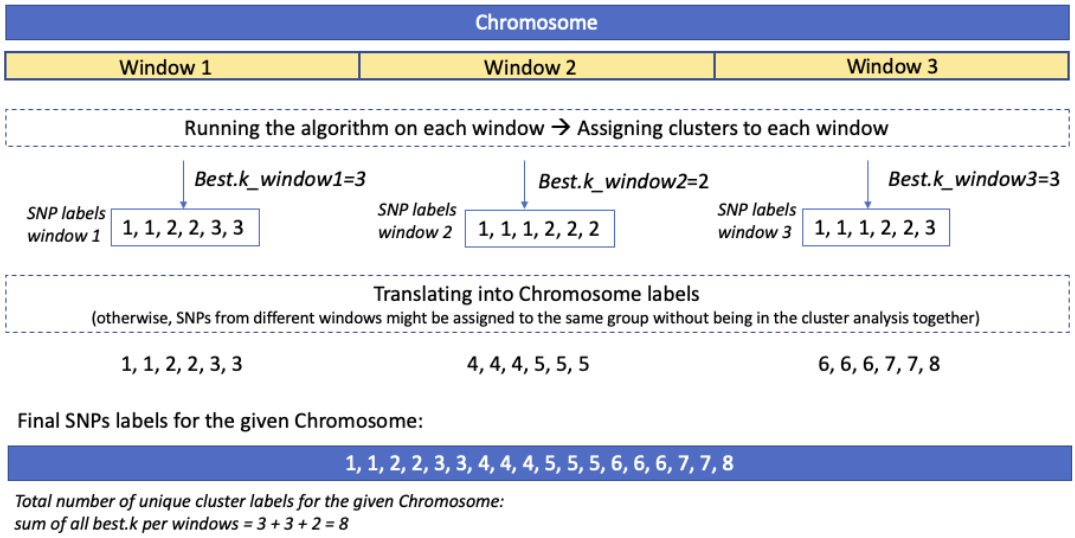
To apply the Sparse group Lasso for feature selection, groups of features entering the analysis have to be inferred. To achieve this, the first step of BALD algorithm uses hierarchical clustering based on Ward's method (Ward Jr, 1963) with LD similarity metric.
To determine the optimal number of clusters, the second step of BALD algorithm uses the modified gap statistic method (Tibshirani et al., 2001). The metric compares the total dispersion within the clusters for the possible cluster numbers K with their expected values under the null reference distribution of the dataset, i.e., the distribution with no obvious grouping among the features. For more details on the algorithm, please refer to the original paper (Dehman et al., 2015).

Workflow

For the SNP clustering analysis, the full FOR2107 sample was used and subject to the data preprocessing (removal of related individuals and SNPs with the minor allele frequency < 0.05). Clustering of SNPs was performed on the full FOR2107 sample of N=2166 individuals. Since LD is an association between SNPs observed on the same chromosome, clustering was performed chromosome-wise. To further reduce the computational complexity, we used

a sliding window approach, which is typically used in LD analysis. A fixed window size (WS) was 500 SNPs, and the windows were non-overlapping. If the last window of the chromosome had less than $\frac{1}{2}$WS = 250 SNPs, the window was merged with the previous one. For each window, Ward algorithm based on LD metric was run, together with the gap statistic. The later was used to determine the best cluster number (best.K$_{window}$) among all possible numbers of clusters ($K = 1, \ldots, p_{window} - 1$, where p_{window} is the number of SNPs in the given window). Hence, each run resulted with cluster labels $\in \{1, \ldots, \text{best.K}_{window}\}$ for the SNPs in the given window. After the cluster analysis of all windows in the chromosome was done, those labels were merged into the final labels for the given chromosome, resulting in all together $\sum_{\text{windows}} \text{best.K}_{\text{window}}$ unique cluster labels. As written before, R package BALD was used for this part of the analysis. For the example and illustration of the workflow, see Figure 3.2.

Figure 3.2: Clustering of SNPs based on LD example
Example of the SNP clustering workflow. Let us assume there are 3 windows on the arbitrary Chromosome and that each window has 6 SNPs. The algorithm is run on each window, resulting in the optimal cluster number per window (best.K_{window}), obtained by gap statistic and the cluster labels for each SNP. After all windows are ran through, final cluster labels for the Chromosome are formed, resulting in the $\sum_{\text{windows}} \text{best.K}_{\text{window}} = 3 + 2 + 3 = 8$ unique cluster labels.

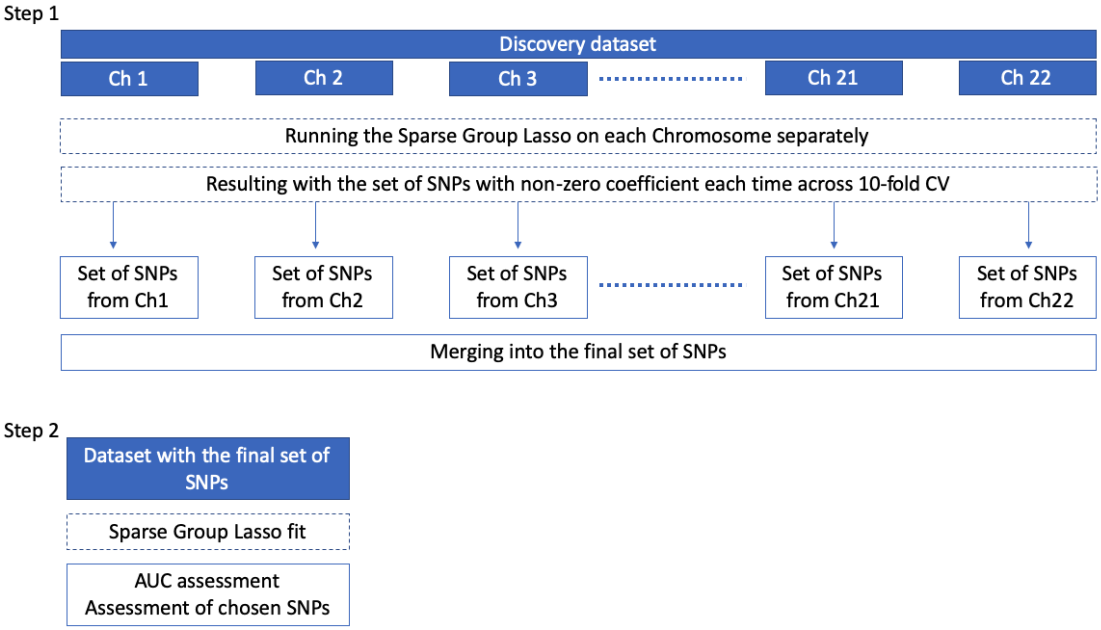


3.3.2.2 Sparse group Lasso for SNP selection

Due to computational complexity, Sparse group Lasso was run chromosome-wise with the 10-fold cross-validation. A pilot study was done testing the different λ parameters for the regularization penalty. Apart from the default ones from the Sparse group Lasso package (λ_1, λ_2 = 0.05), 10 other parameters on the scale from 0 to 0.05 were tested. All parameters except 0.001 resulted either in the unfeasible computational time required to run or as the empty models (0 variables chosen). Hence, with the aim to test the performance of the algorithm, a $\lambda_1 = \lambda_2 = 0.001$ were chosen for controlling the regularization penalty.

In each of the 10 folds (iterations), the model resulted in the chosen set of SNPs, i.e, SNPs with the non-zero coefficient. The final set of SNPs, entering the next step, was obtained by merging the SNPs that were chosen every time for all the chromosomes 1 to 22.
Lastly, the dataset with the final set of SNPs chosen on all chromosomes, was subject to Sparse group Lasso analysis with the 10-fold cross-validation. The average AUC of 10 runs served as a metric of classification and feature selection success. The analysis framework is shown on Figure 3.3.
For this part of the analysis python and the aforementioned Sparse group Lasso package (`https://group-lasso.readthedocs.io/en/latest/maths.html`) was used. ‘

Figure 3.3: Sparse group Lasso analysis steps
Overview of the sparse group Lasso analysis, described above.



Analysis of the sets

For the feature selection methods, it is important to determine if the feature set is stable, meaning that the same set of features is selected even after the data perturbation. We inferred this stability by checking the SNP intersections between runs of 10-fold cross-validation.

To infer the stability across iterations, the percentage of SNPs in common between all pairs of iterations was determined. First, for each pair of iterations i and j, the size of the intersection set, i.e the number of SNPs in common, was determined. In the following, it is marked as $|S_i \cap S_j|$, for $i, j = 1, \ldots, 10$, where S_i and S_j are the sets of SNPs chosen in the iterations i and j, respectively.
The percentage of SNPs in common between two iterations i and j is calculated with respect to the maximum possible number of SNPs in the given intersection, which is the minimum between number of SNPs chosen in the iteration i and the number of SNPs chosen in the iteration j. Hence, the formula:

$$\text{overlap}_{\text{i,j}} = \frac{|S_i \cap S_j|}{min\left\{\text{nr. of SNPs chosen in } i, \text{nr. of SNPs chosen in } j\right\}}, \quad i, j = 1 \ldots 10.$$

For example, if there are 10 SNPs chosen in the iteration 1, 8 SNPs chosen in the iteration 2, and there are 6 SNPs in common (i.e., in the intersection of the two sets), then the percentage of overlap is $\frac{6}{min\{8,10\}} = \frac{6}{8}$ or 75%.

Results 4

4.1 Transdiagnostic subtypes discovery with unsupervised learning

The results from Section 4.1 are based on the peer-reviewed publication by the thesis author Pelin et al., 2021.

4.1.1 Sample characterization

The discovery sample for the clustering analysis consisted of N = 1250 individuals, had an average age of 35 years, and 61% of females. The biggest diagnostic group for the discovery sample was healthy controls, followed by MDD patients.
The replication sample for the clustering consisted of N=622 individuals, had an average age of 36 years, and 65% of females. The biggest diagnostic category for the replication sample was MDD patients, followed by healthy controls.
Age, gender and diagnosis between discovery and replication sample were compared for the significant differences: diagnosis and age differed significantly (p=0.002 and p = 0.01, respectively). Proportions of single diagnostic categories were significantly different for healthy controls (p=0.005) and MDD (p=0.003). Other categories did not show significant differences (BD, p=0.4; SCZ, p=0.1; SZA, p=1; Other diagnosis, p=0.07). Gender was not significantly different (p=0.16). A general description of both samples is shown in Table 4.1, while the mean (SD) of all variables used in the clustering process are shown in Table A5.1 in the Appendix.

Table 4.1: Discovery and replication dataset general description

Variable	Discovery	Replication
N	1250	622
Demographics		
Age, mean (SD)	35.1 (13.0)	36.3 (12.6)
Gender - male, N (%)	483 (39%)	219 (35%)
Years of education, mean (SD)	13.5 (2.6)	13.8 (2.8)
Living with partner, N (%)	277 (28%)	199 (33%)
BMI, mean (SD)	25.3 (5.5)	25.6 (5.4)
Family history (any psychiatric disorder), N(%)	533 (43%)	306 (50%)
Diagnosis		
Age at onset*, mean (SD)	25.2 (11.9)	24.1 (11.7)
Healthy controls, N (%)	590 (47%)	240 (39%)
BD, N (%)	75 (6%)	44 (7%)
MDD, N (%)	477 (38%)	283 (45%)
SCZ, N (%)	53 (4%)	17 (3%)
SZA, N (%)	25 (2%)	13 (2%)
Other, N (%)	30 (2%)	25 (4%)

*Age at onset (AAO) not available for healthy controls

Source: Pelin et al., 2021

4.1.2 Clustering analysis

4.1.2.1 Clustering pipeline results

Our clustering pipeline consisted of four steps: first, to identify the best fitting model regularization; second, to select the optimal number of clusters; third, to get the final labels for each observation, and fourth, to infer the stability of the solution (Section 3.2.3).

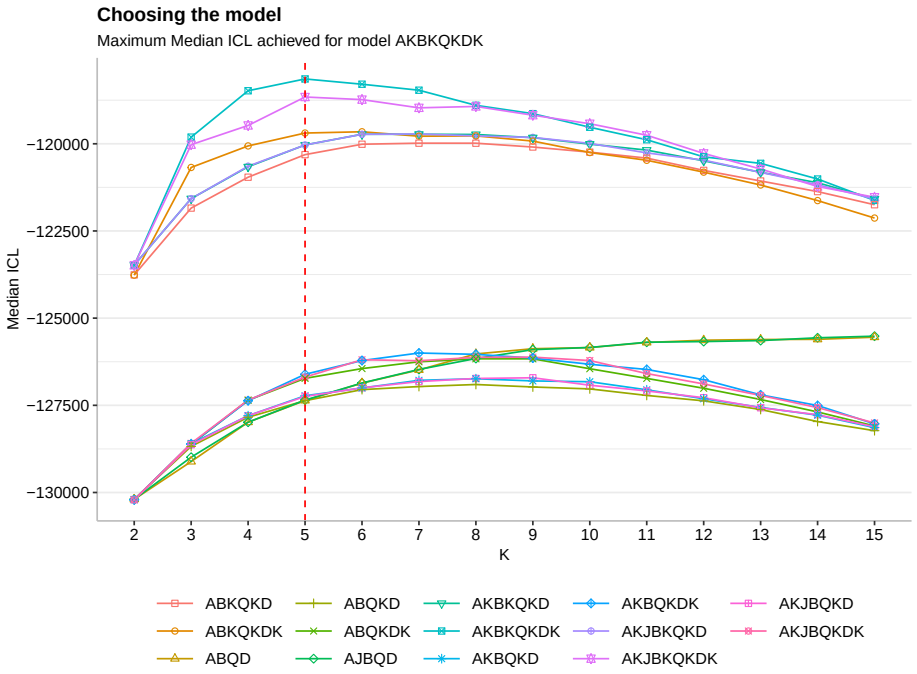
Among the 14 possible model regularizations, there were two groups created (Figure 4.1). Models that allow for modeling of the class-specific noise are in the better fitting group (higher ICL value). Among those, the model regularization $A_k B_k Q_k D_k$ achieved the highest median ICL value, and therefore was chosen as the optimal one and submitted to the next step. The selected model $A_k B_k Q_k D_k$ has the following characteristics, explained in `https://cran.r-project.org/web/`

`packages/HDclassif/HDclassif.pdf`:

A_k: The classes have different parameters but there is only one parameter per class
B_k: Each class has its proper noise
Q_k: All classes have their proper orientation matrix
D_k: The dimensions are free and proper to each class.

In the context of the data, the optimal model that was chosen assumes that there is a cluster-specific variance (A_k), cluster-specific noise (B_k), and that clusters can have different orientations (Q_k). Since the HDDC algorithm is the subspace clustering algorithm based on the assumption that the clusters live in the subspaces of a lower dimension than the initial one, D_k is in this context considered to be the dimension of the cluster-specific subspace. For more theoretical details, see Bouveyron et al., 2007.

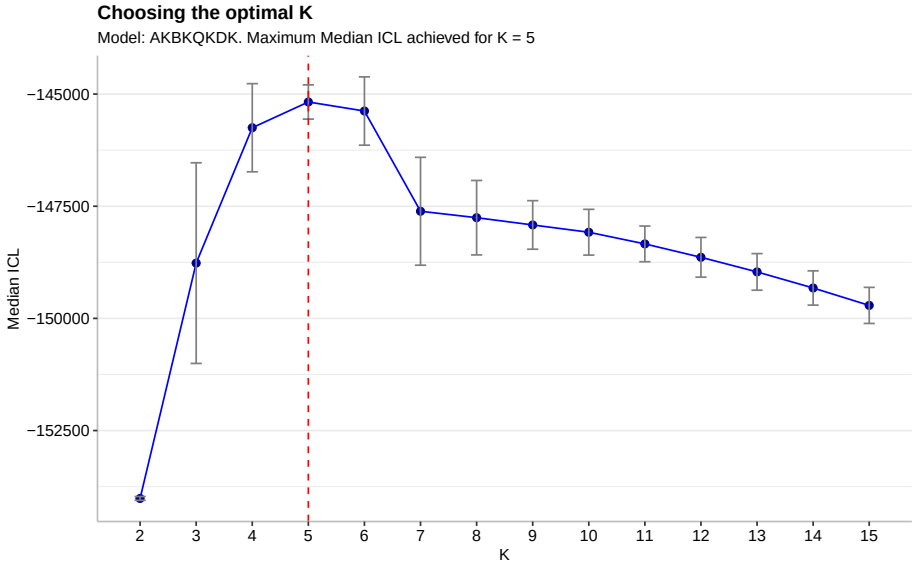
Figure 4.1: Choosing the optimal model regularization
Two groups of models were created (separated by a substantial difference in median ICL). The group achieving higher ICL values consisted of models allowing for modeling of class-specific noise. Among these models, the model $A_k B_k Q_k D_k$ achieved the highest median ICL value and was submitted to the second step



Source: Pelin et al., 2021

The second step of the clustering pipeline consisted of Leave-one-out Jackknife runs, resulting in 1250 solutions for each $K = 2, \ldots, 15$. The highest median ICL was achieved for $K = 5$ (Figure 4.2).

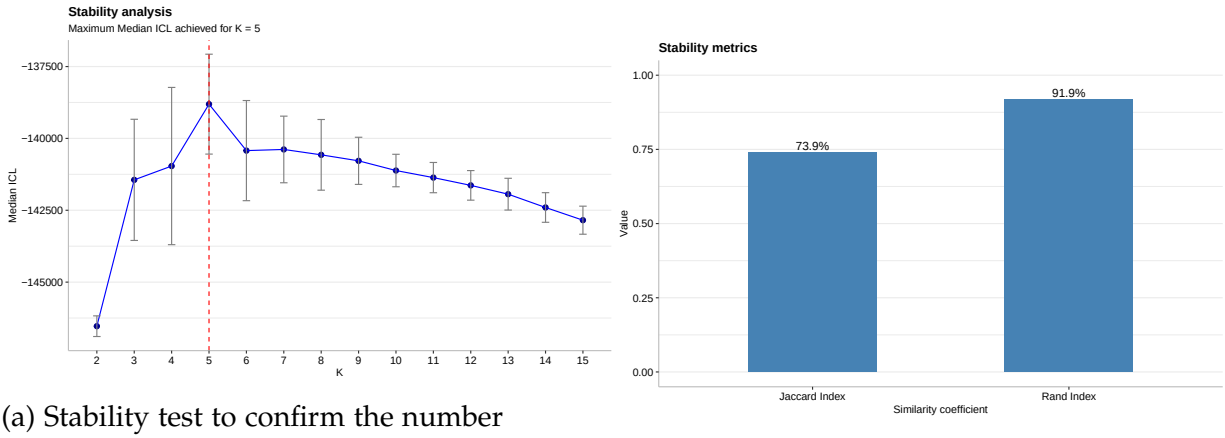
Figure 4.2: Choosing the optimal cluster number
Error bars show median absolute deviations



Source: Pelin et al., 2021

In the third step, the majority voting scheme assigned the labels to all individuals based on 1250 runs and $K = 5$, thus forming the 5 final clusters. The stability test, i.e., the additional 100 runs of fitting the model $A_k B_k Q_k D_k$, also resulted with $K = 5$ as the optimal cluster number (Figure 4.3a). The calculation of similarity between the stability solution and the final one resulted in a Rand Index of 92% and Jaccard Index of 74% (Figure 4.3b).
Importantly, $K = 5$ was selected as the optimal choice in all three steps, which all use different resampling strategies to reduce the risk of overfitting.

Figure 4.3: Stability test



(a) Stability test to confirm the number of clusters. Error bars show median absolute deviations

(b) Comparison of stability solution to the final solution

Source: Pelin et al., 2021

4.1.2.2 Cluster ranking and distribution

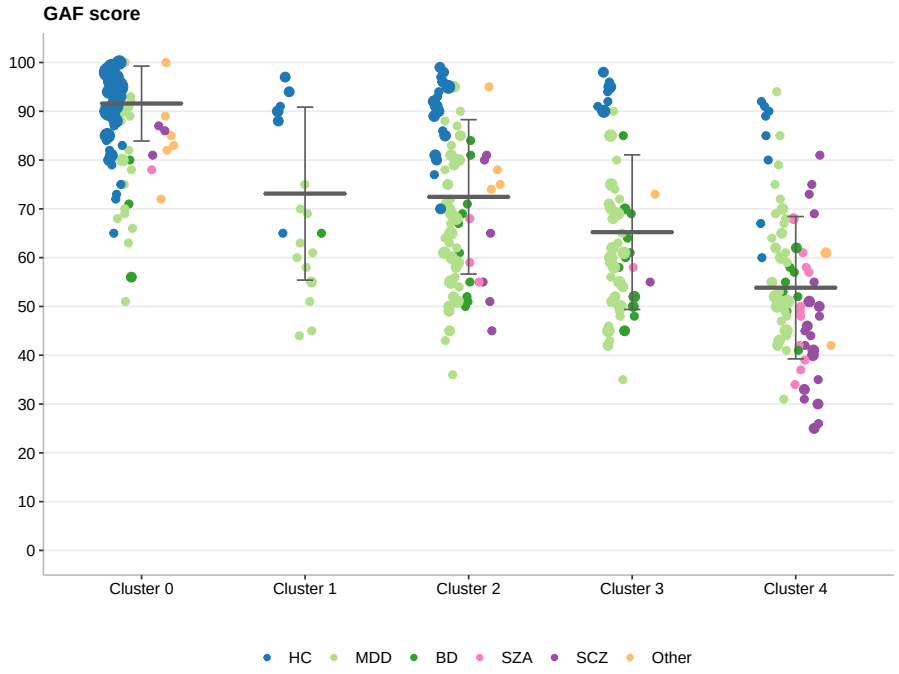
Step 3 of the pipeline resulted in the assignment of the cluster labels for all individuals, forming the $K = 5$ final clusters of different sizes (Table 4.2): Cluster 0 ($n = 535$); Cluster 1 ($n = 38$); Cluster 2 ($n = 266$); Cluster 3 ($n = 215$); Cluster 4 ($n = 196$). The clusters were ordered based on GAF score (Figure 4.4), in order to create the notion of a severity continuum.

Figure 4.5 and Table 4.2 show the distribution of diagnosis across clusters. The biggest Cluster 0 consisted of the majority of healthy controls, while others were diagnostically mixed. Patients diagnosed with SCZ and SZA mostly clustered together into the highest severity Cluster 4, while BD and MDD patients were present in all clusters. The latter induced a question of the existence of different subtypes of these disorders, which was additionally inspected and will be shown later in this chapter.

Table 4.2: Cluster sizes and diagnosis

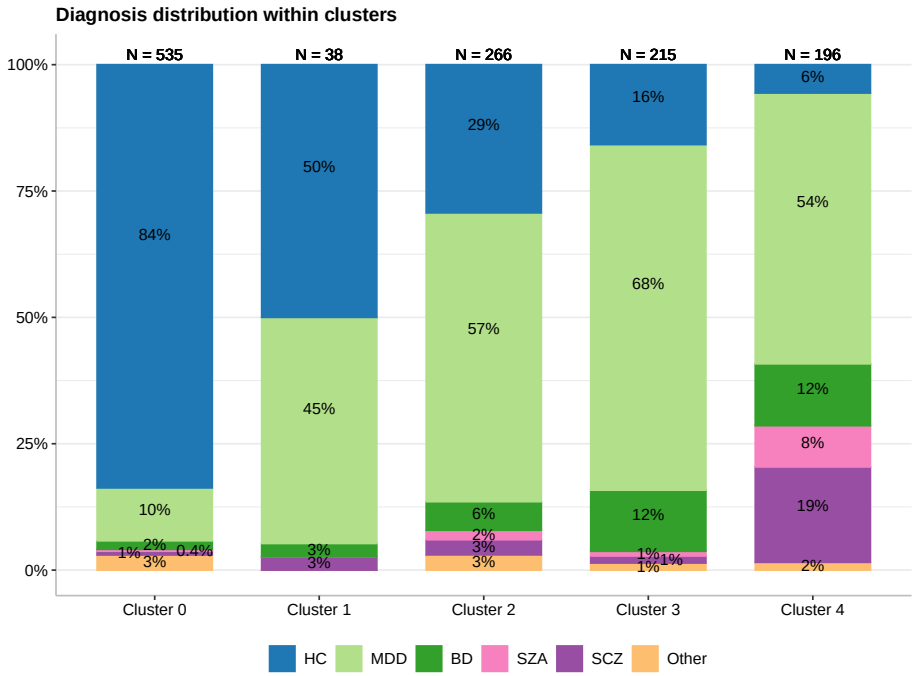
Cluster	HC	BD	MDD	SZA	SZC	Other	Total
Cluster 0	448	9	56	2	4	16	535
Cluster 1	19	1	17	0	1	0	38
Cluster 2	78	15	152	5	8	8	266
Cluster 3	34	26	147	2	3	3	215
Cluster 4	11	24	105	16	37	3	196
Total	590	75	477	25	53	30	1250

Figure 4.4: Ordering of the clusters based on GAF score
The mean is represented by a horizontal line and the standard deviation by error bars. The dot size is relative to the number of individuals having the given value.



Source: Pelin et al., 2021

Figure 4.5: Diagnosis distribution within clusters



Source: Pelin et al., 2021

4.1.3 Cluster characterization

4.1.3.1 Phenotypic characterization of clusters

In order to detect the important features distinguishing each cluster from the rest, HDDA was fit in a *one-vs.-all* fashion using the same 57 variables from the clustering process. The top 10 most important variables are shown in Table 4.3, while the distributions of the top 2 most important variables for each cluster are shown in Figure 4.6.

Variables capturing antecedent events - different types of childhood maltreatment (CTQs) and parental bonding, together with the current quality of life (SF 36) were important for many clusters. As Figure 4.6 shows, Cluster 2 and Cluster 4 showed on average much higher values of emotional neglect and lower parental bonding. Cluster 4 also showed a much higher average score for sexual maltreatment in childhood. In

terms of limitations in daily life, Cluster 3 had lower average values, meaning that their life might be negatively impacted by the disease. Apart from the antecedent events, positive and mania symptoms (SAPS, YMRS) appeared among the top contributing variables, especially for Cluster 4 - these individuals showed higher average values for the respective scores than the rest (Figure 4.6, Table A5.3 in the Appendix).

Table 4.3: Most important features from the HDDA analysis
The top 10 most important variables identified by HDDA, as explained in Section 3.2.4.1.

Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
CTQ Sexual abuse	SF36 Bodily pain	Maternal bonding	SF36 Role physical	SAPS (Positive symptoms)
LEQ Negative Event Score	SF36 Role physical	CTQ Emotional neglect	CTQ Emotional neglect	YMRS
SAPS (Positive Symptoms)	NEO-FFI Agreeableness	Paternal bonding	Maternal bonding	CTQ Physical abuse
CTQ Physical abuse	RSQ Lack of trust	SF36 Physical functioning	CTQ Emotional abuse	CTQ Sexual abuse
SF36 Physical functioning	Maternal bonding	SF36 Role physical	CTQ Physical neglect	LEQ Negative event score
YMRS	STAIS	ACE Sum	ACE Sum	SANS (Negative symptoms)
SANS (Negative symptoms)	SHAPS	RSQ Avoidance of closeness	SF36 Physical functioning	CTQ Physical neglect
SF36 Role physical	SF36 Physical functioning	CTQ Emotional abuse	CTQ Physical abuse	LEQ Positive event score
SCL90R Phobic anxiety	NEO-FFI Neuroticism	CTQ Physical abuse	SF36 Energy	VLMT Sum
CTQ Physical neglect	SAPS (Positive symptoms)	SAPS (Positive Symptoms)	Paternal bonding	CTQ Emotional abuse

Source: Pelin et al., 2021

Figure 4.6: Distributions of the two most important variables per cluster
The figure shows the two most important variables for each cluster, based on the HDDA analysis shown in Table 4.3. The mean is represented by a horizontal line and the standard deviation by error bars. The dot size is relative to the number of individuals having the given value. The dark square on top indicates the cluster for which the variable was among the top two most important variables.

Source: Pelin et al., 2021

In Tables 4.4 and A5.3 in the Appendix, we can further inspect the distributions of the important variables per cluster in order to characterize them in detail. Table A5.3 in the Appendix shows the mean (SD) of all variables used in the clustering process, for the full discovery dataset and each cluster.

Individuals in Cluster 0 were the youngest cluster on average, with the mean (SD) of 31.7 (11.9), and had the highest number of education years (14.4 years on average). In general, this cluster was characterized by the lowest severity in the majority of measures, including the lowest maltreatment scores, positive symptoms, and depression scores. Moreover, it showed the best quality of life metrics and the lowest proportion of individuals with cases of psychiatric disorder in the family (26%). Cluster 4, on the other hand, showed severe impairment in many measures, particularly regarding childhood maltreatment and prevalence of positive, negative, and mania symptoms. In the clustering procedure, only the sum scores of both positive and negative symptoms were used in order not to over-represent the psychotic patients. Their subscales were inspected post-hoc and are shown in Table 4.5. Among the positive symptoms experienced by individuals in Cluster 4, delusions and positive formal thought disorder were the most prominent ones. Moreover, the individuals in Cluster 4 showed the highest severity in other additional variables examined post-hoc (Table A5.2 in the Appendix).

Table 4.4: Characterization of the discovery sample and clusters

Variable	Discovery	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
N	1250	535	38	266	215	196
Demographics						
Age, mean (SD)	35.1 (13.0)	31.7 (11.9)	38.6 (13.9)	35.3 (12.7)	37.9 (13.7)	40.3 (12.8)
Gender - male, N (%)	483 (39%)	205 (38%)	9 (24%)	106 (40%)	74 (34%)	89 (45%)
Years of education, mean (SD)	13.5 (2.6)	14.4 (2.4)	13.3 (2.6)	13.8 (2.6)	13.1 (2.8)	12.1 (2.7)
Living with partner, N (%)	277 (28%)	108 (20%)	7 (18%)	51 (19%)	62 (29%)	49 (25%)
BMI, mean (SD)	25.3 (5.5)	23.7 (4.3)	24.9 (5.1)	24.8 (4.9)	27.0 (6.6)	28.1 (6.4)
Age at onset*, mean (SD)	25.2 (11.9)	24.5 (9.9)	29.6 (13.3)	23.3 (11.5)	27.8 (12.8)	24.4 (11.4)
Family history (any psych. disorder), N (%)	533 (43%)	141 (26%)	16 (42%)	148 (56%)	105 (49%)	123 (63%)
Quality of life (SF36), mean (SD)						
General health	66.0 (23.4)	81.2 (14.0)	57.9 (26.6)	64.7 (19.8)	49.4 (20.7)	45.8 (21.3)
Mental health	64.3 (22.6)	81.4 (9.8)	55.6 (23.0)	59.4 (18.9)	46.4 (19.3)	45.6 (21.3)
Depression and anxiety, mean (SD)						
HAMA sum	7.3 (7.9)	2.2 (2.5)	9.7 (8.6)	7.6 (6.4)	13.1 (8.6)	14 (8.8)
HAMD sum	5.4 (6.6)	1.1 (1.6)	6.3 (6.7)	5.9 (5.7)	9.9 (7.0)	11 (7.5)
BDI sum	10.7 (10.8)	3.2 (3.3)	15.1 (12.8)	12.7 (9.6)	17.6 (10.1)	20.3 (11.7)
Positive, negative and manic symptoms, mean (SD)						
SANS	5.7 (9.9)	0.6 (1.7)	5.2 (8.6)	6.9 (9.6)	8.0 (9.1)	15.6 (14.4)
SAPS	1.4 (5.2)	0.1 (0.6)	0.1 (0.4)	0.6 (1.6)	0.7 (1.8)	6.7 (11.5)
YMRS	1.2 (2.5)	0.5 (1.1)	0.8 (1.4)	1.1 (1.8)	1.3 (2)	2.9 (4.9)
Maltreatment in Childhood and Youth (CTQ), mean (SD)						
Emotional abuse	9.1 (4.7)	6.3 (1.7)	9.9 (4.9)	11.4 (4.2)	8.0 (3.0)	14.6 (5.9)
Emotional neglect	10.7 (5.2)	7.5 (2.6)	11.8 (5.4)	14.1 (4.1)	9.2 (3.5)	16.5 (5.4)
Physical abuse	6.2 (2.6)	5.3 (0.7)	6.4 (2.2)	6.4 (2)	5.5 (1.1)	9.5 (4.6)
Physical neglect	7.2 (2.7)	5.8 (1.4)	7.2 (2.2)	8.0 (2.2)	6.4 (1.6)	10.4 (3.7)
Sexual abuse	5.8 (2.5)	5.1 (0.4)	5.8 (2)	5.8 (1.9)	5.6 (1.7)	8.0 (4.9)

*Age at onset (AAO) not available for healthy controls

Source: Pelin et al., 2021

Table 4.5: Positive and negative symptoms per cluster

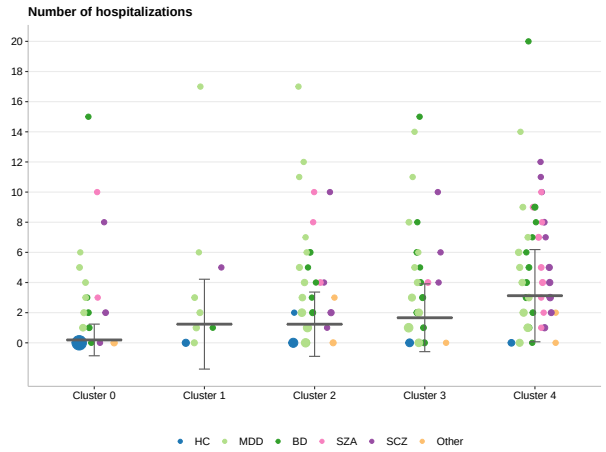
Variable	Discovery	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Positive symptoms (SAPS), mean (SD)						
Hallucinations	0.3 (1.8)	0.03 (0.21)	0 (0)	0.03 (0.23)	0.07 (0.55)	1.49 (4.24)
Bizarre Behavior	0.09 (0.7)	0.01 (0.13)	0 (0)	0.06 (0.31)	0.08 (0.53)	0.41 (1.53)
Positive Formal Thought Disorder	0.5 (1.9)	0.08 (0.43)	0.05 (0.32)	0.29 (1.1)	0.5 (1.4)	2.33 (4.13)
Delusions	0.5 (2.5)	0.02 (0.15)	0.05 (0.23)	0.19 (0.87)	0.09 (0.49)	2.48 (5.75)
Negative symptoms (SANS), mean (SD)						
Anhedonia	1.9 (3.5)	0.1 (0.7)	2.1 (3.5)	2.3 (3.4)	3.1 (4.1)	4.5 (4.9)
Affective blunting	1.5 (3.6)	0.3 (0.9)	1.4 (2.9)	1.83 (3.9)	1.72 (3.26)	4.2 (5.9)
Avolition / Apathy	1.3 (2.5)	0.09 (0.47)	0.8 (1.8)	1.4 (2.4)	1.8 (2.4)	3.8 (3.7)
Alogia	0.5 (1.6)	0.07 (0.4)	0.2 (0.9)	0.7 (1.8)	0.6 (1.5)	1.5 (2.8)

Source: Pelin et al., 2021

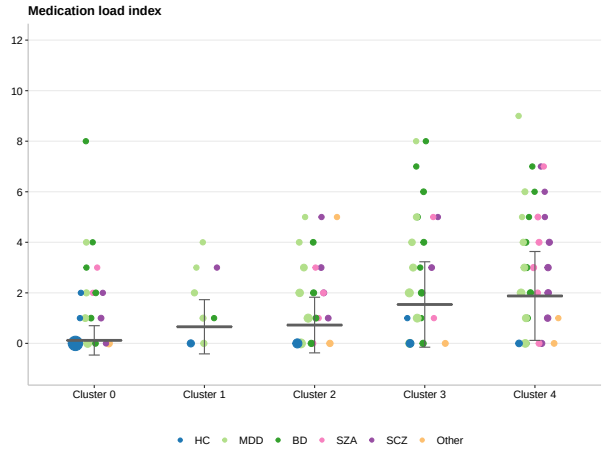
Other clusters ranked between these two extremes. The smallest Cluster 1 showed lower scores of mental health, increased depression and anxiety score, and above average perceived life stress (Table A5.3 in the Appendix). Cluster 2 had average general health ratings but decreased mental health and higher levels of emotional maltreatment in childhood. Cluster 3 contained the highest proportion of affective disorders with elevated anxiety and depression levels, and these individuals showed notably lower mental and general health.

The severity spectrum was also observed when looking at the hospitalization and medication, both not used in the clustering procedure (Table A5.2 in the Appendix). Individuals in Cluster 4 had the highest medication load index (Redlich et al., 2014) and were hospitalized the most. Cluster 0 also showed the lowest severity concerning measures not used in the clustering process (Figure 4.7a, 4.7b). Cluster 3 had the highest antidepressant intake, with 57% of individuals taking it, while Cluster 4 showed the highest antipsychotic intake of 43% (Figure 4.7c).

Figure 4.7: Medication and hospitalization
The mean is represented by a horizontal line and the standard deviation by error bars. The dot size is relative to the number of individuals having the given value.



(a) Number of hospitalizations



(b) Medication load index (Redlich et al., 2014)

Medication	Discovery	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Antidepressant, yes, (%)	28%	4%	32%	35%	57%	51%
Antipsychotic, yes, (%)	15%	2%	11%	15%	21%	43%
Mood stabilizer, yes, (%)	5%	2%	3%	3%	10%	13%
Antidepressant + Antipsychotic, yes, (%)	9%	1%	5%	11%	16%	21%
Antidepressant + Mood stabilizer, yes, (%)	3%	1%	3%	1%	7%	8%
Antipsychotic + Mood stabilizer, yes, (%)	2%	1%	3%	2%	4%	7%

(c) Specific medication group intake

Source of figures (a) and (b): Pelin et al., 2021

4.1.3.2 Genetic characterization of clusters

Lasso analysis

Lasso regression was used to predict the cluster labels in a *one-vs.-all* and *one-vs.-one* fashion using genetic variables (10 PGS and 4 self-reported family history assessments), together with age, gender, and 8 ancestry components as covariates (Section 3.2.4.2, Pelin et al., 2021). Comparisons with prediction AUC >60% are shown in Table 4.6. Lasso models were also applied to the complete discovery sample to assign the final coefficients to the variables (Methods Section 3.2.4.2). The coefficients for each cluster are shown in Figures 4.8-4.10 on the following pages. They are interpreted in the following way: a positive (negative) sign of the coefficient implies that the given predictor is more likely to be higher (lower) in the corresponding cluster, while a coefficient of zero indicates that the respective variable was not important in the respective model.

Two *one-vs.-all* and five *one-vs.-one* comparisons yielded an AUC > 60%. The two extreme clusters, Cluster 0 and Cluster 4, could be best distinguished both when compared to the rest and together. The summary statistics from 1000 runs of Lasso regularized regression models are shown in Table A5.4 in the Appendix.

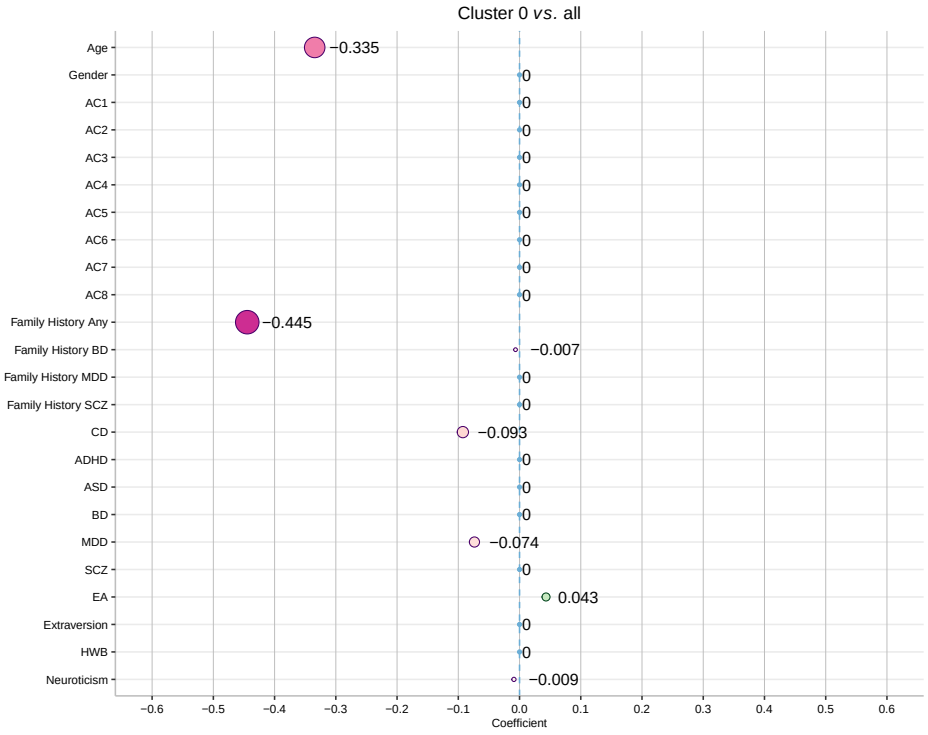
Table 4.6: Metrics of genetic Lasso regularized regression prediction models
An AUC between 50% and 60% is considered random model performance. The table lists models with an AUC >60%, i.e., showing the above-random performance.

Model	AUC	Sensitivity	Specificity
Cluster 0 vs. all	71%	66%	66%
Cluster 4 vs. all	73%	67%	67%
Cluster 0 vs. Cluster 4	81%	75%	75%
Cluster 0 vs. Cluster 2	67%	67%	63%
Cluster 2 vs. Cluster 4	67%	64%	62%
Cluster 0 vs. Cluster 3	66%	63%	63%
Cluster 3 vs. Cluster 4	64%	61%	60%

Source: Pelin et al., 2021

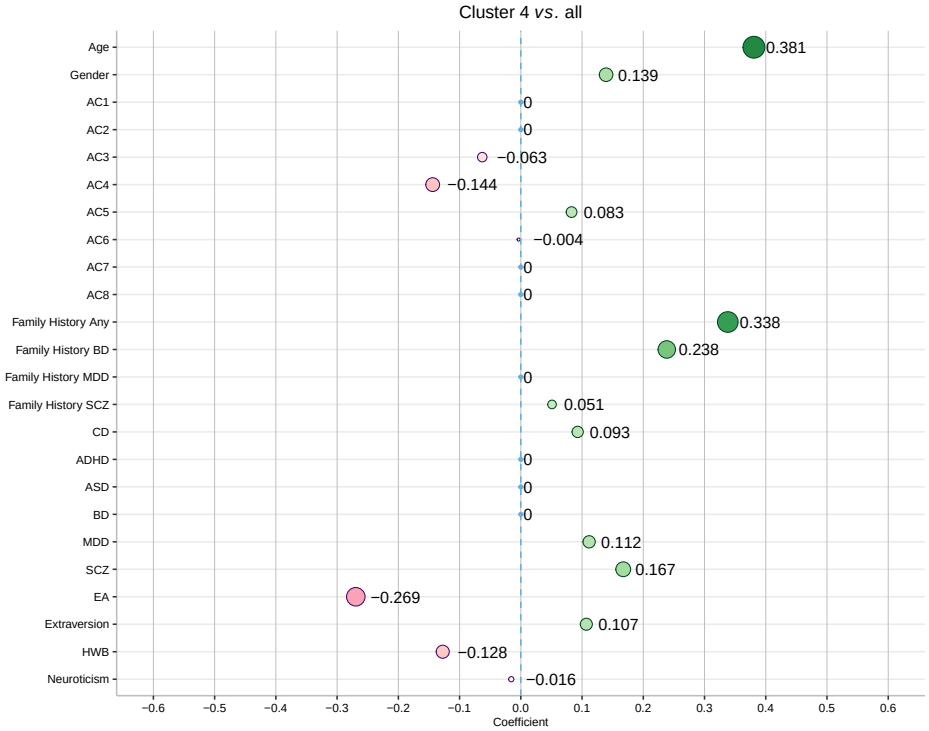
Cluster 0 could be distinguished from the rest with AUC = 71%. The final Lasso model resulted in 7 variables with a non-zero coefficient (Figure 4.8). A genetic variable with the highest absolute coefficient (the strongest effect) was the family history of any disorder and its negative sign indicates that the individuals in Cluster 0 are more likely to have fewer cases of psychiatric disorder in the family, compared to other clusters. Five PGSs had non-zero coefficients and all but the Educational attainment PGS had a negative sign. Hence, individuals in this cluster are more likely to have higher PGS for Educational attainment, while lower for neuroticism, psychiatric cross-disorder, and MDD, compared to others. Based on the coefficients, they are also less likely to have cases of BD in the family. Covariate age also showed a high effect in the negative direction, which suggests that individuals in Cluster 0 are more likely to be younger.

Figure 4.8: Lasso coefficients for *Cluster 0 vs. all* model
The dot size and color gradient (green for coefficients > 0, magenta for coefficients < 0) are proportional to the absolute value of the coefficient



Cluster 4 could be distinguished from the rest with AUC = 73%. The final Lasso model resulted in 16 variables with a non-zero coefficient (Figure 4.9). A genetic variable with the highest absolute coefficient (the strongest effect) was the family history of any disorder, and its positive sign indicates that the individuals in Cluster 4 are more likely to have cases of psychiatric disorder in the family, compared to other clusters. Seven PGS variables had non-zero coefficients - these individuals are likely to have higher PGS for psychiatric cross-disorder, MDD, SCZ, and extraversion. Educational attainment PGS showed the strongest PGS effect, in the negative direction. These individuals are also more likely to have cases of BD and SCZ in the family (positive coefficients for the two variables). Covariates age and gender have higher coefficients, indicating the increased likelihood of being older and male for the individuals in this cluster, compared to the rest.

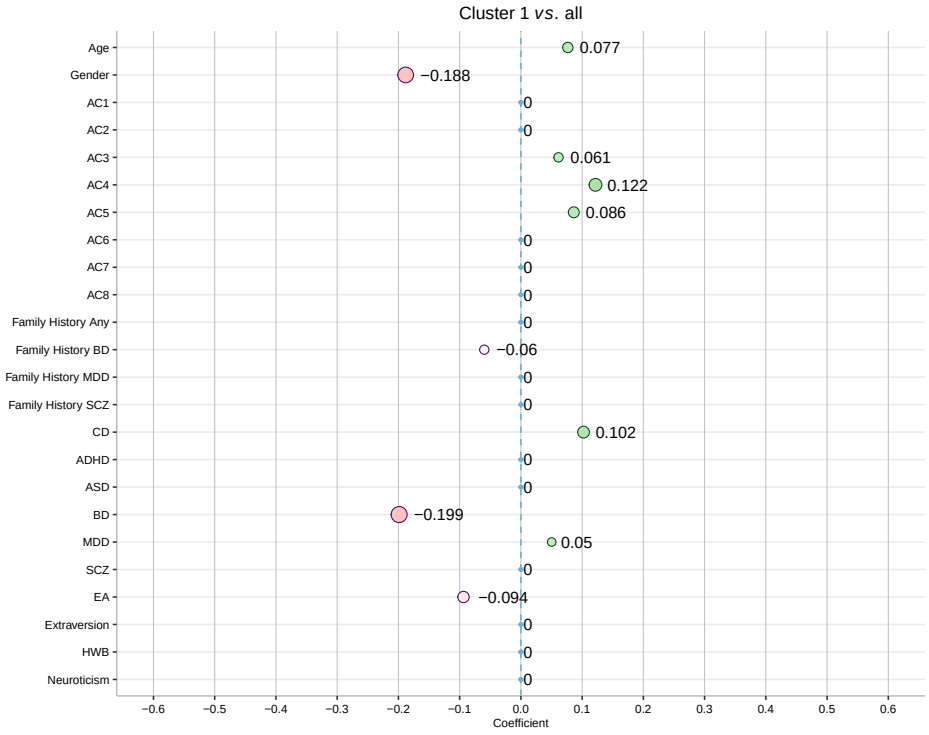
Figure 4.9: Lasso coefficients for *Cluster 4 vs. all* model
The dot size and the color gradient (green for coefficients > 0, magenta for coefficients < 0) are proportional to the absolute value of the coefficient



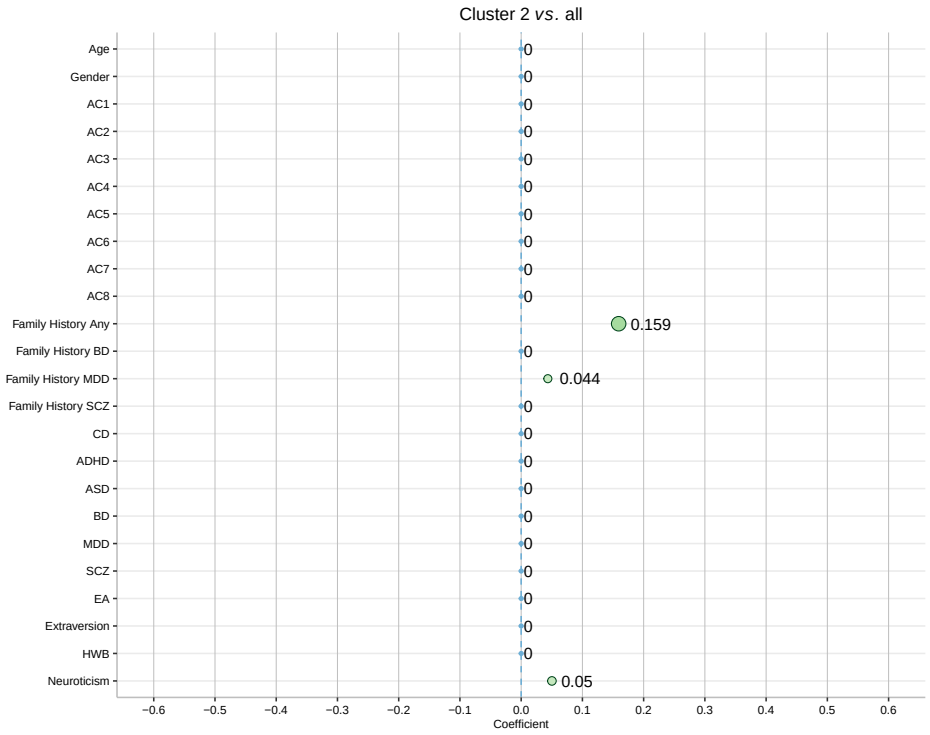
Other *one-vs.-all* Lasso models resulted in low AUC. Nevertheless, the final coefficients were inspected and are shown on the Figure 4.10.

Figure 4.10: Lasso coefficients for other *one-vs.-all* models
The dot size and the color gradient (green for coefficients > 0, magenta for coefficients < 0) are proportional to the absolute value of the coefficient

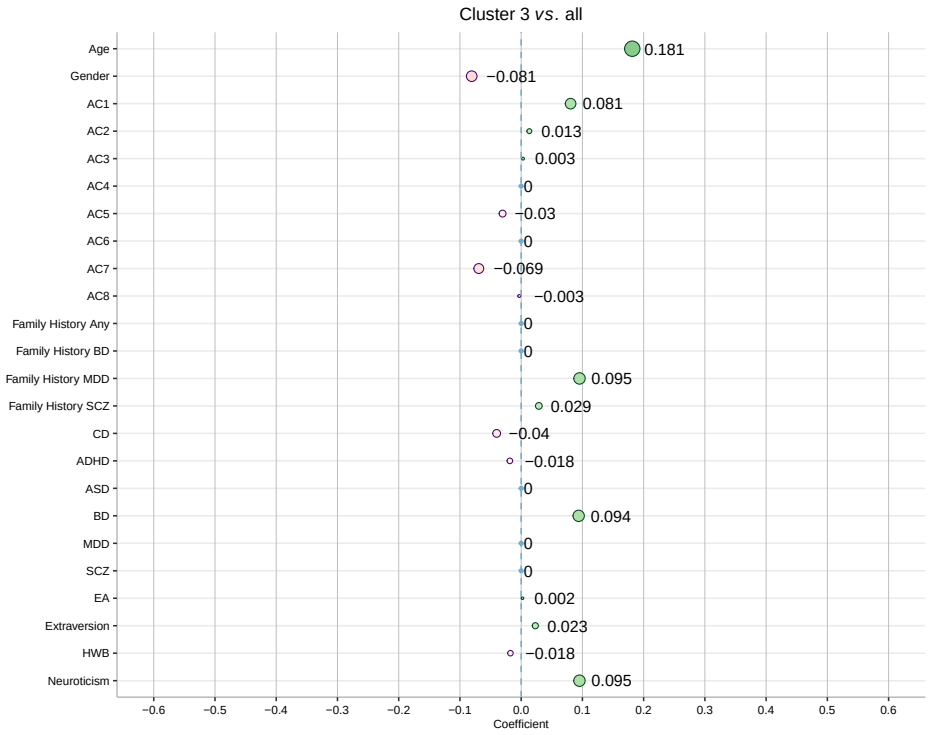
(a) Cluster 1 *vs.* all



(b) Cluster 2 *vs.* all



(c) Cluster 3 *vs.* all



Significance testing - PGS and Family History

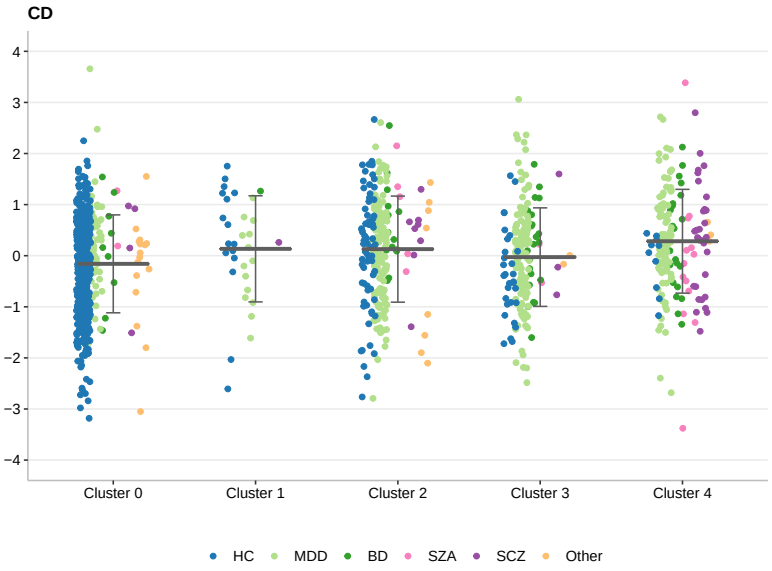
Univariate significance testing was done with the Westfall and Young (W-Y) procedure, as explained in the Section 2.2 of the Methods chapter. Three *one-vs.-all* and five *one-vs.-one* comparisons yielded significant variables (Tables 4.7 and 4.8, for *one-vs.-all* and *one-vs.-one*, respectively).

Figure 4.11 shows four polygenic risk scores that were significant after Bonferroni correction in at least one comparison. PGS for SCZ was significantly different in many comparisons - individuals in Cluster 4 had significantly higher PGS SCZ, and Cluster 0 significantly lower PGS SCZ compared to other clusters (Figure 4.11a). Two other PGS were significantly higher for the individuals in Cluster 4 - PGS for psychiatric cross-disorder and PGS for MDD (Figure 4.11b, 4.11c). PGS for Educational attainment was significantly higher in Cluster 0 and significantly lower in Cluster 4 compared to the other clusters (Figure 4.11d).

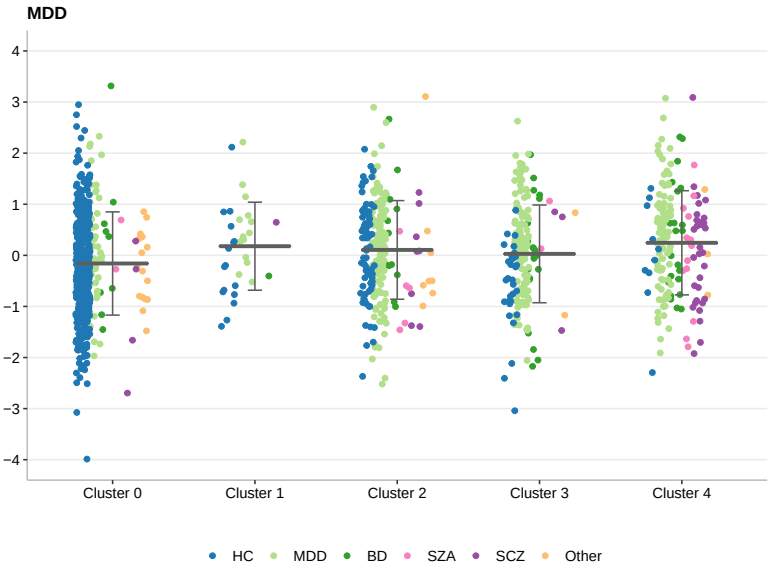
Figure 4.11: Significant Polygenic risk scores
PGS were standardized by Z score transformation, the y-axis unit are standard deviations. The mean is represented by a horizontal line and the standard deviation by error bars. The corresponding *p*-values are shown in Tables 4.7 and 4.8.

(a) PGS Schizophrenia

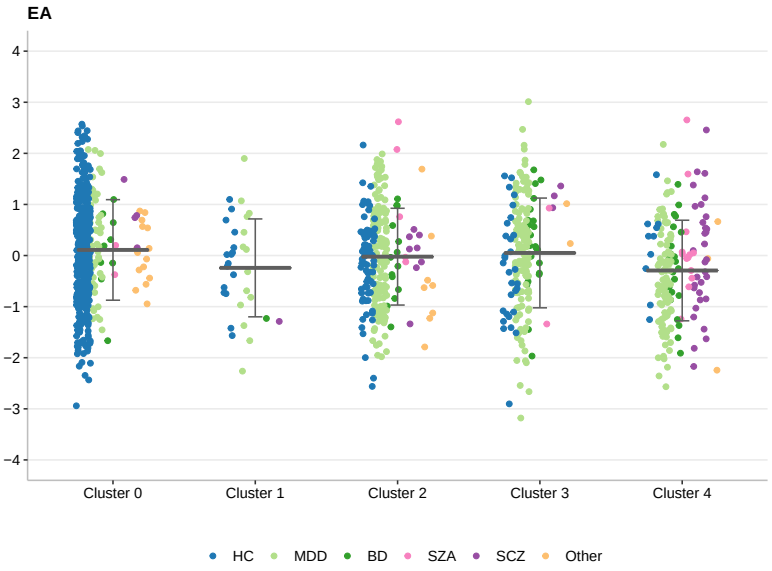
(b) PGS psychiatric cross-disorder



(c) PGS Major depressive disorder



(d) PGS Educational attainment



Source: Pelin et al., 2021

Table 4.7: Significance testing of genetic variables with the W-Y procedure – *one-vs.-all* comparisons
Only the comparisons with significant variables after the Westfall and Young adjustment are shown. *n.s.* stands for *not significant* (adjusted p-value>0.05).

***one-vs-all* Comparison**	**Variable**	**t-statistic**	***p*-value adjusted using W-Y**	***p*-value further adjusted for the number of comparisons** $(N = 5)$
Cluster 0 *vs.* all	Age	-7.9	8×10^{-4}	4×10^{-3}
	Family History Any	-10.2	8×10^{-4}	4×10^{-3}
	Family History BD	-4.2	8×10^{-4}	4×10^{-3}
	Family History MDD	-7.0	8×10^{-4}	4×10^{-3}
	PGS Cross psychiatric disorder	-4.8	8×10^{-4}	4×10^{-3}
	PGS MDD	-4.8	1×10^{-3}	8×10^{-3}
	PGS Schizophrenia	-3.6	9×10^{-3}	4×10^{-2}
	PGS Educational attainment	3.3	2×10^{-2}	$9 \times 10^{-2} (n.s.)$
	PGS Neuroticism	-3.1	3×10^{-2}	$1 \times 10^{-1} (n.s.)$
Cluster 2 *vs.* all	Family History Any	4.6	1×10^{-3}	5×10^{-3}
	Family History MDD	3.9	6×10^{-3}	3×10^{-2}
Cluster 4 *vs.* all	Age	6.1	9×10^{-4}	4×10^{-3}
	Family History Any	6.4	9×10^{-4}	4×10^{-3}
	PGS Cross psychiatric disorder	4.0	3×10^{-3}	1×10^{-2}
	PGS MDD	3.5	7×10^{-3}	4×10^{-2}
	PGS Schizophrenia	3.8	3×10^{-3}	1×10^{-2}
	PGS Educational attainment	-4.3	9×10^{-4}	4×10^{-3}

Source: Pelin et al., 2021

Table 4.8: Significance testing of genetic analyses with the W-Y procedure – *one-vs.-one* comparisons
Only the comparisons with significant variables after the Westfall and Young adjustment are shown. *n.s.* stands for *not significant* (adjusted p-value>0.05).

***One-vs-One* Comparison**	**Variable**	**t-statistic**	***p*-value adjusted using W-Y**	***p*-value further adjusted for the number of comparisons** $(N = 10)$
Cluster 0 *vs.* Cluster 2	Age	-3.9	1×10^{-3}	1×10^{-2}
	Family History Any	-7.7	1×10^{-3}	1×10^{-2}
	Family History MDD	-5.8	1×10^{-3}	1×10^{-2}
	PGS Cross psychiatric disorder	-3.6	8×10^{-3}	$8 \times 10^{-2} (n.s.)$
	PGS MDD	-3.5	1×10^{-2}	$1 \times 10^{-1} (n.s.)$
	PGS Neuroticism	-3.4	1×10^{-2}	$1 \times 10^{-1} (n.s.)$
Cluster 0 *vs.* Cluster 3	Age	-5.2	1×10^{-3}	1×10^{-2}
	Family History Any	-5.3	1×10^{-3}	1×10^{-2}
	Family History MDD	-4.3	1×10^{-3}	1×10^{-2}
Cluster 0 *vs.* Cluster 4	Age	-7.9	8×10^{-4}	8×10^{-3}
	AC4	3.1	3×10^{-2}	$3 \times 10^{-1} (n.s.)$
	Family History Any	-9.1	8×10^{-4}	8×10^{-3}
	Family History BD	-4.2	1×10^{-2}	$1 \times 10^{-1} (n.s.)$
	Family History MDD	-4.4	1×10^{-3}	1×10^{-2}
	PGS Cross psychiatric disorder	-5.0	8×10^{-4}	8×10^{-3}
	PGS Schizophrenia	-4.4	8×10^{-4}	8×10^{-3}
	PGS Educational attainment	4.7	8×10^{-4}	8×10^{-3}
Cluster 2 *vs.* Cluster 4	Age	-3.9	4×10^{-3}	4×10^{-2}
	Family History BD	-3.3	2×10^{-2}	$2 \times 10^{-1} (n.s.)$
Cluster 3 *vs.* Cluster 4	Family History BD	-3.1	4×10^{-2}	$4 \times 10^{-1} (n.s.)$
	PGS Educational attainment	3.2	3×10^{-2}	$3 \times 10^{-1} (n.s.)$

Source: Pelin et al., 2021

Assessing the PGS information gain

The addition of PGSs and ACs (model A) in the multinomial regression prediction model resulted in an R^2=11.7%. increase over a null model containing only age and gender (as explained in Methods Section 3.2.4.3). A model with only family history (model B) yielded an increase of R^2=10.8% over the null model, whereas a model with family history and ACs (model C) resulted in a gain of R^2=13.9%. The full model (model D) containing the PGSs, family history, and ACs increased R^2 by 20.3%. The inclusion of PGSs significantly improved the model with family history and ACs (model

A *vs.* model B, likelihood ratio test p=5 $\times$ 10^{-5}). The table below summarizes the results:

Table 4.9: Assessment of the PGS information gain

	Model A	Model B	Model C	Model D
Metric	**Y** = Age + Gender + PGSs + ACs	**Y** = Age + Gender + Family History	**Y** = Age + Gender + Family History + ACs	**Y** = Age + Gender + PGSs + Family History + ACs
AIC	3076.2	2975.5	2999.1	2991.4
Nagelkerke R^2	11.7%	10.8%	13.9%	20.3%
Likelihood ratio tests				
Model D *vs.* Model A	2×10^{17}			
Model D *vs.* Model B	5×10^{-5}			
Model D *vs.* Model C	2×10^{-5}			
Model C *vs.* Model B	1×10^{-1} ($n.s.$)			

Source: Pelin et al., 2021

4.1.4 Potential disorder subtypes analysis

Patients diagnosed with MDD and BD distributed across all identified clusters, suggesting that different stages or subtypes of the disorders were identified. Since MDD was the largest diagnostic group in the sample, the secondary analysis was conducted to assess the heterogeneity of MDD patients across clusters (Table 4.10). The HDDA classification analysis was done in a *one-vs.-all* fashion to assess the most important features for MDD patients only (Table 4.11).

MDD patients in Cluster 0 had similar clinical presentations as healthy controls. They showed the lowest severity in many areas and 80% of them were classified as in remission of either a single episode or recurrent depression at the moment of assessment. MDD patients in Cluster 1 had high levels of somatization, lower energy, and higher life stress scores, as well as higher age of MDD onset. MDD patients in Cluster 2 showed the lowest average age of onset. They also showed higher levels of maltreatment and emotional neglect in childhood which might be anticipated as the external stressors and the triggers for the disease. However, these individuals also showed a high predisposition for MDD, with 48% having a case of MDD in their families. Similar to Cluster 0, MDD patients in Cluster 3 had low negative environmental factors scores and higher parental bonding. Nonetheless, their disease seems to have had a negative effect on their quality of life - they showed low levels of energy and reported having difficulties in their daily activities due to emotional and physical health issues. In accordance with the high proportion of SCZ patients in Cluster 4, MDD (and BD) patients in this cluster experienced depression with psychotic characteristics, with higher antipsychotic intake and higher positive symptoms (Table 4.12). When looking at the genetics, only MDD patients in this cluster showed a significant difference - they had significantly higher PGS for ADHD and lower PGS for Educational attainment than MDD patients in other clusters (Figure 4.12). Interestingly, ADHD PGS was not significant in the statistical testing analysis including all individuals from the sample (Table 4.7).

Table 4.10: MDD subtypes analysis

This table shows phenotypic characteristics of MDD patients, constituting a secondary, descriptive analysis to assess the heterogeneity within MDD patients. The significance of genetic variables was tested using the Westfall and Young procedure. Here, only the significant variables are shown. The distributions of the two significant PGS per cluster are shown in Figure 4.12.

Variable	Discovery MDD	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
N (MDD diagnosis)	477	56	17	152	147	105
Age at onset, mean (SD)	25.8 (12.5)	25.3 (11.1)	29.1 (13.2)	23.7 (12.0)	28.1 (13)	25.3 (12.7)
SCL - Global Severity Index, mean (SD)	0.98 (0.6)	0.21 (0.12)	1.22 (0.6)	0.79 (0.42)	1.16 (0.53)	1.38 (0.58)
Family history any psychiatric disorder, N (%)	284 (60%)	26 (46%)	7 (44%)	96 (64%)	82 (57%)	73 (71%)
Family History of MDD, N (%)	207 (43%)	18 (32%)	4 (24%)	73 (48%)	64 (44%)	48 (46%)
CTQ sum score, mean (SD)	46.11 (15.7)	32.4 (5.7)	50.9 (8.3)	47.8 (9.8)	35.8 (7.2)	64.6 (16.9)
Quality of life - mental health, mean (SD)	48.1 (21.2)	78.0 (9.1)	37.4 (19.6)	53.3 (18.5)	40.4 (16.8)	36.9 (17.3)
Quality of life - general health, mean (SD)	52.9 (22.5)	75.8 (15.3)	42.9 (26.4)	60.5 (19.9)	45.2 (19.5)	41.9 (19.7)
Quality of life – Energy, mean (SD)	34.4 (20.8)	61.1 (15.6)	20.3 (16.4)	40.1 (17.3)	26.2 (16.8)	25.9 (18.5)
SCL – Somatization, mean (SD)	10.8 (8.2)	3.1 (2.5)	15.0 (7.6)	7.0 (4.8)	14.1 (7.98)	15.1 (9.1)
Life stress, mean (SD)	29.9 (10.3)	16.1 (5.9)	34 (8.7)	28 (8.8)	34 (8.7)	33.8 (9.2)
NEO-FFI Neuroticism, mean (SD)	29.3 (9.0)	18.1 (6.5)	33.4 (7.7)	29.5 (8.1)	30.8 (8.1)	32.4 (8.1)
Maternal bonding, mean (SD)	20.8 (8.6)	27.4 (5.5)	13.9 (5.1)	18.7 (7.3)	26.3 (6.2)	13.7 (7.3)
Paternal bonding, mean (SD)	18.7 (8.3)	24.2 (5.9)	16.9 (8.8)	16.6 (7.2)	21.9 (7.5)	14.6 (8.6)
Positive Symptoms, mean (SD)	0.79 (2.4)	0.11 (0.41)	0.18 (0.52)	0.41 (1.19)	0.63 (1.57)	2.0 (4.4)
Antidepressants, yes, N (%)	292 (61%)	18 (32%)	12 (71%)	84 (55%)	106 (72%)	72 (69%)
Antipsychotic, yes, N (%)	84 (18%)	4 (7%)	2 (12%)	26 (17%)	28 (19%)	24 (23%)
Mood stabilizer, yes, N (%)	18 (4%)	2 (4%)	1 (6%)	1 (0.7%)	5 (4%)	9 (9%)

Variable	Discovery MDD	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Significantly different PGSs (*one-vs-all*)						
EA	*p*-value W-Y adjusted					1×10^{-3}
	Bonferroni-corrected (N=5)					5×10^{-3}
ADHD	*p*-value W-Y adjusted					1.9×10^{-3}
	Bonferroni-corrected (N=5)					9.5×10^{-3}

Source: Pelin et al., 2021

Table 4.11: *One-vs.-all* HDDA classification analysis using only MDD-diagnosed patients

This table shows the most important clinical variables for each cluster when analyzing only MDD patients. The variables were identified by the *one-vs.-all* HDDA classification analysis using n=477 MDD patients and their respective cluster labels. The importance was calculated based on the average AUC drop, as explained in the Methods section 3.2.4.1.

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
CTQ Sexual abuse	Maternal bonding	SF36 Physical functioning	Maternal bonding	CTQ Physical abuse
SAPS (Positive symptoms)	SF36 Bodily pain	SF36 Role physical	CTQ Emotional neglect	CTQ Sexual abuse
CTQ Physical neglect	SF36 Role physical	CTQ Physical abuse	CTQ Emotional abuse	SAPS (Positive Symptoms)
LEQ Negative event score	YMRS	CTQ Emotional neglect	SF36 Role Physical	CTQ Emotional abuse
CTQ Physical abuse	SF36 Energy	RSQ Avoidance of closeness	CTQ Physical abuse	ACE Sum
ACE Sum	SHAPS	CTQ Emotional abuse	CTQ Physical neglect	LEQ Positive event score
SCL Positive symptoms	Social support	ACE Sum	ACE Sum	SANS (Negative symptoms)
SF36 Social functioning	Verbal IQ	Maternal bonding	SF36 Bodily pain	SCL90R Phobic fear
SANS (Negative symptoms)	SAPS (Positive symptoms)	CTQ Sexual abuse	SF36 General health	VLMT Sum
SF36 Mental health	CTQ Emotional neglect	SCL90R Somatization	SF36 Role emotional	CTQ Physical neglect

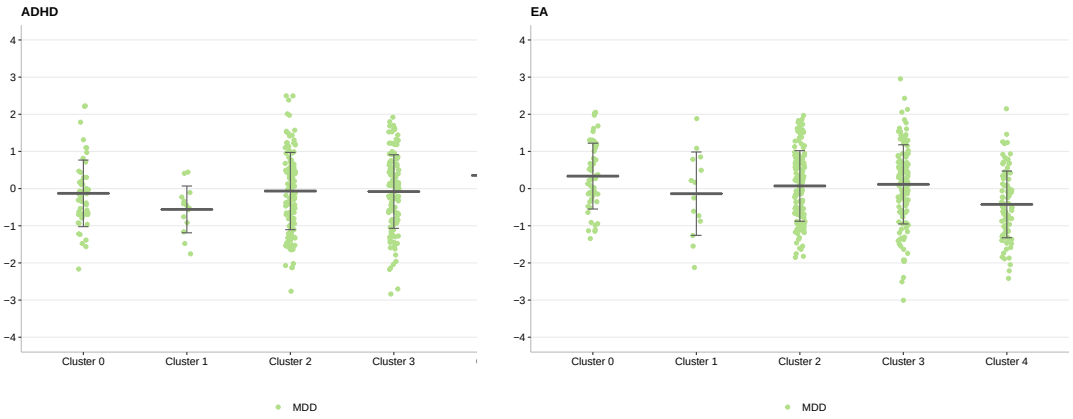
Source: Pelin et al., 2021

Table 4.12: Assessing psychotic symptoms of MDD and BD patients in Cluster 4
Comparison of positive symptoms scores of MDD and BD patients in Cluster 4 to MDD/BD patients in the other clusters. The table shows mean (SD) values for the respective variables. Statistical significance was assessed using one-sided t-tests with the hypothesis that Cluster 4 patients show more positive symptoms. *n.s.* stands for *not significant* ($p < 0.05$).

	MDD			BD		
Variable	**Cluster 4**	**Clusters 0-3**	**t-test (one-sided)**	**Cluster 4**	**Clusters 0-3**	**t-test (one-sided)**
N	105	372		24	51	
Positive Symptoms	2.0 (4.4)	0.4 (1.3)	t = 3.6, p=0.0003	5.4 (7.2)	1.4 (2.4)	t = 2.6, p=0.007
Delusions	0.5 (1.9)	0.09 (0.5)	t = 2.4, p = 0.01	0.7 (1.6)	0.1 (0.5)	t = 1.7, p=0.05
Hallucinations	0.3 (1.4)	0.03 (0.2)	t = 1.8, p = 0.04	0.1 (0.3)	0.02 (0.1)	t = 1.0, p=0.1 (*n.s.*)
Bizarre Behavior	0.2 (0.9)	0.06 (0.4)	t = 1.5, p = 0.06 (*n.s.*)	0.5 (1.2)	0.04 (0.3)	t = 1.9, p=0.03
Positive Formal Thought Disorder	1.0 (2.7)	0.3 (1.0)	t = 2.7, p=0.004	4.1 (5.6)	1.3 (2.2)	t = 2.4, p=0.01

Source: Pelin et al., 2021

Figure 4.12: Significant Polygenic risk scores for MDD patients only
PGS were standardized by Z score transformation, the y-axis unit are standard deviations. The mean is represented by a horizontal line and the standard deviation by error bars.



(a) PGS ADHD for MDD patients

(b) PGS EA for MDD patients

Source: Pelin et al., 2021

4.1.5 Characterization of healthy controls

The percentage of healthy controls in the clusters decreased with cluster severity - they made 84% of the Cluster 0, 50% of Cluster 1, 29% of Cluster 2, 16% of Cluster 3, and 6% of Cluster 4 (Figure 4.5). Healthy controls in Clusters 1-4 showed some symptoms resembling the ones of the psychiatric patients in these clusters, which can be seen in Table 4.13. Genetic analyses were performed by observing only healthy controls across clusters and revealed a nominal significance of ADHD PGS, similar to the analysis of MDD subtypes. These healthy individuals could be either at risk for developing a disorder or their symptoms were not sufficiently severe to satisfy the criteria for the diagnosis. These aspects should be further inspected in the follow-up assessments.

Table 4.13: Clinical assessment of healthy controls

The table shows the phenotypic characteristics of healthy controls, constituting a secondary, descriptive analysis to assess the heterogeneity within healthy controls. The significance of genetic variables was tested using the Westfall and Young procedure. Here, only the significant variables are shown. Importantly, the number of healthy controls differed strongly between clusters, leading to a large class imbalance and resulting in a lack of power for the analysis of healthy controls on their own.

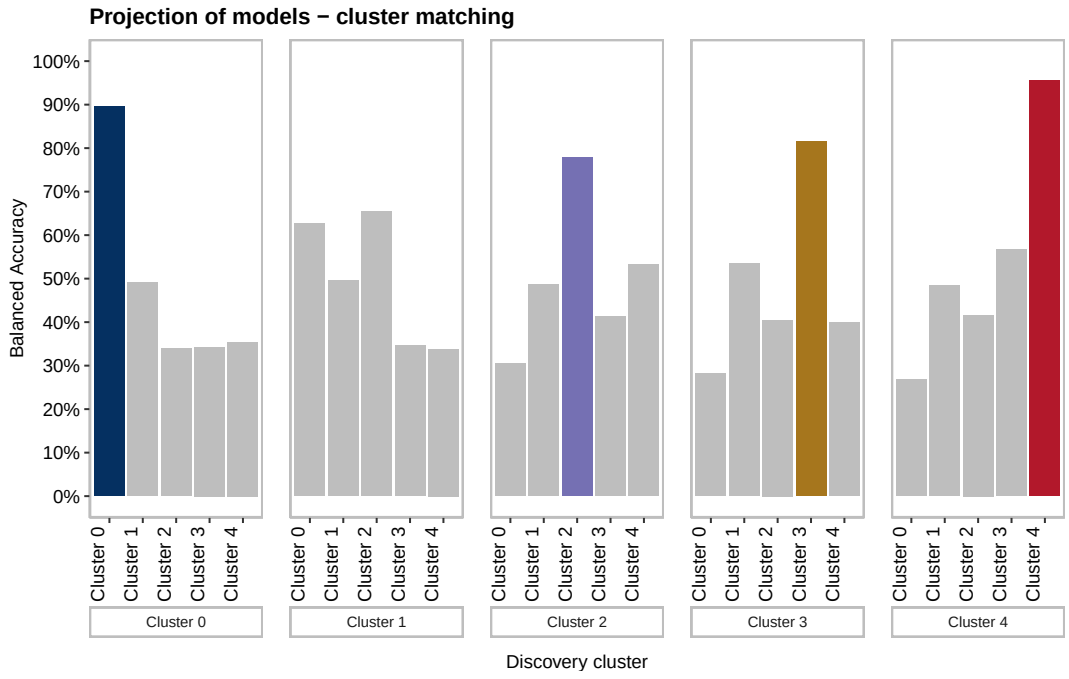
Variable	**Discovery HC**	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
N (Healthy controls)	590	448	19	78	34	11
STAI-S	34.4 (7.9)	32.8 (6.2)	42.9 (8.6)	37.7 (9.5)	42.5 (11.4)	35.4 (9.8)
STAI-T	33.6 (8.0)	31.9 (6.7)	40.7 (8.7)	37.1 (8.9)	41.6 (9.5)	36.3 (11.4)
SCL - Global Severity Index, mean (SD)	0.2 (0.2)	0.2 (0.1)	0.4 (0.3)	0.3 (0.2)	0.5 (0.3)	0.3 (0.2)
Family history any psychiatric disorder, N (CTQ sum score, mean (SD)	32.1 (8.2)	29.5 (4.2)	34.0 (9.3)	43.2 (9.3)	31.2 (5.9)	56.6 (17.9)
Quality of life - mental health, mean (SD)	79.3 (12.0)	81.9 (9.8)	70.1 (13.5)	72.8 (13.4)	65.4 (16.1)	78.9 (11.0)
Quality of life - general health, mean (SD)	79.5 (15.6)	82.1 (13.8)	70.1 (20.3)	74.0 (17.8)	67.1 (17.5)	67.3 (15.0)
Quality of life – Energy, mean (SD)	64.5 (15.1)	67.3 (12.6)	53.9 (17.5)	58.8 (18.5)	46.0 (16.0)	65.9 (18.0)
SCL – Somatization, mean (SD)	3.5 (3.4)	2.8 (2.3)	6.6 (4.9)	4.9 (4.1)	7.8 (5.9)	5.9 (6.5)
Life stress, mean (SD)	16.3 (6.9)	14.8 (6.1)	21.3 (8.7)	20.3 (7.1)	22.5 (8.1)	20.3 (6.3)
NEO-FFI Neuroticism, mean (SD)	15.4 (7.2)	14.2 (6.6)	21.5 (7.7)	18.4 (7.6)	20.5 (8.0)	17.3 (5.4)
Maternal bonding, mean (SD)	28.4 (6.5)	30.2 (4.6)	27.3 (8.4)	20.2 (7.2)	29.1 (6.2)	14.5 (6.4)
Paternal bonding, mean (SD)	25.6 (7.7)	27.4 (6.3)	26.5 (6.8)	16.7 (7.4)	25.7 (7.4)	11.1 (8.1)
Positive Symptoms, mean (SD)	0.2 (0.8)	0.1 (0.5)	0.1 (0.2)	0.2 (0.7)	0.2 (0.8)	1.7 (4.0)
Negative Symptoms, mean (SD)	0.6 (1.7)	0.5 (1.4)	0.6 (1.1)	1.1 (3.0)	0.4 (1.0)	2.3 (3.4)
Significantly different PGSs (*one-vs.-all*)						
ADHD	*p*-value W-Y adjusted	2×10^{-2}		5×10^{-2}		
	Bonferroni-corrected (N=5)	1×10^{-1} $(n.s)$		5×10^{-1} $(n.s)$		

Source: Pelin et al., 2021

4.1.6 Replication analysis

Replication analysis was performed according to Section 3.2.5 in Methods chapter. HDDA models trained on the discovery and fitted on the replication sample matched all but the smallest Cluster 1 (Figure 4.13).

Figure 4.13: Discovery-stage HDDA models projected to the replication sample
Matched discovery-stage and replication clusters achieving the prediction>70% are represented with colored bars.
The balanced accuracy was used as the evaluation metric for the prediction. This metric evaluates a binary classifier by accounting for the imbalance in classes. It is calculated as the average of the proportion of correctly classified individuals from each class: (true positive rate + true negative rate)/2.

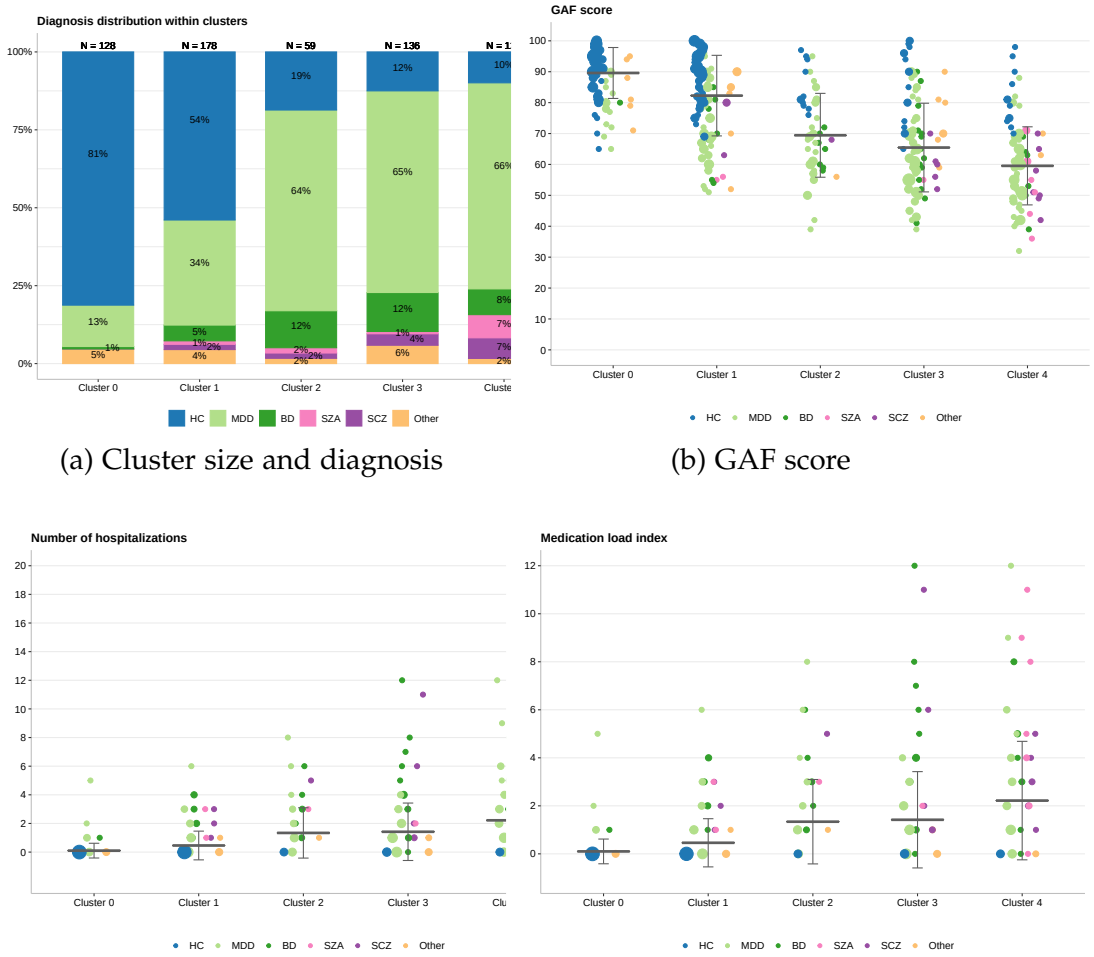


Source: Pelin et al., 2021

The matched replication clusters had the same severity ranking based on GAF score (4.14b) as their paired discovery clusters, and a lot of other variables had comparable severity patterns (Figures 4.14c-d, Table A5.5 and A5.6 in the Appendix). Moreover, the percentage of healthy controls was decreasing with the severity, and most of the SCZ

and SZA patients clustered in the highest severity Cluster 4 (Figure 4.14a), a pattern observed in the discovery-stage analysis.

Figure 4.14: Replication clusters characterization
In the figures b) - d) The mean is represented by a horizontal line and the standard deviation by error bars. The dot size is relative to the number of individuals having the given value.



(a) Cluster size and diagnosis

(b) GAF score

(c) Hospitalization

(d) Medication index load

Source: Pelin et al., 2021

The genetic Lasso models trained on the discovery sample and applied to the replication sample as explained in Methods Section 3.2.5, resulted in AUC of 68% for *Cluster 4 vs. all*, and AUC of 63% for *Cluster 0 vs. all*, similar to the metrics in the discovery sample. All projection models that resulted in AUCs >60% and are shown in Table 4.14.

Table 4.14: Metrics of genetic Lasso regularized regression prediction models in the replication clusters
After the discovery and replication clusters had been matched, the Lasso model parameters trained on the discovery sample were used for Lasso regularized regression analyses of the genetic variables in the replication sample. An AUC between 50% and 60% is considered random model performance. The table lists models with an AUC >60%, i.e., showing above-random performance.

Model	AUC	Sensitivity	Specificity
Cluster 0 vs. all	63%	60%	60%
Cluster 4 vs. all	68%	67%	66%
Cluster 0 vs. Cluster 4	75%	72%	72%
Cluster 2 vs. Cluster 4	69%	72%	67%
Cluster 3 vs. Cluster 4	61%	60%	60%
Cluster 0 vs. Cluster 2	60%	59%	59%
Cluster 0 vs. Cluster 3	60%	70%	53%

Source: Pelin et al., 2021

Significance testing with genetic data on the replication sample was performed with W-Y process, just as in the discovery analysis.

The replication-stage Cluster 0 showed the significantly lower PGS for SCZ and psychiatric cross-disorder (adjusted p = 0.005 and p = 0.03, respectively), as observed in the discovery sample (Figures 4.15a,b, Table 4.15). The findings for the discovery-stage Cluster 4 were confirmed in the replication Cluster 4 for the significantly higher MDD PGS and significantly lower EA PGS (p=0.01, p=0.005, respectively). Also, PGS for psychiatric cross-disorder and SCZ replicated as higher in Cluster 4, but these scores did not pass the final Bonferroni correction (Table 4.15). In the pairwise analyses, the replicated associations included the PGS for psychiatric cross-disorder, SCZ, and Educational attainment for the Cluster 0 *vs.* 4 comparison (Table 4.16). Replication analysis for MDD individuals resulted in a significantly lower Educational attainment PGS for the MDD patients in Cluster 4, confirming the finding from the discovery

analysis (Figure 4.16b). The association for ADHD PGS in MDD patients in Cluster 4 was not replicated (Figure 4.16a).

Table 4.15: Replication - Significance testing of genetic variables with the W-Y procedure – *one-vs.-all* comparisons
Only the comparisons with significant variables after the Westfall and Young adjustment are shown. *n.s.* stands for *not significant* (adjusted p-value>0.05).

***one-vs-all* Comparison**	**Variable**	**t-statistic**	***p*-value adjusted using W-Y**	***p*-value further adjusted for the number of comparisons** $(N = 5)$
Cluster 0 *vs.* all	Age	-3.9	3×10^{-2}	$1 \times 10^{-1} (n.s.)$
	PGS Cross psychiatric disorder	-3.7	7×10^{-3}	3×10^{-2}
	PGS Schizophrenia	-3.2	1×10^{-3}	5×10^{-3}
Cluster 2 *vs.* all	Age	-3.8	1×10^{-2}	5×10^{-2}
Cluster 4 *vs.* all	PGS Cross psychiatric disorder	3.4	1.6×10^{-2}	$8 \times 10^{-2} (n.s.)$
	PGS MDD	3.7	3×10^{-3}	1×10^{-2}
	PGS Schizophrenia	3.1	2×10^{-2}	$1 \times 10^{-1} (n.s.)$
	PGS Educational attainment	-4.8	1×10^{-3}	5×10^{-3}
	PGS Neuroticism	3.5	6×10^{-3}	3×10^{-2}

Source: Pelin et al., 2021

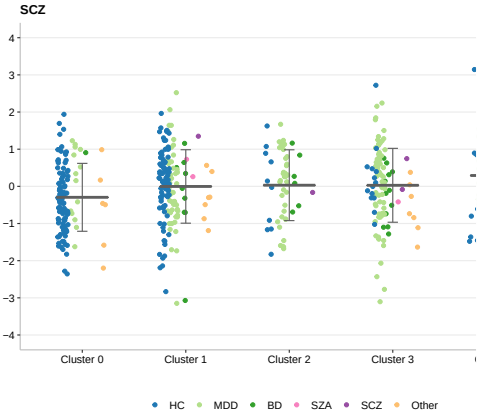
Table 4.16: Replication - Significance testing of genetic variables with the W-Y procedure – *one-vs.-one* comparisons
Only the comparisons with significant variables after the Westfall and Young adjustment are shown. *n.s.* stands for *not significant* (adjusted p-value>0.05).

***one-vs-one* Comparison**	**Variable**	**t-statistic**	***p*-value adjusted using W-Y**	***p*-value further adjusted for the number of comparisons** $(N = 10)$
Cluster 0 *vs.* Cluster 1	Age	-3.7	5×10^{-3}	5×10^{-2}
Cluster 0 *vs.* Cluster 4	PGS Cross psychiatric disorder	-4.5	1×10^{-3}	1×10^{-2}
	PGS MDD	-3.8	9×10^{-3}	$9 \times 10^{-2} (n.s.)$
	PGS Schizophrenia	-4.4	1×10^{-3}	1×10^{-2}
	PGS Educational attainment	4.2	2×10^{-3}	2×10^{-2}
Cluster 1 *vs.* Cluster 2	Age	4.3	2×10^{-3}	2×10^{-2}
Cluster 1 *vs.* Cluster 4	PGS MDD	-3.4	1×10^{-2}	$1 \times 10^{-1} (n.s.)$
	PGS Educational attainment	4.3	2×10^{-3}	2×10^{-2}
	PGS Neuroticism	-3.9	4×10^{-3}	4×10^{-2}
Cluster 2 *vs.* Cluster 3	Age	-3.3	2×10^{-2}	$2 \times 10^{-1} (n.s.)$
Cluster 2 *vs.* Cluster 4	Age	-3.6	1×10^{-2}	$1 \times 10^{-1} (n.s.)$
	PGS Educational attainment	3.6	1×10^{-2}	$1 \times 10^{-1} (n.s.)$
Cluster 3 *vs.* Cluster 4	PGS Educational attainment	3.2	4×10^{-2}	$4 \times 10^{-1} (n.s.)$

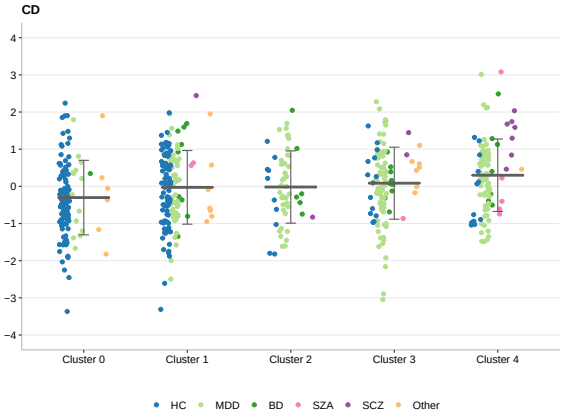
Source: Pelin et al., 2021

Figure 4.15: Replication - Significant Polygenic risk scores
PGS were standardized by Z score transformation, the y-axis unit are standard deviations. The mean is represented by a horizontal line and the standard deviation by error bars.

(a) PGS Schizophrenia

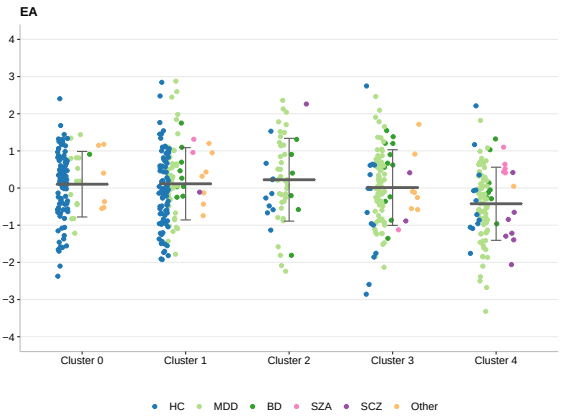


(b) PGS Psychiatric cross-disorder



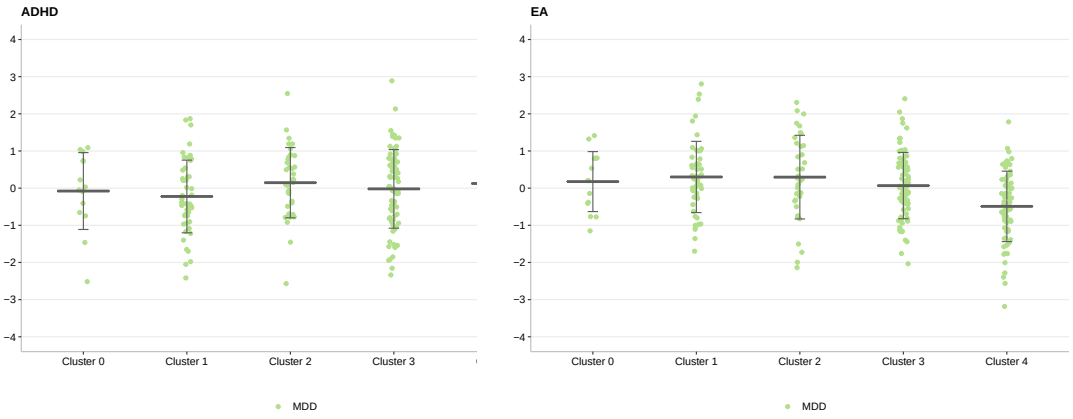
(c) PGS Major depressive disorder

(d) PGS Educational attainment



Source: Pelin et al., 2021

Figure 4.16: Replication - Polygenic risk scores for MDD patients only
PGS were standardized by Z score transformation, the y-axis unit are standard deviations. The mean is represented by a horizontal line and the standard deviation by error bars.
PGS EA was significantly lower for MDD in Cluster 4, as in the discovery-stage analysis (Figure 4.12). Association of PGS ADHD for MDD in the replication sample did not replicate.



(a) PGS ADHD for MDD patients

(b) PGS EA for MDD patients

Source: Pelin et al., 2021

4.1.7 Severity continuum and the Principal component analysis

The aim of this follow-up analysis was to assess whether a simple severity component can explain the identified clusters. To this end, we conducted a Principal component analysis (PCA) using the same 57 variables and individuals entering the clustering analysis. Please note, the PCA was not done prior to the clustering analysis, as the HDDC algorithm was developed to deal with high dimensions and with correlated features (Bouveyron et al., 2007).

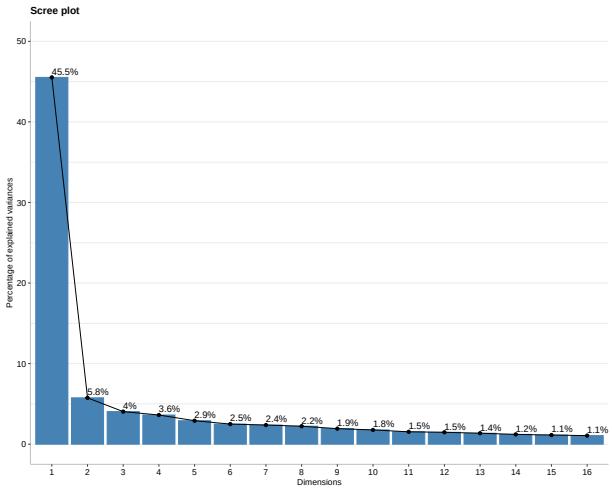


Figure 4.17: Variance explained by the first 16 PCs

The total of 80% of the variance was explained by the first 16 PCs (Figure 4.17), with the first component (PC1) explaining 45.5% of the variance. As shown in the Table 4.17, the SCL90R global severity index was the top-contributing variable for this component. Moreover, the PC1 correlated with the cluster labels with Spearman[1] $\rho = 0.75$. However, when removing the "healthy-like" Cluster 0, the correlation of the PC1 to Clusters 1-4 drops to $\rho = 0.38$. Hence, the PCA predominantly captured the distinction between health and disease, instead of the overall severity gradient.

[1]Spearman Rank Correlation Coefficient, "Spearman Rank Correlation Coefficient" 2008

Table 4.17: PCA with variables used for clustering
The table below shows the results of the Principal component analysis performed on variables used for clustering with the purpose of exploring to which degree severity was captured by PC1.

PC	Variance explained	Top five contributing variables				
1	45.5%	SCL90R global severity index	SCL90R depression	SCL90R Positive Symptom Total	STAIT	BDI Sum
2	5.8%	CTQ Emotional neglect	Maternal bonding	CTQ Physical neglect	CTQ Emotional abuse	CTQ Physical abuse
3	4.0%	VLMT Sum	Letter number span test	Corsi block-tapping test	Positive symptoms	NEOFFI Openness to experience
4	3.6%	NEOFFI Extraversion	Social support	NEOFFI Agreeableness	VLMT Sum	NEOFFI Openness to experience
5	2.9%	LEQ Positive Event Score	RSQ Fear of rejection or abandonment	SF36 Physical Functioning	IQ	YMRS

Source: Pelin et al., 2021

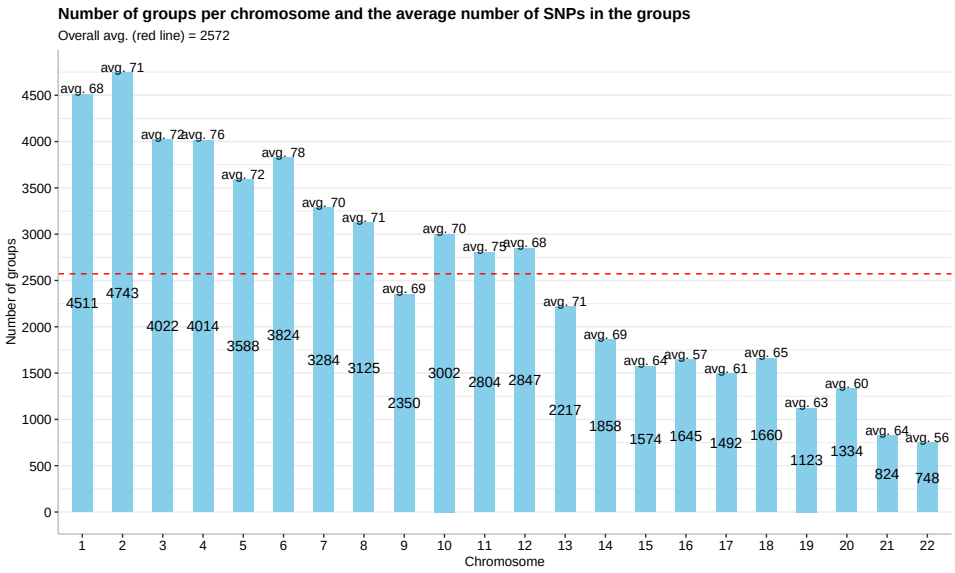
Additionally, *SigClust* (Huang et al., 2015) was used to further investigate the hypothesis that a severity continuum could best explain our results and data. However, this method did not find a statistical support for the hypothesis that the data is coming from a single continuous Gaussian distribution (p<0.05). Hence, the five categorical clusters did not entirely coincide with a severity continuum but ranked along with it.

4.2 Feature selection with genetic data

4.2.1 SNPs clustering

Ward's hierarchical clustering method with LD as a similarity metric was used to cluster SNPs on each chromosome, as explained in the Methods Section 3.3.2. The process resulted in a total of 56 589 clusters of SNPs. Figure 4.18 shows the number of groups found on each chromosome, and the average number of SNPs in each group. The obtained SNP cluster labels were used in the all subsequent analyses with the Sparse group Lasso algorithm.

Figure 4.18: Number of SNP clusters per Chromosome
The plot shows the number of different clusters of SNPs for each chromosome, detected by Ward's hierarchical clustering



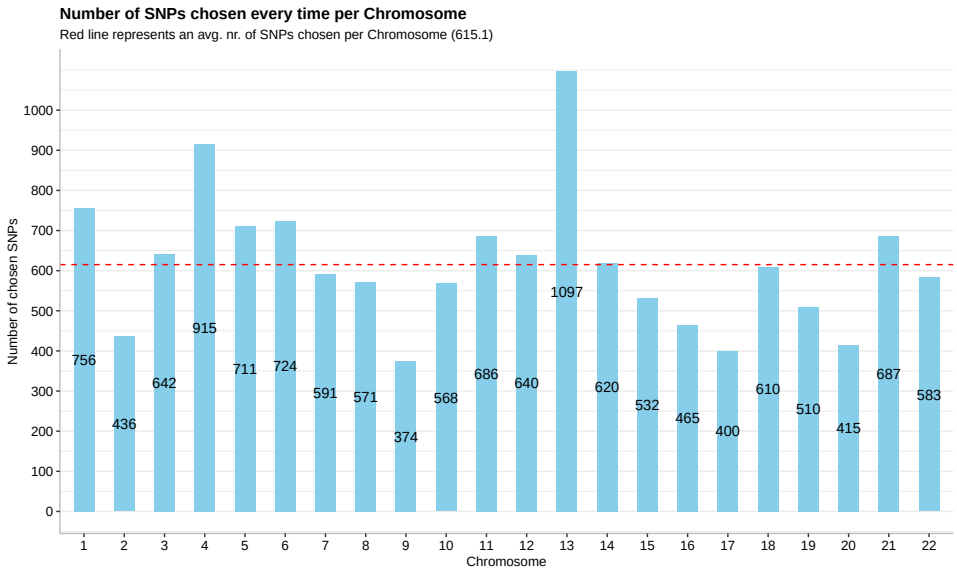
4.2.2 Phenotype prediction with Sparse group Lasso

4.2.2.1 Cluster 0 *vs.* all

Sparse group Lasso was run chromosome-wise, resulting in 13 533 SNPs that had the non-zero coefficient every time during the 10-fold cross-validation. Figure 4.19 shows

the number of SNPs chosen every time per chromosome, with an average of 615 SNPs chosen.

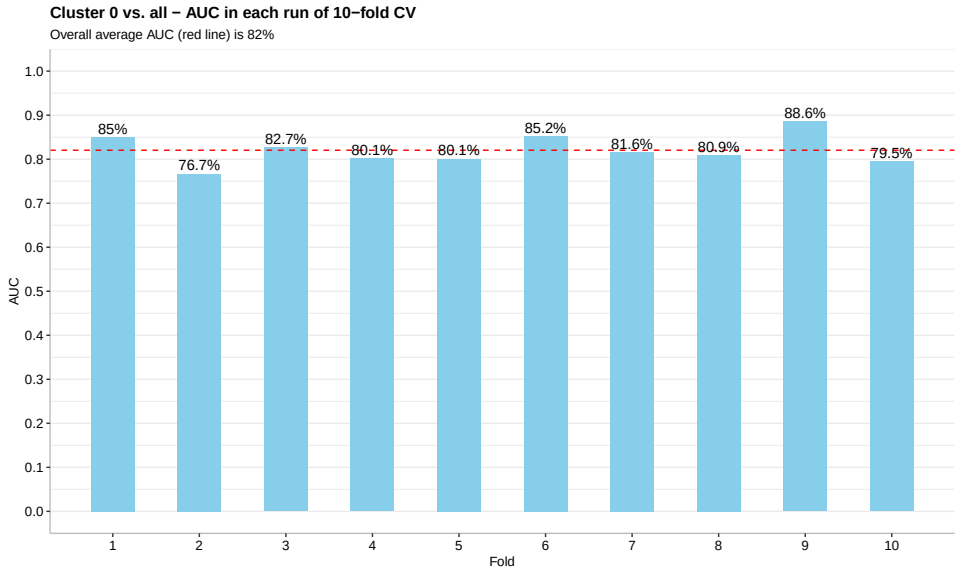
Figure 4.19: Number of SNPs chosen every time per Chromosome



Then, the Sparse group Lasso model was fit on the dataset with the chosen SNPs, i.e., on the dataset containing 1120 individuals and 13 533 SNPs. Figure 4.20 shows the AUC achieved on the dataset through 10 runs of cross-validation and the average AUC across folds, which is at a high 82%.

Due to the high AUC value and a pre-selection of SNPs on the same dataset, the chance of model overfitting occurring is very likely and has to be considered. Usually, to control for that, the prediction algorithm has to be tested on the validation dataset. Due to higher number of individuals in the full FOR2107 cohort, we decided to apply this strategy on the next task, where MDD patients and healthy controls were observed.

Figure 4.20: Cluster 0 *vs.* all - AUC across 10-fold cross validation
The total of 13 533 SNPs chosen in the chromosome wise Lasso runs were included in the final dataset. The model resulted with high AUC of 82% when fitted on the dataset. However, due to high AUC and a pre-selection of SNPs, we have to consider the model overfitting.



4.2.2.2 Healthy controls *vs.* MDD

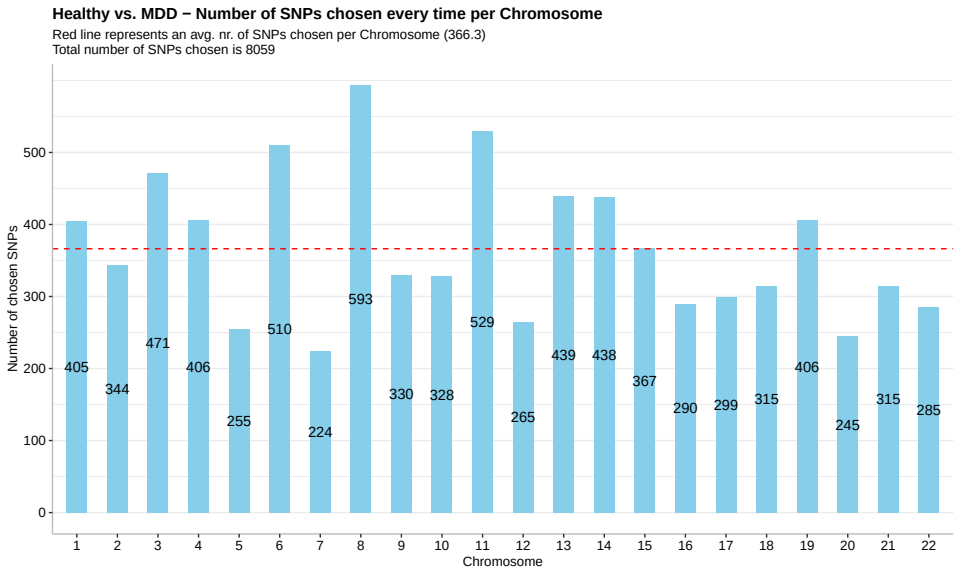
Since the phenotype variable using the subtype obtained from the subtype discovery analysis in Section 4.1 of this chapter possibly resulted in overfitting, we wanted to check if this would be the case with the formal labels. We contrasted here the healthy controls and MDD cases, as in the regular GWAS settings, to be able to compare the differences and check if the SNPs, that are already known to be associated with MDD from the GWA studies, will appear in our model.

In general, the unbiased measure of feature selection success and test for the overfitting is obtained when the model is fit on the previously unseen, validation dataset. Since the sample size in this analysis was higher (containing healthy controls and MDD patients from the full FOR2107 sample), we could apply a slightly different strategy, by splitting the dataset into the discovery and validation datasets. The split was stratified based on the binary phenotype variable with 1 representing a case (MDD), and 0

control (HC). Both datasets contained n=888 individuals. Apart from the discovery and validation split at the beginning, the workflow was the same as in the *Cluster 0 vs. all* analysis (shown in Figure 3.3). The difference to it was the additional model performance assessment on the validation sample in the last step (shown in Figure A1 in the Appendix).

First, Sparse group Lasso was run chromosome-wise. SNPs having the non-zero coefficient every time in the 10-fold cross-validation process were selected for the next step. The total number of SNPs selected for the next step was 8059. Figure 4.21 shows the number of SNPs chosen every time per chromosome.

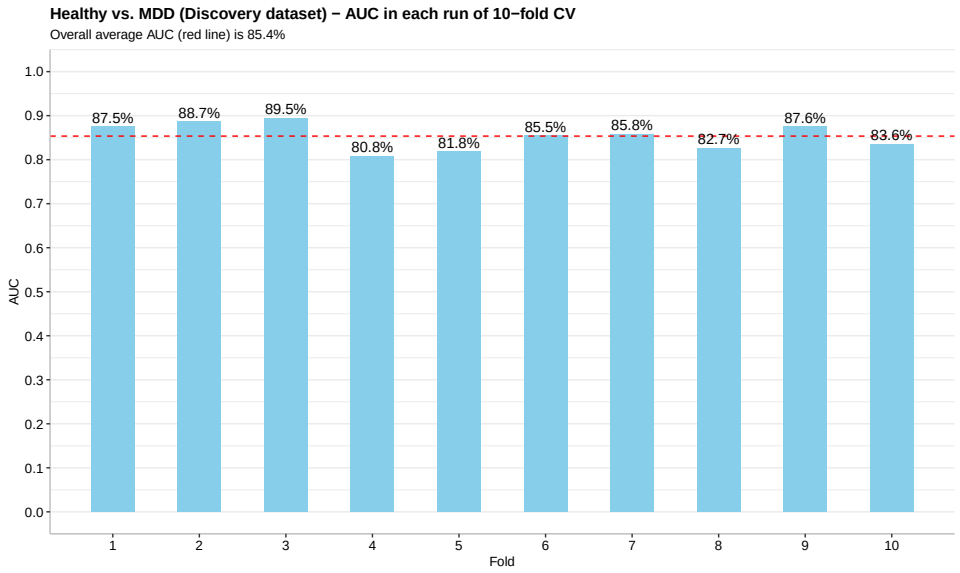
Figure 4.21: Number of SNPs chosen every time per Chromosome



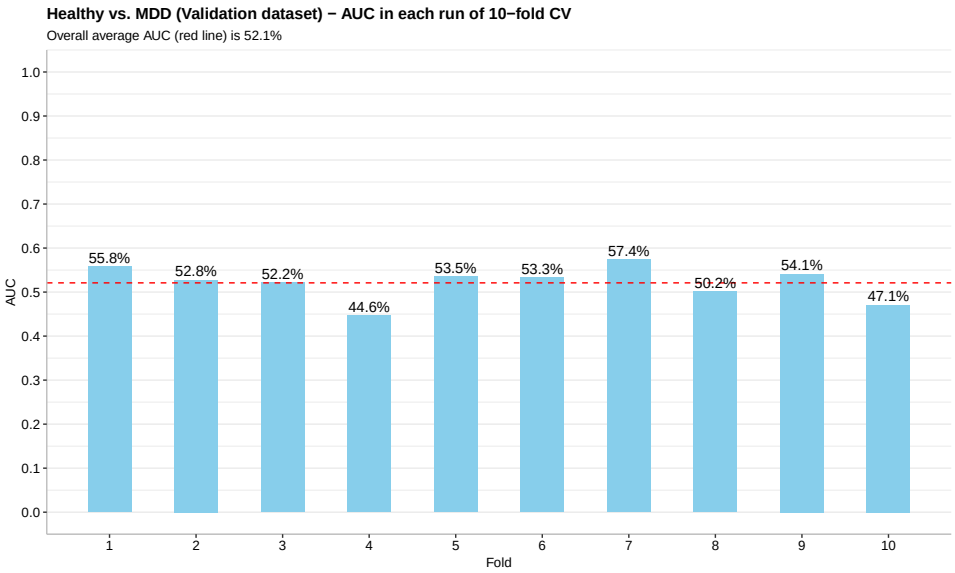
Then, the Sparse group Lasso model was fit on the discovery dataset with the chosen SNPs, i.e., on the dataset of 888 individuals and 8059 SNPs. Figure 4.22a shows the AUC achieved on the discovery dataset through 10 runs of cross-validation and the average AUC across folds, which is at high 85%. This, however, could be a sign of overfitting since the features were chosen on the same dataset, as we observed in the *Cluster 0 vs. all* settings. Finally, the Sparse group Lasso was fitted on the previously unseen, validation dataset, consisting of the same 8059 SNPs chosen in the discovery-stage analysis as predictors. As Figure 4.22b shows, the drop of AUC to the 50% is notable,

hence the model was indeed overfitting.

Figure 4.22: Healthy *vs.* MDD - AUC across 10-fold cross validation
The total of 8 059 SNPs chosen in the discovery-stage chromosome wise analysis were included in both final datasets. The model resulted with high AUC of 85% when fitted on the discovery dataset and low AUC of 52% when fitted on the validation dataset. Thus, model was overfitting and showed poor generalizability.



(a) Discovery dataset



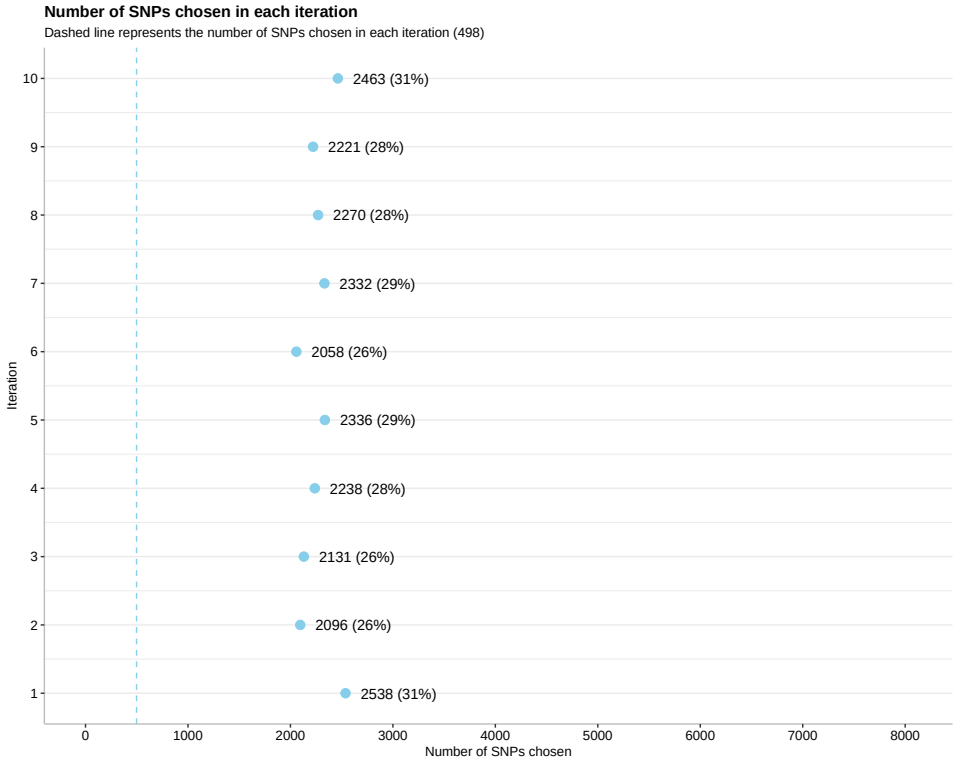
(b) Validation dataset

Stability of selected feature sets

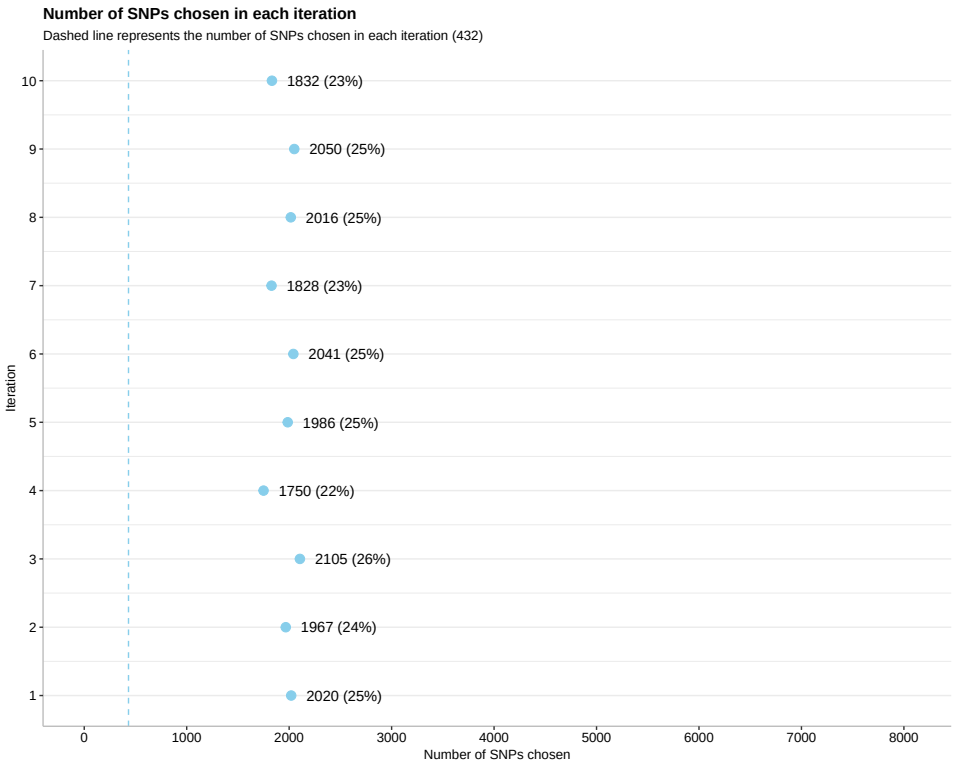
Even though the model was overfitting, we could still infer how stable it was in terms of SNPs chosen across the runs. We tested the stability of sets for Healthy *vs.* MDD analysis, as it enables the discovery/validation comparison. The stability was inferred by checking the SNP intersections between a) runs of 10-fold cross-validation, for discovery and validation fit separately, and b) final sets of SNPs chosen on both datasets.

Figure 4.23 shows the number of SNPs chosen in each run. We see that for both discovery and validation, the number of SNPs chosen is fairly uniformly distributed, with all iterations choosing between 25% and 31% of SNPs.

Figure 4.23: Healthy *vs.* MDD - number of SNPs chosen across iterations
Number of SNPs chosen, i.e. having the non-zero coefficient, across iterations of 10-fold cross validation. Percentage is calculated based on the total number of SNPs in the dataset (8059). The dashed line represents the number of SNPs chosen in each iteration (498 in the discovery and 432 in the validation).



(a) Discovery dataset

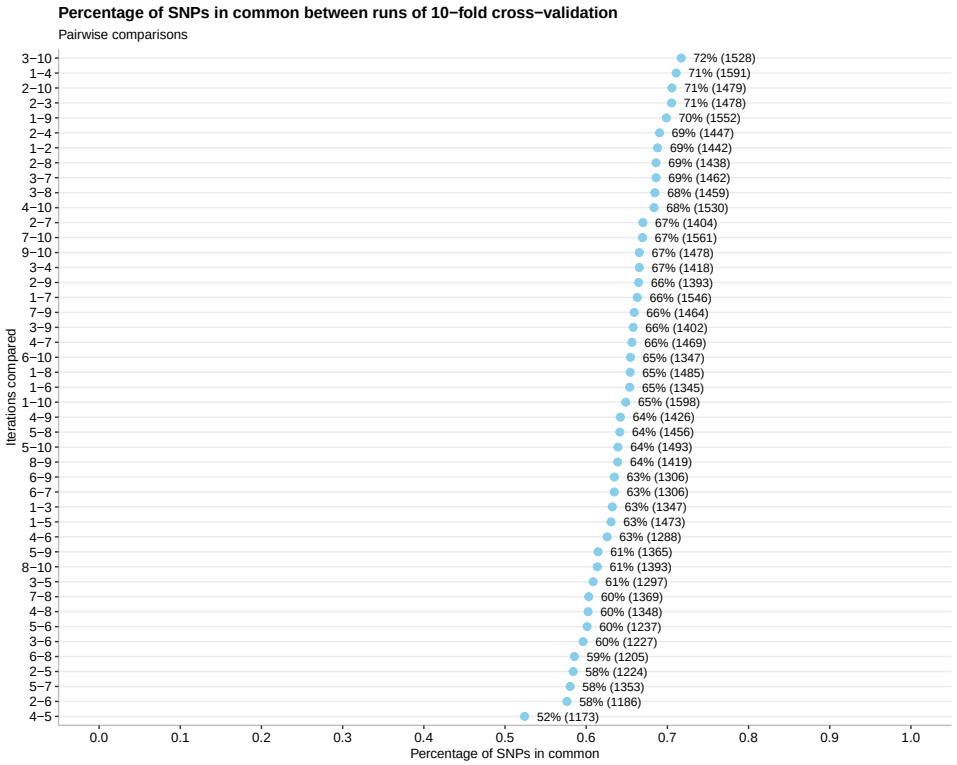


(b) Validation dataset

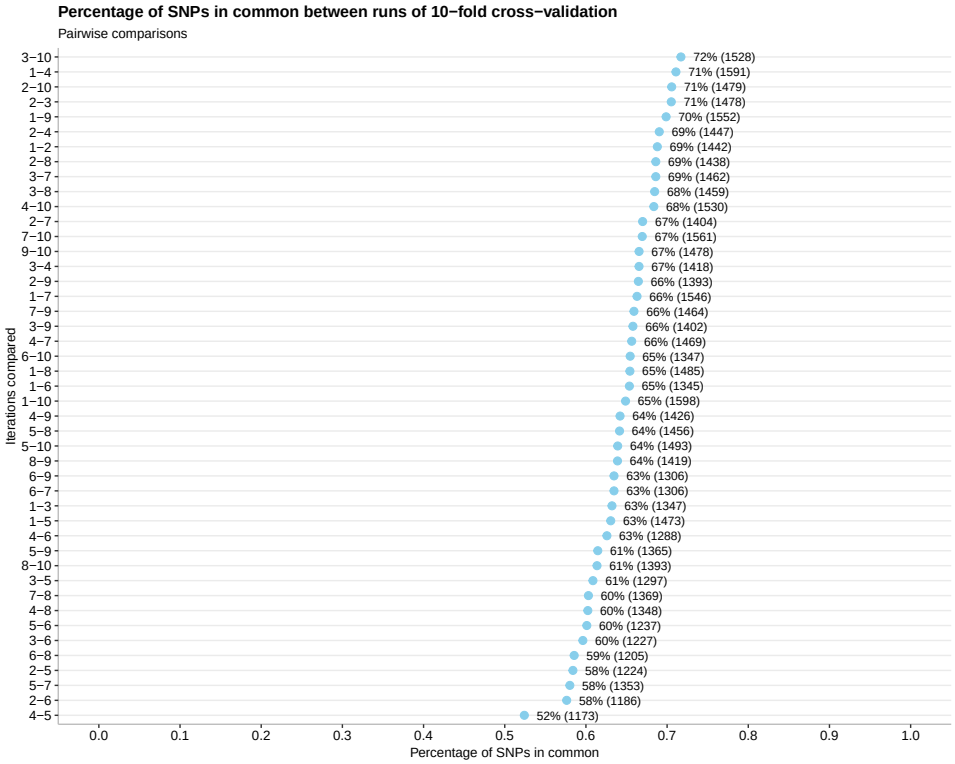
There were 498 SNPs in the intersection of 10 sets of SNPs chosen in each iteration in the discovery sample, while in the validation there were 432 SNPs.

In order to infer the stability across iterations, the percentage of SNPs in common between all pairs of iterations was determined, as explained in Section 3.3.2.2. In Figure 4.24, we see that the percentage of SNPs in common for all pairwise combinations was between 50% and 70%. When comparing the SNPs in common between the final set of SNPs chosen in the discovery and the replication, 3 SNPs in common were found. Therefore, we can not conclude that the algorithm was stable in terms of selection of features.

Figure 4.24: Healthy *vs.* MDD - stability of feature sets
The number in the parentheses represents the size of the intersection set between to iterations i and j ($|S_i \cap S_j|$)



(a) Discovery dataset



(b) Validation dataset

Cross-referencing with known associated SNPs from GWAS

We wanted to check if any of the 8059 SNPs chosen chromosome-wise was among SNPs that were previously reported to have the association to MDD. To do that, we cross-referenced our list of SNPs with the database in GWAS Catalog[2]. Six SNPs were found to have been previously reported in association with MDD or other psychiatric disorders:

[2]GWAS Catalog was founded by the National Human Genome Research Institute in 2008. It provides the database of SNP-trait associations from the eligible published GWA studies. Studies are identified by literature search, extracting the reported trait and SNP-trait association, which are mapped onto the human genome by chromosomal location and displayed on the human karyotype. Source: https://www.ebi.ac.uk/gwas/docs/about

SNP	Trait (GWAS catalog)
rs1002656	Depression,Neuroticism, BD, MDD, General factor of neuroticism,anorexia nervosa, ADHD, ASD, OCD, SZC, or Tourette syndrome (pleiotropy)
rs4491452	MDD
rs11070670	MDD
rs9297357	ASD, ADHD, BD, MDD, ADHD, BD, MDD, SCZ (combined)
rs1579282	Depressive symptoms x independent stressful life events interaction (2df test)
rs12457996	MDD

Apart from the SNPs associated with MDD, there were some reported to be associated with ADHD (rs7164335, rs438259, rs7448069), a disorder which is associated with the development of MDD or SCZ, according to several studies (Hamshere et al., 2013; Dalsgaard et al., 2014; Rubino et al., 2009). However, none of those SNPs appeared among the 498 ones that were chosen every time, i.e., that were in the final discovery set.

5 Discussion and Outlook

5.1 Discussion

Psychiatric disorders are highly heterogeneous with respect to their clinical representation and disease trajectory. The current psychiatric taxonomies, DSM and ICD, have emerged as the standard systems that describe how we diagnose psychiatric disorders worldwide. However, accumulating research evidence shows that the existing diagnostic groups do not suffice to capture the heterogeneity in etiology and symptomatology of mental health. As a result, an increasing focus has been set on a revision and reformulation of the current system.

In this thesis, we applied computational methods to generate new knowledge in the field of classification of psychiatric diseases and new subtypes discovery. A transdiagnostic sample was analyzed, going beyond the existing boundaries not only between different disorders but also between health and disease. The clustering results presented in this thesis were based on the results published in Pelin et al., 2021.
Additionally, we explored a multivariate way of selecting single SNPs in the high-dimensional genetic data to depict the smaller and sufficiently informative subset of SNPs relevant for the phenotype of interest.

The High-dimensional data clustering algorithm (HDDC) was used to identify the clusters of psychiatric disorders in the transdiagnostic sample. The sample consisted of healthy controls and patients diagnosed with major depressive disorder, bipolar disorder, schizoaffective disorder, schizophrenia, and other disorders such as social phobia. Significance testing and classification were used to characterize the clusters with genetic variables (PGS and family history). The combination of clustering and classification was used for the detection of important subsets of SNPs for the two outcome variables.

Transdiagnostic clusters identification

In this thesis, as well as in the publication Pelin et al., 2021, five diagnostically mixed clusters were identified. For the purpose of creating the notion of a severity continuum, we ordered the clusters based on the GAF measure, our severity proxy. The numeration 0-4 was used, with 0 representing the lowest disease severity level and 4 the highest severity level.
Cluster 0, mostly contained healthy controls and had in general good health and mental well-being. It showed the lowest depression levels, positive symptoms as well as maltreatment in childhood, whereas their quality of life was the highest. Cluster 4, on the other end of the scale, was characterized as the cluster that was highly impaired in many areas, with an emphasis on maltreatment in childhood and youth, and a prevalence of positive symptoms. This cluster had the highest proportion of SCZ and SZA diagnosed patients. Clusters 1-3 ranked between the two extremes and could be differentiated mostly by different depression scores, levels of childhood maltreatment, parental bonding, and daily functioning.

We tested if the five categorical clusters exactly followed the severity continuum, i.e., if a simple severity component could explain the clustering, but could not find statistical support for this hypothesis (PCA and SigClust analysis in Section 4.1.7). Hence, the five categorical clusters did not entirely coincide with a severity continuum but ranked along with it.

Replication analysis in a smaller independent sample resulted in cluster replications of four out of five clusters, where only the smallest Cluster 1 did not replicate. This, together with the characteristics of the replication-stage clusters in terms of the severity scale and replication associations in some genetic features, indicates that the cluster solution was stable and that the algorithm did not overfit in the discovery-stage analysis. However, even though the replication sample consisted of independent individuals whose data was acquired subsequently, these individuals were recruited as part of the same study as those from the discovery sample. Moreover, the proportions of MDD patients and healthy controls were different for the two samples, limiting their comparability. Furthermore, the replication sample consisted of fewer individuals than the discovery sample, which had a weakening effect on the statistical power. Hence, future studies with higher sample sizes and more independent discovery and replication samples could provide more information and explain the discrepancies.

To the best of our knowledge, this is the first study to cluster and examine the multidomain clinical profiles in a sample including both patients diagnosed with psychiatric disorders and healthy controls. Nonetheless, the cluster characteristics and the severity spectrum are partly consistent with the findings from some previous studies. A transdiagnostic study (Grisanzio et al., 2018) also identified a cluster with a high share of healthy individuals and with the lowest scores in the measured symptoms, similar to the Cluster 0 identified in this work. The severe psychosis subtype identified in another study (Dwyer et al., 2020) may correspond to our highest severity Cluster 4 when observing its high share of SCZ patients, significantly lower PGS for Educational attainment, and low general functioning. Furthermore, a single-disorder subtyping study (Maglanoc et al., 2018) identified five MDD clusters, with one subgroup exhibiting a lack of many symptoms, comparable to our Cluster 0. Moreover, our findings highlight the association of adverse experiences, childhood trauma, and lack of support with hospitalizations, positive symptoms, disease severity, and the need for a more intense treatment, all supported by several prior studies (Carbone et al., 2019; Varese et al., 2012; Misiak et al., 2017; Li et al., 2015; Janssen et al., 2004).

Potential subtypes discovery and identification of individuals at risk

Compared to the formal DSM diagnostic categories, the cluster solution resulting from this thesis (and Pelin et al., 2021) exceeded the diagnostic boundaries predominantly for BD and MDD diagnosis, while the SCZ and SZA diagnosed patients mostly grouped together, in the highest severity Cluster 4. These findings confirm the etiological similarities between the two affective disorders, BD and MDD, differentiating them from the generally psychotic disorders (Levey et al., 2020; Coleman et al., 2020). The higher number of SCZ patients might have resulted in better discrimination of psychotic subtypes, as the previous studies showed (Dwyer et al., 2020; Bansal et al., 2018).

All five clusters contained MDD patients, potentially representing different MDD subtypes or disease stages. Around 80% of MDD in Cluster 0 were coded as in remission of either single episode or recurrent depression at the moment of assessment. Therefore, their clinical picture resembles the characteristics of healthy individuals. MDD in Cluster 4, on the other hand, exhibited some psychotic features and had high negative symptoms. Interestingly, these MDD patients showed a different signal in genetic analysis, by having significantly higher PGS for ADHD compared to the MDD patients in Clusters 0-3. Previous studies have reported associations between childhood ADHD and the development of or susceptibility to other severe psychiatric

disorders later in life (Hamshere et al., 2013; Dalsgaard et al., 2014; Rubino et al., 2009). A retrospective evaluation of ADHD symptoms in childhood, which is currently unavailable, might provide further insight on this correlation. Moreover, MDD (and BD) patients in Cluster 4 had significantly more psychotic characteristics than MDD (BD) patients in Clusters 0-3. MDD patients in Cluster 1 had a high level of somatization, a neurotic personality, low energy levels, high perceived life stress, and a higher age of onset. They could represent a reactive depression subtype, with burnout characteristics. MDD cases in Cluster 2 showed prominent external factors regarding maltreatment and emotional neglect in childhood and had the lowest age at onset (23 years on average). Therefore, they may suffer from an exogenous depression caused by external stressors. MDD patients in Cluster 3 did not show a strong influence of negative environmental factors, but the contrary - they reported high support, high parental bonding, and low maltreatment levels, similar to Cluster 0. Nevertheless, they showed the limitations in daily activities due to health problems, hence their life has been impacted negatively by depression. (Pelin et al., 2021)

Healthy controls in Clusters 1-4 showed some symptoms resembling the ones of the psychiatric patients in these clusters. However, it appears that these symptoms were insufficient or not severe enough for a formal diagnosis of a mental illness. These healthy individuals might have only exhibited the short-term symptoms, such as those caused by a recent adverse life event. They may, alternatively, still develop a psychiatric disorder later during their lifetime. Similar to the MDD subtypes analysis, ADHD PGS appeared as (nominally) significant between healthy controls grouped to different clusters. Follow-up assessments within the same cohort may reveal which share of these individuals stays healthy over time.

Combining PGS with family history

Analyses with PGS and family history showed limited or non-existent significant differences between Clusters 1-3. For the extreme Clusters 0 and 4, ranked on the two opposite sides of the severity scale, both the univariate differences and the prediction scores of classification models were the strongest. These findings go in line with the PGS distribution shown in Figure 2.4 - PGS still might have a good predictive value only for individuals in the lowest and highest risk groups. As mentioned in the Background chapter, this imperfect accuracy and limited power of stand-alone PGS are in general expected since genetic components are not the only risk factors for diseases, and the PGS can currently explain only part of the genetic component

of a given condition (Wray et al., 2020). Therefore, combining PGS with other risk information of an individual is of great importance (Hujoel et al., 2021; Murray et al., 2020; Bigdeli et al., 2016) and was confirmed also in this thesis - the model with both family history and PGSs significantly improved the variance explained.

Even though the significance of PGS is not so strong and mostly expressed in the extreme clusters, some findings correspond to the effects identified in the previous studies (Lee et al., 2019b; Howard et al., 2019; Coleman et al., 2020; Pardiñas et al., 2018; Wray et al., 2018): schizophrenia, MDD, and psychiatric cross-disorder PGS were significantly higher in the comparison of Cluster 4 and Cluster 0, while the PGS for educational attainment was lower.

Moreover, the genetic analyses with PGS and family history show the importance of focusing both on univariate and multivariate statistical approaches, as the latter can uncover the relationships between variables and their interactions. For example, the univariate analysis in form of significance testing found 6 significant variables for *Cluster 4 vs. all* comparison, while the Lasso model for *Cluster 4 vs. all* assigned non-zero coefficients to 16 variables. PGS that did not show the association in the univariate testing (e.g. Neuroticism, Hedonic well-being) might not be informative enough on their own, but are contributing when observed in combination with other PGSs and family history. Hence, it is beneficial to focus on both types of analyses to determine the possible links between the variables.

Importance of transdiagnostic approaches and severity assessment

The clustering results in this thesis (and Pelin et al., 2021) demonstrate the importance of transdiagnostic clustering approaches, stratifying the sample consisting of different psychiatric disorders. Individuals diagnosed with the same psychiatric disorder can experience very heterogeneous symptoms and have different levels of impairment. Hence, they might require different treatment regimes or clinical interventions. Moreover, their symptoms may partially overlap with symptoms occurring in patients formally diagnosed with different diagnoses, emphasizing the need for symptom-specific rather than diagnosis-specific treatment (Pelin et al., 2021). Results also show that inclusion of healthy controls in such studies is beneficial, as it may either detect individuals in remission (psychiatric patients in Cluster 0) or individuals that are at risk for developing a psychiatric disorder if they continue to exhibit the symptoms of the specific group they are in (healthy controls in Clusters 1-4). These findings go in line with the

assumption that the symptom space is dimensional and that mental health could be conceptualized along continuous dimensions, with mental well-being on the one end and severe disease on the other. Clinically, such an approach, if supported by future studies, might help to decide whether or not to clinically intervene, and if the clinical intervention is needed - which level of the treatment is required given the symptom's severity (Dalgleish et al., 2020).

Computational approaches in psychiatric research

A very common approach in dealing with high-dimensional data and correlated features is to apply the PCA to reduce the complexity before submitting the data further into the clustering algorithm. In this work, however, we took the advantage of a more complex subspace clustering algorithm that is optimized for handling a higher number of correlated features and therefore, does not require any kind of prior application of global dimensionality reduction methods. Indeed, we showed that using more complex algorithms like HDDC could be beneficial when working with the data coming from a sample of patients diagnosed with complex psychiatric disorders since it can capture more than the distinction between health and disease (Section 4.1.7). However, the algorithm used in this work is a type of clustering algorithm that relies on a previously given number of clusters and a discrete categorization. With an assumption of the existence of a symptom continuum spanning from health to severe mental illness, future studies might examine the methods which incorporate the concept of continuum into their objective function (Shah and Koltun, 2018; Pelin et al., 2021).

The results of feature selection analysis showed poor generalizability of the prediction model and unstable sets of SNPs chosen across different runs, both when predicting the identified Cluster 0, as well as the formal labels. This can lead us to conclude that for the analysis including a high number of SNPs, the univariate approaches, such as GWAS, appear to be more promising to date (Saeys et al., 2007). Still, the multivariate approaches have received attention and could produce significant advancements with the development of new methods, standards, validation, and transparency (Bracher-Smith et al., 2020; Qian et al., 2020). Moreover, with the ultimate goal of facilitating disease association studies, it is necessary to apply these kinds of methods to larger datasets. This is especially important for psychiatric disorders, which show substantially smaller single SNP effect sizes in the studies, compared to the common chronic diseases (Zhang et al., 2018a). Therefore, to test our approach for feature selection, it would be beneficial to apply it to the larger sample, potentially from publicly available datasets.

5.2 Outlook

Data-driven approaches in psychiatry offer great advantages to the field, by uncovering the patterns and relations hidden in the data available nowadays. Such methods may contribute to gain a deeper understanding of the underlying mechanisms of the disorders and improve diagnostics and treatment. We have demonstrated that transdiagnostic clustering approaches might help to better understand the heterogeneity between and within psychiatric disorders. Moreover, if applied to other cohorts, they might identify the groups of patients that share the clinical symptoms and, hence, could benefit from similar treatments. However, the path to the paradigm shift in psychiatric nosology is still long and strewn with challenges.

To date, there are many clustering studies with different approaches with respect to the scope of psychiatric disorders observed, data modalities, or machine learning algorithms used. These various methodologies between studies result in many different taxonomic solutions and further research is needed to investigate similarities as well as discrepancies between them to ultimately reach a consensus.

There is no doubt that psychiatric disorders are extremely complex and heterogeneous in their nature, and therefore, data-driven methods chosen to deal with such data have to be carefully considered. Challenges posed to the clustering algorithms, such as the choice of similarity metric and high-dimensionality, all discussed in the Background chapter, are playing a big role in the resulting differences in taxonomic solutions. Two possible solutions could be considered to decrease the discrepancies coming from this methodological part.
First, researchers might develop richer clustering models that incorporate domain knowledge into the process and guide the algorithm toward clinically relevant variation (Marquand et al., 2016). Domain knowledge of medical doctors may include the data features they specifically focus on while assessing the clinical picture and deciding on the treatment regime. These features should not only be based on the current symptoms, but also on the historical data such as past disorder developments and treatment responses, and the expected trajectory of the disease progression. This type of data could be collected, summarized, and validated by a large number of practitioners over years, based on a large number of cases. Clearly, the collection of such information is not an easy endeavor as it requires a big collaborative effort of like-minded researchers and medical practitioners.

Second, longitudinal clustering approaches may help refine psychiatric nosology, by identifying the individual disease trajectories and forming the groups of heterogeneous disease development pathways (Fountain et al., 2012; Rhebergen et al., 2012). This way, the further course of disease development may be better understood and predicted, and clinical management choices might be better informed. For example, the individuals in the same disease trajectory cluster might profit from similar treatment, irrespective of the formal diagnostic label. However, longitudinal studies are not easy to establish and can be very expensive as they take a lot of time during which the sample size often decreases as the participants withdraw from the study.

This work is part of the FOR2107 study which is intended to be a longitudinal study, however, at the time of writing, only the baseline time point was available and therefore, the cross-sectional clustering was performed. The future work within this cohort may reveal many interesting relationships and new insights. For example, it may be beneficial to determine the cluster changes over time, assess whether healthy controls in the higher severity clusters are indeed at risk to develop a psychiatric disorder, or apply longitudinal clustering methods. Moreover, it would be interesting to conduct a follow-up project which would include genetic variables in the clustering process to determine the differences to the clusters identified with the clinical variables and possible relationships between symptoms and genetic underpinnings.

To conclude, today's technologies together with the amounts of data generated present enormous opportunities for researchers aiming for an improvement of the current psychiatric nosology. These opportunities and big collaborative efforts may indeed lead us to the personalized treatments and, thus, improved lives of the ones suffering from mental illnesses.

Bibliography

Abraham, Gad, Adam Kowalczyk, Justin Zobel, and Michael Inouye (2013). "Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease." In: *Genetic epidemiology* 37.2, pp. 184–195.

Akaike, Hirotogu (1998). "Information theory and an extension of the maximum likelihood principle." In: *Selected papers of hirotugu akaike*. Springer, pp. 199–213.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders : DSM-5*. English. 5th ed. American Psychiatric Association Arlington, VA, xliv, 947 p. ;

Andlauer, Till FM and Markus M Nöthen (2020). "Polygenic scores for psychiatric disease: from research tool to clinical application." In: *Medizinische Genetik* 32.1, pp. 39–45.

Assent, Ira (July 2012). "Clustering high dimensional data." In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2, pp. 340–350.

Balding, David J, Martin Bishop, and Chris Cannings (2008). *Handbook of statistical genetics*. John Wiley & Sons.

Bansal, Vikas et al. (2018). "Genome-wide association study results for educational attainment aid in identifying genetic heterogeneity of schizophrenia." In: *Nature communications* 9.1, pp. 1–12.

Barch, Deanna M. (2020). "What Does It Mean to Be Transdiagnostic and How Would We Know?" In: *American Journal of Psychiatry* 177.5. PMID: 32354269, pp. 370–372.

Baselmans, Bart ML and Meike Bartels (2018). "A genetic perspective on the relationship between eudaimonic–and hedonic well-being." In: *Scientific reports* 8.1, pp. 1–10.

Beijers, Lian, Klaas J. Wardenaar, Hanna M. van Loo, and Robert A. Schoevers (June 2019). "Data-driven biological subtypes of depression: systematic review of biological approaches to depression subtyping." In: *Molecular Psychiatry* 24.6, pp. 888–900. ISSN: 1476-5578.

Bell, Morris, Silvia Corbera, Jason Johannesen, Joanna Fiszdon, and Bruce Wexler (Oct. 2011). "Social Cognitive Impairments and Negative Symptoms in Schizophrenia: Are There Subtypes With Distinct Functional Correlates?" In: *Schizophrenia bulletin* 39.

Bellman, Richard (Nov. 1954). "The theory of dynamic programming." In: *Bull. Amer. Math. Soc.* 60.6, pp. 503–515.

Berg, StÚphanie M van den et al. (2016). "Meta-analysis of genome-wide association studies for extraversion: findings from the genetics of personality consortium." In: *Behavior genetics* 46.2, pp. 170–182.

Bergé, Laurent, Charles Bouveyron, and Stéphane Girard (2012). "HDclassif: An R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data." In: *Journal of Statistical Software, Articles* 46.6, pp. 1–29. ISSN: 1548-7660.

Bermingham, Mairead L et al. (2015). "Application of high-dimensional feature selection: evaluation for genomic prediction in man." In: *Scientific reports* 5.1, pp. 1–12.

Biernacki, Christophe, Gilles Celeux, and Gérard Govaert (2000). "Assessing a mixture model for clustering with the integrated completed likelihood." In: *IEEE transactions on pattern analysis and machine intelligence* 22.7, pp. 719–725.

Bigdeli, Tim B et al. (2016). "Genome-wide association study reveals greater polygenic loading for schizophrenia in cases with a family history of illness." In: *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 171.2, pp. 276–289.

Blashfield, Roger, Jared Keeley, Elizabeth Flanagan, and Shannon Miles (Mar. 2014). "The cycle of classification: DSM-I through DSM-5." In: *Annual review of clinical psychology* 10, pp. 25–51.

Bouveyron, Charles, Stéphane Girard, and Cordelia Schmid (2005). "High-Dimensional Discriminant Analysis." In: *Communications in Statistics - Theory and Methods* 36, pp. 2607 –2623.

– (2007). “High-dimensional data clustering.” In: *Comput. Stat. Data Anal.* 52, pp. 502–519.

Bracher-Smith, Matthew, Karen Crawford, and Valentina Escott-Price (2020). “Machine learning for genetic prediction of psychiatric disorders: a systematic review.” In: *Molecular Psychiatry*, pp. 1–10.

Calafato, Maria Stella et al. (2018). “Use of schizophrenia and bipolar disorder polygenic risk scores to identify psychotic disorders.” In: *The British Journal of Psychiatry* 213.3, pp. 535–541.

Carbone, Elvira Anna et al. (Aug. 2019). “Adverse childhood experiences and clinical severity in bipolar disorder and schizophrenia: A transdiagnostic two-step cluster analysis.” In: *Journal of Affective Disorders* 259.

Chan, Chi, Megan Shanahan, Luz Ospina, Emmett Larsen, and Katherine Burdick (May 2017). “Premorbid Social and Academic Adjustment Trajectories in Schizophrenia and Bipolar Disorder: A Transdiagnostic Cluster Analysis.” In: *Biological Psychiatry* 81, S135–S136.

Cheng, Yuqi et al. (Nov. 2014). “Delineation of Early and Later Adult Onset Depression by Diffusion Tensor Imaging.” In: *PloS one* 9, e112307.

Chiu, Derek S. and Aline Talhouk (Jan. 2018). “diceR: an R package for class discovery using an ensemble driven approach.” In: *BMC Bioinformatics* 19.1, p. 11. ISSN: 1471-2105.

Coleman, Jonathan RI et al. (2020). “The genetics of the mood disorder Spectrum: genome-wide association analyses of more than 185,000 cases and 439,000 controls.” In: *Biological psychiatry* 88.2, pp. 169–184.

Consortium, International Schizophrenia (2009). “Common polygenic variation contributes to risk of schizophrenia that overlaps with bipolar disorder.” In: *Nature* 460.7256, p. 748.

Crouse, Jacob et al. (Mar. 2020). “Transdiagnostic neurocognitive subgroups and functional course in young people with emerging mental disorders: a cohort study.” In: *BJPsych Open* 6.

Dalgleish, Tim, Melissa Black, David Johnston, and Anna Bevan (Mar. 2020). “Transdiagnostic approaches to mental health problems: Current status and future directions.” In: *Journal of consulting and clinical psychology* 88, pp. 179–195.

Dalsgaard, S, PB Mortensen, M Frydenberg, CM Maibing, M Nordentoft, and PH Thomsen (2014). "Association between Attention-Deficit Hyperactivity Disorder in childhood and schizophrenia later in adulthood." In: *European Psychiatry* 29.4, pp. 259–263.

De Maturana, Evangelina López et al. (2014). "Next generation modeling in GWAS: comparing different genetic architectures." In: *Human genetics* 133.10, pp. 1235–1253.

Dehman, Alia, Christophe Ambroise, and Pierre Neuvial (2015). "Performance of a blockwise approach in variable selection using linkage disequilibrium information." In: *BMC bioinformatics* 16.1, pp. 1–14.

Demontis, Ditte et al. (2019). "Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder." In: *Nature genetics* 51.1, pp. 63–75.

Dias, Taciana G. Costa et al. (2015). "Characterizing heterogeneity in children with and without ADHD based on reward system connectivity." In: *Developmental Cognitive Neuroscience* 11, pp. 155 –174.

Dickinson, Dwight et al. (Mar. 2017). "Attacking Heterogeneity in Schizophrenia by Deriving Clinical Subgroups From Widely Available Symptom Data." In: *Schizophrenia bulletin* 44.

Drysdale, Andrew T. et al. (Jan. 2017). "Resting-state connectivity biomarkers define neurophysiological subtypes of depression." In: *Nature Medicine* 23.1, pp. 28–38. ISSN: 1546-170X.

Dubitzky, Werner, Martin Granzow, and Daniel P Berrar (2007). *Fundamentals of data mining in genomics and proteomics*. Springer Science & Business Media.

Dwyer, Dominic et al. (Feb. 2018). "Brain Subtyping Enhances The Neuroanatomical Discrimination of Schizophrenia." In: *Schizophrenia Bulletin* 44.

Dwyer, Dominic B. et al. (May 2020). "An Investigation of Psychosis Subgroups With Prognostic Validation and Exploration of Genetic Underpinnings: The PsyCourse Study." In: *JAMA Psychiatry* 77.5, pp. 523–533. ISSN: 2168-622X.

Dy, Jennifer G and Carla E Brodley (2000). "Feature subset selection and order identification for unsupervised learning." In: *ICML*. Citeseer, pp. 247–254.

Forbush, Kelsie, Kelsey Hagan, Rachel Salk, and Jennifer Wildes (Mar. 2017). "Concurrent and prognostic utility of subtyping anorexia nervosa along

dietary and negative affect dimensions." In: *Journal of Consulting and Clinical Psychology* 85, pp. 228–237.

Fountain, Christine, Alix S. Winter, and Peter S. Bearman (2012). "Six Developmental Trajectories Characterize Children With Autism." In: *Pediatrics* 129.5, e1112–e1120. ISSN: 0031-4005.

Fried, Eiko I and Donald J Robinaugh (2020). *Systems all the way down: embracing complexity in mental health research*.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." In: *Journal of Statistical Software, Articles* 33.1, pp. 1–22. ISSN: 1548-7660.

Fullerton, Janice and John Nurnberger (July 2019). "Polygenic risk scores in psychiatry: Will they be useful for clinicians?" In: *F1000Research* 8, p. 1293.

Fusar-Poli, Paolo et al. (June 2019). "Transdiagnostic psychiatry: a systematic review." In: *World Psychiatry* 18, pp. 192–207.

Gates, Kathleen M., Peter C. M. Molenaar, Swathi P. Iyer, Joel T. Nigg, and Damien A. Fair (Mar. 2014). "Organizing heterogeneous samples using community detection of GIMME-derived resting state functional networks." eng. In: *PloS one* 9.3. 24642753[pmid], e91322–e91322. ISSN: 1932-6203.

Ge, Tian, Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng, and Jordan W. Smoller (Apr. 2019). "Polygenic prediction via Bayesian regression and continuous shrinkage priors." In: *Nature Communications* 10.1, p. 1776. ISSN: 2041-1723.

Ge, Youngchao, Sandrine Dudoit, and Terence P Speed (2003). "Resampling-based multiple testing for microarray data analysis." In: *Test* 12.1, pp. 1–77.

Geisler, Daniel et al. (Aug. 2015). "Brain structure and function correlates of cognitive subtypes in schizophrenia." In: *Psychiatry research* 234.

Giambattista, Concetta, Patrizia Ventura, Paolo Trerotoli, Mariella Margari, Roberto Palumbi, and Lucia Margari (Jan. 2019). "Subtyping the Autism Spectrum Disorder: Comparison of Children with High Functioning Autism and Asperger Syndrome." In: *Journal of Autism and Developmental Disorders* 49.

Goldstein-Piekarski, A. N., L. M. Williams, and K. Humphreys (June 2016). "A trans-diagnostic review of anxiety disorder comorbidity and the impact of

multiple exclusion criteria on studying clinical outcomes in anxiety disorders." In: *Translational Psychiatry* 6.6, e847–e847. ISSN: 2158-3188.

Gould, Ian, Alana Shepherd, Kristin Laurens, Murray Cairns, Vaughan Carr, and Melissa Green (Dec. 2014). "Multivariate neuroanatomical classification of cognitive subtypes in schizophrenia: A support vector machine learning approach." In: *NeuroImage: Clinical* 6.

Greve, Benjamin, Iris Pigeot, Inge Huybrechts, Valeria Pala, and Claudia Börnhorst (2016). "A comparison of heuristic and model-based clustering methods for dietary pattern analysis." In: *Public health nutrition* 19.2, pp. 255–264.

Grilo, Carlos, Robin Masheb, and G. Wilson (Jan. 2002). "Subtyping binge eating disorder." In: *Journal of consulting and clinical psychology* 69, pp. 1066–72.

Grisanzio, Katherine A., Andrea N. Goldstein-Piekarski, Michelle Yuyun Wang, Abdullah P. Rashed Ahmed, Zoe Samara, and Leanne M. Williams (Feb. 2018). "Transdiagnostic Symptom Clusters and Associations With Brain, Behavior, and Daily Function in Mood, Anxiety, and Trauma Disorders." In: *JAMA Psychiatry* 75.2, pp. 201–209. ISSN: 2168-622X.

Gron, Aurlien (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 1st. O'Reilly Media, Inc. ISBN: 1491962291.

Grove, Jakob et al. (2019). "Identification of common genetic risk variants for autism spectrum disorder." In: *Nature genetics* 51.3, pp. 431–444.

Guyon, Isabelle and André Elisseeff (Mar. 2003). "An Introduction to Variable and Feature Selection." In: *J. Mach. Learn. Res.* 3.null, 1157–1182. ISSN: 1532-4435.

Halldorsson, Bjarni V et al. (2004). "Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies." In: *Genome research* 14.8, pp. 1633–1640.

Hamshere, Marian L et al. (2013). "Shared polygenic contribution between childhood attention-deficit hyperactivity disorder and adult schizophrenia." In: *The British Journal of Psychiatry* 203.2, pp. 107–111.

Haroon, Ebrahim et al. (Dec. 2018). "Increased inflammation and brain glutamate define a subtype of depression with decreased regional homogeneity, impaired network integrity, and anhedonia." In: *Translational Psychiatry* 8.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC. ISBN: 1498712169.

He, Dan, Irina Rish, David Haws, and Laxmi Parida (2015). "MINT: mutual information based transductive feature selection for genetic trait prediction." In: *IEEE/ACM transactions on computational biology and bioinformatics* 13.3, pp. 578–583.

He, Jingwu and Alexander Zelikovsky (2006). "MLR-tagging: informative SNP selection for unphased genotypes based on multiple linear regression." In: *Bioinformatics* 22.20, pp. 2558–2561.

Helmes, Edward and Jhan Landmark (Dec. 2003). "Subtypes of Schizophrenia: A Cluster Analytic Approach." In: *Canadian journal of psychiatry. Revue canadienne de psychiatrie* 48, pp. 702–8.

Helzer, John E, Helena C Kraemer, and Robert F Krueger (2006). "The feasibility and need for dimensional psychiatric diagnoses." In: *Psychological medicine* 36.12, p. 1671.

Howard, David et al. (May 2020). "Genetic stratification of depression in UK Biobank." In: *Translational Psychiatry* 10, p. 163.

Howard, David M et al. (2019). "Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions." In: *Nature neuroscience* 22.3, pp. 343–352.

Huang, Hanwen, Yufeng Liu, Ming Yuan, and JS Marron (2015). "Statistical significance of clustering using soft thresholding." In: *Journal of Computational and Graphical Statistics* 24.4, pp. 975–993.

Hujoel, Margaux Louise Anna, Po-Ru Loh, Benjamin Neale, and Alkes L Price (2021). "Incorporating family history of disease improves polygenic risk scores in diverse populations." In: *bioRxiv*.

Inouye, Michael et al. (2018). "Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention." In: *Journal of the American College of Cardiology* 72.16, pp. 1883–1893.

Jablensky, A. (Sept. 2006). "Subtyping schizophrenia: implications for genetic research." In: *Molecular Psychiatry* 11.9, pp. 815–836. ISSN: 1476-5578.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated. ISBN: 1461471370.

Janssen, Ian et al. (2004). "Childhood abuse as a risk factor for psychotic experiences." In: *Acta Psychiatrica Scandinavica* 109.1, pp. 38–45.

Janssens, A Cecile JW (2019). "Validity of polygenic risk scores: are we measuring what we think we are?" In: *Human molecular genetics* 28.R2, R143–R150.

Johns, Louise and Jim van Os (Dec. 2001). "The Continuity of Psychotic Experiences in the General Population." In: *Clinical psychology review* 21, pp. 1125–41.

Joiret, Marc, Jestinah M Mahachie John, Elena S Gusareva, and Kristel Van Steen (2019). "Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies." In: *BioData mining* 12.1, pp. 1–23.

Kaczkurkin, Antonia et al. (Sept. 2019). "Neurostructural Heterogeneity in Youths With Internalizing Symptoms." In: *Biological Psychiatry* 87.

Kendler, K. S. (Dec. 2009). "An historical framework for psychiatric nosology." eng. In: *Psychological medicine* 39.12. 19368761[pmid], pp. 1935–1941. ISSN: 1469-8978.

Kessler, Ronald C., Wai Tat Chiu, Olga Demler, Kathleen R. Merikangas, and Ellen E. Walters (June 2005). "Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication." eng. In: *Archives of general psychiatry* 62.6. 15939839[pmid], pp. 617–627. ISSN: 0003-990X.

Khoury, Brigitte, Cary Kogan, and Sariah Daouk (2017). "International Classification of Diseases 11th Edition (ICD-11)." In: *Encyclopedia of Personality and Individual Differences*. Ed. by Virgil Zeigler-Hill and Todd K. Shackelford. Cham: Springer International Publishing, pp. 1–6. ISBN: 978-3-319-28099-8.

Kircher, Tilo et al. (Sept. 2018). "Neurobiology of the major psychoses: a translational perspective on brain structure and function—the FOR2107 consortium." In: *European Archives of Psychiatry and Clinical Neuroscience* 269.

Lamers, Femke et al. (Mar. 2011). "Comorbidity Patterns of Anxiety and Depressive Disorders in a Large Cohort Study: the Netherlands Study of Depression and Anxiety (NESDA)." In: *The Journal of clinical psychiatry* 72, pp. 341–8.

Lee, Phil, Verneri Anttila, Hyejung Won, Yen-Chen A. Feng, Jacob Rosenthal, and Zhaozhong Zhu (2019a). "Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders." In: *Cell* 179, pp. 1469–1482.

Lee, Phil H et al. (2019b). "Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders." In: *Cell* 179.7, pp. 1469–1482.

Lee, Phil Hyoun and Hagit Shatkay (2006). "BNTagger: improved tagging SNP selection using Bayesian networks." In: *Bioinformatics* 22.14, e211–e219.

Levey, Daniel F et al. (2020). "GWAS of depression phenotypes in the million veteran program and meta-analysis in more than 1.2 million participants yields 178 independent risk loci." In: *medRxiv*.

Lewandowski, Kathryn, Sarah Sperry, Bruce Cohen, and Dost Öngür (Apr. 2014). "Cognitive variability in psychotic disorders: A cross-diagnostic cluster analysis." In: *Psychological Medicine*.

Li, Jundong and Huan Liu (2017). "Challenges of feature selection for big data analytics." In: *IEEE Intelligent Systems* 32.2, pp. 9–15.

Li, Leping, Clarice R Weinberg, Thomas A Darden, and Lee G Pedersen (2001). "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method." In: *Bioinformatics* 17.12, pp. 1131–1142.

Li, Xian-Bin, Qi-Yong Li, Jin-Tong Liu, Liang Zhang, Yi-Lang Tang, and Chuan-Yue Wang (2015). "Childhood trauma associates with clinical features of schizophrenia in a sample of Chinese inpatients." In: *Psychiatry research* 228.3, pp. 702–707.

Lichtenstein, Paul et al. (Jan. 2009). "Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study." eng. In: *Lancet (London, England)* 373.9659. 19150704[pmid], pp. 234–239. ISSN: 1474-547X.

Long, Nanye, Daniel Gianola, Guilherme JM Rosa, Kent A Weigel, and Santiago Avendaño (2009). "Comparison of classification methods for detecting associ-

ations between SNPs and chick mortality." In: *Genetics Selection Evolution* 41.1, pp. 1–14.

Luciano, Michelle et al. (2018). "Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism." In: *Nature genetics* 50.1, pp. 6–11.

López-Ratón, Mónica, María Rodríguez-Álvarez, Carmen Cadarso-Suárez, and Francisco Gude-Sampedro (2014). "OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests." In: *Journal of Statistical Software, Articles* 61.8, pp. 1–36. ISSN: 1548-7660.

Maglanoc, Luigi et al. (May 2018). "Data-Driven Clustering Reveals a Link Between Symptoms and Functional Brain Connectivity in Depression." In: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 4.

Maj, Mario (2005). "'Psychiatric comorbidity': an artefact of current diagnostic systems?" In: *British Journal of Psychiatry* 186.3, 182–184.

– (June 2018). "Why the clinical utility of diagnostic categories in psychiatry is intrinsically limited and how we can use new approaches to complement them." In: *World Psychiatry* 17, pp. 121–122.

Marquand, Andre, Thomas Wolfers, Maarten Mennes, Jan Buitelaar, and Christian Beckmann (Apr. 2016). "Beyond Lumping and Splitting: A Review of Computational Approaches for Stratifying Psychiatric Disorders." In: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 1.

Misiak, Błażej, Maja Krefft, Tomasz Bielawski, Ahmed A Moustafa, Maria M Sąsiadek, and Dorota Frydecka (2017). "Toward a unified theory of childhood trauma and psychosis: a comprehensive review of epidemiological, clinical, neuropsychological and biological findings." In: *Neuroscience & Biobehavioral Reviews* 75, pp. 393–406.

Mitchell, Tom M. (1997). *Machine Learning*. New York: McGraw-Hill. ISBN: 978-0-07-042807-2.

Monti, Stefano, Pablo Tamayo, Jill Mesirov, and Todd Golub (2003). "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data." In: *Machine learning* 52.1-2, pp. 91–118.

Mostert, Jeanette C. et al. (Feb. 2018). "Similar Subgroups Based on Cognitive Performance Parse Heterogeneity in Adults With ADHD and Healthy Con-

trols." eng. In: *Journal of attention disorders* 22.3. 26374770[pmid], pp. 281–292. ISSN: 1557-1246.

Murray, Graham K, Tian Lin, Jehannine Austin, John J McGrath, Ian B Hickie, and Naomi R Wray (2020). "Could Polygenic Risk Scores Be Useful in Psychiatry?: A Review." In: *JAMA psychiatry*.

Nagelkerke, Nico JD et al. (1991). "A note on a general definition of the coefficient of determination." In: *Biometrika* 78.3, pp. 691–692.

Okbay, Aysu et al. (2016). "Genome-wide association study identifies 74 loci associated with educational attainment." In: *Nature* 533.7604, pp. 539–542.

Owen, Michael (Nov. 2014). "New Approaches to Psychiatric Diagnostic Classification." In: *Neuron* 84, pp. 564–571.

Palacio-Niño, Julio-Omar and Fernando Berzal (2019). "Evaluation metrics for unsupervised learning algorithms." In: *arXiv preprint arXiv:1905.05667*.

Pardiñas, Antonio F et al. (2018). "Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection." In: *Nature genetics* 50.3, pp. 381–389.

Parsons, Lance, Ehtesham Haque, and Huan Liu (June 2004). "Subspace Clustering for High Dimensional Data: A Review." In: *SIGKDD Explor. Newsl.* 6.1, 90–105. ISSN: 1931-0145.

Pelin, Helena et al. (2021). "Identification of transdiagnostic psychiatric disorder subtypes using unsupervised learning." In: *Neuropsychopharmacology*. ISSN: 1740-634X.

Phuong, Tu Minh, Zhen Lin, and Russ B Altman (2005). "Choosing SNPs using feature selection." In: *2005 IEEE Computational Systems Bioinformatics Conference (CSB'05)*. IEEE, pp. 301–309.

Pollard, KS, S Dudoit, and MJ van der Laan (2004). *Multiple testing procedures: R multtest package and applications to genomics. UC Berkeley Division of Biostatistics Working Paper Series*. Tech. rep. Working Paper 164. http://www. bepress. com/ucbbiostat/paper164.

Porter, Heather F and Paul F O'Reilly (2017). "Multivariate simulation framework reveals performance of multi-trait GWAS methods." In: *Scientific reports* 7.1, pp. 1–12.

Power, Robert A et al. (2017). "Genome-wide association for major depression through age at onset stratification: major depressive disorder working group of the psychiatric genomics consortium." In: *Biological Psychiatry* 81.4, pp. 325–335.

Psychiatric Genomics Consortium, Schizophrenia Working Group of the et al. (2014). "Biological insights from 108 schizophrenia-associated genetic loci." In: *Nature* 511.7510, pp. 421–427.

Qian, Junyang et al. (2020). "A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank." In: *PLoS genetics* 16.10, e1009141.

Quenouille, M. H. (1949). "Approximate Tests of Correlation in Time-Series." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 11.1, pp. 68–84. ISSN: 00359246.

Redlich, Ronny et al. (2014). "Brain morphometric biomarkers distinguishing unipolar and bipolar depression: a voxel-based morphometry–pattern classification approach." In: *JAMA psychiatry* 71.11, pp. 1222–1230.

Rhebergen, D, F Lamers, J Spijker, R de Graaf, ATF Beekman, and BWJH Penninx (2012). "Course trajectories of unipolar depressive disorders identified by latent class growth analysis." In.

Rifkin, Ryan and Aldebaro Klautau (Dec. 2004). "In Defense of One-Vs-All Classification." In: *J. Mach. Learn. Res.* 5, 101–141. ISSN: 1532-4435.

Ring, Howard, Marc Woodbury-Smith, Peter Watson, Sally Wheelwright, and Simon Baron-Cohen (Feb. 2008). "Clinical heterogeneity among people with high functioning autism spectrum conditions: Evidence favouring a continuous severity gradient." In: *Behavioral and brain functions : BBF* 4, p. 11.

Robin, Xavier et al. (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves." In: *BMC Bioinformatics* 12, p. 77.

Rubino, I Alex, Ellen Frank, Roberta Croce Nanni, Daniela Pozzi, Teresa Lanza Di Scalea, and Alberto Siracusano (2009). "A comparative study of axis I antecedents before age 18 of unipolar depression, bipolar disorder and schizophrenia." In: *Psychopathology* 42.5, pp. 325–332.

Ruderfer, Douglas M et al. (2018). "Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes." In: *Cell* 173.7, pp. 1705–1715.

Saeys, Yvan, Inaki Inza, and Pedro Larranaga (2007). "A review of feature selection techniques in bioinformatics." In: *bioinformatics* 23.19, pp. 2507–2517.

Sasikala, S, S Appavu alias Balamurugan, and S Geetha (2015). "A novel feature selection technique for improved survivability diagnosis of breast cancer." In: *Procedia Computer Science* 50, pp. 16–23.

Seow, Lee Seng Esmond et al. (Dec. 2017). "Correct recognition and continuum belief of mental disorders in a nursing student population." In: *BMC Psychiatry* 17.

Shah, Shital C and Andrew Kusiak (2004). "Data mining and genetic algorithm based gene/SNP selection." In: *Artificial intelligence in medicine* 31.3, pp. 183–196.

Shah, Sohil Atul and Vladlen Koltun (2018). "Deep continuous clustering." In: *arXiv preprint arXiv:1803.01449*.

Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding Machine Learning: From Theory to Algorithms*. USA: Cambridge University Press. ISBN: 1107057132.

Shorter, Edward (Mar. 2015). "The history of nosology and the rise of the Diagnostic and Statistical Manual of Mental Disorders." eng. In: *Dialogues in clinical neuroscience* 17.1. 25987864[pmid], pp. 59–67. ISSN: 1958-5969.

Simon, Noah, Jerome Friedman, Trevor Hastie, and Robert Tibshirani (2013). "A sparse-group lasso." In: *Journal of computational and graphical statistics* 22.2, pp. 231–245.

Sklar, Pamela et al. (2011). "Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4." In: *Nature genetics* 43.10, p. 977.

So, Hon-Cheong and Pak C Sham (2017). "Exploring the predictive power of polygenic scores derived from genome-wide association studies: a study of 10 complex traits." In: *Bioinformatics* 33.6, pp. 886–892.

"Spearman Rank Correlation Coefficient" (2008). In: *The Concise Encyclopedia of Statistics*. New York, NY: Springer New York, pp. 502–505. ISBN: 978-0-387-32833-1.

Spitzer, R.L., Janet Williams, M Gibbon, and Michael First (Sept. 1992). "The structured clinical interview for DSM-III-R (SCID). I: history, rationale, and description." In: *Archives of general psychiatry* 49, pp. 624–9.

Stahl, Eli A et al. (2019). "Genome-wide association study identifies 30 loci associated with bipolar disorder." In: *Nature genetics* 51.5, pp. 793–803.

Steel, Zachary et al. (Apr. 2014). "The global prevalence of common mental disorders: a systematic review and meta-analysis 1980-2013." eng. In: *International journal of epidemiology* 43.2. 24648481[pmid], pp. 476–493. ISSN: 1464-3685.

Sun, Huaiqiang et al. (May 2015). "Two Patterns of White Matter Abnormalities in Medication-Naive Patients With First-Episode Schizophrenia Revealed by Diffusion Tensor Imaging and Cluster Analysis." In: *JAMA psychiatry* 72.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1, pp. 267–288. ISSN: 00359246.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie (2001). "Estimating the number of clusters in a data set via the gap statistic." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2, pp. 411–423.

Ullah, Adnan, Usman Qamar, Farhan Hassan Khan, and Saba Bashir (2017). "Dimensionality reduction approaches and evolving challenges in high dimensional data." In: *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*, pp. 1–8.

Varese, Filippo et al. (2012). "Childhood adversities increase the risk of psychosis: a meta-analysis of patient-control, prospective-and cross-sectional cohort studies." In: *Schizophrenia bulletin* 38.4, pp. 661–671.

Veatch, Olivia, Jeremy Veenstra-VanderWeele, Melissa Potter, Margaret Pericak-Vance, and Jonathan Haines (Dec. 2013). "Genetically Meaningful Phenotypic Subgroups in Autism Spectrum Disorders." In: *Genes, brain, and behavior* 13.

Von Luxburg, U. (2010). *Clustering Stability: An Overview*. Foundations and Trends in Machine Learning Series. Now Publishers. ISBN: 9781601983442.

Wagner, Silke and Dorothea Wagner (2007). *Comparing clusterings: an overview*. Universität Karlsruhe, Fakultät für Informatik Karlsruhe.

Waldmann, Patrik, Gábor Mészáros, Birgit Gredler, Christian Fuerst, and Johann Sölkner (2013). "Evaluation of the lasso and the elastic net in genome-wide association studies." In: *Frontiers in genetics* 4, p. 270.

Ward Jr, Joe H (1963). "Hierarchical grouping to optimize an objective function." In: *Journal of the American statistical association* 58.301, pp. 236–244.

Weir, Bruce S (1979). "Inferences about linkage disequilibrium." In: *Biometrics*, pp. 235–254.

Westfall, P. H. and S. S. Young (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley.

Wittchen, H.-U., U. Wunderlich, S. Gruschwitz, and M. Zaudig (1997). *SKID I. Strukturiertes Klinisches Interview für DSM-IV. Achse I: Psychische Störungen. Interviewheft und Beurteilungsheft. Eine deutschsprachige, erweiterte Bearb. d. amerikanischen Originalversion des SKID I.*

Wray, Naomi R et al. (2018). "Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression." In: *Nature genetics* 50.5, pp. 668–681.

Wray, Naomi R et al. (2020). "From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer." In: *JAMA psychiatry*.

Yee, Thomas (2010). "The VGAM Package for Categorical Data Analysis." In: *Journal of Statistical Software, Articles* 32.10, pp. 1–34. ISSN: 1548-7660.

Yu, Chenglong, Mauricio Arcos-Burgos, Julio Licinio, and M-L Wong (May 2017). "A latent genetic subtype of major depression identified by whole-exome genotyping data in a Mexican-American cohort." In: *Translational Psychiatry* 7, e1134.

Yuan, Ming and Yi Lin (2006). "Model selection and estimation in regression with grouped variables." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1, pp. 49–67.

Zachar, Peter and Kenneth Kendler (2017). "The Philosophy of Nosology." In: *Annual Review of Clinical Psychology* 13.1. PMID: 28482691, pp. 49–71.

Zhang, Lei, Yu-Fang Pei, Jian Li, Christopher J Papasian, and Hong-Wen Deng (2009). "Univariate/multivariate genome-wide association scans using data from families and unrelated samples." In: *PloS one* 4.8, e6502.

Zhang, Yan, Guanghao Qi, Ju-Hyun Park, and Nilanjan Chatterjee (2018a). "Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits." In: *Nature genetics* 50.9, pp. 1318–1326.

Zhang, Yu-Hang, Yu Hu, Yuchao Zhang, Lan-Dian Hu, and Xiangyin Kong (2018b). "Distinguishing three subtypes of hematopoietic cells based on gene expression profiles using a support vector machine." In: *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1864.6, pp. 2255–2265.

Appendix

Methods A1

Individual-level quality control

Discovery sample

Of 1623 individuals eligible for the discovery sample, 3 were excluded because they withdrew their consent and one for missing diagnostic information from all subsequent analysis. Next, 47 pairs of genetic relatives were identified in PLINK[1] using the command –genome. Of each pair showing a PI-HAT $\geq$ 12.5, the individual with the higher genotyping rate was kept in the analyses. Finally, 322 individuals with missing data in any of the 57 variables used for clustering were omitted. The final discovery sample consisted of 1250 individuals, including 590 healthy controls, 477 MDD, 75 BD, 25 SZA, and 53 SCZ patients, and 30 individuals with other diagnoses disorders including anxiety disorders, adjustment disorders, and substance use disorders.

Replication sample

Of 855 individuals eligible for the replication dataset, one was excluded because of a withdrawn consent and two for missing diagnostic information. Next, 27 relatives were removed using the same procedure as applied for the discovery sample. Finally, 201 individuals were excluded due to missing data in any of the variables used for clustering. The final replication sample for the analysis consisted of 622 individuals, including 240 healthy controls, 283 MDD, 44 BD, 13 SZA, and 17 SCZ patients, and 25 individuals with other diagnoses including anxiety, adjustment, and substance use disorders.

Source: Pelin et al., 2021

[1] PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.
Source: https://zzz.bwh.harvard.edu/plink/
Purcell, Shaun, et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses." The American journal of human genetics 81.3 (2007): 559-575.

Table A1
Variables used in the clustering procedure

Category	Variable, mean (SD)	Discovery	Replication
Attachment style	RSQ Anxiety of separation	2.7 (0.7)	2.7 (0.7)
	RSQ Avoidance of closeness	2.5 (0.9)	2.5 (0.9)
	RSQ Desire for independence	3.9 (0.8)	3.97 (0.7)
	RSQ Lack of trust	2.3 (0.9)	2.4 (0.9)
Depression and anxiety level	BDI-II Sum	10.7 (10.8)	11.3 (10.3)
	HAMA Sum	7.3 (7.9)	8.4 (8.2)
	HAMD Sum21	5.4 (6.6)	6.1 (6.8)
	STAIS	42.2 (13.2)	43 (12.8)
	STAIT	43.2 (14.2)	44.2 (13.8)
Anhedonia	SHAPS	1.99 (2.9)	1.97 (2.7)
Life events and stress	LEQ Negative Events score	10 (13.8)	10.4 (12.3)
	LEQ Positive Events score	9.8 (9.3)	9.4 (9.3)
	PSS Sum	22.8 (10.8)	23.8 (10.5)
Maltreatment in childhood and youth	ACE Sum	1.6 (1.9)	1.6 (1.8)
	CTQ Emotional abuse	9.1 (4.7)	9.1 (4.5)
	CTQ Emotional neglect	10.7 (5.2)	11.1 (5.2)
	CTQ Physical abuse	6.2 (2.6)	6.1 (2.4)
	CTQ Physical neglect	7.2 (2.7)	6.8 (2.5)
	CTQ Sexual abuse	5.8 (2.5)	5.7 (2.5)
Mania Symp.	YMRS	1.2 (2.5)	1.4 (2.7)
Neg. Symp.	SANS sum score	5.7 (9.9)	4.3 (7.3)
Pos. Symp.	SAPS sum score	1.4 (5.2)	0.7 (2.7)
Personality	NEO-FFI Agreeableness	33.1 (6.0)	33.2 (6.1)
	NEO-FFI Conscientiousness	32 (7.5)	32.2 (7.4)
	NEO-FFI Extraversion	26.4 (8.2)	25.8 (8.2)
	NEO-FFI Neuroticism	22.1 (10.5)	23.1 (10.2)
	NEO-FFI Openness to experience	30.3 (7.0)	30.4 (6.8)
	SPQB (Schizotypy)	5.95 (4.7)	5.97 (4.7)

Protective factors	Maternal bonding	24.7 (8.6)	24.4 (8.5)
	Paternal bonding	22.3 (8.7)	21.2 (8.9)
	RS25 Sum Score (Resilience)	125.4 (27.2)	125.1 (26.9)
	Social support	4.1 (0.8)	4.1 (0.7)
SF Health Survey (Quality of life)	SF36 Bodily pain	76.3 (26.2)	74.7 (26.1)
	SF36 Energy/fatigue	50.2 (23.0)	47.5 (22.1)
	SF36 General health	66 (23.4)	68.1 (23)
	SF36 Mental health	64.3 (22.6)	62.8 (22.1)
	SF36 Physical functioning	89.5 (17.1)	89.8 (15.5)
	SF36 Role emotional	65.5 (42.4)	60.6 (43.2)
	SF36 Role physical	76.6 (36.7)	71.2 (36.1)
	SF36 Social functioning	73.3 (30)	70.4 (30.1)
Symptom checklist	SCL90R Additional Items	4.1 (4.0)	4.4 (3.8)
	SCL90R Anxiety	5.4 (6.7)	5.4 (6.2)
	SCL90R Depression	11.4 (11.9)	12.2 (11.6)
	SCL90R Global severity index	0.6 (0.6)	0.6 (0.5)
	SCL90R Hostility	2.9 (3.7)	3.2 (3.9)
	SCL90R Interpersonal sensitivity	6.8 (7.3)	6.96 (7.0)
	SCL90R Obsessive–compulsive behavior	8.4 (8.1)	8.7 (7.7)
	SCL90R Paranoid ideation	3.5 (4.4)	3.3 (4.0)
	SCL90R Phobic anxiety	2.3 (4.1)	2.3 (3.9)
	SCL90R Positive symptom distress Index	1.5 (0.5)	1.5 (0.5)
	SCL90R Positive symptom total	29.8 (21.0)	31.2 (19.5)
	SCL90R Psychoticism	3.9 (5.2)	3.7 (4.9)
	SCL90R Somatization	6.96 (7.1)	7.1 (6.5)
Neuro-psychology	Verbal IQ	113.99 (13.7)	113.8 (13.4)
	VLMT Sum	56.8 (10.2)	56.1 (9.6)
	Corsi block-tapping test	7.3 (3.4)	17.2 (3.1)
	Letter Number Span test	16.1 (3.2)	16.1 (3.4)

Source: Pelin et al., 2021

Table A2

Characterization of the discovery sample and its clusters regarding variables not used in the clustering process

Variable	Full	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
N	1250	535	38	266	215	196
Age, mean (SD)	35.1 (13.0)	31.7 (11.9)	38.6 (13.9)	35.3 (12.7)	37.9 (13.7)	40.3 (12.8)
Gender - male, N (%)	483 (39%)	205 (38%)	9 (24%)	106 (40%)	74 (34%)	89 (45%)
AAO*, mean (SD)	25.2 (11.9)	24.5 (9.9)	29.6 (13.3)	23.3 (11.5)	27.8 (12.8)	24.4 (11.4)
Years of education, mean (SD)	13.5 (2.6)	14.4 (2.4)	13.3 (2.6)	13.8 (2.6)	13.1 (2.8)	12.1 (2.7)
BMI, mean (SD)	25.3 (5.5)	23.7 (4.3)	24.9 (5.1)	24.8 (4.9)	27.0 (6.6)	28.1 (6.4)
GAF score, mean (SD)	76.5 (19.2)	91.5 (7.7)	73.1 (17.2)	72.5 (15.8)	65.2 (15.8)	53.8 (14.6)
Positive Symptoms (SAPS)						
Hallucinations, mean (SD)	0.3 (1.8)	0.03 (0.21)	0 (0)	0.03 (0.23)	0.07 (0.55)	1.49 (4.24)
Bizarre Behavior, mean (SD)	0.09 (0.7)	0.01 (0.13)	0 (0)	0.06 (0.31)	0.08 (0.53)	0.41 (1.53)
Positive Formal Thought Disorder, mean (SD)	0.5 (1.9)	0.08 (0.43)	0.05 (0.32)	0.29 (1.1)	0.5 (1.4)	2.33 (4.13)
Delusions, mean (SD)	0.5 (2.5)	0.02 (0.15)	0.05 (0.23)	0.19 (0.87)	0.09 (0.49)	2.48 (5.75)
Negative Symptoms (SANS)						
Anhedonia, mean (SD)	1.9 (3.5)	0.1 (0.7)	2.1 (3.5)	2.3 (3.4)	3.1 (4.1)	4.5 (4.9)
Affective blunting, mean (SD)	1.5 (3.6)	0.3 (0.9)	1.4 (2.9)	1.83 (3.9)	1.72 (3.26)	4.2 (5.9)
Avolition / Apathy, mean (SD)	1.3 (2.5)	0.09 (0.47)	0.8 (1.8)	1.4 (2.4)	1.8 (2.4)	3.8 (3.7)
Alogia, mean (SD)	0.5 (1.6)	0.07 (0.4)	0.2 (0.9)	0.7 (1.8)	0.6 (1.5)	1.5 (2.8)

Medication and hospitalization						
Hospitalization - nr. of times, mean (SD)	1.1 (2.2)	0.2 (1.0)	1.2 (2.9)	1.2 (2.1)	1.7 (2.2)	3.1 (3.0)
Medication index load, mean (SD)	0.8 (1.4)	0.1 (0.6)	0.6 (1.1)	0.7 (1.7)	1.5 (1.7)	1.9 (1.8)
Antidepressant, yes, N (%)	350 (28%)	23 (4%)	12 (32%)	94 (35%)	122 (57%)	99 (51%)
Antipsychotic, yes, N (%)	187 (15%)	12 (2%)	4 (11%)	41 (15%)	46 (21%)	84 (43%)
Mood stabilizer, yes, N (%)	65 (5%)	9 (2%)	1 (3%)	7 (3%)	22 (10%)	26 (13%)
Antidepressant + Antipsychotic, yes, N (%)	111 (9%)	3 (1%)	2 (5%)	29 (11%)	35 (16%)	42 (21%)
Antidepressant + Mood stabilizer, yes, N (%)	40 (3%)	4 (1%)	1 (3%)	3 (1%)	16 (7%)	16 (8%)
Antipsychotic + Mood stabilizer, yes, N (%)	31 (2%)	3 (1%)	1 (3%)	4 (2%)	9 (4%)	14 (7%)
Smoking						
No or minimal addiction, N (%)	1071 (86%)	498 (93%)	34 (89%)	228 (86%)	175 (81%)	136 (69%)
Average addiction, N (%)	68 (5%)	21 (4%)	1 (3%)	19 (7%)	8 (4%)	19 (10%)
Strong addiction, N (%)	70 (6%)	10 (2%)	1 (3%)	16 (6%)	22 (10%)	21 (11%)
Very strong addiction, N (%)	41 (3%)	6 (1%)	2 (5%)	3 (1%)	10 (5%)	20 (10%)
Sociodemographic - Type of living						
Alone, N (%)	342 (27%)	105 (20%)	16 (42%)	90 (34%)	64 (30%)	67 (34%)
Marriage/life partner, N (%)	277 (22%)	108 (20%)	7 (18%)	51 (19%)	62 (29%)	49 (25%)
Parents/relatives, N (%)	128 (10%)	45 (8%)	3 (8%)	29 (11%)	27 (13%)	24 (12%)
Non-martial partner, N (%)	167 (13%)	89 (17%)	4 (11%)	29 (11%)	21 (10%)	24 (12%)
Therapeutic facilities, N (%)	12 (1%)	0 (0%)	0 (0%)	0 (0%)	1 (0%)	11 (6%)
Shared flat, N (%)	303 (24%)	185 (35%)	7 (18%)	61 (23%)	31 (14%)	19 (10%)
Other, N (%)	15 (1%)	2 (0%)	1 (3%)	4 (2%)	6 (3%)	2 (1%)
Sociodemographic - Income						
Own work, N (%)	541 (43%)	259 (48%)	16 (42%)	117 (44%)	87 (40%)	62 (32%)
Parent/Partner/other, N (%)	334 (27%)	201 (38%)	10 (26%)	53 (20%)	48 (22%)	22 (11%)

Wage replacement / Sickness pay / Unemployment pay, N (%)	276 (22%)	35 (7%)	11 (29%)	70 (26%)	60 (28%)	100 (51%)
Other, N (%)	86 (7%)	35 (7%)	1 (3%)	24 (9%)	17 (8%)	9 (5%)
Sociodemographic - Social contacts						
Sociodemographic - Social contacts						
1 time per week, N (%)	274 (22%)	97 (18%)	10 (26%)	64 (24%)	60 (28%)	43 (22%)
Several times per week, N (%)	722 (58%)	402 (75%)	14 (37%)	139 (52%)	90 (42%)	77 (39%)
1 time every 14 days, N (%)	112 (9%)	20 (4%)	4 (11%)	31 (12%)	31 (14%)	26 (13%)
1 time in month, including distant acquaintances, N (%)	95 (8%)	10 (2%)	7 (18%)	27 (10%)	22 (10%)	29 (15%)
No social contact apart from meeting at work, N (%)	35 (3%)	4 (1%)	3 (8%)	4 (2%)	11 (5%)	13 (7%)
Meeting friends under no circumstances, N (%)	7 (1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	7 (4%)

Source: Pelin et al., 2021

Table A3

Characterization of the discovery sample and its clusters regarding variables used in the clustering process

Category	Variable, mean (SD)	Full	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
	N	1250	535	38	266	215	196
Attachment style	RSQ Anxiety of separation	2.7 (0.7)	2.5 (0.5)	2.8 (0.8)	2.9 (0.8)	2.9 (0.8)	2.8 (0.8)
	RSQ Avoidance of closeness	2.5 (0.9)	1.98 (0.6)	2.7 (0.9)	2.9 (0.8)	2.5 (0.8)	3.1 (0.9)
	RSQ Desire for independence	3.9 (0.8)	3.8 (0.8)	4.0 (0.7)	4.1 (0.7)	3.8 (0.9)	4.0 (0.8)
	RSQ Lack of trust	2.3 (0.9)	1.8 (0.6)	2.8 (0.9)	2.7 (0.8)	2.5 (0.9)	3.0 (0.9)
Depression and anxiety level	BDI-II Sum	10.7 (10.8)	3.2 (3.3)	15.1 (12.8)	12.7 (9.6)	17.6 (10.1)	20.3 (11.7)
	HAMA Sum	7.3 (7.9)	2.2 (2.5)	9.7 (8.6)	7.6 (6.4)	13.1 (8.6)	14.0 (8.8)
	HAMD Sum21	5.4 (6.6)	1.1 (1.6)	6.3 (6.7)	5.9 (5.7)	9.9 (7.0)	10.95 (7.5)
	STAIS	42.2 (13.2)	33 (6.4)	48.9 (12.2)	44.9 (11.7)	51.7 (13.1)	51.5 (13.1)
	STAIT	43.2 (14.2)	32.6 (6.9)	49.3 (14.1)	47.6 (12.8)	53.1 (12.5)	54.2 (12.5)
Anhedonia	SHAPS	1.99 (2.9)	0.6 (1.2)	2.6 (4.0)	2.3 (2.7)	3.5 (3.5)	3.6 (3.7)
Life events and stress	LEQ Negative Events score	10 (13.8)	3.3 (4.3)	12.4 (11.8)	9.99 (8.9)	14.2 (11.1)	23.4 (24.1)
	LEQ Positive Events score	9.8 (9.3)	9.4 (7.1)	8.1 (7.7)	9.5 (8.4)	8.7 (9.5)	12.5 (14)
	PSS Sum	22.8 (10.8)	15 (6.1)	26.7 (10.8)	25.6 (8.9)	31.1 (9.7)	30.6 (9.6)
Maltreatment in childhood and youth	ACE Sum	1.6 (1.9)	0.5 (0.8)	1.7 (1.6)	2.4 (1.7)	1.3 (1.3)	3.96 (2.4)
	CTQ Emotional abuse	9.1 (4.7)	6.3 (1.7)	9.9 (4.9)	11.4 (4.2)	7.99 (3.0)	14.6 (5.9)
	CTQ Emotional neglect	10.7 (5.2)	7.5 (2.6)	11.8 (5.4)	14.1 (4.1)	9.2 (3.5)	16.5 (5.4)
	CTQ Physical abuse	6.2 (2.6)	5.3 (0.7)	6.4 (2.2)	6.4 (2)	5.5 (1.1)	9.5 (4.6)
	CTQ Physical neglect	7.2 (2.7)	5.8 (1.4)	7.2 (2.2)	8 (2.2)	6.4 (1.6)	10.4 (3.7)
	CTQ Sexual abuse	5.8 (2.5)	5.1 (0.4)	5.8 (2.0)	5.8 (1.9)	5.6 (1.7)	7.98 (4.9)
Mania Symp.	YMRS	1.2 (2.5)	0.5 (1.1)	0.8 (1.4)	1.1 (1.8)	1.3 (2)	2.9 (4.9)
Neg. Symp.	SANS sum score	5.7 (9.9)	0.6 (1.7)	5.2 (8.6)	6.9 (9.6)	7.96 (9.1)	15.6 (14.4)
Pos. Symp.	SAPS sum score	1.4 (5.2)	0.1 (0.6)	0.1 (0.4)	0.6 (1.6)	0.7 (1.8)	6.7 (11.5)

Personality	NEO-FFI Agreeableness	33.1 (6.0)	35.6 (5.1)	30.5 (4.8)	31.1 (5.5)	32.9 (6.3)	29.5 (5.9)
	NEO-FFI Conscientiousness	32 (7.5)	34.9 (6.1)	30.1 (6.9)	29.4 (7.7)	30.7 (7.5)	29.4 (7.8)
	NEO-FFI Extraversion	26.4 (8.2)	30.8 (6.1)	23.7 (9.2)	23.5 (8)	24.2 (7.3)	21.1 (7.9)
	NEO-FFI Neuroticism	22.1 (10.5)	14.7 (6.6)	26.9 (9.5)	26 (9.4)	28.2 (9.1)	29.4 (9.1)
	NEO-FFI Openness to experience	30.3 (7.0)	31.2 (6.6)	29.1 (5.2)	31.1 (6.9)	28.8 (7.4)	28.6 (7.4)
	SPQB (Schizotypy)	5.95 (4.7)	2.9 (2.6)	7.0 (4.7)	7.95 (4.2)	6.6 (4.3)	10.6 (4.8)
Protective factors	Maternal bonding	24.7 (8.6)	29.8 (4.8)	21.6 (9.9)	19.4 (7.4)	27 (6.2)	15.7 (8.7)
	Paternal bonding	22.3 (8.7)	26.95 (6.3)	22.5 (9.2)	16.9 (7.4)	23.2 (7.6)	15.7 (8.7)
	RS25 Sum Score (Resilience)	125.4 (27.2)	140.9 (16.8)	119.9 (25.1)	117.9 (26)	112.6 (27.3)	108.6 (30.1)
	Social support	4.1 (0.8)	4.6 (0.4)	3.8 (1.0)	3.8 (0.8)	4.0 (0.8)	3.5 (0.96)
SF Health Survey (Quality of life)	SF36 Bodily pain	76.3 (26.2)	88.9 (15.6)	53.9 (23.9)	78.9 (22.8)	61.1 (29.4)	59.7 (30)
	SF36 Energy/fatigue	50.2 (23.0)	66.5 (13.0)	40 (25.4)	45.9 (19.3)	31.2 (18.5)	34.1 (22.2)
	SF36 General health	66 (23.4)	81.2 (14.0)	57.9 (26.6)	64.7 (19.8)	49.4 (20.7)	45.8 (21.3)
	SF36 Mental health	64.3 (22.6)	81.4 (9.8)	55.6 (23)	59.4 (18.9)	46.4 (19.3)	45.6 (21.3)
	SF36 Physical functioning	89.5 (17.1)	97.6 (4.9)	81.97 (19.95)	93.3 (9.1)	78.2 (23.2)	76.3 (22.5)
	SF36 Role emotional	65.5 (42.4)	94.6 (16.4)	56.1 (43.2)	58.4 (41.3)	31.5 (39.4)	34.9 (42.3)
	SF36 Role physical	76.6 (36.7)	96.8 (11.5)	52 (39.2)	85.1 (27.9)	44.8 (40.9)	49.5 (43.5)
	SF36 Social functioning	73.3 (30)	95.2 (10)	66.4 (31)	67.6 (26.0)	49.4 (28.2)	48.5 (29.5)

Symptom checklist	SCL90R Additional Items	4.1 (4.0)	1.6 (1.7)	5.1 (3.8)	4.2 (3.2)	6.9 (4.4)	7.7 (4.3)
	SCL90R Anxiety	5.4 (6.7)	1.3 (1.7)	7.6 (6.9)	5.5 (4.96)	9.6 (7.3)	11.3 (8.6)
	SCL90R Depression	11.4 (11.9)	2.7 (2.7)	15.7 (13.8)	13.4 (10.0)	20.4 (11.7)	21.7 (12.4)
	SCL90R Global severity index	0.6 (0.6)	0.2 (0.1)	0.8 (0.6)	0.6 (0.4)	0.98 (0.6)	1.2 (0.6)
	SCL90R Hostility	2.9 (3.7)	0.8 (1.1)	4.1 (4.2)	3.0 (2.7)	4.9 (4.2)	5.8 (5.1)
	SCL90R Interpersonal sensitivity	6.8 (7.3)	1.9 (2.1)	8.6 (8.1)	8.8 (6.4)	10.3 (7.1)	13.8 (8.4)
	SCL90R Obsessive–compulsive behavior	8.4 (8.1)	2.7 (2.4)	10.1 (8.0)	9.5 (6.6)	14.6 (8.3)	15.3 (8.4)
	SCL90R Paranoid ideation	3.5 (4.4)	0.8 (1.4)	4.2 (4.7)	4.1 (3.5)	4.9 (4.6)	8.1 (5.3)
	SCL90R Phobic anxiety	2.3 (4.1)	0.3 (0.8)	2.4 (3.3)	2.1 (2.7)	3.9 (4.6)	6.4 (6.3)
	SCL90R Positive symptom distress Index	1.5 (0.5)	1.1 (0.2)	1.7 (0.5)	1.5 (0.4)	1.9 (0.5)	1.98 (0.6)
	SCL90R Positive symptom total	29.8 (21.0)	12.8 (8.7)	37.0 (20.7)	35.9 (16.7)	44.4 (17.4)	50.4 (17.7)
	SCL90R Psychoticism	3.9 (5.2)	0.7 (1.1)	4.9 (5.2)	4.3 (3.9)	6.2 (5.1)	9.4 (7.0)
	SCL90R Somatization	6.96 (7.1)	2.8 (2.4)	10.3 (7.4)	6.3 (4.6)	12.1 (7.9)	12.8 (9.3)
Neuropsychology	Verbal IQ	114 (13.7)	114.8 (13.6)	116.1 (12.2)	115.7 (13.9)	112.4 (12.8)	110.8 (14.5)
	VLMT Sum	56.8 (10.2)	60.0 (8.6)	57.3 (9.5)	56.95 (8.7)	55.99 (9.9)	48.7 (11.7)
	Corsi block-tapping test	17.3 (3.4)	18.4 (3.2)	17.2 (3.4)	17.1 (3.0)	16.7 (3.2)	15.3 (3.6)
	Letter Number Span test	16.1 (3.2)	16.8 (3.1)	16.4 (2.9)	16.5 (2.8)	15.9 (3)	14 (3.4)

Source: Pelin et al., 2021

Table A4

Genetic lasso regularized regression prediction models – summary statistics from 1000 runs

The table shows summary statistics from 1000 runs of lasso regularized regression models applied to characterize clusters using genetic variables. The numbers in the table represent the number of times the variable had a non-zero coefficient in 1000 runs and, in parentheses, the mean coefficient calculated from 1000 runs. The standard error of the mean is shown in square brackets.

Variable	Cluster 0 vs. all	Cluster 1 vs. all	Cluster 2 vs. all	Cluster 3 vs. all	Cluster 4 vs. all
Age	1000	566	215	928	1000
	-0.3 [0.004]	0.09 [0.004]	0.01 [0.001]	0.15 [0.003]	0.38 [0.002]
Gender	81	672	199	664	953
	0 [0.000]	-0.19 [0.006]	0.01 [0.001]	-0.08 [0.002]	0.14 [0.002]
AC1	68	471	431	633	242
	0 [0.000]	-0.03 [0.003]	-0.03 [0.001]	0.07 [0.002]	0.01 [0.001]
AC2	312	513	175	512	235
	-0.01 [0.001]	0.07 [0.004]	0 [0.001]	0.03 [0.001]	0 [0.001]
AC3	272	535	285	482	773
	0.01 [0.001]	0.13 [0.005]	-0.01 [0.001]	0.02 [0.001]	-0.07 [0.002]
AC4	334	601	154	424	960
	0.02 [0.001]	0.13 [0.004]	0 [0.001]	-0.01 [0.001]	-0.15 [0.002]
AC5	70	574	338	558	836
	0 [0.000]	0.1 [0.004]	-0.02 [0.001]	-0.04 [0.002]	0.09 [0.002]
AC6	64	474	211	438	516
	0 [0.000]	-0.01 [0.003]	0.01 [0.001]	0.02 [0.001]	-0.03 [0.001]
AC7	69	493	458	645	310
	0 [0.000]	-0.05 [0.003]	0.04 [0.002]	-0.07 [0.002]	0.01 [0.001]
AC8	65	457	186	473	248
	0 [0.000]	-0.02 [0.002]	0.01 [0.001]	-0.02 [0.001]	0 [0.001]
Family History Any	1000	466	964	474	1000
	-0.4 [0.003]	0.03 [0.006]	0.15 [0.003]	0.02 [0.002]	0.34 [0.003]
Family History BD	517	529	308	430	1000
	-0.04 [0.002]	-0.32 [0.012]	-0.02 [0.001]	-0.03 [0.002]	0.24 [0.002]
Family History MDD	229	457	630	698	191
	-0.01 [0.001]	-0.06 [0.006]	0.04 [0.002]	0.07 [0.002]	-0.01 [0.001]

Family History SCZ	81	527	465	580	735
	0 [0.000]	-0.09 [0.009]	-0.05 [0.002]	0.03 [0.002]	0.06 [0.002]
PGS Cross psychiatric disorder	766	544	363	491	822
	-0.07 [0.002]	0.34 [0.012]	0.02 [0.001]	-0.07 [0.003]	0.08 [0.002]
PGS ADHD	82	495	156	475	249
	0 [0.001]	-0.15 [0.006]	0 [0.001]	-0.02 [0.001]	0 [0.001]
PGS ASD	115	456	418	420	270
	0 [0.000]	-0.06 [0.004]	0.03 [0.001]	0 [0.001]	0.01 [0.001]
PGS BD	200	667	159	666	203
	-0.01 [0.001]	-0.26 [0.008]	0.01 [0.001]	0.1 [0.003]	-0.01 [0.001]
PGS MDD	737	550	167	352	926
	-0.06 [0.002]	0.09 [0.004]	0 [0.001]	-0.01 [0.001]	0.13 [0.003]
PGS Schizophrenia	405	465	111	351	971
	-0.02 [0.001]	-0.14 [0.006]	-0.01 [0.001]	0.01 [0.001]	0.18 [0.003]
PGS Educational attainment	634	584	138	474	996
	0.05 [0.002]	-0.11 [0.004]	0 [0.000]	0.02 [0.001]	-0.27 [0.003]
PGS Extraversion	314	490	352	530	918
	-0.02 [0.001]	0.06 [0.003]	-0.02 [0.001]	0.04 [0.002]	0.11 [0.002]
PGS Hedonic wellbeing	160	485	368	473	883
	0.01 [0.001]	0.07 [0.004]	0.02 [0.001]	-0.04 [0.002]	-0.13 [0.003]
PGS Neuroticism	494	493	620	672	544
	-0.03 [0.001]	-0.1 [0.004]	0.05 [0.002]	0.1 [0.003]	-0.05 [0.002]

Source: Pelin et al., 2021

Table A5

Characterization of the replication sample and its clusters regarding variables used in the clustering process

Category	Variable, mean (SD)	Full	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
	N	1250	535	38	266	215	196
Attachment style	RSQ Anxiety of separation	2.7 (0.7)	2.4 (0.5)	2.5 (0.6)	3.2 (0.8)	2.7 (0.8)	2.95 (0.9)
	RSQ Avoidance of closeness	2.5 (0.9)	1.9 (0.6)	2.4 (0.7)	3.3 (0.8)	2.4 (0.8)	2.96 (0.9)
	RSQ Desire for independence	3.97 (0.7)	3.9 (0.6)	3.95 (0.7)	3.96 (0.7)	3.9 (0.8)	4.1 (0.7)
	RSQ Lack of trust	2.4 (0.9)	1.7 (0.5)	2.3 (0.6)	3.1 (0.8)	2.3 (0.8)	3 (0.9)
Depression and anxiety level	BDI-II Sum	11.3 (10.3)	2.9 (2.8)	6.6 (5.7)	14.6 (8.4)	15.8 (9.1)	20.4 (12.1)
	HAMA Sum	8.4 (8.2)	2.5 (3.0)	4.8 (4.5)	9.7 (6.8)	11.6 (7.5)	15.7 (10.2)
	HAMD Sum21	6.1 (6.8)	1.3 (2.1)	2.9 (3.2)	7.4 (6.4)	8.8 (6.5)	12.2 (7.9)
	STAIS	43 (12.8)	32.5 (6.0)	37.3 (8.8)	48.0 (9.4)	50.7 (11.8)	51.6 (13.1)
	STAIT	44.2 (13.8)	31.3 (6.5)	38.3 (9.9)	52.8 (9.8)	51.8 (11.3)	53.9 (13.4)
Anhedonia	SHAPS	1.97 (2.7)	0.4 (0.8)	0.9 (1.4)	3.0 (2.4)	3.3 (3.2)	3.2 (3.4)
Life events and stress	LEQ Negative Events score	10.4 (12.3)	3.1 (3.4)	6.1 (5.8)	11.2 (8.9)	11.8 (8.6)	22.6 (19.1)
	LEQ Positive Events score	9.4 (9.3)	8.2 (7.1)	9.6 (7.9)	10.2 (8.3)	7.2 (7.5)	12.7 (13.8)
	PSS Sum	23.8 (10.5)	14.4 (6.2)	19.3 (7.4)	27.9 (7.1)	29.5 (8.7)	31.9 (10.5)
Maltreatment in childhood and youth	ACE Sum	1.6 (1.8)	0.4 (0.6)	1.3 (1.3)	2.9 (1.6)	1.1 (1.3)	3.1 (2.3)
	CTQ Emotional abuse	9.1 (4.5)	6.1 (1.5)	8.2 (2.8)	13.4 (4.8)	7.8 (3.0)	13.2 (5.5)
	CTQ Emotional neglect	11.1 (5.2)	6.9 (2.1)	10.8 (3.7)	16.6 (4.7)	9.4 (3.4)	14.98 (6.1)
	CTQ Physical abuse	6.1 (2.4)	5.2 (0.6)	5.7 (1.3)	6.0 (1.7)	5.5 (1.0)	8.4 (4.1)
	CTQ Physical neglect	6.8 (2.5)	5.4 (1.0)	6.5 (1.4)	8.5 (2.3)	6.1 (1.7)	8.8 (3.6)
	CTQ Sexual abuse	5.7 (2.5)	5 (0.3)	5.3 (1.0)	5.3 (1.0)	5.3 (1.2)	7.8 (4.8)
Mania Symp.	YMRS	1.4 (2.7)	0.3 (0.7)	1.2 (1.7)	1.3 (2.2)	1.3 (1.7)	3.2 (4.6)
Neg. Symp.	SANS sum score	4.3 (7.3)	0.3 (0.9)	1.7 (3.1)	3.96 (5.6)	6.4 (7.3)	10.4 (10.7)
Pos. Symp.	SAPS sum score	0.7 (2.7)	0.04 (0.2)	0.2 (0.6)	0.5 (1.4)	0.4 (1.2)	2.8 (5.3)

Personality	NEO-FFI Agreeableness	33.2 (6.1)	37.4 (4.1)	32.6 (5.3)	30.0 (5.7)	34.5 (5.6)	29.96 (6.6)
	NEO-FFI Conscientiousness	32.2 (7.4)	36.3 (6)	32.7 (6.4)	29.5 (6.9)	30.9 (7.4)	29.9 (8.2)
	NEO-FFI Extraversion	25.8 (8.2)	32 (5.8)	26.7 (6.8)	19.6 (6.3)	24.1 (8.2)	22.7 (8.5)
	NEO-FFI Neuroticism	23.1 (10.2)	13.9 (6.6)	19.5 (8.2)	30.4 (7.3)	26.7 (8.7)	30.7 (8.6)
	NEO-FFI Openness to experience	30.4 (6.8)	32.7 (5.4)	30.7 (6.2)	29.5 (7.6)	29.2 (6.8)	29.3 (8)
	SPQB (Schizotypy)	5.97 (4.7)	2.4 (2.4)	5.1 (3.6)	9.6 (4.9)	5.8 (4.2)	9.4 (4.8)
Protective factors	Maternal bonding	24.4 (8.5)	30.9 (4.1)	24.1 (6.6)	17.4 (8.6)	26.8 (6.9)	18.4 (9.6)
	Paternal bonding	21.2 (8.9)	28.8 (5)	19.7 (7.2)	14.1 (7.9)	23.9 (7.7)	16.1 (9.2)
	RS25 Sum Score (Resilience)	125.1 (26.9)	145.1 (14.3)	134.0 (22.0)	105.3 (22.6)	116.9 (24.7)	109.8 (28.8)
	Social support	4.1 (0.7)	4.7 (0.3)	4.3 (0.5)	3.5 (0.8)	4.1 (0.7)	3.7 (0.9)
SF Health Survey (Quality of life)	SF36 Bodily pain	74.7 (26.1)	90.5 (13.3)	83.1 (18.6)	77.3 (21.7)	66.3 (28.9)	53.9 (27.9)
	SF36 Energy/fatigue	47.5 (22.1)	66.4 (13.1)	58.1 (16.6)	38.98 (15.3)	30.8 (15.1)	34.6 (22.5)
	SF36 General health	68.1 (23)	88.6 (12.8)	75.9 (16.3)	64.5 (16.6)	56.3 (20.8)	49.8 (22.4)
	SF36 Mental health	62.8 (22.1)	83.7 (9.4)	73.4 (13.8)	55.1 (16.8)	48.2 (17.2)	45.5 (22)
	SF36 Physical functioning	89.8 (15.5)	97.7 (4.6)	94.8 (8.6)	92.5 (8.9)	85.6 (16.5)	77.4 (21.9)
	SF36 Role emotional	60.6 (43.2)	95.8 (13.2)	81.1 (31.1)	47.5 (40.7)	27.9 (36.3)	36.4 (44.1)
	SF36 Role physical	71.2 (36.1)	97.1 (10.2)	89.2 (17.8)	70.8 (28.7)	47.1 (37)	44.6 (40.9)
	SF36 Social functioning	70.4 (30.1)	95.7 (9.1)	85.6 (17.6)	63.8 (28.3)	50.9 (26.5)	46.3 (30.1)

Symptom checklist	SCL90R Additional Items	4.4 (3.8)	1.6 (1.6)	2.8 (2.1)	5.6 (3.8)	5.6 (3.4)	7.5 (4.6)
	SCL90R Anxiety	5.4 (6.2)	1.1 (1.4)	2.6 (2.5)	6 (5.6)	7.4 (5.6)	11.4 (8.1)
	SCL90R Depression	12.2 (11.6)	2.7 (3.3)	6.5 (5.8)	16.1 (10.1)	18.0 (9.8)	22.4 (13.4)
	SCL90R Global severity index	0.6 (0.5)	0.2 (0.1)	0.3 (0.2)	0.8 (0.4)	0.8 (0.4)	1.2 (0.7)
	SCL90R Hostility	3.2 (3.9)	0.9 (1.2)	1.7 (1.8)	4.3 (3.4)	3.9 (3.5)	6.6 (5.5)
	SCL90R Interpersonal sensitivity	6.96 (7.0)	1.7 (2.1)	4.1 (3.6)	11.2 (6.8)	8.0 (5.6)	13.6 (8.8)
	SCL90R Obsessive–compulsive behavior	8.7 (7.7)	2.5 (2.6)	4.7 (3.99)	11.3 (5.9)	12.7 (6.5)	15.1 (9.0)
	SCL90R Paranoid ideation	3.3 (4.0)	0.7 (1.1)	1.9 (2.0)	4.7 (3.5)	3.3 (3.0)	7.3 (5.7)
	SCL90R Phobic anxiety	2.3 (3.9)	0.2 (0.7)	0.6 (1.1)	2.5 (2.7)	2.8 (3.1)	6.3 (6.2)
	SCL90R Positive symptom distress Index	1.5 (0.5)	1.1 (0.2)	1.2 (0.2)	1.6 (0.4)	1.7 (0.4)	2.0 (0.6)
	SCL90R Positive symptom total	31.2 (19.5)	12.5 (9.1)	22.7 (11.9)	41.7 (15.9)	40.0 (14.5)	48.4 (19.9)
	SCL90R Psychoticism	3.7 (4.9)	0.5 (1.0)	1.6 (1.8)	5.7 (4.1)	4.4 (3.9)	8.6 (7.1)
	SCL90R Somatization	7.1 (6.5)	2.8 (2.1)	4.2 (3.2)	8.0 (5.9)	9.3 (5.7)	13.0 (8.7)
Neuropsychology	Verbal IQ	113.8 (13.4)	117.4 (13.4)	116.4 (13.8)	110.1 (12.4)	112.5 (12.2)	109.3 (12.9)
	VLMT Sum	56.1 (9.6)	60.6 (6.8)	56.2 (9.5)	54.7 (8.9)	55.6 (9.6)	52.4 (10.9)
	Corsi block-tapping test	17.2 (3.1)	18.6 (2.6)	17.5 (2.7)	17.0 (3.1)	16.9 (3.4)	15.6 (3.1)
	Letter Number Span test	16.1 (3.4)	17.9 (2.6)	16.5 (3.2)	15.7 (3.1)	15.9 (3.3)	14.2 (3.6)

Source: Pelin et al., 2021

Table A6

Significance testing with the Westfall and Young procedure – comparison of discovery-replication pairs

After the discovery and replication clusters were matched, further statistical analysis, i.e., significance testing with the Westfall and Young procedure (described in the Supplementary Methods S6), was carried out to confirm the matching and homogeneity of linked discovery and replication clusters. All 57 variables which entered the clustering process were tested. The table shows only the variables that were significantly different between the linked clusters. A matching cluster assignment was confirmed for four clusters. n.s.: adjusted p>0.05

Discovery-Replication cluster pair	Significant variable	t- statistic	p-value adjusted using Westfall and Young	p-value further adjusted for the number of comparisons ($N = 5$)
Cluster 0 – Cluster 0	SF36 General health	5.7	2×10^{-3}	1×10^{-1}
	NEO-FFI Agreeableness	4.2	4×10^{-3}	2×10^{-2}
	Letter Number Span test	4.1	4×10^{-3}	2×10^{-2}
	Paternal Bonding	3.5	3×10^{-2}	$1 \times 10^{-1} (n.s.)$
	CTQ Physical Neglect	-3.8	4×10^{-2}	$2 \times 10^{-1} (n.s.)$
Cluster 2 – Cluster 2	NEO-FFI Extraversion	-4.1	4×10^{-3}	2×10^{-2}
	NEO-FFI Neuroticism	3.9	1×10^{-2}	$5.7 \times 10^{-1} (n.s.)$
	CTQ Emotional Neglect	3.8	2×10^{-2}	$8 \times 10^{-2} (n.s.)$
	RS25 Sum Score	-3.8	2×10^{-2}	$8 \times 10^{-2} (n.s.)$
	RSQ Avoidance of closeness	3.4	4×10^{-2}	$2 \times 10^{-1} (n.s.)$
	RSQ Lack of trust	3.4	4×10^{-2}	$2 \times 10^{-1} (n.s.)$
Cluster 3 – Cluster 3	SCL90R Paranoid ideation	-4.1	4×10^{-3}	2×10^{-2}
	SCL90R Somatization	-3.9	6×10^{-3}	3×10^{-2}
	SCL90R Psychoticism	-3.8	6×10^{-3}	3×10^{-2}
	SCL90R Global severity index	-3.9	1×10^{-2}	$8 \times 10^{-2} (n.s.)$
	SF36 Physical functioning	3.5	2×10^{-2}	$1 \times 10^{-1} (n.s.)$
	SCL90R Interpersonal sensitivity	-3.3	4×10^{-2}	$2 \times 10^{-1} (n.s.)$
Cluster 4 – Cluster 4	CTQ Physical Neglect	-3.8	7×10^{-3}	3×10^{-2}
	SAPS	-4.5	9×10^{-3}	4×10^{-2}
	SANS	-3.7	1.5×10^{-2}	$7 \times 10^{-2} (n.s.)$

Source: Pelin et al., 2021

Figure A1

Sparse group Lasso analysis steps with Healthy *vs.* MDD phenotype